Wahrscheinlichkeit, Statistik, Induktion

Christian Wurm
cwurm@phil.hhu.de

July 5, 2017

Contents

Kursinhalte und Quellen

Das Ziel soll es sein, Methoden der Wahrscheinlichkeitstheorie, Statistik und des maschinellen Lernens zu verstehen. Insgesamt kann man das Thema umschreiben mit der Frage: wie können wir sinnvolle Schlüsse mit unsicherer und ungenügender Information ziehen? Was wichtig ist: es geht mir nicht um die einfache Anwendung fertiger Methoden (was oft genug sinnlos ist), sondern um Verständnis. Das hat natürlich Vor- und Nachteile, macht die Sache aber insgesamt nicht leichter.

Dieses Skript orientiert sich in Sachen Wahrscheinlichkeitstheorie in weiten Teilen am Skript von Marcus Kracht (zu finden online unter http://wwwhomes.uni-bielefeld.des/mkracht/html/statistics.pdf); das ist für diejenigen, die es ganz genau wissen wollen. Dort finden sich ausführlichere Definitionen und Beweise, die ich hier meist auslasse.

Weiterhin benutze ich Edwin Jaynes' "Probability Theory: The Logic of Science". Jaynes war Physiker und einer der wichtigsten Wegbereiter der Bayesianischen Statistik.

Für den Teil um das maschinelle Lernen benutze hauptsächlich ich von Stuart Russell & Peter Norvig "Artificial Intelligence: A Modern Approach", ein Buch das gleichermaßen breit, gut informiert, gründlich wie leicht zugänglich ist, das also nur empfohlen werden kann.

Ein sehr neues Buch das ich benutze und empfehlen kann ist "Deep Learning" von Bengio et al., hier v.a. die ersten Kapitel. Hier werden einige Dinge sehr präzise umrissen; das meiste spielt aber hier erstmal keine Rolle.

1 Induktion und Lernen – eine Begriffsklärung

In der Literatur ist oft etwas undifferenziert von Lernen und Induzieren die Rede. Dabei gibt es eine klare und sinnvolle Unterscheidung:
"**Lernen**" bedeutet: wir wissen, was gelernt werden soll, und uns interessiert, wie und ob jemand, der das Ziel nicht kennt, dorthin gelangt. In diesem Sinne kann man z.B. sagen: die Kinder haben Arithmetik gelernt, die Studenten haben Algebra gelernt etc.
"**Induktion**" bedeutet: wir möchten eine allgemeine Regel/System erstellen, die normalerweise für eine unendliche Menge von Beobachtungen gilt, für die wir aber nur eine endliche Menge von Beobachtungen haben. Der entscheidende Punkt ist: wir kennen nicht die korrekte Regel, wir wissen nur dass es eine gibt. Was immer wir am Ende haben, ist möglicherweise falsch. Beispiele sind:

- Die Wahrscheinlichkeit eines gewissen Satzes in einer gewissen Sprache (woher sollen wir das wissen)
- Die Theorie der Schwerkraft (kann ja immer noch falsch sein)
- Eine Grammatik (für eine unendliche Sprache) gegeben eine endliche Menge von Sätzen die wir beobachten

Die Beispiele zeigen schon: für praktische und wissenschaftliche Anwendungen ist der Begriff der Induktion tatsächlich viel interessanter und relevanter als der des Lernens. Der Begriff der Induktion ist eng mit dem der Wahrscheinlichkeit verknüpft, insbesondere in der Praxis. Deswegen werden wir uns zunächst damit beschäftigen.

2 Wahrscheinlichkeiten - eine erste Intuition

2.1 Wahrscheinlichkeit als eine Theorie rationalen Handelns

Praktisch alles, was wir in diesem Seminar machen, basiert auf Wahrscheinlichkeitstheorie. Deswegen ist es wichtig, dass wir eine gute Intuition dafür haben, was Wahrscheinlicheit bedeutet. Die Wahrscheinlichkeit ist erstmal ein Maß dafür, wie sicher/unsicher wir sind, dass ein Ereignis eintritt. Dabei bezeichnet man mit 1 die Sicherheit, dass es eintritt, mit 0 die Sicherheit,

dass es nicht eintritt; Wahrscheinlichkeiten sind also Zahlen in [0,1]. Wir schreiben P(A) für die Wahrscheinlichkeit von A, wobei A für ein beliebiges Ereignis steht. Nehmen wir nun 2 Ereignisse A,B; nehmen wir weiterhin an, $P(A) > P(B)$. Dann bedeutet das soviel wie: wir gehen davon aus, A eher eintritt als B. Das hat eine sehr natürliche Interpretation, wenn wir z.B. von Risiken und Rationalität sprechen: nehmen wir an

A=Ein Fahrradfahrer stirbt in einem Unfall, weil er keinen Helm aufhat.

B=Ein Fahrradfahrer stirbt in einem Unfall, weil er den Radweg gegen die Fahrtrichtung fährt.

Nehmen wir weiterhin an, $P(A) < P(B)$ (das lässt sich mit Statistiken verifizieren). In diesem Fall würden wir sagen, ein Radfahrer, der mit Helm den Radweg gegen die Fahrtrichtung fährt, ist irrational (aber nicht unbedingt, wenn er ohne Helm fährt). Im Zusammenhang mit Risiken gilt also: Wahrscheinlichkeiten haben viel mit rationalem handeln zu tun, und in gewissem Sinne ist die Wahrscheinlichkeitstheorie so etwas wie eine **Theorie des rationalen Handelns**.

Um dieses Beispiel Konzept weiter zu klären, nehmen wir ein etwas komplexeres Beispiel. Nehmen wir an, sie wollen auf einen Berg; der Berg ist schön, die Aussicht sicher auch. Allerdings ist der Berg auch steil, wenn man ihn hochklettert kann es sein das man fällt, und wenn man fällt, dann ist man tot. Das allein ist aber noch kein Grund, nicht hochzugehen – sonst dürften Sie ja auch nicht in ein Auto (oder aufs Fahrrad) steigen. Die Frage ist: ist das Risiko akzeptabel, also im Bereich dessen, was sie eingehen würden? Dieses Risiko ist natürlich die Wahrscheinlichkeit, dass Sie runterfallen:

$$P(F) = x$$

Wir suchen also $P(F)$, und diese Größe ist unbekannt. Sie *schätzen* diese Größe aber auf eine gewisse Art und Weise. *Schätzen* ist hier bereits ein technischer Begriff, und wir nennen bezeichnen die geschätzte Wahrscheinlichkeit mit

$$\hat{R}(F) = \hat{x}.$$

Geschätzte Wahrscheinlichkeiten haben also einen Hut auf. Nun wird die Sache aber komplizierter: nehmen wir an, $\hat{x}$ ist ihnen zu groß, d.h. das Risiko

ist Ihnen zu hoch. Nun gibt es noch eine weitere Optionen: bei riskanten Bergtouren geht man meistens am Seil, damit im Falle eines Sturzes Ihr Gefährte Sie halten kann. Allerdings kann er das nicht mit bloßen Händen: sondern nur, wenn das Seil durch einen Haken läuft – und der Haken hält! Nehmen wir nun an, es gibt auf Ihrem Weg alte Haken. Demnach ist es von unten gesehen unmöglich zu sagen, ob sie halten oder nicht: wir haben keine relevante Information.

Wie haben wir also das Risiko zu bewerten? Hier spielen nun zwei Faktoren eine Rolle:

1. Wie ist das Risiko, dass Sie stürzen?
2. Und wie ist die Wahrscheinlichkeit, dass die Haken einem Sturz standhalten?

Hier finden wir unser erstes wichtiges Prinzip: da wir für 2. keine relevante Information haben, sagen wir (H ist das Ereignis dass ein Haken hält):

$$\hat{P}(H) = 0.5$$

Das ist das **Prinzip der Indifferenz**: falls wir keinerlei Information haben ob ein Ereignis E eintritt oder nicht, dann schätzen wir $\hat{P}(E) = 0.5$. Dieses Prinzip muss man oft noch leicht generalisieren (dann wird Formulierung etwas abstrakter):

Prinzip der Indifferenz Sei **E** ein Zufallsexperiment mit n möglichen Ergebnissen, $E_1, ..., E_n$. Wenn wir keinerlei relevante Information haben über **E**, dann gilt für all $i : 1 \leq i \leq n$: $\hat{P}(E_i) = \frac{1}{n}$ (man nehme einen handelsüblichen Würfel, dann haben wir $n = 6$, und das Zufallsexperiment ist ein Wurf).

Noch eine weitere Sache kann man hier sehen: gegeben ein Ereignis E bezeichnen wir mit $\overline{E}$ sein **Komplement**, also die Tatsache dass es (im Rahmen des Zufallsexperimentes) *nicht* stattfindet. Mit E_1E_2 bezeichnen wir kurzerhand die Tatsache, dass zwei Ereignisse E_1 und E_2 stattfinden. Das folgende ist nun leicht zu sehen (erinnern wir uns dass R für dass Runterfallen steht):

$$P(R) = P(\overline{H}S),$$

wobei S für das Ereignis des Stürzens steht. Um also $P(R)$ zu errechnen, müssen wir $P(\overline{H}S)$ errechnen. Und hier kommt das zweite große Prinzip der Wahrscheinlichkeitstheorie: die logischen Operationen von Konjunktion ("und"), Negation ("nicht") etc. müssen wir transformieren in **numerische Operationen**. Denn am Ende wollen wir *eine* Zahl haben, die unser geschätztes Risiko wiedergibt. Genau diese Rechenregeln werden wir als nächstes besprechen. Bei diesen Regeln geht es darum, logische Verknüpfungen von Operationen umzuwandeln in numerische Operationen, anhand derer wir das Risiko quantifizieren können.

2.2 Wahrscheinlichkeit und Induktion

Hier haben wir Wahrscheinlichkeiten beschrieben als ein Mittel, um uns rational zu verhalten. Im Zusammenhang mit Induktion suchen wir etwas anderes, aber sehr ähnliches: nicht die rationalste Verhaltensweise, sondern die rationalste Theorie über die Natur der Dinge. Wir suchen also eine rationale Sicht der Dinge. Das ist in der Tat für uns die geläufigste Anwendung für Wahrscheinlichkeiten: sie sollen uns sagen:

Frage der Induktion Gegeben eine Reihe Beobachtungen, die wir gemacht haben, was ist die plausibelste Theorie der zugrundeliegenden Gesamtheit/Realität?

Mit *plausibel* wird üblicherweise gemeint: hat die höchste Wahrscheinlichkeit. Hierbei spielen normalerweise 2 Faktoren eine Rolle:

1. Wie plausibel sind unsere Beobachtungen unter der Annahme, dass die Theorie richtig ist?

2. Wie plausibel ist unsere Theorie in sich?

Denn es kann sein, dass unsere Beobachtungen sehr wahrscheinlich sind unter der Annahme, dass ein Troll sie mit einer gewissen Absicht generiert; aber diese Theorie ist in sich sehr unwahrscheinlich.

Mit "zugrundeliegender Realität" meinen wir meistens eine Wahrscheinlichkeitsverteilung oder eine zugrundeliegende Gesamtheit (Population), von der wir nur einzelne Stichproben beobachten können. Z.B. eine Fabrik stellt Fernseher her; wir prüfen davon eine Auswahl, z.B. 1000 Stück. Dann möchten wir wissen, wie viele der insgesamt produzierten Fernseher (Population) defekt sind, bzw. wie die Wahrscheinlichkeit ist, dass ein beliebiger produzierter Fernseher defekt ist (Wahrscheinlichkeitsverteilung).

Weiterhin möchten wir oft wissen: mit welcher Sicherheit können wir diesen Schluss ziehen? Natürlich ist jede Annahme dieser Art sicherer, je mehr Fernseher wir prüfen. Mit diesen Themen befasst sich **statistische Inferenz**, und wir werden oft Probleme dieser Art treffen.

2.3 Bedeutung von Wahrscheinlichkeiten

Eine Sache, die man gleich zu Anfang klären sollte, ist: *was bedeuten eigentlich Wahrscheinlichkeiten*? Üblicherweise ist man versucht zu sagen: wenn eine Münze mit einer Wahrscheinlichkeit von 1/2 auf Kopf fällt, dann heißt das, dass sie perfekt symmetrisch ist, und weiterhin: wenn wir sie oft genug werfen, wird sie in ca. der Hälfte der Fälle auf Kopf landen. Die Wahrscheinlichkeit beschreibt also eine physische Eigenschaft und in der Folge ein Verhalten.

Das klingt gut, ist aber problematisch: was ist die Wahrscheinlichkeit, dass es Leben auf dem Mars gibt? Und gegeben dass wir eine Münze finden und werfen, was ist die Wahrscheinlichkeit, dass sie auf Kopf landet? Hier haben wir einen anderen Begriff von Wahrscheinlichkeit: er drückt die Stärke unserer Überzeugung aus. Diese kann – im Falle des Mars – mehr oder weniger informiert sein. Im Falle der Münze sagen wir: die Wahrscheinlichkeit, dass sie auf Kopf landet, ist 1/2, denn wir haben keinerlei wissen, dass uns dahin bringen würde, Kopf oder Zahl vorzuziehen. Diese uniforme Verteilung ist also **Ausdruck unserer Ignoranz**. Darauf beruht das Prinzip der Indifferenz.

3 Grundlagen der Wahrscheinlichkeitstheorie

3.1 Desiderata

Wir haben gesehen dass wir für die Wahrscheinlichkeitstheorie 2 große Desiderata haben:

1. Wir wollen (aussagen)logische Operationen für Ereignisse; und
2. wir möchten die logischen Operationen *numerisch interpretieren*, d.h. in numerische Funktionen verwandeln.

3.2 Boolesche Algebren

Logische Operationen können wir in Booleschen Algebren interpretieren:

Definition 1 *Sei M eine Menge. Ein Mengensystem $\mathcal{M} \subseteq \wp(M)$ ist eine Boolesche Algebra über M, falls*

1. *$M \in \mathcal{M}, \emptyset \in \mathcal{M}$;*
2. *falls $N \in \mathcal{M}$, dann ist auch $\overline{N} := M - N \in \mathcal{M}$;*
3. *falls $N_1, N_2 \in \mathcal{M}$, dann sind auch $N_1 \cup N_2 \in \mathcal{M}$.*

NB: die Definition impliziert dass falls $N_1, N_2 \in \mathcal{M}$, dann ist auch $N_1 \cap N_2 \in \mathcal{M}$, da $N_1 \cap N_2 = \overline{\overline{N_1} \cup \overline{N_2}}$. Unsere Definition betrifft eigentlich nur einen Spezialfall von Booleschen Algebren, nämlich solchen über Mengensystemen. Allerdings kann jede endliche Boolesche Algebra auf diesen Spezialfall reduziert werden.

Übung: Mengenlehre und Partitionen

- $M \cap N = \overline{\overline{M} \cup \overline{N}}$ (Interdefinierbarkeit 1)
- $M \cup N = \overline{\overline{M} \cap \overline{N}}$ (Interdefinierbarkeit 2)
- $\overline{\overline{M}} = M$ (doppeltes Komplement)
- $(M \cup N) \cap O = (M \cap O) \cup (N \cap O)$ (de Morgan)

Eine Partition einer Menge M ist eine Menge $X \subseteq \wp(M)$, d.h. $X = \{N_1, ..., N_i\}$ (im endlichen Fall), und es gilt:

1. $N_1 \cup ... \cup N_i = M$
2. für alle $N_i, N_j \in X$, entweder $N_i = N_j$ oder $N_i \cap N_j = \emptyset$.

3.3 Einige Beobachtungen

Wir haben bereits gesagt, dass Wahrscheinlichkeiten Zahlen in $[0, 1]$ sind, wobei wir die Korrespondenz haben

$0 \cong$ Unmöglichkeit
$1 \cong$ Sicherheit

Nun haben wir, aus logischen Gründen folgendes:

(1) $P(A) \leq P(A \text{ oder } B)$ und $P(B) \leq P(A \text{ oder } B)$

(In Zukunft schreiben wir: $P(A \cup B)$). Das ist klar: wann immer A eintritt, tritt auch A oder B ein, also ist das Ereignis wahrscheinlicher etc. Ebenso klar ist:

(2) $P(A \text{ und } B) \leq P(A)$ und $P(A \text{ und } B) \leq P(B)$

(In Zukunft schreiben wir: $P(A \cap B)$ oder einfach $P(AB)$). Das ist klar: die Wahrscheinlichkeit, dass sie bei Ihrer nächsten Radfahrt angefahren werden ist größer als die, dass sie angefahren werden und im Zuge dessen 50euro finden.

Gleichzeitig haben wir folgendes: sei $\bot$ ein Ereignis, das vollkommen unmöglich ist, z.B. Sie würfeln (mit einem handelsüblichen Würfel) eine 7. Dann haben wir natürlich:

(3) $P(A \cap \bot) = 0$; $P(A \cup \bot) = P(A)$

Also, in Worten: $\bot$ ist *absorbierend für Konjunktion* und *neutral für Disjunktion*.

Umgekehrt, sei $\top$ ein Ereignis, dessen Eintritt sicher ist, z.B. dass Sie eine Zahl zwischen 1 und 6 würfeln. Dann haben wir

(4) $P(A \cap \top) = P(A)$; $P(A \cup \top) = 1$

Also gilt: $\top$ ist absorbierend für Disjunktion, und neutral für Konjunktion. Nun haben wir, nach Annahme:

(5) $\quad P(\top) = 1;\ P(\bot) = 0$

Wir suchen also Operationen, für die 1, 0 jeweils neutral bzw. absorbierend sind. Das wird erfüllt von den Operationen + und $\cdot$:

(6) $\quad n + 0 = n \qquad n \cdot 0 = 0$

Ebenso haben wir:

(7) $\quad n \cdot m \leq n \qquad$ und $\qquad n \cdot m \leq m \qquad$, für $n, m \in [0, 1]$,

sowie:

(8) $\quad n \cdot 1 = n \qquad n + 1 \geq 1$

sowie:

(9) $\quad n \leq n + m$ und $m \leq n + m$, für $n, m \in [0, 1]$

Wir haben also folgende Korrespondenz:

Konjunktion $\cong \cdot$
Disjunktion $\cong +$

Das Problem ist, dass sich in dem einfachen Fall die Wahrscheinlichkeiten *nicht* auf 1 aufsummieren. Wir haben eine Korrespondenz, aber das ist noch zu einfach gedacht. Das sieht man auch an folgendem Beispiel:

(10) $\quad P(A \cap A) = P(A) \neq P(A) \cdot P(A)$

sowie

(11) $\quad P(A \cup A) = P(A) \neq P(A) + P(A)$

Konjunktion und Disjunktion sind also **idempotent**, im Gegensatz zur Addition und Multiplikation. Die Materie ist also durchaus komplex; es gibt allerdings eine wunderbar elegante Lösung, die uns mit allen nötigen Rechenregeln versorgt.

3.4 Definition von Wahrscheinlichkeitsräumen

Folgende Definition stammt von Kolmogorov, und ist das Ergebnis langer Überlegungen und Dispute.

Definition 2 *Ein Wahrscheinlichkeitsraum ist ein Tripel* $(\Omega, \mathfrak{A}, P)$*, wobei* $\mathfrak{A} \subseteq \wp(\Omega)$ *eine Boolesche Algebra ist, und* $P : \mathfrak{A} \to [0, 1]$ *eine Wahrscheinlichkeitsfunktion, so dass*

1. $P(\Omega) = 1$*;*
2. $P(\emptyset) = 0$*, und*
3. *falls* $A_1, A_2, ..., A_n$ *paarweise disjunkt sind, dann ist*

$$P(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i)$$

Zur Erklärung: mit **paarweise disjunkt** meint man: für alle i, j so dass $1 \leq i, j \leq n$, falls $i \neq j$, dann ist $A_i \cap A_j = \emptyset$.

Die Bedingung der Booleschen Algebra ist wie folgt zu verstehen: falls $A, B \subseteq \Omega$ Ereignisse sind, die eine Wahrscheinlichkeit haben, dann haben auch die Ereignisse $A \cup B$ (d.h.: A oder B trifft ein), $A \cap B$ (d.h. beide A und B treffen ein) und $\overline{A}$ (d.h. A trifft nicht ein) eine Wahrscheinlichkeit.

3.5 Ereignisse und Ergebnisse

Wir nennen eine Menge $A \subseteq \Omega$ ein **Ereignis**; wir nennen $a \in \Omega$ ein **Ergebnis**. Meistens entspricht ein Ergebnis a einem Ereignis $\{a\}$. Aber nicht immer ist das intuitiv: nehmen wir an, wir würfeln mit zwei Würfeln, wobei unsere Ergebnisse die Form haben

$$\langle m, n \rangle$$

Nun ist "der erste Wurf ist eine 2" kein Ergebnis, sondern ein Ereignis, nämlich das Ereignis

$$\{\langle 2, 1 \rangle, ..., \langle 2, 6 \rangle\}$$

Daher weisen wir Wahrscheinlichkeiten normalerweise Ereignissen zu, nicht Ergebnissen.

3.6 Die Komplement-Regel

Wir kommen nun zu den Rechenregeln. Die Regel für die Berechnung des Komplementes $P(\overline{A})$ aus $P(A)$ lautet wie folgt:

(1) $P(\overline{A}) = 1 - P(A)$

Das lässt sich sehr einfach ableiten: wir haben

1. $P(A \cup \overline{A}) = P(\Omega) = 1$ und
2. $A \cap \overline{A} = \emptyset$;

also:

$$\begin{aligned} 1 &= P(A \cup \overline{A}) \\ &= P(A) + P(\overline{A}) \\ \Leftrightarrow 1 - P(A) &= P(\overline{A}) \end{aligned}$$

3.7 Die Summenregel

Die Summenregel erlaubt es uns, die logische Disjunktion rechnerisch aufzulösen. Die Summenregel lautet:

(2) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Intuitiv bedeutet das: um die Wahrscheinlichkeit einer Disjunktion zu berechnen, reicht es die Wahrscheinlichkeiten zu addieren, wenn man nur die Wahrscheinlichkeitsmasse abzieht, die auf die Konjunktion beider Ereignisse entfällt (illustrierbar mittels Venn-Diagramm). Das lässt sich wie folgt ableiten aus den Axiomen:

$$\begin{aligned} P(A \cup B) &= P(A \cup (B \cap \overline{A})) && \text{(Mengenlehre)} \\ &= P(A) + P(B \cap \overline{A}) && \text{(Disjunkte Mengen)} \\ &= P(A) + P(B \cap (\overline{A} \cup \overline{B})) && \text{(Mengenlehre)} \\ &= P(A) + P(\overline{\overline{B \cap (\overline{A \cap B})}}) && \text{(Mengenlehre)} \\ &= P(A) + P(\overline{\overline{B} \cup (A \cap B)}) && \text{(Mengenlehre)} \\ &= P(A) + (1 - P(\overline{B} \cup (A \cap B))) && \text{(Subtraktionsregel)} \end{aligned}$$

$$
\begin{aligned}
&= P(A) + (1 - P(\overline{B}) + P(A \cap B)) && \text{(Disjunkte Mengen)} \\
&= P(A) + (1 - (1 - P(B)) + P(A \cap B)) && \text{(Disjunkte Mengen)} \\
&= P(A) + (1 - (1 - P(B))) - P(A \cap B)) && \text{(Arithmetik)} \\
&= P(A) + (1 - 1) + P(B) - P(A \cap B)) && \text{(Arithmetik)} \\
&= P(A) + P(B) - P(A \cap B)) && \text{(Arithmetik)}
\end{aligned}
$$

3.8 Die Produktregel

Um die Konjunktion sinnvoll zu interpretieren, brauchen wir die Definition der *bedingten Wahrscheinlichkeit*. Wir definieren

$$(3) \quad P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Nehmen Sie nun an wir suchen die Wahrscheinlichkeit $P(A \cap B)$; wir bekommen sie durch eine ganz simple Termumformung:

$$(4) \quad P(A \cap B) = P(A|B)P(B)$$

Da $P(A \cap B) = P(B \cap A)$ ($\cap$ ist kommutativ), bekommen wir also die Produktregel:

$$(5) \quad P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$$

Intuitiv ist das wie folgt zu verstehen: wenn A und B eintreffen, dann bedeutet dass das 1. A eintrifft und 2. unter dieser Voraussetzung B eintrifft (oder umgekehrt). Wichtig ist: $P(A|B)$ sagt nichts über zeitliche Reihenfolge! So ist die Formel intuitiv richtig. Wir werden später noch mehr zum Konzept der bedingten Wahrscheinlichkeit erfahren.

Diese Umformung mag einem auf Anhieb nicht sehr hilfreich erscheinen, allerdings ist sie eines der zentralen Gesetze. Denn in der Praxis kennen wir oft bedingte Wahrscheinlichkeiten besser als unbedingte, so dass wir uns das Gesetz leicht zunutze machen können. Es gibt übrigens noch eine allgemeinere Form der Produktregel:

$$(6) \quad P(A \cap B|X) = P(A|BX)P(B|X) = P(B|AX)P(A|X)$$

Das generalisiert die letzte Formel, da X beliebig (auch leer) sein kann.

3.9 Das sog. Bayessche Gesetz und seine Bedeutung

Das Bayessche Gesetz bzw. Theorem ist im Prinzip auch nichts anderes als eine Term-Umformung, vorausgesetzt alle Definitionen soweit. Es sieht wie folgt aus:

(7)

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A \cap B)}{P(B)} \frac{P(A)}{P(A)} = \frac{P(B \cap A)}{P(A)} \frac{P(A)}{P(B)} = P(B|A) \frac{P(A)}{P(B)}$$

Die Bedeutung ist folgende: wir haben Wahrscheinlichkeitstheorie eingeführt als ein Werkzeug, um uns rational zu verhalten. Noch häufiger werden wir sie benutzen, um eine rationale Sicht der Dinge zu bekommen. Die Frage wird sein: gegeben unsere Beobachtungen, was ist die wahrscheinlichste Annahme über die Natur der Dinge? Seien also B unsere Beobachtungen, H eine Hypothese (über zugrundeliegende Wahrscheinlichkeiten); wir suchen also $P(H|B)$. Das lässt sich aber nicht ohne weiteres errechnen; wir bekommen normalerweise nur $P(B|H)$! Das Bayessche Gesetz erlaubt uns aber, von $P(B|H)$ zu $P(H|B)$ zu gelangen – mit einen Seitenannahmen, doch dazu später mehr.

3.10 Einige Beispiele von Wahrscheinlichkeitsräumen

3.10.1 Laplace-Räume

In einem Laplace-Raum gilt folgendes: wir haben $\mathfrak{A} = \wp(\Omega)$, das heißt zunächst, jedes mögliche Ereignis bekommt eine Wahrsccheinlichkeit. Außerdem haben wir, für alle $A \in \wp(\Omega)$, $P(A) = |A|/|\Omega|$. Das bedeutet soviel wie: alle *Ergebnisse*, also alle "atomaren" Ereignisse, sind gleich wahrscheinlich. Das beste Beispiel für einen Laplace Raum ist ein fairer Würfel mit n Zahlen (n ist beliebig, muss aber endlich sein!). Natürlich bedeutet das nicht, dass alle Ereignisse gleich wahrscheinlich sind, denn wenn wir einen handelsüblichen Würfel mit 6 Zahlen nehmen, dann ist das Ereignis $\{2, 4, 6\}$ eines geraden Ergebnisses natürlich wahrscheinlicher als das Ereignis $\{2\}$ dass wir eine 2 werfen.

3.10.2 Bernoulli-Räume

Ein Bernoulli Raum hat nur zwei Ergebnisse: wir haben $\Omega = \{1, 0\}$, außerdem haben wir wie vorher: $\mathfrak{A} = \wp(\Omega)$, und $P(1) = 1 - P(0)$. Das typische

Beispiel für einen Bernoulli-Raum ist der Wurf einer Münze, die möglicherweise auch unfair ist.

3.10.3 Diskrete Wahrscheinlichkeitsräume

Diskrete Wahrscheinlichkeitsräume sind eine Generalisierung von Laplace und Bernoulli-Räumen. Ein Wahrscheinlichkeitsraum ist diskret, falls $\mathfrak{A} = \wp(\Omega)$, also wenn jedes denkbare Ereignis eine Wahrscheinlichkeit hat.

Ein wichtiges Ergebnis ist das folgende (das wir hier nur informell erklären): jeder endliche Wahrscheinlichkeitsraum kann als ein diskreter Raum "aufgefasst werden". Mit der Wendung "aufgefasst werden" meinen wir soviel wie: kann darauf abgebildet werden, ohne dass wir irgendwelche Information verlieren.

3.11 Produkträume

Produkträume sind eine intuitiv sehr einfache Erweiterung von Wahrscheinlichkeitsräumen. Nehmen wir an wir haben einen Würfel und kennen seine Wahrscheinlichkeiten. Wir möchten jetzt aber Wahrscheinlichkeiten wissen dafür, dass wir mit demselben Würfel zweimal in Folge eine gewisse Zahl Würfeln; uns interessiert also beispielsweise das Ereignis $\langle 2, 3\rangle$ (die spitzen Klammern stehen hier für geordnete Paare, also hier für das Ereignis: erster Wurf 2, zweiter Wurf 3). Das Ereignis $\{\langle 2, 3\rangle\}$ ist allerdings *kein* Element unseres Wahrscheinlichkeitsraums. Wie geben wir ihm dennoch eine Wahrscheinlichkeit?

Hier hilft uns das Produkt zweier Räume, oder, in diesem konkreten Fall, das Produkt eines Raumes mit sich selbst. Wir nehmen zwei Räume $(\Omega_1, \mathfrak{A}_1, P_1)$ und $(\Omega_2, \mathfrak{A}_2, P_2)$. Die möglichen Ergebnisse des Produktraumes sind einfach definiert als $\Omega_1 \times \Omega_2$, das kartesische Produkt der beiden Ergebnismengen. Im obigen Beispiel wäre das also die Menge $\{1, 2, 3, 4, 5, 6\} \times \{1, 2, 3, 4, 5, 6\}$. Die Menge der Ereignisse stellt uns allerdings vor einige technische Schwierigkeiten, denn das kartesische Produkt zweier Booleschen Algebren ist nicht notwendig eine Boolesche Algebra. Wir brauchen also eine etwas kompliziertere Definition:

$$(8) \quad \mathfrak{A}_1 \bigotimes \mathfrak{A}_2 := \{\bigcup_{i=1}^{p} A_i \times B_i : \text{für alle } i,\ A_i \in \mathfrak{A}_1, B_i \in \mathfrak{A}_2\}$$

Wahrscheinlichkeiten im Produktraum werden wie folgt definiert:

(9) $(P_1 \times P_2)(A \times B) := P_1(A) \cdot P_2(B)$

Natürlich muss $(P_1 \times P_2)$ alle Bedingungen einer Wahrscheinlichkeitsfunktion erfüllen (s.o.). Dann lässt sich mit einiger Mühe folgendes zeigen:

Lemma 3 *Seien $\mathcal{P}_1 = (\Omega_1, \mathfrak{A}_1, P_1)$ und $\mathcal{P}_2 = (\Omega_2, \mathfrak{A}_2, P_2)$ zwei Wahrscheinlichkeitsräume. Dann ist $\mathcal{P}_1 \times \mathcal{P}_2 := (\Omega_1 \times \Omega_2, \mathfrak{A}_1 \times \mathfrak{A}_2, P_1 \times P_2)$, der **Produktraum** der beiden, auch ein Wahrscheinlichkeitsraum.*

3.12 Unabhängige Ereignisse

Zwei Ereignisse sind unabhängig von einander, falls in unserem Wahrscheinlichkeitsraum gilt: $P(A|B) = P(A)$. (das impliziert übrigens dass $P(B|A) = P(B)$. Warum?). Daraus wiederum können wir mithilfe der Definition der bedingten Wahrscheinlichkeiten direkt ableiten:

$$(10)\quad P(A|B) = \frac{P(A \cap B)}{P(B)} \Leftrightarrow\ P(A \cap B) = P(A|B) \cdot P(B) = P(A) \cdot P(B).$$

Wir können also die Wahrscheinlichkeit von $A \cap B$, falls A, B unabhängig sind, mittels $P(A) \cdot P(B)$ berechnen.

Ein typisches Beipiel für zwei unabhängige Ereignisse ist folgendes: wir werfen einen Würfel zweimal, und uns interessiert die Wahrscheinlichkeit dass wir beim ersten Wurf eine 1, beim zweiten Wurf eine 2 werfen. Woher wissen wir dass die beiden Ereignisse unabhängig sind? Zunächst betrachten wir unseren Wahrscheinlichkeitsraum. Sei $\mathcal{W} = (\Omega, \mathfrak{A}, P)$ der Wahrscheinlichkeitsraum (Bernoulli-Raum) eines einfachen Wurfes eines (gerechten) Würfels. Uns interessiert dann der Produktaum $\mathcal{W} \otimes \mathcal{W}$. Was sind die beiden Ereignisse A=erster Wurf 1, B=zweiter Wurf 2 in diesem Wahrscheinlichkeitsraum? Zunächst gilt: unsere Ergebnisse, d.h. atomare Ereignisse, sind geordnete Paare, und Ereignisse sind Teilmengen von $\Omega \times \Omega$. Daher gilt: $A = \{1\} \times \Omega$, und $B = \Omega \times \{2\}$; die Ereignisse sind also jeweils das kartesische Produkt einer 1-elementigen Menge mit der Menge Ω, wobei Ω einmal zur linken, einmal zur rechten Seite steht. (Warum?)

Wenn wir davon ausgehen, dass die beiden Ereignisse unabhängig sind, können wir leicht deren Wahrscheinlichkeit berechnen: $P(A) = P \times P(\{1\} \times$

$\Omega) = P(\{1\} \cdot 1$; $P(B) = P \times P(\Omega \times \{2\}) = P(\{1\} \cdot 1$. Woher wissen wir, dass die beiden Ereignisse unabhängig sind in unserem Produktraum $\mathcal{W} \times \mathcal{W}$? Wir können das kurz prüfen:

(11)

$$P(\{1\}\times\Omega|\Omega\times\{2\}) = \frac{P(\{1\} \times \Omega) \cap (\Omega \times \{2\})}{\Omega \times \{2\}} = \frac{P(\langle 1, 2\rangle}{\Omega \times \{2\}} = \frac{\frac{1}{36}}{\frac{1}{6}} = \frac{1}{6} = P(\{1\}\times\Omega)$$

NB: wir zeigen hier blo Dinge, die nach unserer Intuition offensichtlich sind. Allerdings ist es wichtig zu wissen, dass der Formalismus mit unseren Intuitionen übereinstimmt.

3.13 Bedingte Wahrscheinlichkeit

Nehmen wir an, Hans hat drei Kinder, und die Wahrscheinlichkeit, einen Jungen zu haben ist $\frac{1}{2}$. Die Wahrscheinlichkeit, dass Hans genau einen Jungen hat, ist $\frac{3}{8}$. (Warum?) Angenommen aber, wir wissen dass Hans eine Tochter hat, wie ist dann die Wahrscheinlichkeit dass er genau einen Sohn hat? Gleich sollte sie nicht sein, denn wir haben die Menge der möglichen Ereignisse reduziert - es ist unmöglich, dass er drei Söhne hat! Also hat sich die Menge der möglichen Ergebnisse geändert, statt 8 Ergebnissen finden wir nur noch 7. Wir nehmen an, dass die Wahrscheinlichkeiten weiterhin gleich verteilt sind. Außerdem gilt nach wie vor: in 3 der 7 Ereignisse hat Hans genau einen Sohn. Also schließen wir: sei A das Ereignis: genau ein Sohn; B das Ereignis: mindestens eine Tochter. Dann ist die Wahrscheinlichkeit von A gegeben B, geschrieben $A|B$, $\frac{3}{7}$.

Das war eine sehr intuitive Art Rechnung. Etwas genauer ist es wie folgt. Wenn wir zwei Ereignisse A, B betrachten, dann gibt es vier die beiden zu kombinieren: (1) A und B treffen ein, (2) A trifft ein, B nicht, (3) B trifft ein, A nicht, (4) keines von beiden trifft ein. Wenn wir nun nach der Wahrscheinlichkeit von $A|B$ fragen, dann haben wir bereits Möglichkeiten (2) und (4) eliminiert, es bleiben also nur (1) und (3). Wir verringern also den Raum der Möglichkeiten; diese sind: $P(A \cap B)$ und $P((\overline{A}) \cap B)$. Wir bekommen also als Wahrscheinlichkeit:

$$(12) \quad P(A|B) = \frac{P(A \cap B)}{P(A \cap B) + P((\overline{A}) \cap B)} = \frac{P(A \cap B)}{P(B)}$$

Die letzte Gleichung folgt, da $(A \cap B) \cup ((\overline{A}) \cap B) = B$, und $(A \cap B) \cap ((\overline{A}) \cap B)) = \emptyset$. Dies definiert die bedingte Wahrscheinlichkeit, und ist bekannt als **Bayes Gesetz der bedingten Wahrscheinlichkeit**.

Bedingte Wahrscheinlichkeiten sind von großer Wichtigkeit nicht nur für die Stochastik, sondern auch für die Statistik. Eine wichtige Konsequenz ist die folgende: wir können die Wahrscheinlichkeit eines Ereignisses $A \cap B$ errechnen durch

(13) $P(A \cap B) = P(A|B)P(B)$.

Weiterhin haben wir natürlich $A = (A \cap B) \cup (A \cap (\overline{B})$. Da $(A \cap B) \cap (A \cap (\overline{B}) = \emptyset$, gilt also auch $P(A) = P(A \cap B) + P(A \cap \overline{B}$. Daraus folgt:

(14) $P(A) = P(A|B)P(B) + P(A|\overline{B})P(\overline{B})$

Das bedeutet, leicht verallgemeinert, wenn wir eine Partition M von Ω haben, dann müssen wir nur die bedingten Wahrscheinlichkeiten $A|B_i : B_i \in M$ kennen, um die Wahrscheinlichkeit von A zu berechnen.

Der Grund warum bedingte Wahrscheinlichkeiten eine so große Rolle für die Statistik spielen ist der sogenannte **Satz von Bayes**. Oftmals ist unser Ziel, die Ordnung von bedingten Wahrscheinlichkeiten umzukehren. Was wir leicht berechnen können ist die Wahrscheinlichkeit eines Ereignisses in einem gegebenen Wahrscheinlichkeitsraum. In der Statistik verhält es sich aber umgekehrt: wir haben nur ein gewisses Ereignis, und wir möchten Rückschlüsse auf zugrundeliegende Wahrscheinlichkeiten machen. Wir möchten also von P(Ereignis|Wahrscheinlichkeitsraum) zu P(Wahrscheinlichkeitsraum|Ereignis). Der Satz von Bayes gibt uns dazu die Möglichkeit:

(15)
$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A \cap B)}{P(B)} \cdot \frac{P(A)}{P(A)} = \frac{P(B \cap A)}{A} \cdot \frac{P(A)}{P(B)} = P(B|A)\frac{P(A)}{P(B)}.$$

3.14 Verbundwahrscheinlichkeiten und Marginalisierung

Eine wichtiger und grundlegender Begriff der Wahrscheinlichkeitstheorie sind die sog. **marginalen Wahrscheinlichkeiten**. Die marginale Wahrscheinlichkeit von A ist einfach $P(A)$. Das Konzept ist sehr einfach, das Problem

ist aber dass wir oft nur die Wahrscheinlichkeit von 2 (oder mehr) gleichzeitig eintretenden Ereignissen oder bedingten Wahrscheinlichkeiten beobachten können. Z.B.:

- Die Wahrscheinlichkeit, dass unser Besuch zu spät kommt, wenn er mit der Bahn kommt ($\cong$ Wahrscheinlichkeit, dass die Bahn Verspätung hat)
- Die Wahrscheinlichkeit, dass unser Besuch zu spät kommt, wenn er mit dem Auto kommt ($\cong$ Wahrscheinlichkeit dass Stau ist)

Angenommen, wir haben (mehr oder weniger korrekte) Wahrscheinlichkeiten für die beiden Ereignisse $P(V|B)$ und $P(V|A)$. Wie kommen wir zur marginalen Wahrscheinlichkeit dass unser Besuch zu spt kommt?

Hierfür brauchen wir noch etwas, nämlich die Wahrscheinlichkeit dass unser Besuch das Auto/die Bahn nimmt, nämlich $P(A)$ und $P(B)$. Nun können wir folgende Tatsache nutzen:

1. A, B schließen sich gegenseitig aus; und
2. eines von beiden muss der Fall sein (nehmen wir mal an).

Das bedeutet: A, B **partitionieren** Ω. Das bedeutet aber auch: $V \cap A$ und $V \cap B$ partitionieren V, also:

$$P(V) = P(V \cap A) + P(V \cap B) \tag{16}$$

(Axiom 3 der Wahrscheinlichkeitsräume!) Das bedeutet, im Fall einer Partition vereinfacht sich die Summenregel in wesentlich, so dass wir einfach eine Addition bekommen. Mit der Multiplikationsregel können wir den Term folgendermaßen auflösen:

$$P(V) = P(V \cap A) + P(V \cap B) = P(V|A)P(A) + P(V|B)P(B) \tag{17}$$

Da wir diese Werte (nach Annahme) kennen, können wir also die Wahrscheinlichkeit berechnen. Das funktioniert, solange wir Partitionen des Wahrscheinlichkeitsraumes bilden (endlich und sogar unendlich viele); so bekommen wir die allgemeine Form der Marginalisierung:

$$P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i), \text{ vorausgesetzt } B_1, ..., B_n \text{ partitionieren} A \tag{18}$$

Das funktioniert übrigens sogar mit reellwertigen Parametern, nur brauchen wir dann Integrale.

3.15 Wahrscheinlichkeitsgesetze – allgemeine Form

Wir haben oben die wichtigsten Regeln der Wahrscheinlichkeitstheorie eingeführt. Zusammen mit dem Begriff der bedingten Wahrscheinlichkeit kann man sie noch allgemeiner formulieren. Zu Übersichtszwecken fügen wir hier nochmal alle zusammen:

Komplementregel: $P(\overline{A}|X) = 1 - P(A|X)$

Summenregel: $P(A \cup B|X) = P(A|X) + P(B|X) - P(A \cap B|X)$

Produktregel: $P(A \cap B|X) = P(A|X)P(B|A,X) = P(B|X)P(A|B,X)$

Bayes Gesetz: $P(A|B,X) = P(B|A,X)\frac{P(A|X)}{P(B|X)}$

Marginalisierung $P(A|X) = \sum_{i=1}^{n} P(A|B_i,X)P(B_i,X)$,
vorausgesetzt $B_1, ..., B_n$ partitionieren A

Übungsaufgabe 1

Nehmen wir Mensch-ärgere-dich-nicht; Sie haben alle Männchen draußen. Wie ist die Wahrscheinlichkeit, dass wir mit 3 Würfen mindestens eine 6 werfen, also ein Männchen ins Spiel bekommen?

Hier gibt es jetzt verschiedene Rechenwege!

Übungsaufgabe 2

Reinhold Messner muss einen steilen Eishang unter einem hängenden Gletscher queren. Die Wahrscheinlichkeit, dass sich während der Dauer seiner Querung von oberhalb eine Schneemasse löst und ihn in die Tiefe reißt, schätzt er auf 1/4. Die Wahrscheinlichkeit, dass er selbst (als erfahrener Eisgeher) bei der Querung ausgleitet und abstürtzt, schätzt er auf 1/20.

1. Wie schätzt er also seine Überlebenschancen für den Fall einer Querung ein? (Vorsicht: bilden Sie die richtigen Partitionen!)

2. Messner hat in seinem Leben 100mal einen vergleichbaren Eishang unter einem vergleichbaren Hängegletscher gequert und hat überlebt.

Gleichzeitig beträgt die Wahrscheinlichkeit, auf einer Himalaya-Expedition den Yeti zu sehen, nach Messners Einschätzung $0,000001 = \frac{1}{1.000.000}$. Was war wahrscheinlicher (immer nach Messners Einschätzung) – dass Messner seine 100 Eisquerungen überlebt oder dass er auf einer seiner 25 Himalaya-Expeditionen den Yeti sieht? Begründen Sie!

3. Wir als außenstehende sagen: die Wahrscheinlichkeit, dass Messners Einschätzung bezüglich der Yeti-Wahrscheinlichkeit stimmt, beträgt ebenfalls nur $\frac{1}{1.000.000}$, während es mit einer Wahrscheinlichkeit von $\frac{999.999}{1.000.000}$ den Yeti gar nicht gibt. Was ist also für uns die Wahrscheinlichkeit, dass Messner auf seinen 25 Expeditionen den Yeti wirklich gesehen hat?

Lösung

1. $(\frac{57}{80})$ (auf zwei Wegen: 1-Todeswahrscheinlichkeit, oder P(keine Lawine)·P(kein Ausgleiten))

2. Überlebenswahrscheinlichkeit: $(\frac{57}{100})^{100} \approx 1.89 \cdot 10^{-15}$. Yeti-Wahrscheinlichkeit (nach Messner): $1 - (\frac{999.999}{1.000.000})^{25} \approx 2.5 \cdot 10^{-5}$.

3. Yeti-Wahrscheinlichkeit (nach uns) ist Messners Yeti-Wahrscheinlichkeit mal $\frac{1}{1.000.000}$, also $2.5 \cdot 10^{-5} \cdot 10^{-6} = 2.5 \cdot 10^{-11}$. Also immer noch wahrscheinlicher als sein Überleben!

Hausaufgabe 1 - Beim Metzger

Abgabe bis zum 2.5.2017 *vor dem Seminar*, egal ob digital/analog und auf welchem Weg.

Nehmen Sie an, Sie haben Hackfleich vom Metzger im Kühlschrank, das noch gut aussieht, aber Sie wissen nicht mehr, wann Sie es gekauft haben. Der Metzger weiß es auch nicht mehr, aber sagt Ihnen, dass die Wahrscheinlichkeit, dass man von leicht verdorbenem Hackfleich (also solchem, das noch gut aussieht) Bauchweh kriegt, bei 1/3 liegt. Er sagt aber auch, dass Hackfleisch, das noch gut aussieht, allgemein nur in 1/100 aller Fälle (leicht) verdorben ist, und davon abgesehen auch der Verzehr von unverdorbenem Hackfleich in 1/50 aller Fälle zu Bauchweh führt. Sie lassen es sich also schmecken.

1. Wie groß ist die Wahrscheinlichkeit, dass Sie Bauchweh bekommen?

2. Nehmen Sie an, prompt nach dem Essen bekommen Sie Bauchschmerzen. – Wie hoch ist die Wahrscheinlichkeit, dass das Hackfleisch tatsächlich verdorben war?

Hausaufgabe 2 - Eine Krankheit

Abgabe bis zum 2.5.2017 *vor dem Seminar*, egal ob digital/analog und auf welchem Weg.

Es geht um eine Krankheit, die durchschnittlich einen von 100.000 Menschen trifft. Um die Krankheit zu diagnostizieren gibt es einen Test. Der Test liefert ein positives Resultat (sagt also aus, dass die Testperson die Krankheit hat) mit einer Wahrscheinlichkeit von 0,98, wenn die Testperson krank ist. Auch wenn die Testperson gesund ist, kommt es mit einer Wahrscheinlichkeit von 0,007 zu einem positiven Resultat.

Sie lassen diesen Test machen und das Ergebnis ist positiv. Wie groß ist die Wahrscheinlichkeit, dass Sie tatsächlich krank sind?

4 Zufallsvariablen

4.1 Definition

Erinnern Sie sich dass für eine Funktion f wir mit f^{-1} soviel meinen wie die Umkehrfunktion. Da die einfache Umkehrung einer Funktion nicht unbedingt eine Funktion ist (wegen fehlender Eindeutigkeit), ist die formale Definition wie folgt:

(19) $f^{-1}(a) = \{b : f(b) = a\}$

$f^{-1}(x)$ ist also immer eine Menge, und falls es kein b gibt, so dass $f(b) = a$, dann gilt

(20) $f^{-1}(a) = \emptyset$

Mit diesem Wissen und dem Wissen dass $\emptyset$ in jedem Wahrscheinlichkeitsraum enthalten ist, werden Sie die folgende Definition besser verstehen.

Sei $\mathcal{P} = (\Omega, \mathfrak{A}, P)$ ein Wahrscheinlichkeitsraum, und

$$X : \Omega \to \mathbb{R}$$

eine Funktion. X ist eine **Zufallsvariable** falls für alle $x \in \mathbb{R}$, $X^{-1}(x) \in \mathfrak{A}$. In einem *diskreten* Wahrscheinlichkeitsraum ist jede Funktion $X : \Omega \to \mathbb{R}$ eine Zufallsvariable. Das bedeutet:

$$P(X^{-1}(x)) \in [0, 1]$$

ist eine definierte Wahrscheinlichkeit; wir schreiben das oft einfach $P(X = x)$, und sagen: die Wahrscheinlichkeit dass X den Wert x annimmt.

NB: eine Zufallsvariable ist keine Variable, sondern eine Funktion; der irreführende Name wurde aus dem Englischen *random variable* rück-übersetzt. Der eigentliche Deutsche Begriff Zufallsgröße ist aber (meines Wissens) nicht mehr gebräuchlich.

Zufallsvariablen werden oft benutzt, um Wahrscheinlichkeitsräume zu vereinfachen. Nehmen wir das obige Beispiel mit den Verspätungen in Auto und Bahn: es können viele Dinge geschehen mit einer gewissen Wahrscheinlichkeit. Wir können nun eine Zufallsvariable definieren, die alle Ereignisse auf die Verspätung abbilden (=Zahl der Minuten), in der sie resultieren. $P(X = 30)$ wäre dann die Wahrscheinlichkeit, dass unser Besuch 30min Verspätung hat.

4.2 Erwartungswert

Zufallsvariablen wecken gewisse Erwartungen. Der **Erwartungswert** über einer Zufallsvariablen ist wie folgt definiert:

$$\mathbf{E}(X) := \sum_{x \in \mathbb{R}} x \cdot P(X^{-1}(\{x\})) \tag{21}$$

Statt $P(X^{-1}(\{x\}))$ schreibt man meist $P(X = x)$, d.h. die Wahrscheinlichkeit, mit der X den Wert $x \in \mathbb{R}$ zuweist. Wenn wir beispielsweise die Werte von X als Geldwerte auffassen, die wir in einem Spiel gewinnen (oder im Fall von negativen Zahlen verlieren), dann ist der Erwartungswert soviel wie der Geldbetrag, den wir in einem durchschnittlichen Spiel gewinnen/verlieren (gemittelt sowohl über den Betrag als auch die Wahrscheinlichkeit des Gewinns/Verlustes!).

Wenn wir eine diskrete Wahrscheinlichkeitsfunktion haben, also jedes Ergebnis (nicht Ereignis!) eine Wahrscheinlichkeit bekommt, dann gibt es eine wesentlich einfachere Definition:

$$\mathbf{E}(X) := \sum_{\omega \in \Omega} X(\omega) \cdot P(\omega) \tag{22}$$

4.3 Erwartungswerte bei Wetten – quantifizierte Gewißheit

Stellen Sie sich vor, Sie wetten um einen Euro über einen gewissen Sachverhalt, z.B.: Sie sagen, `Al Pacino war der Hauptdarsteller im Paten`, Ihr Gegenüber sagt: `es war Robert de Niro`. Nun sind Sie nicht zu 100% sicher, aber zu 90%. Wir haben:

- Einen diskreten Wahrscheinlichkeitsraum mit drei Ergebnissen, A(l),R(obert),K(einer der beiden).
- Ihre subjektive Wahrscheinlichkeitsverteilung $P(A) = 0.9$, $P(R) = 0.05$, $P(K) = 0.05$.
- Die Variable $X : \{A, R, K\} \to [0, 1]$, mit der Definition $X(A) = 1$, $X(R) = -1$, $X(K) = 0$.

In diesem Fall gilt: der Erwartungswert ist das, was Sie zu gewinnen erwarten, ihr erwarteter Gewinn bzw. Verlust. Hier sehen wir, dass Wahrscheinlichkeiten subjektiv sind, denn Ihr Gegenüber wird natürlich ganz andere Wahrscheinlichkeiten verteilen.

Man kann dasselbe Szenario noch ausbauen und Erwartungswerte nutzen, um **Gewißheiten zu quantifizieren**. Z.B. könnte Ihr gegenüber sagen: Die Wette, wie wir sie oben besprochen haben, gehe ich nicht ein. Aber wenn Du so sicher bist, dann können wir ja folgende Wette machen:

- Falls A stimmt, gebe ich Dir einen Euro;
- falls R stimmt, gibst Du mir 10;
- andernfalls bleiben wir bei 0.

Die Frage ist: ist das eine gute Wette? Wir haben

$$0.05 \cdot -10 + 0.05 \cdot 0 + 0.9 \cdot 1 = -0.5 + 0.9 = 0.4 \tag{23}$$

Die Antwort wäre also: ja, wir erwarten einen Gewinn von 0.4 Euro. Alternativ gäbe es folgende Wette:

- Falls A stimmt, gebe ich Dir einen Euro;
- falls A nicht stimmt, gibst Du mir 10;

In diesem Fall gilt:

$$0.1 \cdot -10 + 0.9 \cdot 1 = -0.1 \tag{24}$$

Die Wette wäre also nicht mehr gut für uns, denn unsere Sicherheit beträgt 1:9.

Was wir hier sehen ist lehrreich aus folgendem Grund: wir sind oft schlecht darin, unsere subjektive (Un)sicherheit in Zahlen auszudrücken. Mittels des Wettmodels sind wir aber in der Lage zu quantifizieren:

Meine (subjektive) Gewißheit, dass Eregnis A eintritt, ist $> n/m$ (wobei $n < m$, genau dann wenn ich mir bei einer Wette vom letzten Typ, bei einer Verteilung $X(A) = n$, $X(\overline{A}) = -(m - n)$, einen Gewinn erwarten (= positiver Erwartungswert).

4.4 Ein Beispiel: Erwartete Länge von Wörtern im Text

Nehmen wir an wir haben eine Sprache mit endlich vielen Wörtern (im Gegensatz zum Deutschen), also etwa das Englische. Nehmen wir ebenfalls an, wir kennen für jedes englische Wort w die Wahrscheinlichkeit, mit der w in irgendeinem zufälligen Text an einer zufälligen Stelle auftritt; wir haben also eine diskrete Wahrscheinlichkeitsverteilung, gegeben durch die diskrete Verteilung $P : \Sigma^* \to [0, 1]$. Was uns interessiert ist folgendes: wenn wir immer wieder einen zufälligen Text aufschlagen und ein zufälliges Wort heraussuchen, wie *lang* wird dieses Wort im Durchschnitt sein? Oder anders gesagt, wie lang ist das durchschnittliche englische Wort?

Um diese Frage zu beantworten, brauchen wir zunächst die Funktion $|-| : \Sigma^* \to \mathbb{N}$, wobei $|w|$ die Länge von w denotiert. Eine erste Antwort auf die Frage nach der erwarteten Länge eines durchschnittlichen englischen Wortes wäre wie folgt: denotieren wir das englische Lexikon mit L; erinnern Sie sich außerdem dass für Mengen (statt Ketten) $|-|$ die Kardinalität denotiert.

$$\frac{\sum_{w \in L} |w|}{|L|} \tag{25}$$

Wir summieren also alle Längen von den verschiedenen Wörtern auf, und Teilen sie durch die Anzahl der Wörter. Das gibt uns die durchschnittliche Länge der Worte in L, aber nicht die durchschnittliche Länge der Worte im Text, denn es beachtet nicht die unterschiedliche Wahrscheinlichkeit, mit der die einzelnen Worte im Text verteilt sind. Um die zu berücksichtigen, müssen wir $P(w)$ in unsere Formel einbauen:

$$\sum_{w \in L} |w| \cdot P(w) \tag{26}$$

Wir müssen in diesem Fall nicht mehr durch $|L|$ dividieren, da eine ähnliche Funktion bereits von $P(w)$ übernommen wird; denn $\sum_w \in LP(w) = 1$. Wie sie vielleicht schon erraten haben, ist $|--|$ eine Zufallsvariable, und die Formel in (26) ist nichts anderes als ihr Erwartungswert.

4.5 Würfeln - mal wieder

Nehmen wir an, wir werfen zwei faire Würfel. Das führt zu einem Produktraum zweier Laplace-Räume, der wiederum ein Laplaceraum ist. Wir definieren nun eine Zufallsvariable X auf unserem Produktraum durch

$$X(\langle x, y\rangle) = x + y.$$

D.h. z.B. $X(\langle 3, 4\rangle) = 7$, wobei $\langle 3, 4\rangle$ das Ergebnis "erster Wurf 3, zweiter Wurf 4" darstellt. X entspricht einem Spiel, indem nur die Summe der Würfel eine Rolle spielt.

Die Zufallsvariable X eröffnet uns jetzt einen neuen Wahrscheinlichkeitsraum, nämlich

$$(X[\Omega], \wp(X[\Omega]), P \circ X^{-1}).$$

$X[-]$ ist die punktweise Erweiterung von X auf eine Menge, z.B.

$$X[\{a, b\}] = \{X(a), X(b)\}.$$

D.h. also in unserem Beispielfall $X[\Omega] = \{1, 2, 3, ..., 12\}$. Mit $P \circ X^{-1}$ meinen wir die Komposition der beiden Funktionen, also

$$P \circ X^{-1}(x) = P(X^{-1}(x)).$$

Das sieht komplizierter aus als es ist. Was ist bespielsweise die Wahrscheinlichkeit von 2 in unserem neuen Raum? Nun, wir haben $X^{-1}(2) = \{\langle 1, 1\rangle)\}$, und $P(\langle 1, 1\rangle) = \frac{1}{36}$. Was ist die Wahrscheinlichkeit von 5? Intuitiv sollte die höher sein, denn es gibt ja einige Möglichkeiten mit zwei Würfeln 5 Augen zu werfen. Und tatsächlich haben wir

$$P \circ X^{-1}(5) = P(\{\langle 1, 4\rangle, \langle 2, 3\rangle, \langle 3, 2\rangle, \langle 4, 1\rangle\}) = 4 \cdot \tfrac{1}{36} = \tfrac{1}{9}.$$

Wir sehen also: der neue Wahrscheinlichkeitsraum ist kein Laplace-Raum!

Was ist der Erwartungswert? Auch diesmal können wir (zum Glück!) die einfachere Formel benutzen, da wir ja im alten Wahrscheinlichkeitsraum das kleine p haben - jedes atomare Ergebnis hat ja die Wahrscheinlichkeit $\frac{1}{36}$. Wir bekommen also:

$$\begin{aligned} \mathbf{E}(X) &= \sum_{\omega \in \Omega} X(\omega) \cdot p(\omega) \\ &= \frac{27 + 33 + 39 + 45 + 51 + 57}{36} \\ &= 7 \end{aligned} \tag{27}$$

Das bedeutet, wir erwarten im Schnitt mit einem Wurf 7 Augen zu bekommen.

4.6 Varianz und Standardabweichung

Man muss sich darüber klar sein, dass der Erwartungswert nicht zwangsläufig ein Wert sein muss, der überhaupt vorkommt (ebenso wie etwa der Durchschnittswert). Wenn wir eine faire Münze haben, $X(K) = 1, X(Z) = -1$, dann ist $\mathbf{E}(X) = 0$ – also kein Wert, der irgendeinem Ergebnis entspricht. Es gibt noch einen weiteren Punkt, der sehr wichtig ist. Der Erwartungswert gibt uns eine Art Mittelwert im Hinblick auf die Wahrscheinlichkeit. Wir wissen aber nicht, wie die Ergebnisse um den Erwartungswertverteilt sind: sie können sich zum Erwartungswert hin häufen (sie das Beispiel der zwei Würfel); sie können sich aber auch auf beiden Seiten des Erwartungswertes häufen: siehe das letzte Beispiel der Münze.

Der Erwartungswert ist also für gewisse wichtige Fragen nicht informativ. Hier brauchen wir das Konzept der **Varianz**. Die Definition der Varianz einer Zufallsvariable ist zunächst nicht sehr erhellend:

$$\mathbb{V}(X) = \mathsf{E}((X - \mathsf{E}(X))^2) \tag{28}$$

Was bedeutet diese Definition? Um sie zu verstehen, muss man zunächst wissen dass für zwei Zufallsvariablen X, Y, $X+Y$, definiert durch $X+Y(\omega) = X(\omega) + Y(\omega)$, und $X \cdot Y$ definiert durch $X \cdot Y(\omega) = X(\omega) \cdot Y(\omega)$, wiederum Zufallsvariablen sind. Also ist $X - \mathsf{E}(X)$ eine Zufallsvariable, und dann ebenso $(X - \mathsf{E}(X))^2$, und dementsprechend können wir wiederum deren Erwartungswert bestimmen. Die Zufallsvariable $X - \mathsf{E}(X)$ bildet ein Ergebnis ω auf die Differenz $X(\omega) - \mathsf{E}(X)$; es sagt uns also, wie weit ein Ergebnis von der Erwartung abweicht. Als nächstes wird dieses Ergebnis quadriert zu $(X(\omega) - \mathsf{E}(X))^2$, um alle Werte positiv zu machen (uns interessiert nur die Abweichung, nicht die Richtung der Abweichung). Wir haben also eine Zufallsvariable, die uns die Abweichung eines Ergebnisses vom Erwartungswert von X im Quadrat liefert. Die Varianz ist schließlich der Erwartungswert dieser Variable. In einem Satz, die Varianz ist die erwartete Abweichung der Zufallsvariablen von ihrem Erwartungswert im Quadrat.

Dementsprechend ist die **Standardabweichung** $\sigma(X)$ einer Zufallsvari-

able X die Wurzel der Varianz:

$$\sigma(X) = \sqrt{\mathtt{V}(X)} \tag{29}$$

Die Standardabweichwung gibt also die durchschnittliche Abweichung eines Ergebnisses (unter der Zufallsvariable) vom Erwartungswert. Es gibt ein sehr wichtiges Ergebnis für die Standardabweichung, mittels dessen seine Bedeutung sofort klar wird:

Für eine Zufallsvariable X mit Erwartungswert $\mathsf{E}(X)$ und Standardabweichung σ gilt immer: für alle $t \in \mathbb{R}$ und die Wahrscheinlichkeitsmasse R, die zwischen $\mathsf{E}(X) - t\sigma$ und $\mathsf{E}(X) + t\sigma$ liegt, also

$$R = P(X \in [\mathsf{E}(X) - t\sigma, \mathsf{E}(X) + t\sigma]) \tag{30}$$

Dann gilt:

$$R \geq 1 - \frac{1}{t^2} \tag{31}$$

Z.B. $t = 2$ bedeutet, dass $3/4$ der Wahrscheinlichkeitsmasse in $[\mathsf{E}(X) - 2\sigma, \mathsf{E}(X) + 2\sigma]$. Also 2 Standardabweichungen decken $3/4$ der Wahrscheinlichkeit ab usw.

Übungsaufgabe 3

Abgabe bis 16.5.2017 vor dem Seminar.

Im Spielcasino gilt folgendes: Spiele entsprechen Zufallsexperimenten, Wetten entsprechen Zufallsvariablen. Der Ausspruch "Die Bank gewinnt immer" soviel wie: für jedes Spiel hat die Bank einen positiven Erwartungswert. Nun gibt es aber folgende Strategie: ich setze beim Roulette 1euro auf rot. Wenn ich gewinne, fange ich die Strategie von vorne an. Wenn ich verliere, setze ich den doppelten Betrag auf rot, immer weiter, bis ich irgendwann gewinne. Wenn ich schlielich gewinne, war mein Einsatz immer höher als meine Verluste bisher, ich mache also unterm Strich Gewinn.

1. Wie passt das zusammen damit, dass die Bank immer gewinnt? Wie entwickelt sich der Erwartungswert im Lauf des Spiels?

2. Unter welchen Annahmen könnten Sie tatsächlich das Casino sprengen (d.h. beliebig Gewinn machen)?

5 Wichtige Wahrscheinlichkeitsverteilungen

Im letzten Beispiel war unser Wahrscheinlichkeitsraum der Raum zweier Würfe mit einem fairen Würfel. Wir haben gesehen dass die Zufallsvariable $X : \Omega \to \mathbb{R}$,

(32) $X(\langle i, j\rangle) = i + j$

aus einem Wahrscheinlichkeitsraum $\mathcal{P}_1 = (\Omega, \wp(\Omega), P)$ einen neuen Wahrscheinlichkeitsraum macht, nämlich den Raum

$$\mathcal{P}_2 = (X[\Omega], \wp(X[\Omega]), P \circ X^{-1})$$

Beide Räume sind diskret, aber der Raum $\mathcal{P}_1$ hat eine wichtige Eigenschaft, die $\mathcal{P}_2$ fehlt: er ist Laplace, d.h. alle Ergebnisse sind gleich wahrscheinlich. $\mathcal{P}_2$ ist natürlich nicht Laplace; dennoch sieht man ihm auf gewisse Weise an, dass er aus einem Laplace-Raum entstanden ist. Wir werden uns zunächst mit den sog. Binomialverteilungen beschäftigen.

Definition 4 *Binomialverteilungen sind Verteilungen, die aus einem (vielfachen) Produkt eines Bernoulli-Raums mit sich selbst konstruiert werden mittels einer additiven Zufallsvariable wie in (32).*

Danach werden wir uns den allgemeineren Multinomialverteilungen zuwenden, für die unser Würfelraum ein Beispiel liefert.

Wir haben gesagt dass Zufallsvariablen Funktionen in die rellen Zahlen sind. Eine wichtige Konsequenz ist, dass wir, gegeben einen Wahrscheinlichkeitsraum mit Wahrscheinlichkeitsfunktion P und eine Zufallsvariable X, eine Funktion

$$f_X : \mathbb{R} \to \mathbb{R}$$

bekommen, die definiert ist durch

(33) $f_X(x) = P(X = x) = P(X^{-1}(x))$

(letztere Gleichung qua unserer Konvention; verwechseln Sie nicht das große X und das kleine x!). Diese Funktion ist die **Wahrscheinlichkeitsverteilung** von X. NB: die Wahrscheinlichkeitsverteilung ist eindeutig definiert durch den Wahrscheinlichkeitsraum und die Zufallsvariable. Deswegen wird oft von

Wahrscheinlichkeitsfunktionen P gesprochen als wären sie eine Wahrscheinlichkeitsverteilungen, und umgekehrt. Das kann manchmal zu Verwirrung führen, denn es ist ja nicht gesagt dass die Ergebnisse in Ω reelle Zahlen sind, und daher kann man von keiner Verteilung für P selbst sprechen. Falls aber $\Omega \subseteq \mathbb{R}$, dann ist die Identitätsfunktion id, wobei

$$\text{f.a. } x \in \mathbb{R},\ id(x) = x,$$

eine Zufallsvariable. Und da

$$P \circ id = P \circ id^{-1} = P,$$

kann man auch von einer Wahrscheinlichkeitsverteilung von P sprechen.

Eine Wahrscheinlichkeitsverteilung f_X heißt **diskret**, wenn es nur endlich oder abzählbar unendlich viele $x \in \mathbb{R}$ gibt, so dass $f_X(x) \neq 0$ (erinnern Sie sich: falls $X^{-1}(x) = \emptyset$, dann ist $P(X^{-1}(x)) = 0$, also $f_X(x) = 0$.

5.1 n über k

Die Formel $\binom{n}{k}$ (sprich n über k) ist von zentraler Bedeutung für die Wahrscheinlichkeitstheorie und Statistik. Sie ist wie folgt definiert:

(34)
$$\binom{n}{k} = \frac{n}{1} \cdot \frac{n-1}{2} \cdot \ldots \cdot \frac{n-(k-1)}{k} = \frac{n \cdot (n-1) \cdot \ldots \cdot (n-k+1)}{k!} = \frac{n!}{k!(n-k)!}$$

Die letze Gleichung gilt nur unter der Voraussetzung dass n, k positive ganze Zahlen sind, und $n \geq k$. In unseren Beispielen wird diese Voraussetzung immer erfüllt sein. Die intuitive Bedeutung dieser Formel ist die folgende:

- nehmen wir an, wir haben eine Menge M, so dass $|M| = n$.
- $\binom{n}{k}$ ist die Anzahl von verschiedenen Mengen $N \subseteq M$, so dass $|N| = k$.

Warum brauchen wir diese Formel? Nehmen wir einen Raum, der das n-fache Produkt eines Wahrscheinlichkeitsraumes darstellt; etwa: ein n-facher Münzwurf. Wir möchten nun die Wahrscheinlichkeit des Ereignisses: k-mal Kopf. Dieses Ereignis umfasst alle Ergebnisse (Ergebnisse sind n-tupel), von

denen k-Komponenten Kopf sind. Wieviele Ereignisse sind das? Die Antwort ist $\binom{n}{k}$. Diese Formel ist also sehr wichtig um Wahrscheinlichkeiten von Ereignissen der Art zu berechnen: k von n Ergebnissen sind x (x irgendein Ergebnis), egal welche.

5.2 Binomiale Verteilungen

Zur Erinnerung: ein Bernoulli-Raum ist ein Wahrscheinlichkeitsraum mit $|\Omega| = 2$. Wir setzen kanonisch

1. $\Omega = \{0, 1\}$ (denn die Bezeichnung der Ereignisse ist natürlich willkürlich); außerdem
2. $p = P(1)$, $q = (1 - p)$

Nehmen wir Einfachheit halber an, dass $\mathcal{P}$ Bernoulli und Laplace ist, z.B. der Raum zum Wurf einer fairen Münze. Wir denotieren das Ereignis "Kopf" mit 0, "Zahl" mit 1. Da also unsere Ereignisse reelle Zahlen sind, nehmen wir kurzerhand die Zufallsvariable id, d.i. die Identitätsfunktion. Wir erweitern jetzt den Raum zu einem n-fachen Produktraum, d.h. zu dem Raum eines n-fachen Münzwurfes; und wir nehmen eine Zufallsvariable

$$X : \{0, 1\}^n \to \mathbb{R},$$

so dass

$$X(\langle \omega_1, .., \omega_n \rangle) = \sum_{i=1}^{n} \omega_i;$$

d.h. nichts anderes als dass uns X für irgendein Ergebnis sagt wie oft wir Zahl geworfen haben, unabhängig von der Reihenfolge der Ergebnisse.

Wir wissen bereits, wie wir die Wahrscheinlichkeit für das Ereignis ausrechnen, dass wir von den n Würfen k-mal Zahl werfen; beachten Sie, dass in der neuen Terminologie wir dieses Ereignis mit $X^{-1}(k)$ bezeichnen können!

$$X^{-1}(k) = \binom{n}{k} p^k q^{n-k} \tag{35}$$

(p ist die Wahrscheinlichkeit von Zahl, q die Wahrscheinlichkeit von Kopf.) Wenn wir nun die Wahrscheinlichkeitsverteilung haben wollen für das n-fache

Produkt des Bernoulli-Raumes und unserer Variable X, dann kriegen wir folgende Funktion $f : \mathbb{R} \to \mathbb{R}$:

$$(36) \quad f_X(x) = \begin{cases} \binom{n}{x} p^x q^{n-x}, \text{ falls } x \in \{0, 1, ..., n\} \\ 0 \text{ andernfalls} \end{cases}$$

Dies ist die Formel für die sogenannte **Binomialverteilung**, die wohl wichtigste diskrete Wahrscheinlichkeitsverteilung. Diese Verteilung ist symmetrisch genau dann wenn $p = 0.5$, $p = 1$ oder $p = 0$. In beiden letzten Fällen gibt die Funktion für alle Eingaben bis auf eine 0 aus, wie Sie leicht prüfen können. In allen anderen Fällen ist die Funktion asymmetrisch.

Die Binomialverteilung, wie wir sie geschrieben haben, ist eine Funktion, d.i. eine Abbildung von reellen Zahlen in die reellen Zahlen. Eigentlich handelt es sich aber um eine *Familie* von Funktionen, da wir für p und n uns nicht allgemein festlegen müssen (aber mit p auch q festlegen!). Die Funktion ändert sich aber je nach den Werten die p und q nehmen, daher sagt man p und q sind die **Parameter** der Funktion. Wir schreiben also die Familie der Binomialverteilungen als

$$(37) \quad \mathbf{B}(x|p, n) = \begin{cases} \binom{n}{x} p^x q^{n-x}, \text{ falls } x \in \{0, 1, ..., n\} \\ 0 \text{ andernfalls} \end{cases}$$

Hier können wir p, n entweder als zusätzliche Argumente der Funktion betrachten, oder als konkrete Instanziierungen für ein Element der Familie von Funktionen. Wichtig ist aber dass $0 \leq p \leq 1$, und $n \in \mathbb{N}$, sonst ist die Funktion (bis auf weiteres) nicht definiert. Wir haben folgende Konvention: wir sagen Binomialverteilung, wenn wir die ganze Familie von Funktionen meinen, und Binomialfunktion, wenn wir eine konkrete Funktion betrachten. Eine wichtige Eigenschaft der Binomialverteilung ist die folgende:

Lemma 5 *Für den Erwartungswert einer Binomialfunktion gilt immer* $\mathbf{E}(\boldsymbol{B}(x|p, n)) = pn$

Den Beweis lasse ich an dieser Stelle aus, da er an vielen Stellen nachgelesen werden kann. Ein berühmter und wichtiger Satz ist der Satz von Moivre-Laplace, der besagt dass für $n \to \infty$ (also für immer öfter Würfeln) die Binomialverteilung gegen die Gauss'sche Normalverteilung konvergiert.

5.3 Kategoriale Wahrscheinlichkeitsräume und Multinomiale Verteilungen

Die Generalisierung von $|\Omega| = 2$ auf $|\Omega| = n : n \in \mathbb{N}$, also von Bernoulli-Räumen auf beliebige endliche Räume, sind *kategoriale* Räume und Wahrscheinlichkeitsfunktionen. Ebenso wie Binomialverteilungen aus der Iteration von Bernoulli-Räumen entstehen (d.h. durch ein endliches Produkt eines Bernoulli Raumes $\mathcal{P}$ mit sich selbst, auch $\mathcal{P}^k$ geschrieben), entstehen **Multinomialverteilungen** durch ein endliches Produkt eines kategorialen Raumes mit sich selbst. Multinomialverteilungen sind komplizierter als Binomialverteilungen aus folgendem Grund: nehmen wir an, $|\Omega| = n$, und als Konvention

$$\Omega = \{0, 1, ..., n-1\}.$$

Wir notieren

$$P(i) = p_i.$$

Für die Multinomialverteilung ist nun jedes

$$p_i : 0 \leq i \leq n-1$$

ein Parameter. Auch die Kombinatorik dieser Räume ist wesentlich komplizierter, weswegen es (meines Wissens nach) keine geschlossene Formel für Multinomialfunktionen gibt. Im Grenzwert (für $n \to \infty$) konvergiert aber auch die Multinomialverteilung auf die Gauss'sche Normalverteilung. Das ist eine Folge des Zentralen Grenzwertsatzes, der wiederum eine Generalisierung des Satzes von Moivre-Laplace darstellt. Das bedeutet also: wenn wir mit n Würfeln spielen und die Verteilung für die Summe der Augen suchen, dann wird diese Verteilung immer ähnlicher der Normalverteilung, je größer n ist. Das zeigt Ihnen auch, wie außerordentlich wichtig die Normalverteilung ist für Stochastik und Statistik - auch wenn Sie sie noch nicht kennen.

5.4 Normal-Verteilungen und der Zentrale Grenzwertsatz

Wir werden Normalverteilungen nur sehr kurz anreißen, weil deren Funktion ziemlich kompliziert ist, und sie in der statistischen Sprachverarbeitung keine herausragende Rolle spielen. Wenn wir sie dennoch kurz besprechen,

liegt das an der herausragenden Rolle die sie in der gesamten Statistik spielen, und insbesondere ihrer Bedeutung für die beiden zuletzt besprochenen Binomial- und Multinomialverteilungen. Die Normalverteilung ist eine Familie von Funktionen mit zwei Parametern, dem **Mittelwert** μ und der **Standardabweichung** σ; deren Formel ist

$$(38) \quad f(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} = \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{1}{2}(\frac{x-\mu}{\sigma})^2\right)$$

Die Normalverteilung ist eine stetige Funktion über reelle Zahlen, im Gegensatz zu den anderen Verteilungen die wir hier betrachten. Ich werde diese Funktion hier nicht erklären, aber es ist wichtig zu wissen dass die Normalverteilung *die* statistische Verteilung schlechthin ist. Ihre Bedeutung versteht man vielleicht am besten aus dem **zentralen Grenzwertsatz**, den ich hier auch nur informell beschreibe: nehmen wir an, wir haben einen Wahrscheinlichkeitsraum, über dem wir n unabhängige, gleich verteilte Zufallsvariablen definieren können (z.B. n-mal Münze werden/würfeln etc., wobei jede Zufallsvariable X_i uns das Ergebnis des i-ten Wurfes liefert). Wir nennen wir diese Zufallsvariablen also

$$X_i : 1 \leq i \leq n.$$

Wir definieren nun eine neue Zufallsvariable

$$(39) \quad Y = X_1 + X_2 + ... + X_n$$

(erinnern Sie sich wie die Addition von Funktionen definiert ist: $f + g(x) := f(x) + g(x)$.

Der zentrale Grenzwertsatz Der zentrale Grenzwertsatz besagt: je größer n ist, desto stärker gleicht sich Y an die Normalverteilung an.

Das ist aus mindestens zwei Gründen wichtig: 1. die Binomialfunktion ist für große n gar nicht mehr berechenbar; wir können sie aber, je größer n, desto genauer mit der Normalverteilung approximieren. 2. Fehler in komplizierten Messungen oder Berechnungen, oder allgemeiner gesagt: Reihen von zufälligen Prozessen, verhalten sich genau so wie unsere Multinomialverteilungen; sie können also durch die Normalverteilung modelliert werden. Insbesondere bedeutet das: Reihen von Meßfehlern (etwa in der Physik, Astronomie) summieren sich *nicht* auf!

5.5 Potenzgesetze

Bisher haben wir von Verteilungen (realwertigen Funktionsgraphen) gesprochen, die von gewissen Wahrscheinlichkeitsräumen und Zufallsvariablen induziert werden. Wir können aber auch aus einer anderen Perspektive von Verteilungen sprechen: nämlich als der Verteilung von gewissen Daten, die wir tatsächlich in der Realität beobachtet haben. In diesem Fall kennen wir die Werte (natürlich nur endlich viele), aber wissen nichts über die zugrundeliegenden Wahrscheinlichkeiten. Die erste Perspektive ist die Perspektive der Stochastik, die letztere ist die Perspektive der Statistik. Die **Zipf-Verteilung** ist sehr wichtig für die Linguistik, weil wir sie sehr häufig beobachten; aus diesem Grund werden wir jetzt die statistische Perspektive einnehmen.

Nehmen wir an, wir haben einen Datensatz, der aus Paaren von reellen Zahlen besteht, oder sich daraufhin auflösen lässt. Paare von reellen Zahlen sind deswegen so wichtig, weil Funktionen *extensional* betrachtet nichts anderes sind als Paare von Zahlen

$$(x, f(x)).$$

Man nennt diese Zahlenpaare auch den **Graphen** von f.

Nehmen wir beispielsweise eine Menge von Wörtern, wie sie in einem Text vorkommen (z.B. *Die Wahlverwandtschaften*). Eine wichtige Unterscheidung, an die wir uns zunächst gewöhnen müssen, ist die von *type* und *token*. Als *type* bezeichnet man ein Wort als abstraktes Objekt (aber durchaus die konkrete Form, also nicht das Lemma/Lexem!). Als *token* bezeichnet man jedes Vorkommen dieses Objektes. Wenn ich also in einem Abschnitt zweimal das Wort isst finde, dann ist es derselbe *type*, aber zwei verschiedene *token*. Uns interessieren zunächst die *types*, die in dem Text vorkommen. Das sind keine Zahlenpaare; aber wir ordnen jedem Wort (*type*) ein Zahlenpaar zu: die erste Zahl gibt an, das wievielte Wort es ist in einer Liste, in der alle Worte (*types*) unseres Textes nach ihrer Häufigkeit (Anzahl der *tokens*) im Text geordnet sind, also etwa 1 wenn es das häufigste Wort ist. Die zweite Zahl gibt an, wie viele *token* von diesem Type es in unserem Text gibt. Die erste Zahl nennen wir den *Rang* des Wortes, die zweite die Häufigkeit. Wir haben also einen Datensatz

$$D \subseteq \mathbb{R} \times \mathbb{R}$$

aus Paaren von Zahlen (die Worte selbst kommen nicht mehr vor).

Wir nehmen nun an, diese Paare sind eine Teilmenge des Graphen einer Funktion; aber wir wissen natürlich nicht welche! Unsere Aufgabe ist es nun, eine Funktion zu finden, die gute Eigenschaften hat (z.B. einfach ist), aber dennoch unsere Daten gut *approximiert*. Potenzgesetze findet man dann, wenn es eine Polynomfunktion gibt, die unsere Daten beschreibt.

Wir sagen also der Datensatz D folgt einem **Potenzgesetz**, wenn es ein Polynom

$$a_1 \cdot x^b + a_2 \cdot x^{b-1} + ...$$

gibt, so dass für alle $(x, y) \in D$,

(40) $y \approx a_1 \cdot x^b + a_2 \cdot x^{b-1} + ... + a_b,$

wobei $\approx$ eine näherungsweise Gleichheit bedeutet. Wichtig für das Polynom ist dass b der größte Exponent ist; alle Terme bis auf $a_1 \cdot x^b$ werden dann weggelassen, und man schreibt:

(41) $y \propto a \cdot x^b,$

was bedeutet dass die beiden miteinander korrelieren. Der Datensatz, den wir betrachtet haben, folgt tatsächlich einem Potenzgesetz, und noch genauer gesagt einer Zipf-Verteilung.

5.6 Zipfs Gesetz

In unserem Fall ist klar, dass Rang und Häufigkeit miteinander invers korrelieren: je niedriger der Rang eines Wortes ist, desto größer ist seine Häufigkeit, denn 1 ist der Rang des häufigsten Wortes etc. **Zipfs Gesetz** ist eigentlich kein Gesetz, sondern eine empirische Beobachtung, die aber durch ihre Regelmäßigkeit fast den Status eines Gesetzes hat; denn sie bestätigt sich für alle Arten von Texten. Wir kürzen den Rang eines Wortes mit $r(w)$ ab; seine Häufigkeit bezeichnen wir mit $f(w)$ (das f kommt von Frequenz). Zipfs Gesetz ist ein Potenzgesetz, und in seiner einfachsten Fassung besagt es:

(42) $f(w) \propto \frac{1}{r(w)}$

Das ist ein Potenzgesetz, da $\frac{1}{x} = x^{-1}$. Was besagt diese Formel? Beachten Sie dass durch das Zeichen $\propto$ wir einen weiteren Term weglassen können; aber dieser Term darf keinen Exponenten haben. Was die Formel also besagt ist: es gibt eine Zahl k, so dass

$$(43) \quad f(w) \approx a_0(r(w))^{-1} + a_1 = a_0\frac{1}{r(w)} + a_1$$

Durch einfache Termumformung erfahren wir, dass es a_0, a_1 gibt, so dass

$$(44) \quad f(w) \cdot r(w) \approx a_0 + a_1 r(w)$$

für alle Worte w, die in unserem Korpus vorkommen. Wenn wir das ganze etwas vereinfachen und $a_1 = 0$ setzen (d.h. wir gehen von $\approx$ zu $\propto$), sehen wir dass

$$(45) \quad f(w) \cdot r(w) \propto a_0,$$

d.h. Rang und Frequenz eines Wortes sind genau **invers proportional** zueinander. Z.B. werden wir das 10-häufigste Wort in etwa doppelt so oft finden wie das 20-häufigste Wort, und 10 mal häufiger als das 100-häufigste Wort. Das häufigste Wort wird sogar 100 mal häufiger sein als das 100-häufigste, etc.

Die Bedeutung von Zipfs Gesetz für die Computerlinguistik ist immens. Um zu sehen warum, betrachten wir ein Beispiel: nehmen wir an in unserem Text kommen 10000 Wörter (*types*) vor. Das häufigste Wort kommt 2000mal vor. Das bedeutet dann, dass das 2000-häufigste Wort nur einmal vorkommen sollte - und das wiederum heißt dass 8000 (von 10000!) Wörtern (*types*) überhaupt nur einmal vorkommen! Diese Wörter nennt man auch *hapax legomena* ("einmal gelesen"), und diese machen in den meisten Texten tatsächlich die große Mehrheit der *types* aus. Umgekehrt können wir daraus schließen, dass wenn wir die 100 häufigsten Wörter (*types*) abgedeckt haben, wir bereits den größten Teil der tokens abgedeckt haben!

Es gibt also zwei wichtige Konsequenzen für die Computerlinguistik: wenn wir beispielsweise ein Lexikon zur Worterkennung oder Übersetzung schreiben möchten, dann bekommen wir schon relativ gute Abdeckungen wenn wir nur die häufigsten Wörter abdecken. Wenn wir aber umgekehrt mit statistischen Methoden Informationen über Wörter erfahren wollen, z.B. in welchen Kontexten sie vorkommen können, dann haben wir für die allermeisten Wörter

ein Problem, denn *fast alle* Wörter sind selten, und wenn ein Wort selten vorkommt, dann ist es schwierig mit statistischen Mitteln etwas zuverlässiges darüber zu erfahren.

5.7 Zipfs Gesetz und Wortlänge

Wir haben bereits die Funktion $|-| : \Sigma^* \to \mathbb{N}$ besprochen, die einem Wort seine Länge zuweist. Zipf hat ebenfalls beobachtet, dass es eine inverse Korrelation gibt von Wortlänge zu Worthäufigkeit. Wir haben also

$$f(w) \propto \frac{1}{|w|}. \tag{46}$$

Anders gesagt, je länger ein Wort, desto seltener ist es, und ein Wort mit Länge 5 sollte etwa 3-mal häufiger sein als ein Wort mit Länge 15 (sehr grob gesprochen). Zipf maßseinen Beobachtungen eine sehr große Bedeutung bei, und er führte sie alle zurück auf das Prinzip der kleinsten Anstrengung, die wir im Hinblick auf ein Ziel hin aufbringen möchten (*principle of least effort*), welches er allem menschlichen handeln zugrunde legte. Während seine Beobachtungen allgemein anerkannt sind, sind seine Hypothesen über die Ursachen der Beobachtungen weitestgehend zurückgewiesen worden.

Tatsächlich gibt es gute Argumente gegen seine Hypothesen. Erinnern Sie sich an die zufälligen Texte, von denen wir gesprochen haben. Ein solcher Text ist eine Zeichenkette über

$$(\Sigma \cup \square)^*,$$

wobei $\square$ für das Leerzeichen steht. In diesem Text hat jeder Buchstabe (und das Leerzeichen) eine Wahrscheinlichkeit, und diese ist vollkommen unabhängig von der Umgebung, in der er steht. Wir haben also beispielsweise

$$P(a) = 0.1, P(b) = 0.2, \ldots, P(\square) = 0,05$$

Ein Wort in diesem Text ist eine maximale Teilkette, die kein $\square$ enthält; d.h. eine Teilkette, die kein $\square$ enthält, aber links und rechts von $\square$ begrenzt ist.

Nehmen wir also an, wir generieren einen rein zufälligen Text nach unseren Wahrscheinlichkeiten. Eine merkwürdige Tatsache ist nun, dass wir auch in diesem rein zufälligen Text eine Zipf-Verteilung finden werden! D.h.

$$f(w) \propto \frac{1}{|w|} \tag{47}$$

gilt auch für die rein zufälligen Worte in unserem rein zufälligen Text. Diese Verteilung scheint also weniger durch besondere Eigenschaften natürlicher Sprache bedingt, sondern eine Folge allgemeinerer mathematischer Regelmäßigkeiten. Aber welche sind das? Nun, wir haben bereits einmal ausgerechnet, wie man die Wahrscheinlichkeiten von Worten in einem solchen Zufallstext berechnet. Die Wahrscheinlichkeit, dass wir irgendein Wort mit k Buchstaben treffen ist

$$P(\square)^2(1-P(\square))^k \tag{48}$$

Es ist klar, dass diese Zahl kleiner wird, je größer k wird. Daraus folgt, dass die Wahrscheinlichkeit von Worten immer weiter abnimmt, je länger sie werden, ganz unabhängig von den einzelnen Buchstaben aus denen sie bestehen und deren Wahrscheinlichkeiten. Wir haben also notwendig eine inverse Korrelation von Länge und Wahrscheinlichkeit, und mit einiger Mühe lässt sich zeigen, dass das eine Zipf-Verteilung ergibt.

5.8 Anmerkungen zu Zipf

Zipf-Verteilungen sind nicht nur in der Sprache allgegenwärtig. Sie gelten z.B. auch für Städte (Rang nach größe und Einwohnerzahl), Einkommensverhältnisse (zumindest in Italien, siehe Pareto-Verteilung) und viele andere Dinge.

6 Hypothesen prüfen

6.1 Verteilungen und Vertrauensgrenzen in R

R ist eine mächtige Programmiersprache, die für statistische Analysen ausgelegt ist. Dementsprechend sind bereits viele wichtige Funktionen eingebaut und müssen nicht erst mühsam definiert werden. Das umfasst z.B. die Funktion $\binom{n}{k}$, die geschrieben wird mit `choose(n,k)`:

```
> n <- 10
> k <- 6
> choose(n,k)
[1] 210
```

Das erlaubt uns beispielsweise, die Bimialverteilung zu definieren:

```
> bin.vert <- function(k, n, p) {
choose(n,k) * p^ k * (1-p)^ (n-k)
}
```

Das liefert uns z.B.

```
> bin.vert(40,150,0.75)
[1] 2.631372e-35
```

wobei $e-35$ soviel bedeutet wie *mal* 10^{-35}, d.h. wir müssen das Komma um 35 Stellen nach links verschieben, um den richtigen Wert zu bekommen. Die Binomialverteilung ist übrigens auch schon eingebaut in R, wir hätten uns die Arbeit also auch sparen können; sie wird abgerufen als `dbinom(k,n,p)`.

Wir werden jetzt einen einfachen Fall von statistischer Inferenz betrachten. Es folgt aus den grundlegenden Eigenschaften der Binomialverteilung und des Erwartungswertes, dass

$$(49)\quad argmax_{p\in[0,1]}\texttt{dbinom}(\texttt{k},\texttt{n},\texttt{p}) = \frac{k}{n}$$

D.h. für gegebene k, n nimmt die Funktion ihr Maximum in $\frac{k}{n}$. Umgekehrt gilt natürlich auch folgendes:

$$(50)\quad argmax_{0\leq i\leq n}\texttt{dbinom}(\texttt{i},\texttt{n},\frac{\texttt{k}}{\texttt{n}}) = k$$

D.h. für eine Gegebene Wahrscheinlichkeit $\frac{k}{n}$ und gegebene Anzahl von Iterierungen n nimmt die Funktion ihr Maximum für $i = k$ (erster Parameter). Nun kann man aber folgendes beobachten: je größer ich n, k wähle (bei gleichbleibendem $\frac{n}{k}$), desto kleiner wird dieses Maximum:

```
dbinom(4,10,(4/10))
[1] 0.2508227
```

```
dbinom(40,100,(4/10))
[1] 0.08121914
```

```
dbinom(400,1000,(4/10))
[1] 0.02574482
```

Der Grund hierfür ist ganz einfach: wir haben eine diskrete Funktion (nur endlich viele Werte > 0), die sich insgesamt auf 1 summieren, und je größer wir n, k wählen, desto mehr Werte sind > 0, während ihre Gesamtsumme gleich bleibt, d.h.

$$(51) \quad \sum_{1 \leq k \leq n} \texttt{dbinom(k,n,p)} = 1).$$

Also müssen die Werte kleiner werden (man sagt: die Wahrscheinlichkeitsmasse wird aufgeteilt unter diesen Werten). Das bedeutet aber auch: je öfter wir ein Experiment iterieren, desto unwahrscheinlicher wird das wahrscheinlichste Ergebnis, und je öfter wir ein Bernoulli Experiment wiederholen (mit Anzahl n), desto unwahrscheinlicher wird es, dass wir tatsächlich den "wahren" Parameter $\frac{n}{k}$ treffen, d.h. k-mal Ergebniss 1 haben. Das widerspricht zunächst unserer Intuition, da wir denken: je öfter wir ein Experiment iterieren, desto mehr nähern sich die Ergebnisse der "wahren" Wahrscheinlichkeit an.

Dieses Problem ist kompliziert und löst sich im "Gesetz der großen Zahlen" auf. Wir umgehen das erstmal ganz praktisch, indem wir anstelle einzelner Werte die sog. *Vertrauensgrenzen* oder *Konfidenzintervalle* benutzen. Intervalle werden in R mittels Vektoren gehandhabt:

```
> 0:5
[1] 0 1 2 3 4 5
> x <- 0:6
> x[3:5]
[1] 2 3 4
> sum(x)
[1] 21
```

Die letzte Zeile ist die Summe $1 + 2 + \ldots + 6$. Wir definieren jetzt eine Wahrscheinlichkeitsfunktion, die Intervalle berechnet:

```
> p <- 1/2
> n <- 40
>int <- dbinom(0:n,n,p)
```

Diese Funktion berechnet eine Liste `dbinom(0,40,1/2),dbinom(1,40,1/2), dbinom(2,40,1/2)` etc. Hierbei gibt es zu beachten dass `dbinom(0,40,1/2)=int[1]`, `dbinom(40,40,1/2)=int[41]`! Das wahrscheinlichste Ergebnis für k ist – nach allem was wir wissen –

```
> int[21]
[1] 0.1253707
```

Das ist relativ niedrig. Was wir aber jetzt machen können ist auf Intervalle von Ergebnissen zugreifen:

```
> int[19:23]
[1] 0.1031187 0.1194007 0.1253707 0.1194007 0.1031187
```

Was wir sehen ist folgendes: 1. die Verteilung ist symmetrisch (denn $p = 0.5$), 2. sie hat ihr Maximum bei $k = 20$ (entspricht `int[21]`!) Es gibt aber noch dritte wichtige Beobachtung:

```
> sum(int[19:23])
[1] 0.5704095
```

D.h.: wenn wir die Werte für die Ergebnisse $k = 18$ bis $k = 22$, also die 5 wahrscheinlichsten Werte addieren, dann entfällt auf diese Werte bereits *die Hälfte* der Wahrscheinlichkeitsmasse! Wir werden diese Prozedur jetzt leicht generalisieren. Dazu müssen wir noch wissen, dass für einen Vector wie `vec<- 1:n` wir den k-ten Wert mit `vec[k]<- i` ändern können.

```
> mittel <- 21
> interval <- 1:20
> for (i in 1:20) { indices <- seq(mittel-i, mittel+i) ; interval[i]
<- sum(int[indices]) }
```

Was wir hier bekommen ist folgendes: `interval[4]` ist `sum(int[21-4:21+4])`, also die Summe der 9 wahrscheinlichsten Ergebnisse.

```
>interval[5]
[1] 0.9193095
```

Diese machen also bereits 90% der Wahrscheinlichkeitsmasse aus! Damit

wir diese Zahlen etwas anschaulicher machen, setzen wir sie in eine Tabelle.

```
> vertrauen <- data.frame(grenze = rep(1:20), wahrscheinlichkeit
= interval)
> vertrauen[1:6,1:2]
    grenze  wahrscheinlichkeit
 1  1       0.3641720
 2  2       0.5704095
 3  3       0.7318127
 4  4       0.8461401
 5  5       0.9193095
 6  6       0.9615227
```

Hier sehen wir ein fundamentales Prinzip der Statistik, das eigentlich willkürlich ist: man legt normalerweise die **Vertrauensgrenze** bei 95% fest. Das heißt: wenn wir p als Parameter eines Bernoulli-Raumes nicht kennen, nehmen wir erstmal an dass $p = 0.5$ (das ist die sog. uniforme Verteilung, die unseren Mangel an Wissen widerspiegelt). Man nennt das auch die **Nullhypothese**. Wir nehmen nun also diesen Parameter als gegeben an. Dann zeigt uns unsere Funktion `int` dass unser Ergebnis k aus 40 Iterierungen des Experiments mit einer Wahrscheinlichkeit von über 0.95 im Interval [21-6,21+6] liegen muss. Wenn das Ergebnis darin liegt, dann finden wir die Nullhypothese noch akzeptabel, wenn das Ergebnis außerhalb der Vetrauensgrenzen liegt, dann weisen wir die Nullhypothese zurück: sie ist zu unplausibel. Unsere Vertrauensgrenze liegt also bei einer Abweichung von 6 vom Erwartungswert; wenn unser Ergebnis innerhalb der Grenzen liegt, haben wir nichts gewonnen; wenn es außerhalb liegt, lehnen wir die Nullhypothese ab. Wir stellen das ganze nun grafisch dar:

```
> plot(vertrauen$grenze, vertrauen$wahrscheinlichkeit, type="b",
xlab="Grenze", ylab="Wahrscheinlichkeit")
> segments(0,0.95,5.7,0.95)
> segments(5.7,0,5.7,0.95)
```

Wir sehen also wie mit wachsender Größe des Intervalls die Wahrscheinlichkeitsmasse steil wächst und letztlich langsam gegen 1 konvergiert.

Hier kann man nun auch sehen, wie sich unsere vorige Paradoxie auflöst. Beim jetztigen Beispiel liegen die Vertrauensgrenzen bei einer Abweichung

von 6 vom Mittelwert bei einer maximal möglichen Abweichung von 20. Wir rechnen nun dasselbe Beispiel nochmal mit $n = 400$ durch.

```
> n = 400
> sum(int2[(201-60):(201+60)])
[1] 1
```

Wir haben – proportional gesehen, die Grenzen genauso gesetzt wie vorher, diesmal bei 60 von 200. Wir sehen aber, dass der Wert schon so nahe an 1 ist, dass R ihn nicht mehr unterscheidet. Dass heißt bei einer Iterierung von $n = 400$ eine Proportional Abweichung von 3/10 um ein vielfaches unwahrscheinlicher ist! In diesem Sinne gibt uns eine häufigere Iteration ein besseres Abbild der tatsächlichen Wahrscheinlichkeit.

Der Sinn der Vetrauensgrenzen ist eigentlich folgender: wir nehmen eine zugrunde liegende (normale) Verteilung als **Nullhypothese** an; falls unser tatsächlich beobachtetes Ergebnis außerhalb dieser Grenzen liegt, weisen wir die Nullhypothese zurück. Wir haben also ein einfaches Mittel, eine Hypothese zurückzuweisen.

6.2 Der Bayesianische Ansatz

Nehmen wir an, wir haben eine Münze. Wir werfen sie 10 mal, und wir erhalten 5-mal Kopf, 5-mal Zahl; nennen wir dieses Ereignis $\mathcal{E}$. Was ist die Wahrscheinlichkeit, dass das passiert? Vorsicht: wir können das natürlich nicht wir; wir kennen ja gar nicht die Wahrscheinlichkeit, mit der die Münze Kopf/Zahl gibt; insbesondere wissen wir gar nicht, ob die Münze fair ist oder nicht. Diese Situation ist im "wirklichen Leben" wesentlich häufiger als die dass wir die Wahrscheinlichkeitsverteilung kennen. Was wir in dieser Situation meistens wollen ist folgendes: wir würden gerne wissen wie wahrscheinlich eine gewisse *Wahrscheinlichkeitsverteilung* ist, gegeben das Ergebnis dass wir beobachtet haben. Also in unserem Fall: gegeben dass wir aus 10 Würfen 5-mal Kopf haben, wie wahrscheinlich ist es, dass die Münze fair ist?

Das ist nicht ganz einfach, und ohne weitere Annahmen sogar unmöglich. Wir nähern uns dem Problem zunächst wie folgt. Nimm an wir haben zwei Verteilungen, die eine gegeben durch

$$p_1(K) = p_1(Z) = 0.5$$

, die andere durch

$$p_2(K) = 0.4,\ p_2(Z) = (0.6)$$

(K ist das Ereignis Kopf, Z ist Zahl). Nenne die erste Verteilung F (wie fair), die zweite U (wie unfair). Wir können nun sehr einfach die Wahrscheinlichkeit von $\mathcal{E}$ gegeben die Verteilung F ausrechnen:

$$P(\mathcal{E}|F) = \binom{10}{5} 0.5^5 \cdot 0.5^5 \approx 0.246 \tag{52}$$

Ebenso leicht lässt sich die Wahrscheinlichkeit von $\mathcal{E}$ gegeben die Verteilung U ausrechnen:

$$P(\mathcal{E}|U) = \binom{10}{5} 0.6^5 \cdot 0.4^5 \approx 0.201 \tag{53}$$

Beachten Sie dass an diesem Punkt Wahrscheinlichkeitsverteilungen selbst Ereignisse geworden sind! Was wir möchten sind nun die Wahrscheinlichkeiten $P(F|\mathcal{E})$ und $P(U|\mathcal{E})$, also die Wahrscheinlichkeiten der Wahrscheinlichkeitsverteilungen *gegeben* unser Würfelergebnis. Uns allen ist klar, dass wir hier mit dem Satz von Bayes arbeiten müssen.

Nun kommt allerdings der Punkt wo wir einige Annahmen machen müssen, die etwas willkürlich, aber in der einen oder anderen Form unvermeidbar sind. Die erste Annahme ist folgende: wir nehmen an, dass entweder U oder U der Fall ist, d.h. $P(U) + P(F) = 1$. Das ist natürlich willkürlich, denn es gibt noch (unendlich) viele andere denkbare Wahrscheinlichkeitsverteilungen für unsere Münze. Allerdings müssen wir die Möglichkeiten in irgendeiner Form einschränken, um an dieser Stelle weiter zu kommen. Die zweite Annahme die wir machen müssen ist: wir müssen $P(U)$ und $P(U)$ bestimmte Werte zuweisen. Der Grund ist folgender. Wir haben

$$P(\mathcal{E}|F) = 0.246, P(\mathcal{E}|U) = 0.201 \tag{54}$$

Wir suchen jetzt $P(F|\mathcal{E})$. Nach Bayes Theorem gilt

$$P(F|\mathcal{E}) = P(\mathcal{E}|F) \cdot \frac{P(F)}{P(\mathcal{E})} = 0.246 \cdot \frac{P(F)}{P(\mathcal{E})} \tag{55}$$

Wir sehen jetzt: wir kommen nicht weiter ohne $P(F)$. Nehmen wir also einfach an, dass $P(F) = P(U) = 0.5$. Wir brauchen aber noch die Wahrscheinlichkeit $P(\mathcal{E})$; allerdings kennen wir nur $P(\mathcal{E}|F)$ und $P(\mathcal{E}|U)$! Hier rettet uns die Annahme, dass $P(F \cup U) = 1$ (es gibt keine dritte mögliche Wahrscheinlichkeitsverteilung), und die Tatsache dass $F \cap U = \emptyset$; das bedeutet $\{F, U\}$ ist eine Partition der Ergebnismenge! Also gilt:

$$\begin{aligned} P(\mathcal{E}) &= P((\mathcal{E} \cap F) \cup (\mathcal{E} \cap U) \\ &= P(\mathcal{E} \cap F) + P(\mathcal{E} \cap U) \\ &= P(\mathcal{E}|F)P(F) + P(\mathcal{E}|U)P(U) \\ &= 0.246 \cdot (0.5) + 0.201 \cdot (0.5) \\ &= 0.2235 \end{aligned} \tag{56}$$

Da wir nun die Wahrscheinlichkeit $P(\mathcal{E})$ haben, können wir auch $P(F|\mathcal{E})$ und $P(U|\mathcal{E})$ ausrechnen:

$$P(F|\mathcal{E}) = P(\mathcal{E}|F) \cdot \frac{P(F)}{P(\mathcal{E})} = 0.246 \cdot \frac{0.5}{0.2235} \approx 0.55 \tag{57}$$

Daraus folgt dass

$$P(U|\mathcal{E}) \approx 1 - 0.55 = 0.45$$

(wir können das natürlich auch einfach nachprüfen, indem wir in der letzte Gleichung U statt F verwenden.)

Das ist ein einfaches Beispiel von sogenannter **Bayesianischer Statistik**. Bayesianische Statistik ist sehr elegant und liefert uns genau die Informationen die wir suchen. Es gibt allerdings einige Probleme: das größte sind die beiden Annahmen, die wir auf dem Weg machen mussten (NB: in komplexeren Beispielen ist es noch viel schwieriger, plausible Annahmen, die sogenannten *priors*, die *a priori* Wahrscheinlichkeiten zu finden; und selbst in unserem sehr einfachen Beispiel wird man kaum sagen können dass unsere Annahmen sehr plausibel waren). Wir haben z.B. angenommen dass

$$P(F) = P(U) = 0.5.$$

Das Problem ist: wenn wir etwas anderes angenommen hätten, z.B.

$$P(F) = 0.8,\ P(U) = 0.2,$$

dann hätten sich auch unsere *a posteriori* Wahrscheinlichkeiten für $P(F|\mathcal{E})$ und $P(U|\mathcal{E})$ geändert! Und wenn wir statt der zwei Wahrscheinlichkeitsverteilungen noch eine dritte zugelassen hätten, etwa

$$p_3(K) = 0.3,\ p_3(Z) = 0.7,$$

dann hätte auch das unser Ergebnis radikal verändert (das können Sie selbst nachprüfen); die Rechnung bleibt essentiell dieselbe, nur einige der Faktoren ändern sich. Daneben gibt es noch eine Reihe technischer Probleme, die in kompizierteren Beispielen entstehen (insbesondere bei stetigen Wahrscheinlichkeitsverteilungen).

Noch zwei Anmerkungen sollte ich machen: 1. Trotz dieser Probleme ist Bayesianische Statistik wesentlich informativer als alle klassische Statistik: denn wir haben die Wahrscheinlichkeit von Hypothesen (d.h. Wahrscheinlichkeitsverteilungen) gegeben eine Menge von Daten, und das ist mehr als uns die klassische Statistik liefern wird.

2. Für den Bayesianer sind Wahrscheinlichkeiten keine Grenzwerte von relativen Häufigkeiten (der sog. Frequentismus), sondern sie quantifizieren Glauben. D.h. eine Wahrscheinlichkeit von 1 heißt: ich bin vollkommen überzeugt, nichts wird mich von meinem Glauben abbringen; eine Wahrscheinlichkeit von 0 bedeutet: nichts wird mich davon überzeugen das etwas wahr ist. Man sieht das auch am obigen Beispiel: wenn ich die *a priori* Wahrscheinlichkeit auf $P(U) = 0$ setze, dann wird die *a posteriori* Wahrscheinlichkeit $P(U|\mathcal{E})$ ebenfalls immer 0 sein.

6.3 Sequentielle Überprüfung von Hypothesen 1

Eigentlich ist dieser Abschnitt eher eine Fußnote, er ist aber wichtig um dem Mißbrauch der präsentierten Methoden vorzubeugen, und um ein besseres Verständnis für ihre Natur zu bekommen. Das Problem ist folgendes: nehmen wir an, wir machen ein Experiment (100 Münzwürfe), allerdings ist das Ergebnis nach unserer Auffassung nicht konklusiv – es erlaubt keine definitive Schlussfolgerung darüber, ob die Münze fair ist oder nicht. Also wiederholen wir das Experiment, und prüfen das Ergebnis usw. Irgendwann haben wir dann ein zufriedenstellendes Ergebnis erreicht. Diese Herangehensweise ist leider gängige Praxis, stellt aber in den meisten Fällen einen groben Mißbrauch dar.

Betrachten wir die Methode der Vertrauensgrenzen, und nehmen wir das Procedere sieht in der Praxis wie folgt aus:

1. Wir machen das Experiment (100 Würfe), schauen ob das Ergebnis innerhalb unser Vertrauensgrenzen liegt – und die Antwort ist positiv.

2. Aus irgendeinem Grund – vielleicht liegt es am Rande, vielleicht haben wir einen Verdacht – befriedigt uns das nicht.

3. Wir wiederholen das Experiment (100 Würfe), und schauen ob das Gesamtergebnis ($n \cdot 100$ Würfe) innerhalb der Vertrauensgrenzen liegt.

4. Nach n Durchgängen der vorigen Punkte 2. und 3. (z.B. $n = 5$) liegt das Ergebnis außerhalb der Vertrauensgrenzen. "Aha", sagen wir, "wußt ich es doch. Zum Glück habe ich nicht aufgegeben!"

Wo liegt der Fehler in dieser Vorgehensweise?

Schauen wir uns zwei Ereignisse an:

> E_{500}: Das Ergebnis liegt nach 500 Würfen innerhalb der Vertrauensgrenzen.

Dem gegenüber steht ein anderes Ereignis:

> E^5_{100}: Das Ergebnis liegt sowohl nach 100,200,300,400 also auch nach 500 Würfen innerhalb der Vetrauensgrenzen.

Haben wir

$$P(E_{500}) = P(E_{100}^5)?$$

Diese Frage ist leicht zu beantworten, denn unsere Ereignisse konsitutieren sich als Mengen von Ergebnissen (Folgen in $\{0,1\}^{500}$). Nun ist es leicht zu sehen, dass

$$E_{100}^5 \subseteq E_{500}$$

nach der Definition der beiden Ereignisse. Nun nehmen wir aber folgendes Ergebnis:

$e := \langle 0^{250}, 1^{250} \rangle$, d.h. erst 250mal Kopf, dann 250mal Zahl.

Wir haben natürlich $e \in E_{500}$ – denn das Ergebnis liegt genau am Erwartungswert. Wir haben aber $e \notin E_{100}^5$, denn nach den ersten 200 Würfen (alle Kopf!) liegt unser Ergebnis mit Sicherheit außerhalb jeglicher Vertrauensgrenze. Daraus folgt:

$$E_{100}^5 \subsetneq E_{500}$$

und folgereichtig, da $P(e) > 0$,

$$P(E_{100}^5) < P(E_{500}).$$

Aber wie stark ist dieser Effekt? Unser Beispiel e ist derart unwahrscheinlich, dass wir es vernachlässigen können. Das Problem ist aber, dass die Vertrauensgrenzen relativ mit wachsender Zahl von Ergebnissen relativ immer enger werden, es werden also immer mehr Ergebnisse ausgeschlossen!

Betrachten wir einmal genauer passiert: die Eregnisse

$$E_{100},\ E_{100}^2,\ E_{200}$$

sind analog zu den obigen definiert. Für E_{100}^2 (2 Iterationen) können wir die folgende Rechnung aufmachen:

$$P(E_{100}^2) = P(E_{100} \cap E_{200}) = P(E_{200}|E_{100})P(E_{100}) \tag{58}$$

Wir können diese Rechnung leicht verallgemeinern; es handelt sich nämlich um eine sogenannte **Markov-Kette**: in einem Satz bedeutet das:

$$P(E_{300}|E_{200}, E_{100}) = P(E_{300}|E_{200})$$

etc., also zählt immer nur das letzte Ergebnis. Also gilt:

$$(59)\quad P(E_{100}^5) = P(E_{500}|E_{400})P(E_{400}|E_{300})P(E_{300}|E_{200})P(E_{200}|E_{100})P(E_{100})$$

oder etwas allgemeiner ausgedrückt:

$$(60)\quad P(E_{100}^{n+1}) = P(E_{100(n+1)}|E_{100n})P(E_{100}^n)$$

Damit ist also klar sichtbar, dass wir Wahrscheinlichkeit mit sinkendem n immer kleiner wird. Im Gegensatz dazu vergleiche das per Definition gilt:

$$(61)\quad P(E_{500}) \approx P(E_{100}) \approx P(E_m) \approx c,$$

wobei $m \in \mathbb{N}$ beliebig und c unsere Vertrauenskonstante ist, z.B. 0.95. Hier müssen wir also schon sehen, dass das vorgehen der iterierten Tests äußerst problematisch ist – mir messen etwas ganz anderes als was wir vorgeben! Es kommt aber noch besser: mit etwas Überlegung und etwas komplizierterer Mathematik ist es nicht sonderlich schwer zu sehen dass

$$(62)\quad lim_{n\to\infty} P(E_{100}^n) = 0$$

anders gesagt: egal wie weit/eng unsere Vertrauensgrenzen sind, wenn wir sie Methode oben nur oft genug iterieren, werden wir mit mathematischer Sicherheit irgendwann ein Resultat finden, dass außerhalb unserer Grenzen liegt! Wenn sie also diese Methode als legitim erachten, können wir mit mathematischer Sicherheit jede Nullhypothese "widerlegen".

Wohlgemerkt : Sie fragen sich warum sollte jemand eine Münze so oft werfen? Stellen Sie sich folgendes vor: es wird ein Medikament getestet; die Nullhypothese ist, dass es keine Wirkung hat (das ist etwas komplizierter, am im Prinzip ähnlich). Sie haben einige Jahre hart gearbeitet, Tiere gequält, und sind völlig überzeugt von der Wirksamkeit des Medikaments. Sie machen nun eine Testreihe an Menschen; wenn die Testreihe gut läuft, dann verdient die Firma viel Geld, Sie steigen in der Karriereleiter auf. Wenn die Tests ergebnislos verlaufen, dann haben Sie viel Zeit, die Firma viel Geld in den Sand gesetzt, Ihr Chef ist sauer, Ihre Frau enttäuscht etc.

Sie machen die Testreihe mit 100 Teilnehmern, und das Ergebnis liegt gerade am Rand der Vertrauensgrenzen (aber innerhalb!). Ihr Chef sagt: "Dann probieren Sie halt in Gottes Namen den Test mit noch 100 Teilnehmern!"

6.4 SÜH 2 – Unabhängig

Wir können das Problem auch anders angehen: anstatt dass wir unsere Fallzahlen aufaddieren, wiederholen wir einfach das Experiment von 0 an, und lassen das alte Experiment z.B. einfach in der Schublade verschwinden, tun also so, als hätte es nie stattgefunden. Beim 5 Durchlauf haben wir endlich das gewünschte Ergebnis. Ist das in Ordnung? Wir haben allen Grund mißtrauisch sein: wenn wir unsere Vertrauensgrenze bei $c := 0.95$ festsetzen, dann gibt es immerhin eine Wahrscheinlichkeit von $\frac{1}{20}$, dass wir rein zufällig außerhalb landen. Wie ist also die Wahrscheinlichkeit, dass wir mit 5 Experimenten 1 Ergebnis erzielen, dass außerhalb der Vertrauensgrenzen liegt? Das ist nun einfach, denn die Experimente sind nach unserer Annahme unabhängig. D.h. die Wahrscheinlichkeit, dass wir bei 5 Durchgängen immer ein Ergebnis innerhalb der Vetrauensgrenzen finden, liegt unter Annahme von H_0 bei

$$0.95^5 = 0.7737809375 \tag{63}$$

D.h. die Wahrscheinlichkeit unter dieser Methode ein Ergebnis zu finden, bei dem wir H_0 zurückweisen, ist

$$1 - 0.7737809375 \approx 0.23 \tag{64}$$

also bereits bei fast $\frac{1}{4}$! Hier steigt die Wahrscheinlichkeit also rapide, und es ist offensichtlich, dass die Wahrscheinlichkeit, bei n Experimenten immer innerhalb der Vertrauensgrenzen zu landen, gegen 0 geht.

Die Methode, die wir hier betrachtet haben, würde kein Forscher, der kein bewußter Betrüger ist, anwenden (im Gegensatz zu der obigen!). Aber dennoch ist sie fast noch gefährlicher: nehmen wir an, es gibt eine Hypothese H (wie die Unfairness des Würfels), die aber die Eigenschaft hat, dass sie

1. relativ nahe liegt/populär ist/zu den Dingen gehört die wir alle gerne hören; und

2. wer sie statistisch belegen kann, der kann sich eines sehr positiven Echos sicher sein.

Dementsprechend gibt es viele Forscher, die ähnliche Experimente machen (sagen wir 5). 4 von ihnen haben keine guten Ergebnisse (sie liegen innerhalb der Vertrauensgrenze von H_0). Das will niemand hören, und verschwindet

im Schreibtisch. Der fünfte aber hat "gute" Ergebnisse (beim ersten Experiment!), und macht sie natürlich publik (mit bestem Gewissen!). Wir sehen aber natürlich sofort: die Situation ist genau wie oben, denn wer das Experiment ausführt, ist dem Zufall egal!

Wir haben also eine sehr kritische Situation, da wir nur den Teil der Experimente sehen, die einen wünschenswerten Ausgang haben! Das hat (vermutlich) dazu geführt, dass sich viele wichtige experimentelle Ergebnisse der Psychologie in den letzten Jahre also falsch bzw. artifiziell herausgestellt haben. Das entscheidende ist daher, dass Ergebnisse **replizierbar** sind, also wir bei wiederholten Experimenten immer das gleiche Ergebnis haben.

6.5 SÜH 3 – Bayesianisch

Das grösse Problem, das hier zugrunde liegt, ist das wir eine **Asymmetrie** haben: wir nutzen unsere Experimente nur, um eine Hypothese (üblicherweise die Nullhypothese) zu *falsifizieren*; wir verifizieren aber grundsätzlich nichts. Das ist anders im Bayesianischen Ansatz: im obigen Beispiel hatten wir zwei Hypotheses

$$U, F,$$

und deren jeweilige Wahrscheinlichkeiten *apriori* und *aposteriori*. Dieses Szenario ist symmetrisch zwischen den beiden Hypothesen, und es ist tatsächlich so, dass in diesem Fall die sequentielle Heransgehensweise durchaus legitim ist. Das sieht in diesem Fall wie folgt aus: wir haben unsere bekannten *apriori* Wahrscheinlichkeiten

$$P(F) = P(U) = 0.5$$

Zur Erinnerung:

$$P(K|U) = 0.4 \qquad P(K|F) = 0.6$$

Nun werfen wir Kopf (K). Daraus bekommen wir unsere neuen Wahrscheinlichkeiten

$$P(F|K);\ P(U|K)$$

Das lässt sich nach obigem Muster leicht ausrechnen; eine kurze Überlegung liefert uns:

$$P(F|K) > P(U|K)$$

denn $P(K|F) > P(K|U)$. Nun können wir das iterieren:

> Berechne $P(F|K_1K_2)$; $P(U|K_1K_2)$
> Berechne $P(F|K_1K_2Z_3)$; $P(U|K_1K_2Z_3)$
> ...

Es ist klar dass jedes K die Wahrscheinlichkeit von F erhöht, jedes Z die Wahrscheinlichkeit von U. Weiterhin legen wir eine Gewissheits-Konstante $c \in [0,1]$ fest, so dass für uns gilt: falls wir eine Folge $\vec{W} \in \{K,Z\}^*$ beobachten, so dass

$$(\#) \text{ entweder } P(F|\vec{W}) > c \text{ oder } P(U|\vec{W}) > c,$$

dann hören wir auf und akzeptieren die Hypothese, deren Wahrscheinlichkeit größer c ist (das macht natürlich nur Sinn, falls c nahe bei 1 liegt, z.B. $c = 0.99$). Dieses Vorgehen einwandfrei – warum? Weil wir beide Hypothesen gleichermaßen berücksichtigen! Unsere vorige Herangehensweise wäre vergleichbar mit: wir legen eine Konstante $c \in [0,1]$ fest, so dass falls wir eine Folge $\vec{W} \in \{K,Z\}^*$ beobachten, dass

$$(\#') \ P(U|\vec{W}) > c$$

dann hören wir auf und akzeptieren U. Nun ist vollkommen klar dass auf diese Art und Weise die Hypothese U unfair bevorzugt wird, denn alle beide können an einer gewissen Stelle sehr wahrscheinlich sein.

Übungsaufgabe 4

Abgabe bis zum 16.5.2017 *vor dem Seminar*, egal ob digital/analog und auf welchem Weg.

Berechnen Sie, bei welchen Ergebnissen die Vertrauensgrenzen für die folgenden Binomialverteilungen liegen:

a. $n = 50, p = 0.5$, Vertrauenskonstante $c = 0.95$

b. $n = 200, p = 0.5$, Vertrauenskonstante $c = 0.99$

c. Nehmen wir an, $p = 0.6$. Wie würden Sie nun die Vertrauensgrenzen ausrechnen, wo liegt das Problem?

Aufgabe 5

Abgabe bis zum 17.5.2017 *vor dem Seminar*, egal ob digital/analog und auf welchem Weg.

Nehmen wir folgendes an: Sie führen ein Experiment aus (Einfachheit halber Munzwürfe), sie wollen die Münze 100mal werfen, und schauen ob das Ergebnis innerhalb der Vertrauensgrenze ($c = 0.95$) der Nullhypothese liegt (um zu bestimmen, ob die Münze fair ist). Nach 50 Würfen haben Sie den starken Verdacht, dass die Münze unfair ist, da sie bis dahin ein sehr unausgewogenes Ergebnis haben. Also sagen Sie: "Ich spare mir die 50 restlichen Würfe, prüfe das Ergebnis jetzt an dieser Stelle. Ist ja auch Wurst ob ich ursprünglich 100 oder 1000 oder 50mal werfen wollte." Tatsächlich liegt das Ergebnis außerhalb der Vertrauensgrenzen; Sie weisen also H_0 zurück. Nun die Frage: ist Ihr Vorgehen legitim?

7 Sequentielle Bayesianische Hypothesenprüfung

Das rechnerische Problem bei der Methode in Aufgabe 5 ist, dass wir praktisch keine Berechnungen "recyclen" können, d.h. wir müssen immer wieder neu von vorne anfangen. Zunächst schreiben wir etwas allgemeiner:

$$P(F|X_1...X_i),\ P(U|X_1...X_i),$$

wobei $X_1, ..., X_i$ jeweils Variablen sind für ein beliebiges Ereignis $K_1/Z_1, ..., K_i/Z_i$. Nun können wir etwas mathematischer sagen: $P(F|X_1...X_{i+1})$ lässt sich nicht auffassen als Funktion

$$P(F|X_1...X_{i+1}) = f(P(F|X_1...X_i))$$

für ein einfaches f.

Dem entspricht ein theoretisches Problem: wir möchten Hypothesen prüfen, indem wir der Reihe nach

$P(F|K_1K_2);\ P(U|K_1K_2)$
$P(F|K_1K_2Z_3);\ P(U|K_1K_2Z_3)$
...

errechnen, und warten dass eine der beiden Wahrscheinlichkeiten einen (vorher festgelegten) Grenzwert überschreitet). Allerdings rechnen wir in jedem Schritt weiterhin mit den *apriori* Wahrscheinlichkeiten $P(F), P(U)$, obwohl wir es eigentlich besser wissen, also eigentlich bereits

$$P(F|K_1),\ P(F|K_1K_2) \text{ etc.}$$

kennen. Das verstößt gegen einen zentralen Grundsatz der Wahrscheinlichkeitstheorie und Statistik:

> **wir dürfen niemals relevante Information, die uns bekannt ist, außer Acht lassen**.

Denn sonst wäre ja der Willkür Tür und Tor geöffnet, Informationen zu ignorieren, die relevant ist, aber nicht das von uns gewünschte Ergebnis unterstützt. Im Grunde liegt also auch hier ein Missbrauch vor! Der Ansatz in Aufgabe 5 war also sowohl ziemlich kümmerlich in mathematischer Hinsicht als auch missbräuchlich in theoretischer, da er relevante Information, die fertig vorliegt, in weiteren Rechenschritten außer Acht lässt.

Das Problem ist also das folgende: die Berechnung von $P(F|K_1)$ hängt ab von der *apriori*-Wahrscheinlichkeit $P(F)$. Wenn nun das Ereignis K_1 gegeben ist, ändert sich unsere Einschätzung der Wahrscheinlichkeit von F und U. Das wiederum führt dazu, dass z.B.

$$P(K_2|K_1) \neq P(K_2), \tag{65}$$

etwas allgemeiner: lassen wir $X_1, X_2, ...$ als Variablen stehen, so dass X_i den Wert K_i oder Z_i annehmen können. Dann gilt:

$$P(X_2|X_1) \neq P(X_2); \quad P(X_3|X_2X_1) \neq P(X_3) \text{ etc.} \tag{66}$$

Warum ist das so? Wir haben oben festgestellt das z.B.

$$P(X_2) = P(X_2|F)P(F) + P(X_2|U)P(U) \tag{67}$$

Wir nutzen nun eine allgemeinere Formulierungen unserer Regeln (wir zeigen das exemplarisch am Beispiel $X = K$):

$$P(K_2|K_1) = P(K_2|FK_1)P(F|K_1) + P(K_2|UK_1)P(U|K_1) \tag{68}$$

Überlegungen des gesunden Menschenverstandes sagen uns, dass

$$\begin{array}{l} P(K_2|FK_1) = P(K_2|F) \\ P(K_2|UK_1) = P(K_2|U) \end{array} \tag{69}$$

Das Problem ist:

$$\begin{array}{l} P(F|K_1) \neq P(F) \\ P(U|K_1) \neq P(U) \end{array} \tag{70}$$

Wir können das jetzt noch etwas genauer formulieren, denn wir wissen:

$$(71) \quad \begin{array}{l} P(F|K_1) > P(F) \\ P(U|K_1) < P(U) \end{array}$$

Dasselbe gilt natürlich in die andere Richtung für $P(F|Z_1)$ etc. Das bedeutet wenn wir wirklich sequentiell die Hypothese prüfen, dann müssten wir eigentlich auch jedesmal die Wahrscheinlichkeit von F und U *updaten* (d.h. neu errechnen), und mittels dieser Wahrscheinlichkeit die Wahrscheinlichhkeit der Ergebnisse K, Z neu berechnen.

Wir definieren nun eine neue Variable Y, die die Werte F, U annehmen kann, also

$$Y \in \{F, U\}$$

Damit lassen sich die folgenden Überlegungen allgemeiner formulieren. Wir definieren nun die Wahrscheinlichkeitsfunktion P_{seq} der sequentiellen Prüfung wie folgt:

1. $P_{seq}(Y|X_1) = P(Y|X_1)$
2. $P_{seq}(Y|X_1...X_{i+1}) = P(X_{i+1}|Y)\frac{P_{seq}(Y|X_1...X_i)}{P_{seq}(X_{i+1}|X_1...X_i)}$

Das sieht schon wesentlich besser aus: wir benutzen hier sämtliche Information die uns zur Verfügung steht. Auch die Berechnung wird wesentlich einfacher:

$$P(X_{i+1}|Y)$$

(das eigentlich nur eine Kurzschreibweise für $P(X_{i+1}|YX_1...X_i)$ ist) ist eine konstante c_Y;

$$P_{seq}(Y|X_1...X_i)$$

mussten wir (nach Annahme) bereits ohnehin ausrechnen; bleibt noch der Term

$$P_{seq}(X_{i+1}|X_1...X_i);$$

das lässt sich wie folgt berechnen:

(72)
$$P_{seq}(X_{i+1}|X_1...X_i) = P(X_{i+1}|F)P_{seq}(F|X_1...X_i)+P(X_{i+1}|U)P_{seq}(U|X_1...X_i)$$

Das vereinfacht sich zu

$$P_{seq}(X_{i+1}|X_1...X_i) = c_F P_{seq}(F|X_1...X_i) + c_U P_{seq}(U|X_1...X_i) \quad (73)$$

Am Ende bekommen wir also:

$$P_{seq}(Y|X_1...X_{i+1}) = P_{seq}(Y|X_1...X_i)\frac{P(X_{i+1}|Y)}{P_{seq}(X_{i+1}|X_1...X_i)} \quad (74)$$

Das ist ein relativ zufriedenstellendes Ergebnis (das man mit etwas komplexeren Methoden noch besser ausgestalten kann). Wir haben nun unser Ziel erreicht:

$$P(Y|X_1...X_{i+1}) = f(P(Y|X_1...X_i)) \quad (75)$$

wobei f eine relativ einfache Funktion ist. Außerdem benutzen wir in jedem Schritt alle relevanten Informationen. Aber aus diesem Ergebnis lassen sich noch mehr interessante Folgerungen ableiten. Auf den ersten Blick lässt sich folgendes sagen: Wir haben

$$\begin{aligned} P_{seq}(Y|X_1...X_{i+1}) &> P_{seq}(Y|X_1...X_i) \\ &\Leftrightarrow \\ \frac{P(X_{i+1}|Y)}{P_{seq}(X_{i+1}|X_1...X_i)} &> 1 \quad (76) \\ &\Leftrightarrow \\ P_{seq}(X_{i+1}|X_1...X_i) &< P(X_{i+1}|Y) \end{aligned}$$

Nun ist in unserem Fall leicht zu sehen dass immer gilt:

$$P_{seq}(K_{i+1}|X_1...X_i) < P(K|F) \quad (77)$$

und

$$P_{seq}(Z_{i+1}|X_1...X_i) > P(Z|F) \quad (78)$$

d.h. wir haben die Bestätigung dafür, dass jeder Wurf von Kopf die Wahrscheinlichkeit von F erhöht, umgekehrt jeder Wurf von Zahl die Wahrscheinlichkeit von U.

Eine wichtige Frage, die wir hier offengelassen haben, ist was es *bedeutet*, also wie sich P_{seq} von P unterscheidet. Das lässt sich wie folgt beantworten: es gilt:

$$\begin{aligned} P(X_{i+1}|F) = P_{seq}(X_{i+1}|X_1...X_i) \\ \Leftrightarrow \\ P(U|X_1...X_i) = 0 \\ \Leftrightarrow \\ P(F|X_1...X_i) = 1 \end{aligned} \tag{79}$$

Dass wir allerdings tatsächlich

$$P(F|X_1...X_i) = 1$$

haben ist theoretisch ausgeschlossen, denn es gibt kein Ergebnis, das mit $P(U)$ wirklich inkompatibel wäre. Dennoch bedeutet dass:

$$\underset{P(F|X_1...X_i)}{argmin} \frac{P(X_{i+1}|F)}{P(X_{i+1}|X_1...X_i)} = 1 \tag{80}$$

(es soll uns nicht stören dass $P(F|X_1...X_i)$ nicht explizit vorkommt in dem Term; wir wissen ja es ist implizit vorhanden) Umgekehrt sieht man aus demselben Grund dass

$$\underset{P(F|X_1...X_i))}{argmax} \frac{P(X_{i+1}|F)}{P(X_{i+1}|X_1...X_i)} = 0 \tag{81}$$

Das bedeutet in Worten:

> Je unwahrscheinlicher F (bzw. K) ist gegeben unsere bisherige Beobachtungen, desto stärkere relative Evidenz liefert eine Beobachtung von K (bzw. Z) für F (bzw. K).

Dasselbe gilt natürlich auch andersrum:

> Je wahrscheinlicher F (bzw. K) ist gegeben unsere bisherige Beobachtungen, desto schwächere relative Evidenz liefert eine Beobachtung von K (bzw. Z) für F (bzw. K).

Das kann man sich intuitiv wie folgt klar machen: wenn wir F bereits für sehr wahrscheinlich halten, dann liefert uns eine Evidenz für F nur geringe neue Information, eine Evidenz für U aber deutlich mehr. Wir haben hier übrigens eine typische *invers exponentielle Sättigungskurve*; in der logarithmischen Transformation wird das also zu einer einfachen Addition (siehe Jaynes: Probability Theory, Kapitel 4)!

Aufgabe 6

Abgabe bis zum 23.5.2017 *vor dem Seminar*, egal ob digital/analog und auf welchem Weg.

Nehmen Sie an, Sie befinden sich im Urlaub einer großen Stadt. Sie haben (am Abreisetag) Ihre Koffer im Hotel deponiert, sind noch in der Stadt unterwegs. Sie müssen also an einem gewissen Punkt erst ins Hotel, dann zum Bahnhof, um Ihren Zug zu bekommen. Dazu müssen Sie *5mal* umsteigen; die Zeiten für Fahrten, den Weg von einem Gleis zum andern etc. sind Ihnen bekannt; die einzige unbekannte sind die genauen Fahrtzeiten: sie wissen nur dass die Bahnen einheitlich alle *10 Minuten* fahren, so dass Sie also schlimmstenfalls 50min reine Wartezeit für diese Reise zu erwarten haben. Natürlich möchten Sie die Abfahrt so lang als möglich hinauszögern. Aber wieviel Zeit veranschlagen Sie als reine Wartezeit auf Anschlusszüge für Ihre Reise *jetziger Aufenthalt*→*Hotel*→*Bahnhof*?

1. 50min sind natürlich das Maximum, aber es ist sehr unwahrscheinlich dass Sie so lange warten müssen. Stattdessen beschließen Sie folgendes: Sie möchten mit einer Wahrscheinlichkeit von 0.95 Ihren Zug erwischen. Wieviel Zeit veranschlagen Sie unter dieser Prämisse für das Warten auf Anschlußzüge? Der Einfachheit halber nehmen wir hier immer an, dass Zeit diskret im Minutentakt abläuft.

2. Nehmen Sie an, statt 5mal Umsteigen mit einer Bahn alle 10min müssen Sie *10mal* Umsteigen, aber die Bahnen fahren alle *5min*. Mit ansonsten denselben Prämissen wie in 1., wie verändert sich Ihre Einschätzung, veranschlagen Sie mehr oder weniger Zeit für eine Sicherheit von 0.95? Begründen Sie! Sie können das natürlich explizit ausrechnen; es gibt aber auch eine einfache gute Begründung, wenn Sie unsere bisherigen Erkenntnisse zu Vertrauensgrenzen, Varianz etc. nutzen.

8 Einseitige Tests

8.1 Die Hausaufgabe - manuelle Lösung

Nehmen wir das Beispiel aus Aufgabe 6. Hier geht es nicht um Abweichungen vom Erwartungswert in beliebige Richtungen: denn in unserem Fall wäre der Erwartungswert 30min Wartezeit (symmetrische Verteilung zwischen 10 und 50, oder anders je nach Annahme), und die oben beschriebene Methode würde uns die Wahrscheinlichkeit geben, eine gewisse maximale Abweichung vom Erwartungswert in *beide* Richtungen zu bekommen. Unsere Sorge besteht aber darin, vom Erwartungswert in *eine* Richtung abzuweichen, nämlich eine längere Wartezeit zu bekommen. In gewissen Sinne ist das aber noch einfacher: den in diesem Fall müssen wir nur den **rechten Rand** der Verteilung ausrechnen, und sobald dieser Rand eine Wahrscheinlichkeitsmasse von > 0.05 hat, sind wir fertig. Wir suchen also

$$(82) \quad max_n P(W = 50) + P(W = 50 - 1) + ... + P(W = 50 - n) < 0.05$$

Das ist also konzeptuell sehr einfach, aber rechnerisch nicht ganz trivial. Erstmal folgendes: wir nehmen an, unsere Wartezeiten sind 1,2,..,10; die 0 fällt also weg (wir müssen ja umsteigen). Die Wahrscheinlichkeit von 50min Wartezeit ist klar:

$$(83) \quad P(W = 50) = \frac{1}{10}^5 = \frac{1}{100,000}$$

Ebenfalls leicht zu berechnen ist

$$(84) \quad P(W = 49) = 5 \cdot \frac{1}{10}^5 = \frac{5}{100,000} = \frac{1}{20,000}$$

Denn es gibt 5 Möglichkeiten, 49min Wartezeit zu haben. Wie wird das bei 48? Hier gibt es wieder 5 mit einmal 8min Wartezeit, und $\binom{5}{2}$. Für $W = 47$ haben wir $5 + \binom{5}{2} + \binom{5}{3}$.

$$(85) \quad P(W = 48) = 5 \cdot \binom{5}{2} \frac{1}{10}^5 = \frac{50}{100,000} = \frac{1}{2,000}$$

Und weiter gehts:

$$(86) \quad P(W = 47) = 5 \cdot \binom{5}{2} \cdot \binom{5}{3} \frac{1}{10}^5 = \frac{500}{100,000} = \frac{1}{200}$$

Jetzt wird es etwas komplizierter:

(87)
$$P(W=46) = 5{\cdot}2\binom{5}{2}\cdot\binom{5}{3}\cdot\binom{5}{4}\frac{1}{10}^{5} = 5^2{\cdot}20{\cdot}10{\cdot}\frac{1}{100,000} = \frac{5,000}{100,000} = \frac{1}{20}$$

Hier sind wir also oberhalb unserer Grenze – vorher nicht. Die Antwort ist also: wenn wir 46min Zeit einplanen, dann verpassen wir mit einer Wahrscheinlichkeit von > 0.05 unseren Zug. Zum Glück haben wir unseren Wert schnell erreicht, denn ab hier wird es schnell komplizierter: denn wieviel Möglichkeiten gibt es für $W = 45, 44$? n über k ist hier nicht mehr ausreichend, denn wir müssen nun zwei Dinge berücksichtigen:

1. Auf wieviele Arten wir n auf verschiedene Summanden aufteilen können; das sagt uns die sog. **Partitionsfunktion**, die exponentiell wächst und erst kürzlich eine berechenbare Form bekommen hat. Allerdings haben wir höchstens 5 Summanden, und die dürfen nicht größer als 10 sein!

2. Auf wieviele Arten können wir die Summanden auf die einzelnen Züge verteilen? Das ist ein klassisches kombinatorisches Problem.

Hier bekommen wir also eine sehr komplexe Kombinatorik **Kombinatorik**, die wir nicht mehr ohne weiteres in eine algemeine Form bringen können. Daran wird aber deutlich, warum der zentrale Grenzwertsatz so wichtig ist, denn er sagt uns, wie wir mit Normalverteilungen das Problem approximieren können.

8.2 Zweiter Teil

Die Antwort sollte klar sein: höhere Anzahl von Iterationen $\Longrightarrow$ geringere Varianz $\Longrightarrow$ Wahrscheinlichkeitsmasse ist enger um der Erwartungswert verteilt, der in diesem Fall derselbe ist. Dementsprechen entfällt weniger Masse auf die Ränder, und unsere Sicherheit wird größer. Numerisch kann man sich das so erklären: wo wir früher mit

(88) $$k\frac{1}{10^5}$$

gerechnet haben - k ist die kombinatorische Zahl der Möglichkeiten - rechnen wir jetzt mit

(89) $$k\frac{1}{5^{10}} = k\frac{1}{9,765,625}$$

D.h. die Grundwahrscheinlichkeit sind um ca. das 100fache niedriger; dementsprechend auch die Wahrscheinlichkeiten $P(W = 50), ..., P(45)$ – denn hier ist die Kombinatorik noch (fast) dieselbe.

8.3 Ein zweites Beispiel – Fehlerquoten

Im vorigen Beispiel ging es um *Sicherheit* – wir wollen sicher sein (bis zu einem gewissen Punkt), dass wir den Zug erwischen. Einen ähnlichen Fall haben wir im folgenden Beispiel, auch wenn die Dinge etwas anders gelagert sind: nehmen wir an, wir produzieren einen Gegenstand, z.B. Achsen für ein Auto. Wir möchten sicher sein, dass es eine Fehlerquote ϵ gibt, so dass gilt:

$$\frac{\text{Anzahl der fehlerhaft produzierten Achsen}}{\text{Anzahl der insgesamt produzierten Achsen}} < \epsilon \tag{90}$$

Den linken Term nennen wir die **Fehlerquote** f. Das ist eine wichtige Fragestellung, den fehlerhafte Teile können immer produziert werden; was entscheidend ist ist wissen um deren Quote. Die genaue Quote können wir nicht kennen, da Materialtests aufwändig sind und teilweise das Material zerstören. Also suchen wir folgendes:

> Wir legen 2 Konstanten p, α fest, und möchten verifizieren mit einer Sicherheit $s > q$ die Fehlerquote $f < \epsilon$ ist.

Das bedeutet: die Wahrscheinlichkeit, dass $f \geq \epsilon$ ist, soll $< 1 - q := p$ sein. Wir haben hier also ein klassisches Problem von p-Werten und Vertrauensgrenzen. Dem ganzen liegt in diesem Fall auch eine einfache, wenn auch asymmetrische Binomialverteilung zugrunde, von daher haben wir eine einfache arithmetische Formel für unsere Rechnungen. Z.B. legen wir fest:

$$\epsilon = 0.001,\ q = 0.95,\ \text{also } p = 0.05$$

Nun machen wir eine **Stichprobe** $\mathfrak{S}$ von sagen wir 10,000 Objekten, und davon sind 3 defekt. Wir machen nun folgendes: wir nehmen als H_0 an, dass $f = 1/1,000$ (wir werden das später hinterfragen). Da unser Erwartungswert (Binomialverteilung!) für eine Probe der Größe 10,000 nun bei 10 liegt, ist das erstmal ein gutes Ergebnis. Aber ist es gut genug?

Wir berechnen (wie immer bei p-Werten) die Wahrscheinlichkeit, dass unser Ergebnis wie gehabt *oder noch extremer* ausfällt, also für uns: wie ist die Wahrscheinlichkeit, 3 oder noch weniger defekte Teile auf 10,000 zu finden? Dafür haben wir eine einfache Formel:

$$\sum_{i=0}^{3} \binom{10,000}{3} (\frac{999}{1,000})^{10,000-i} \cdot (\frac{1}{1,000})^{i} \tag{91}$$

In R lässt sich das einfach berechnen:

```
> int=0:3
> for(i in 0:3){int[i+1] <- choose(10000,i)*
(999/1000)^(10000-i)*(1/1000)^i}
> int
[1] 4.517335e-05 4.521856e-04 2.262965e-03 7.549258e-03
> sum(int)
[1] 0.01030958
```

Das bedeutet also: unser Ergebnis ist (nach unseren Annahmen) signifikant; die Wahrscheinlichkeit unserer Stichprobe (oder einer noch extremeren Probe) liegt bei 0.01. Das heißt: das Ergebnis liegt außerhalb der Vertrauensgrenzen, wir weisen H_0 zurück.

Aber: was ist das für ein Nullhypothese? Sie ist ja in keinem besonderen Sinne neutral, und wir können ja nicht beliebig Hypothesen annehmen, um sie dann zurückzuweisen. Die Rechtfertigung für diese Methode ist die folgende: gegeben unsere Annahmen war H_0 diejenige Hypothese, die die Daten *am wahrscheinlichsten* macht; denn wir wollten sicher gehen dass $f < 1/1,000$, und das Ergebnis unserer Stichprobe war jenseits des Erwartungswertes in diesem Fall. Für jede andere Wahrscheinlichkeit H_0', die besagt dass $P(f) > 1/1,000$, wäre unsere Stichprobe noch unwahrscheinlicher gewesen, unser Ergebnis noch signifikanter ausgefallen. Wir haben also H_0 angenommen als diejenige Hypothese, die die Daten am wahrscheinlichsten macht, und wenn wir diese Hypothese zurückweisen, dann weisen wir auch jede andere Hypothese zurück dass die Fehlerquote $> 1/1,000$ ist. Wir weisen also H_0 stellvetretend für alle Hypothesen zurück, nach denen die Fehlerquote zu hoch ist. Dementsprechend haben wir gezeigt, was wir zeigen wollen.

NB: das funktioniert natürlich nur, wenn in unserer Stichprobe $\mathfrak{S}$ die Fehlerquote *niedriger* ist als der Erwartungswert unter H_0 – sonst wäre das ganze offensichtlich Unsinn. In diesem Fall kann man Fall kann man p-Werte auch bei asymmetrischen Verteilungen nutzen, man muss aber vorsichtig sein, dass das Vorgehen sinnvoll bleibt.

Die Vertrauensgrenze liegt übrigens bei 4: wenn wir mehr als vier fehlerhafte Teile finden, dann können wir H_0 nicht mit Signifikanz p zurückweisen (für 4 gilt $p = 0.02919595$), für 5 haben wir $p = 0.06699137$). Das lässt sich leicht mit obiger Formel errechnen, es reicht i entsprechend einzusetzen.

9 Statistiken und Tests - Abstrakt

Eine typische Situation in der Statistik ist die folgende: wir haben einen gewissen Datensatz; nehmen wir z.B. an, wir haben einen gewissen Text (Datensatz) D.

> MDNGHRKENGNSKRNSHREHWEJFVBNBFJSKEWRNDSJXYHD
> NWIDHEJKDNXHWKJDHJAJDWREHFKVJCJFHRNENDKXMDJ
> EYUHWNDJFD...

Das kann für etwas beliebiges stehen; wir können auch annehmen, dass es sich um einen sprachlichen Text handelt (mit Wörtern), wobei z.B. B für das Leerzeichen steht. Gleichzeitig haben wir zwei **Sprachmodelle** M_0, M_1. Beide weisen dem Text T eine gewisse Wahrscheinlichkeit zu:

$$M_0(D),\ M_1(D)$$

Wir sollen nun entscheiden, welches Modell besser ist. Die einfache Lösung wäre folgende: wir nehmen einfach

$$max(M_0(D), M_1(D)),$$

und wählen die Hypothese entsprechend. Wenn wir beliebige Hypothesen zulassen, dann enden wir mit einer Art maximum-likelihood Schätzung. Das ist aber in vielen Fällen inadädquat, da durch diese Herangehensweise die Hypothesen *zu stark* durch die konkreten Daten bestimmt werden. Das gilt besonders dann, wenn gewisse Hypothesen stark unabängig motiviert sind, und wir nicht beliebige Hypothesen zur Auswahl haben. In unserem Zusammenhang können wir etwa folgendes Beispiel benutzen:

- M_0 ist ein Modell, in dem alle Wortwahrscheinlichkeiten unabhängig voneinander sind;
- M_1 ist ein Markov-Modell, in dem Wortwahrscheinlichkeiten sich wechselseitig beeinflussen (über einen beschränkten Raum hinweg).

Was uns also interessiert: welche Art der Modellierung ist plausibler?

Bevor wir diese Frage beantworten, müssen wir uns über eine grundlegende Asymmetrie zwischen M_0 und M_1 klarwerden. Es ist erstens klar, dass M_1 bessere Ergebnisse erzielt, wenn wir die Parameter so anpassen, dass sie genau auf D passen; aber das ist soz. trivial und nicht empirisch:

wir möchten eine allgemeine Aussage treffen, von der wir davon ausgehen dass sie auch für andere Texte ihre Gültigkeit hat; wir würden also gerne mit Parametern arbeiten, die allgemein und unabhängig von D geschätzt sind. Wir sollten also davon ausgehen dass die Parameter von M_0, M_1 beide auf unabhängigen Texten geschätzt wurden. Die obige Tatsache aber nur ein Sympton einer tieferliegenden Asymmetrie:

> M_1 ist *spezifischer* als M_0, es spezifiziert mehr Parameter, oder anders gesagt: die Ereignisse sind stärker voneinander abhängig als in M_0.

Aus dieser Tatsache wird – wissenschaftlichen Grundsätzen wie Ockhams Rasiermesser folgend (*entia non sunt multiplicanda*) – das bis auf weiteres M_0 gegenüber M_1 vorzuziehen ist. Wenn wir dennoch sagen, dass M_1 besser ist als M_0, dann brauchen wir dafür gute Gründe. Hier haben wir nun alle Zutaten eines klassischen statistischen Problems beisammen: wir haben zwei voll ausgearbeitete Hypothesen, die sich einteilen lassen in

1. eine **Nullhypothese** M_0 – üblicherweise H_0 geschrieben, und
2. eine alternative Hypothese M_1, üblicherweise H_1 geschrieben.

H_0 ist also die Hypothese, dass Worte im Text unabhänging voneinander sind, H_1 die Hypothese dass sie sich wie Markov-Ketten verhalten. Wir haben nun einen Datensatz D, und möchten uns **entscheiden**, gegeben D, welche der beiden Hypothesen im Allgemeinen vorzuziehen ist. Eine solche Entscheidungsfunktion nennt man einen **Test**.

Hierbei gibt es natürlich folgendes zu beachten: uns interessiert eine zugrunde liegende Wahrscheinlichkeitsverteilung, die erstmal nichts ausschließt. Das bedeutet wir können nicht mit Sicherheit die richtige Antwort finden; es kann immer sein dass unsere Daten D *zufällig* so aussehen, als ob sie von einer Markov-Kette generiert wurden (oder umgekehrt). Wir können das nie ausschliessen; der Trick ist aber: wir können das so unwahrscheinlich wir möglich machen. Zunächst müssen wir folgende Definitionen und Unterscheidungen machen.

Definition 6 *Sei Ω eine Menge von Datensätzen, $P_0, P_1 : \Omega \to [0,1]$ zwei Wahrscheinlichkeitsfunktionen. Sei $H_i : i \in \{0,1\}$ die Annahme, dass P_i die zugrundeliegende Wahrscheinlichkeitsverteilung ist, die D erzeugt hat. Ein* ***Test*** *ist eine Funktion $t : \Omega \to \{H_0, H_1\}$; $t^{-1}(H_1)$ ist der sog.* ***kritische Bereich*** *von t.*

Nehmen wir weiterhin an, wir haben guten Grund H_0 als Nullhypothese zu bezeichnen (im obigen Sinne; es ist etwas kompliziert dieses Konzept formal auszudrücken).

Definition 7 *Ein Test T macht einen* ***Typ I*** *Fehler, falls er H_1 wählt, obwohl H_0 korrekt ist; er macht einen* ***Typ II*** *Fehler, falls er H_0 wählt, obwohl H_1 korrekt ist.*

Im allgemeinen möchte man Typ I Fehler eher vermeiden als Typ II Fehler; das bedeutet, wir möchten eher konservativ sein. Das spiegelt die Tatsache wieder dass die Nullhypothese aus methodologischen Gründen vorzuziehen ist. Praktisch bedeutet dass: wir möchten eher, dass ein Medikament nicht zugelassen wird, da seine Wirkung nicht ausreichend belegt ist (aber womöglich vorhanden), als dass es zugelassen, aber womöglich unwirksam ist.

Wir wissen natürlich nie, ob wirklich ein Fehler vorliegt. Wir können allerdings über die Wahrscheinlichkeit sprechen, mit der ein bestimmter Test bestimmte Fehler macht. Sei T ein Test.

- Die Wahrscheinlichkeit das T *keinen* Typ I Fehler macht, ist seine **Signifikanz**;
- die Wahrscheinlichkeit dass er *keinen* Typ II Fehler macht, ist seine **Mächtigkeit**.

Das bedeutet: je signifikanter ein Test, desto sicherer sind wir die Nullhypothese nicht zu unrecht zu verlassen; je mächtiger er ist, desto sicherer sind wir, nicht zu unrecht bei der Nullhypothese zu bleiben. Ein Test T ist **maximal signifikant**, falls jeder andere Test T', der signifikanter ist als T, echt weniger mächtig ist; T ist **maximal mächtig**, falls jeder Test T' der mächtiger ist, echt weniger signifikant ist.

Sei $p = P(H_0)$ die a priori Wahrscheinlichkeit von H_0; wir nehmen an dass $P(H_1) = 1 - p$, es also keine weitere Hypothesen gibt. Dann haben wir, für den Fall dass wir H_1 wählen,

(92) Wahrscheinlichkeit eines Typ I Fehlers: $F1 := \dfrac{1}{1 + \frac{1-p}{p}\frac{P(D|H_1)}{P(D|H_0)}}$

und für den Fall dass wir H_0 wählen,

(93) Wahrscheinlichkeit eines Typ II Fehlers: $F2 := \dfrac{1}{1 + \frac{p}{1-p}\frac{P(D|H_0)}{P(D|H_1)}}$

Die Ableitung dieser Ergebnisse soll uns hier nicht bekümmern; wir stellen nur folgende Grenzfälle fest:

- für $\frac{1-p}{p}\frac{P(D|H_1)}{P(D|H_0)} \mapsto 1$ haben wir $F1 \mapsto 1/2$ – das ist einleuchtend, da wir in diesem Fall keine Evidenz für beide Hypothesen haben.
- für $(1-p)P(D|H_1) \mapsto 0$ haben wir $F1 \mapsto 1$ – also wie H_1 unplausibel wird, wird die Wahrscheinlichkeit eines Fehlers sicherer.
- umgekehrt gilt für $(1-p)P(D|H_1) \mapsto 0$ dass $F2 \mapsto 0$ – die Wahl von H_0 wird immer wahrscheinlicherer korrekt.

Bevor wir wirkliche Tests einführen können, brauchen wir **Statistiken**.

Definition 8 *Sei Ω ein Wahrscheinlichkeitsraum, $n \in \mathbb{N}$. Eine Statistik S ist eine Funktion auf Ω^n, dem n-fachen Produktraum.*

Statistiken sind also ein sehr allgemeiner Begriff.

Definition 9 *Sei $\mathbb{P} = \{P_\theta : \theta \in \Theta\}$ eine Menge von Wahrscheinlichkeitsfunktionen auf Ω^n. Eine Statistik S ist **ausreichend** für $\mathbb{P}$, falls für alle $\theta, \theta' \in \Theta$, $\omega \in \Omega^n$ gilt:*

$$P_\theta(\omega | S = s\}) = P_{\theta'}(\omega | S = s\})$$

wobei $s = S(\omega)$.

Eine Statistik ist ausreichend, falls sie alle Informationen enthält, die wir brauchen um Wahrscheinlichkeiten zu bestimmen. Wir definieren nun folgende Statistik, gegeben zwei Hypothesen H_0, H_1:

$$(94) \quad R(\omega) := \frac{P(\omega|H_0)}{P(\omega|H_1)} \qquad \text{(Sonderfall für } P(\omega|H_1) = 0)$$

Diese Statistik heißt das likelihood-Verhältnis. Sie ist ausreichend, d.h. enthält alle Information die wir brauchen; sie ist darüber hinaus auch minimal im Sinne dass sie keine nicht-relevante Information enthält. Im Hinblick auf Definition 9 müssten wir schreiben: $\Theta = \{0, 1\}$, und

$$(95) \quad R(\omega) := \frac{P_0(\omega)}{P_1(\omega)}$$

Das ist nur eine andere Schreibweise; das Ergebnis ist trotzdem nicht offensichtlich; wir können es hier nicht zeigen.

Ein **Schwellentest** S_t mit Wert t ist ein Test, der sich für H_0 entscheidet falls $R(\omega) > t$, und für H_1 andernfalls, also:

$$(96) \quad S_t(\omega) = \begin{cases} H_0, \text{ falls } R(\omega) > t \\ H_1 \text{ andernfalls.} \end{cases}$$

Folgender Satz ist von fundamentaler Wichtigkeit:

Satz 10 *Für jeden Wert t ist der Schwellentest S_t maximal signifikant und maximal mächtig.*

Das bedeutet, in gewissem Sinn ist jeder Wert optimal. Da wir aber die Nullhypothese *a priori* bevorzugen, wählt man üblicherweise einen Wert wie $t = 0.05$, mit hoher Signifikanz und geringerer Mächtigkeit.

10 Tests in der Praxis

10.1 Vorspiel: Parameter schätzen

Sei also $\mathfrak{T}$ ein Text der Form

$$w_1 w_2 w_3 w_4 w_5 w_6 ...$$

wobei das Subskript nur etwas über die Position des Wortes sagt, nicht seine Identität. Der Test liefert uns viele Information, z.B. für alle vorkommenende *types* die Anzahl der *token*. Wir listen die Types mit

$$L = \{l_1, l_2, l_3, ...\}$$

mit l wie Lexikon. NB:

$$S : \mathfrak{T} \to L \times \mathbb{N}$$

wobei jedem Wort seine Häufigkeit zugewiesen wird, ist nach unserer Definition eine Statistik (aber es ist nicht gesagt, dass sie ausreichend ist!).

Wir nehmen folgende Konventionen:

- $|\mathfrak{T}|$ ist die Länge des Textes (Gesamtzahl der token).
- $|\mathfrak{T}|_w$ die Haüfigkeit des Wortes w im Text (Anzahl der token dieses types).
- Wir schreiben auf $f(w)$ anstatt $|\mathfrak{T}|_w$, und
- $f(w|v)$ für $|\mathfrak{T}|_{vw}$. Das zählt also: wie oft folgt w auf v.

Mit diesen Zahlen können wir Wahrscheinlichkeiten als relative Häufigkeiten schätzen.

Unabhängige Wahrscheinlichkeit:

$$\hat{P}_0(w) = \frac{|\mathfrak{T}|_w}{|\mathfrak{T}|} \tag{97}$$

Abhängige Wahrscheinlichkeit:

$$\hat{P}_1(w|v) = \frac{|\mathfrak{T}|_{vw}}{|\mathfrak{T}|_v} \tag{98}$$

Das sind die geschätzten Wahrscheinlichkeiten in unserem Text.

H_0 wäre: $\hat{P}_0$ (oder etwas in dieser Art) ist die zugrundeliegende Verteilung, also:

(99) $P_0(w_1w_2w_3...) = \hat{P}_0(w_1) \cdot \hat{P}_0(w_2) \cdot \hat{P}_0(w_3) \cdot ....$

Hingegen ist H_1 die Hypothese, dass wir Markov-Abhängigkeiten erster Stufe haben, also:

(100) $P_1(w_1w_2w_3...) = \hat{P}_1(w_1|\#) \cdot \hat{P}_1(w_2|w_1) \cdot \hat{P}_1(w_3|w_2) \cdot ....$

Es ist natürlich klar dass normalerweise

(101) $P_0(\mathfrak{T}) < P_1(\mathfrak{T})$

denn P_1 ist spezifischer

10.2 p-Werte in der Praxis

Der p-Wert ist etwas anders als die übrigen Testverfahren: normalerweise interessiert uns, wie Wahrscheinlich die Daten sind gegeben die Nullhypothese. Das ist aber oftmals nicht sehr informativ: gerade wenn wir eine realwertige Verteilung haben, dann interessiert uns nicht ein Punkt, sondern ein Integral. Hier hilft uns der p-Wert:

> Der p-Wert gibt die Wahrscheinlichkeit, dass die Daten so sind wie sie sind *oder noch extremer*, gegeben dass die Nullhypothese wahr ist.

Das ist also wieder das Prinzip der Vertrauensgrenzen. Dieses "*oder noch extremer*" ist aber im Allgemeinen ein Problem: was genau soll das heißen? Hier braucht man Statistiken: wir müssen unsere Ergebnisse so transformieren, dass diese Aussage Sinn macht!

Mann kann sich das sehr einfach mit dem Würfelbeispiel klarmachen.

Wir würfeln 100 mal, und haben 63 Zahl.

Nennen wir dieses Ergebnis ω. H_0 ist, dass der Würfel fair ist. Natürlich können wir sehr einfach $P(\omega|H_0)$ ausrechnen, aber das ist natürlich nicht wirklich informativ: bei vielen Würfen wird jedes Ergebnis sehr unwahrscheinlich. Andererseits, wenn wir H_0 unter ω zurückweisen, dann weisen wir es auch unter jedem ω' zurück, bei dem wir noch öfter Zahl geworfen haben. Außerdem sollten wir, da wir ein rein symmetrisches Experiment haben, H_0 ebenfalls zurückweisen unter ω', falls ω' in 63 oder mehr Würfen von Kopf besteht. Was wir also machen möchten ist: wir fassen alle diese Ergebnisse zu *einem* Ereignis zusammen, und schauen wie wahrscheinlich dieses Ereignis unter H_0 ist. Das können wir natürlich einfach ausrechnen nach den üblichen Regeln; etwas formaler *transformieren wir unseren Wahrscheinlichkeitsraum mit Hilfe von Zufallsvariablen.* Das geht so: zunächst nehmen wir

$$X(\text{Zahl}) = 1,\ X(\text{Kopf}) = 0.$$

Als nächstes nehmen wir die übliche Additionsvariable:

$$Y(\langle x_1, ..., x_i\rangle) = \sum_{j=1}^{i} x_j.$$

Damit sind unsere Ergebnisse Zahlen zwischen 1 und 100, und nicht mehr Laplace-verteilt. Dann nehmen wir eine dritte Variable:

$$Z(x) = |50 - x|.$$

Warum diese Variable? Nun, 50 ist der Mittelwert und Erwartungswert unserer Variable Y. Daher liefert uns $Z(x)$ den Wert der Abweichung von dem Erwartungswert. Was uns interessiert ist jetzt:

(102) $P(Z(x) \geq 13|H_0)$

Das ist der p-Wert von H_0 gegeben ω; normalerweise sagt man: falls $p < 0.05$, dann wird H_0 zurückgewiesen (alternativ: 0.01, 0.001). Wie ist das in unserem Fall? Wir haben

$$(103)\quad P(Z(x) \geq 13|H_0) = 2\sum_{i=63}^{100} \binom{100}{i} (\frac{1}{2})^{100}$$

Falls dieser Wert unterhalb der (vorher!) festgelegten Schwelle liegt, weisen wir H_0 zurück; wir sagen dann, das Ergebnis war **signifikant**. Aber Vorsicht: ein p-Wert $< x$ bedeutet

1. *nicht*, dass die Wahrscheinlichkeit von H_0 gegeben die Daten $< x$ ist!
2. *nicht*, dass H_0 falsch ist!
3. *nicht*, dass irgendein H_1 richtiger ist!

Insbesondere ist zu beachten: wenn mein Schwellenwert 0.05 ist, dann sage ich damit: ich möchte die Wahrscheinlichkeit eines Typ I Fehlers unter 1/20 drücken. Das bedeutet aber umgekehrt: in einem von 20 Experimenten habe ich durchschnittlich einen Typ I Fehler. Und was noch drastischer ist: angenommen, ich mache zwanzig Experimente, eines davon mit signifikantem Ergebnis – dann heißt das überhaupt nichts! Das Problem ist nun: wenn ein Wissenschaftler/Konzern eine Studie mit signifikantem Ergebnis veröffentlicht, woher weiß ich, wie viele andere Studien ohne signifikantes Ergebnis in der Schublade liegen? Die Aussagekraft des p-Wertes hängt sehr stark von Faktoren ab, die außerhalb der Studie selbst liegen. Aus genau diesem Grund gibt es momentan eine starke Bewegung von Statistikern, die gegen die Verwendung von p-Werten argumentiert.

Wir können nun zurück zu unserem Beispieltext $\mathfrak{T}$ in einer unbekannten Sprache. Nehmen wir einmal an, H_0 ist richtig: die Verteilung der Wörter

ist zufällig, d.h. die Wahrscheinlichkeit eines Wortes hängt nicht von seinen Nachbarworten ab.

In diesem Fall können wir folgendes machen: wir nehmen an, $\mathfrak{T}$ in seiner Gesamtheit stellt eine **Population** dar, aus der wir eine **Stichprobe** entnehmen. Mit Population meinen wir, dass sie die "richtigen Verhältnisse" hat, also die zugrundeliegende Wahrscheinlichkeitsverteilung widerspiegelt. Diese Wahrscheinlichkeitsverteilung nennen wir wie oben

$$\hat{P}_0$$

Dass dies die korrekte Verteilung ist kann man natürlich nie wissen, sondern nur annehmen, um daraus gewisse Schlussfolgerungen zu ziehen. Aus der Population können wir nun eine Stichprobe ziehen. Bedingung ist, dass sie zufällig ausgewählt ist; und unter dieser Bedingung sollte die Stichprobe die Wahrscheinlichkeitsverteilung der Population widerspiegeln. Wir nehmen nun als Stichprobe die Menge aller Worte, die auf das Wort w folgen. Da unter H_0 w keinen Einfluss hat auf seinen Nachfolger, ist das qua Annahme eine legitime Auswahl. Wir müssen nur sicherstellen, dass w häufig genug auftritt, sonst haben unsere nachfolgenden Untersuchungen geringe Aussagekraft. Wir benennen

$$\mathfrak{T}_w = \text{Stichprobe der Nachfolger von } w$$

Nachdem wir w gewählt haben, können wir uns die Frage stellen:

$$\text{Hat } \mathfrak{T}_w \text{ dieselbe Verteilung wie } \mathfrak{T}?$$

Vermutlich nicht! Weiterhin können wir fragen: wie wahrscheinlich ist $\mathfrak{T}_w$ unter $\hat{P}_0$? Aber das ist natürlich wieder eine unbefriedigende Fragestellung, denn wenn $\mathfrak{T}_w$ groß genug ist, wird $\hat{P}_0(\mathfrak{T}_w)$ immer sehr klein sein. Was uns also wieder interessiert ist die Frage:

> Wie groß ist, gegeben H_0, die Wahrscheinlichkeit von $\mathfrak{T}_w$ oder eines (gleichgroßen Datensatzes, der *noch unwahrscheinlicher* ist unter $\hat{P}_0$?

Und die nächste Frage ist: *wie implementieren wir dieses Konzept formal*? Die Antwort ist dieses mal etwas abstrakter als mit den Würfeln: wir generieren alle *möglichen* $\mathfrak{T}'$ (derselben Größe), so dass gilt: $\hat{P}_0(\mathfrak{T}') \leq \hat{P}_0(\mathfrak{T}_w)$.

Dass sind natürlich nur endlich viele, wir können also folgende Summe theoretisch ausrechnen:

$$(104) \quad \sum_{\mathfrak{T}':\hat{P}_0(\mathfrak{T}')\leq\hat{P}_0(\mathfrak{T}_w)} \hat{P}_0(\mathfrak{T}')$$

Das ist ein Ereignis und liefert eine Wahrscheinlichkeit, und das ist der p-Wert für H_0 gegeben $\mathfrak{T}_w$. Falls dieser Wert kleiner als unser Schwellwert ist (z.B. 0.05), dann weisen wir die Nullhypothese zurück. Diese Hypothese war: die Wahrscheinlichkeiten von aufeinanderfolgenden Wörtern sind unabhängig.

Wir haben also eine hübsche, allgemeine Form, die der Computer relativ gut berechnen kann (bzw. approximieren). Wir als Menschen haben allerdings kaum eine Chance diese Formel in eine allgemeine Form zu bringen, die wir effektiv berechnen können. Die Berechenbarkeit von 104 beruht wiederum auf der Annahme von H_0: wir können einfach lokal jeweils Buchstaben durch weniger wahrscheinlichere Buchstaben ersetzen, und wir bekommen einen unwahrscheinlicheren Datensatz. Wir können also einfach Buchstaben nacheinander lokal ersetzen. Zusammen mit den Distributivgesetzen in der arithmetischen Entsprechung lässt sich das ganz gut ausrechnen.

10.3 Schwellentest in der Praxis

Im Schwellentest müssen wir H_0, H_1 komplett ausbuchstabieren. Hierzu müssen wir Parameter schätzen (wie oben), und bilden das likelihood Verhältnis

$$(105) \quad R(\mathfrak{T}) = \frac{P_0(\mathfrak{T})}{P_1(\mathfrak{T}}$$

Als nächstes können wir einfach unseren Schwellentest anwenden:

$$(106) \quad S_{0.05}(\mathfrak{T}) = \begin{cases} H_0, \text{ falls } R(\mathfrak{T}) \leq 0.05 \\ H_1 \text{ andernfalls.} \end{cases}$$

Beispiel durchrechen: $\mathfrak{T} = abababaababab$.

10.4 t-test in der Praxis

Der t-Test beruht darauf, dass wir Mittelwerte miteinander vergleichen, und zwar für

- Eine normalverteilte Gesamtheit und eine nach Annahme zufällig ausgewählte, normalverteilte Stichprobe daraus
- oder zwei nach Annahme zufällig ausgewählten normalverteilten Stichproben aus einer Gesamtheit.

In diesem Fall ist unser Datensatz ziemlich schwierig auf diese Art und Weise zu behandeln; es wäre besser einen Datensatz von reellen Zahlen zu haben, also:

$$\mathfrak{T}_2 = x_1 x_2 x_3 ... x_n, \text{ wobei } x_i \in \mathbb{R}$$

Wir können nun den Mittelwert dieses Satzes nehmen:

$$\mu(\mathfrak{T}_2) = \sum_{i=1}^{n} x_n \cdot \frac{1}{n} \tag{107}$$

also die Summe aller Werte geteilt durch die Anzahl der Werte. Nun nehmen wir eine zufällige Stichprobe; da nach Annahme von H_0 die Werte voneinander unabängig sind, ist

$$\mathfrak{S} = y_1 y_2 ... y_j : y_i = x_{i'}, x_{i'-1} \in [a, b]$$

Wir nehmen uns also die Folge aller Werte aus $\mathfrak{T}_2$, für die gilt: der Vorgängerwert in $\mathfrak{T}_2$ liegt im Intervall $[a, b]$, was ein beliebig gewähltes Intervall ist. Wir nehmen hier eher ein Intervall als einen Wert, sonst bekommen wir zuwenig Datenpunkte. Nun gilt: $\mathfrak{S}$ ist zufällig gewählt unter Annahme von H_0. Das bedeutet aber auch, wir sollten haben:

$$\mu(\mathfrak{S}) \approx \mu(\mathfrak{T}_2) \tag{108}$$

Wenn die beiden nun sehr unterschiedlich sind, dann spricht das gegen die Annahme, dass H_0 korrekt ist – wir haben also Evidenz, H_0 zurückzuweisen. Andersrum, wenn 109 erfüllt ist, bedeutet dass, das zumindest für $[a, b]$ nichts gegen H_0 spricht.

Allerdings gibt es eine wichtige Sache zu beachten: das setzt voraus dass die Daten in $\mathfrak{T}_2$ einigermaßen *normalverteilt* sind. Um das zu sehen, bedenke

man folgendes: nimm an, $\mathfrak{T}_2$ liegt eine Zipf-Verteilung zugrunde, d.h. es gibt sehr wenige Punkte, die sehr viel Masse auf sich vereinen. Nun kann es gut sein, dass rein zufällig $\mathfrak{S}$ diese (wenigen) Punkte nicht enthält, und daraus folgt natürlich:

(109) $\mu(\mathfrak{S}) \ll \mu(\mathfrak{T}_2)$

($\ll$ bedeutet: deutlich kleiner). Dasselbe kann auch bei beliebigen Stichproben gelten. Das bedeutet: einem t-Test sollte eine Normalverteilung zugrunde liegen.

Kehren wir zurück zu unserem Datensatz D. Das Problem in unserem Beispiel war folgendes: H_1, sobald sie ausbuchstabiert ist, ist viel zu spezifisch. Wir möchten eher eine allgemeinere Hypothese H_1, nämlich dass die Wahrscheinlichkeiten von einzelnen Worten abhängig voneinander sind. Hier können wir den sog. **t-test** benutzen, der

- für einen gegebenen Datensatz, und
- zwei Stichproben daraus

bestimmt, ob zwei Faktoren *unabhängig* voneinander sind. Bevor wir damit anfangen können, müssen wir zunächst unsere Daten etwas präparieren.

Zur Erinnerung: unsere beiden Hypothesen waren:

> H_0, alle Worte als Ereignisse sind unabhängig voneinander;
>
> H_1, die Wahrscheinlichkeit eines Wortauftretens ist abhängig vom vorhergehenden Wort.

Wir brauchen also, für jedes Wort v in unserem Text, eine ganze Reihe Parameter:

1. seine absolute Häufigkeit, geteilt durch die Anzahl der Worte (*token*) im Text (also relative Häufigkeit)
2. die Häufigkeit des Vorkommens von v nach einem gegebenen Wort w, geteilt durch die Häufigkeit von w, für jedes Wort w das im Text auftritt.

Wir kriegen also zu jedem Wort (eine Zeile in einer Tabelle) eine Reihe von Zahlen (Spalten in einer Tabelle).

- Wir nennen die Spalte mit den allgemeinen relativen Häufigkeiten S_1,
- die Spalte mit den relativen Häufigkeiten für Vorgänger w nennen wir S_w.

Es handelt sich also um Vektoren (Listen) von Zahlen.

Was wir als nächstes brauchen ist das Konzept des Mittelwertes: gegeben eine endliche Menge (oder Liste) von Zahlen

$$\mathbf{X} = \{x_1, ...x_n\} \subseteq \mathbb{R}$$

bezeichnen wir den **Mittelwert** von $\mathbf{X}$ mit

$$\mu(\mathbf{X}) = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Wir können nun beispielsweise, indem wir etwas liberal im Umgang mit Listen und Mengen sind, einfach

$$\mu(S_1),\ \mu(S_w) \text{ etc.}$$

berechnen. Wie machen wir das? Wir summieren alle Zahlen der Spalte auf, und dividieren durch die Anzahl der Zeilen darin. NB: die Anzahl der Zeilen ist die Anzahl der *types* in unserem Text. Wir dividieren also zunächst absolute Häufigkeiten durch Anzahl der token, addieren die Ergebnisse und dividieren durch Anzahl der types.

Was bedeutet dieser Wert? Er sagt uns, wie oft ein beliebiges Wort geteilt durch die Anzahl der Worte (*token*) im Text durchschnittlich auftritt. Wegen der arithmetischen Distributivgesetze können wir diese Transformation umkehren: nehmen wir an,

- unser Text $\mathfrak{T}$ enthält k *token* , und sei
- $\mu(S_1) = x$.
- Dann kommt ein beliebiges Wort durchschnittlich kx im Text vor.

Nehmen wir an, jedes Wort kommt nur einmal vor. k ist die Anzahl der *token*; sei $|L|$ die Anzahl der *types* im Text. Wir bekommen dann also als Wert

$$\mu(S_1) = \frac{1}{k},$$

da in diesem Fall $l = k$ ist. Im allgemeinen Fall lässt sich leicht zeigen, dass das Ergebnis immer

$$\mu(S_1) = x = \frac{1}{l} \tag{110}$$

lautet, also die Anzahl der token. Die Frage ist nur: Wie sind die Häufigkeiten verteilt? Auch hierfür haben wir eine gute Antwort: die Verteilung wird normalerweise eine Zipf-Verteilung sein.

Dasselbe können wir nun auch für die anderen Spalten machen; bsp. für das Wort w_1. Während für S_1 das Ergebnis in gewissem Sinne trivial war, ist es nun alles andere als trivial. Was wir dann bekommen ist

$$\mu(S_{w_1}),$$

d.i. die durchschnittliche Häufigkeit eines beliebigen Wortes nach w_1, geteilt durch die Anzahl der Vorkommen von w_1. Hier sehen wir, warum wir die Division machen:

- dadurch werden etwa $\mu(S_1)$ und $\mu(S_{w_1})$ *vergleichbar*.

Für $\mu(S_{w_1})$ gibt es allerdings etwas sehr wichtiges zu beachten: wir dürfen nicht durch die Gesamtlänge der Spalte dividieren, sondern nur durch die Anzahl von Kästchen, die einen positiven Eintrag haben. Warum? Weil wenn wir Dinge anfangen, Dinge zu berücksichtigen, die gar nicht vorkommen, dann müssten wir ja Äpfel und Birnen etc. berücksichtigen.

Was erwarten wir uns also, wenn wir diesen Wert berechnen? Nun, nehmen wir beispielsweise an, in $\mathfrak{T}$ folgt auf das Wort w_1 *immer* dass Wort w_2. Was wäre dann $\mu(S_{w_1})$? Wir bekommen dann tatsächlich den Wert

$$\mu(S_{w_1}) = 1 \tag{111}$$

denn wir haben als Summe aller Zahlen in 1, und da wir in S_{w_1} nur in einem Kästchen einen positiven Eintrag finden, dividieren wir durch 1.

Nehmen wir an wir machen diese Beobachtung. Das ist natürlich Evidenz gegen H_0. Aber ist diese Evidenz stark? Das hängt natürlich davon ab, wie häufig w_1 ist! Denn wenn es nur einmal vorkommt, dann war unsere Beobachtung trivial. Auf der anderen Seite, wenn w_1 sehr häufig ist, dann sollten wir erwarten, dass $\mu(S_{w_1})$ dem Wert $\mu(S_1)$ eher ähnlich ist. Wenn wir uns aber ein bestimmtes, häufiges Wort aussuchen, dann laufen wir natürlich Gefahr, durch diese Auswahl wiederum die Gültigkeit unserer Beobachtung

einzuschränken. Außerdem lassen wir den größten Teil unserer Information ungenutzt. Was können wir also tun?

Wir können uns helfen, indem wir folgendes machen: wir betrachten nicht nur $\mu(S_{w_1})$, sondern wir **generalisieren** die ganze Prozedur, so dass wir den Mittelwert über *alle* Mittelwerte bilden; wir berechnen also

$$\mu(\{\mu(S_w) : w \in \mathfrak{T}\}) \tag{112}$$

Wir betrachten also wiederum alle diese Werte, und mitteln über sie. Was sagt uns das? Das hängt wiederum davon ab, wie die Häufigkeiten verteilt sind: wenn jedes Wort nur einmal vorkommt im Text, dann wird auch dieser Wert sehr uninformativ sein. Wenn aber alle Worte sehr haüfig sind, dann würden wir folgendes erwarten:

$$\mu(\{\mu(S_w) : w \in \mathfrak{T}\} \approx \mu(S_1) = \frac{1}{l} \tag{113}$$

– unter Annahme der Nullhypothese!

Das Problem ist nun folgendes: wir haben gesagt dass wir wahrscheinlich eine Zipf-Verteilung haben, das bedeutet: die überwiegende Mehrheit der Worte taucht nur einmal auf. Es wird also auf jeden Fall einen verzerrten Wert geben, denn die Mehrzahl der Worte wird eben einen sehr hohen Wert haben. Wie kommen wir um dieses Problem herum? Der Trick ist:

> wir nehmen eine **Stichprobe** $\mathfrak{S}$ aus unseren Daten, die **normalverteilt** ist; d.h. mit wenig sehr seltenen und wenig sehr häufigen und vielen mittelhäufigen Wörtern.

Dann berechnen wir wiederum daraus den Mittelwert, und nun können wir die beiden Mittelwerte vergleichen.

Nun nehmen wir an, H_0 ist wahr. In diesem Fall sollten die beiden Mittelwerte in etwa gleich sein, also

$$\mu(\{\mu(S_w) : w \in \mathfrak{S}\} \approx \mu(S_1) \tag{114}$$

Das bedeutet, die Sicherheit, mit der wir wissen dass ein Nachfolger nach einem gewissen Vorgänger kommt, ist nicht wesentlich größer als die Sicherheit, das er überhaupt auftritt (sie wird natürlich immer größer sein, da wir immer gewisse Artefakte haben.

Das bedeutet also: je *weniger* der Wert

$$\mu(\{\mu(S_w) : w \in \mathfrak{T}\})$$

höher sein wird als

$$\mu(S_1),$$

desto *weniger* Evidenz haben wir gegen die Nullhypothese; umgekehrt, je mehr der Wert nach oben abweicht, desto eher können wir die Nullhypothese ablehnen.

Wir wissen natürlich immer noch nicht, ab wann wir H_0 ablehnen können; allerdings haben wir unsere Daten nun so zurecht gelegt, dass sie nur in *eine* Richtung abweichen, fallst H_0 falsch ist; alles was wir brauchen ist ein Schwellenwert, ab dem wir H_0 ablehnen. Um diesen Wert zu finden, brauchen wir natürlich zusätzliche Erwägungen.

Irgendwie bleibt das aber alles unbefriedigend, denn wir haben kein Maß dafür, inwieweit ein Wort für ein nachfolgendes Wort **informativ** ist. Das werden wir als nächstes betrachten.

11 Entropie, Kodierung, und Anwendungen

11.1 Definition

Das Konzept der Entropie formalisiert die *Unsicherheit* in einem System. Die Definition ist wie folgt: wir haben eine Wahrscheinlichkeitsfunktion P und ein Ereignis ω. Die Entropy von ω (nach P), geschrieben $H_P(\omega)$, ist

(115) $H_P(\omega) := P(\omega) \cdot -log(P(\omega))$

Die Entropie eines einzelnen Ereignisses ist normalerweise weniger interessant als die Entropie einer ganzen Verteilung P (über einen diskreten Raum Ω, geschrieben $H(P)$:

(116) $H(P) := -\sum_{\omega \in \Omega} P(\omega) log(P(\omega))$

Es ist leicht zu sehen dass das einfach die Summe der Entropie der Ereignisse ist; wir haben nur das minus ausgeklammert. Als Faustregel lässt sich sagen: in einem Raum mit n Ergebnissen ist die Entropie *maximal*, wenn alle Ereignisse die gleiche Wahrscheinlichkeit $1/n$ haben; sie wird minimal (geht gegen 0), falls es ein Ereignis gibt dessen Wahrscheinlichkeit gegen 1 geht. Das deckt sich mit unseren Intuitionen: je größer die Entropie, desto weniger Sicherheit haben wir, wie das Ergebnis sein wird. Z.B.: nehmen wir das Beispiel eines fairen Würfels; wir können die Entropie des zugehörigen Wahrscheinlichkeitsraumes wie folgt ausrechnen:

```
> x = 0 : 5
> for(i in 1 : 6){
+x[i] < −1/6 ∗ log(1/6)}
> sum(x)
[1] − 2.584963
```

(Wir verzichten darauf, die Entropie ins positive zu wenden). Wenn wir hingegen annehmen, 5 Seiten haben die Wahrscheinlichkeiten 1/10 und die 6 hat eine Wahrscheinlichkeit 1/2, dann bekommen wir:

```
> x = 0 : 5
> for(i in 1 : 5){
+x[i] < −1/10 ∗ log(1/10)}
> x[6] = 1/2 ∗ log(1/2)
> sum(x)
[1] − 2.160964
```

Andersrum gesagt: je größer die Entropie (einer Wahrscheinlichkeitsverteilung für ein Zufallsexperiment), desto größer der Informationsgewinn, der darin besteht das Ergebnis zu erfahren. Wichtig ist: Entropie ist immer unabhängig von den einzelnen Ergebnissen, es spielt also keine Rolle ob die 1 oder die 6 eine erhöhte Wahrscheinlichkeit hat. Alles was zählt ist eben die Ungewissheit; wir können das mit einem weiteren Versuch nachrechnen:

```
> x = 0 : 5
> for(iin1 : 3){
+x[i] < −1/10 ∗ log(1/10)}
> x[4] = (1/20) ∗ log(1/20)
> x[5] = (3/20) ∗ log(3/20)
> x[6] = 1/2 ∗ log(1/2)
> sum(x)
[1] − 1.777507
```

Die Entropie ist also weiter gesunken, denn wir haben die Wahrscheinlichkeiten weiter ungleich aufgeteilt zwischen 2 Ergebnissen: während also die Entropie für 1,2,3,6 gleich geblieben ist, ist sie für 4,5 lokal gesunken, also ist sie auch global gesunken. Man kann auch umgekehrt sagen: da die uniforme Wahrscheinlichkeitsverteilung für uns den *Mangel* an relevanter Information bezeichnet, gibt es die Korrelation

$$\text{maximale Entropie} \approx \text{maximale Unwissenheit}$$

Darauf basiert eine wichtige Methode der Wahrscheinlichkeitstheorie, die sog. Maximum Entropie Schätzung. Die basiert auf dem Grundsatz:

> In Ermangelung sicherer Information ist es besser, möglichst wenig Sicherheit anzunehmen, als falsche Sicherheit die es nicht gibt (es ist besser zu wissen dass man etwas nicht weiß)

Das bedeutet effektiv: wir sollten die Wahrscheinlichkeitsverteilung annehmen, die

1. mit unserem Wissen kompatibel ist,
2. ansonsten aber die Entropie maximiert.

Man definiert die Entropie auch oft für Zufallsvariablen:

$$H(X) := -\sum_{x \in X} P(X = x) log(P(X = x)) \tag{117}$$

11.2 Kodierungstheorie und Entropie

Der Zusammenhang von Entropie und Kodierung (z.B. im Alphabet $\{0,1\}$) beruht darauf, dass wir ein Symbol, was häufig ist ($\cong$ hohe Wahrscheinlichkeit) möglichst kurz kodieren, während relativ seltene Symbol eher lange Kodes bekommen. Auf diese Art sind unsere Kodes von Texten im Normalfall möglichst kurz.

Seien Σ, T zwei Alphabete. Ein Kode (von Σ in T) ist Paar

$$(\phi, X),$$

wobei

$$X \subseteq T^*, \text{ und } \phi : \Sigma \to X$$

eine Bijektion ist, so dass die homomorphe Erweiterung von

$$\phi : \Sigma^* \to X^*$$

weiterhin eine Bijektion ist.

Ein Kode ist **präfixfrei**, falls es kein $x, y \in X$ gibt so dass

$$xz = y, \text{ wobei } z \in T^+.$$

Wir sind in der Informatik meist in Kodes über $\{0,1\}^*$ interessiert, und wir möchten üblicherweise Alphabete kodieren, die mehr als zwei Buchstaben enthalten. Es stellt sich die Frage, wie man das am besten macht. Intuitiv ist unser Ziel: wir möchten, dass jede Kodierung eines Textes möglichst kurz wird. Das ist natürlich trivial, sofern wir nur die Buchstaben Σ haben. Aber nehmen wir an, wir haben eine Wahrscheinlichkeitsverteilung über Σ, und weiterhin, dass die Wahrscheinlichkeiten der Worte unabhänging voneinander sind. Das bedeutet:

- wenn ein Buchstabe sehr wahrscheinlich ist, dann wollen wir ihn kürzer kodieren,
- wenn er unwahrscheinlich ist, dann länger.

Sei $w \in \Sigma^*$. Wir bauen uns eine Zufallsvariable X, so dass $X(a) = |\phi(a)|$ (die Länge des Wortes). Was wir möchten ist: wir möchten den Erwartungswert von X möglichst klein machen. Wir haben

$$\mathcal{E}(X) = \sum_{a \in \Sigma^*} |\phi(a)| \cdot P(a) \tag{118}$$

Jeder Buchstabe im Ausgangsalphabet Σ hat Länge 1; er wird – nach Erwartungswert – in der Kodierung im Schnitt mit $\mathcal{E}(X)$ Symbolen ersetzt. Deswegen nennen wir die *Inversion*

$$\frac{1}{\mathcal{E}(X)} \text{ den } \textbf{Kompressionsfaktor}$$

der Kodierung. Ein wichtiger Punkt ist nun:

$\mathcal{E}(X)$ kann niemals kleiner sein als die Entropie $H(P)$.

Das bedeutet wir müssen jedes Symbol im Schnitt mit mindestens $H(P)$ zeichen kodieren.

Wir möchten im Allgemeinen den Erwartungswert minimieren, d.h. den Kompressionsfaktor maximieren. Es gibt einen einfachen Algorithmus, den sogenannten **Huffman code**, der folgendes liefert:

- Eingabe: ein beliebiges Alphabet Σ mit einer zugehörigen Wahrscheinlichkeitsfunktion $P : \Sigma \to [0, 1]$
- Ausgabe: eine Kodierung von Σ in $\{0, 1\}^*$ in einem Präfix-freien Kode mit maximalen Kompressionsfaktor (es gibt aber immer mehrere solcher Kodierungen).

Auch wenn das Thema nicht wirklich relevant ist, ist der Algorithmus ein Modell im Kleinen für das, was viele Lernalgorithmen machen.

Ein Beispiel Nehmen wir an, $\Sigma = \{a, b, c, d\}$, mit folgenden Wahrscheinlichkeiten (bzw. Häufigkeiten):

- $P(a) = 0.1$
- $P(b) = 0.2$
- $P(c) = 0.3$
- $P(d) = 0.4$

Wir fangen damit an, das Buchstabenpaar zu nehmen, das am seltensten vorkommt. Das ist natürlich

$$\{a, b\} \text{ mit } P(\{a, b\}) = 0.3.$$

Wir ersetzen nun

$$\{a, b\} \mapsto \{x_1\},$$

so dass unser neues Alphabet ist

$$\{x_1, c, d\}, \text{ wobei } P(x_1) = 0.3.$$

Nun machen wir ebenso weiter: im neuen Alphabet ist das Buchstabenpaar mit der geringsten Wahrscheinlichkeit $\{x_1, c\}$, also ersetzen wir

$$\{x_1, c\} \mapsto \{x_2\}$$

mit dem resultierenden Alphabet

$$\{x_2, d\}, \text{ wobei } P(x_2) = P(\{x_1, c\}) = 0.6.$$

Nun machen wir den Schritt ein letztes Mal: das resultierende Alphabet ist

$$\{x_3\} \text{ mit } P(x_3) = 1.$$

Nun "entpacken" wir das ganze wieder. Wir nehmen an, x_3 wird kodiert durch das leere Wort ϵ. ϵ steht dann aber eigentlich für 2 Buchstaben: x_2 und d. Das erste ist wahrscheinlicher, also kodieren wir x_2, indem wir eine 0 an unser Kodewort hängen, d mit einer 1. Nun steht x_2 (bzw. 1) wiederum für zwei Buchstaben, und wir bekommen $x_1 = 00, c = 01$ (in diesem Fall ist es egal, die Wahrscheinlichkeiten sind gleich). Nun dasselbe mit x_1 (bzw. 00); in diesem Fall bekommen wir 000 für b, 001 für a. Wir bekommen also:

- $\phi(a) = 001$
- $\phi(b) = 000$
- $\phi(c) = 01$
- $\phi(d) = 1$

Wir nehmen nun X wie oben, und bekommen:

(119) $\mathcal{E}(X_\phi) = 0.1 \cdot 3 + 0.2 \cdot 3 + 0.3 \cdot 2 + 0.4 \cdot 1 = 1.9$

Der Kompressionsfaktor ist also $\frac{1}{1.9}$. Natürlich kommt dasselbe raus, wenn wir im Kode einfach 0 und 1 vertauschen. Wenn wir das vergleichen mit dem folgenden Block-Kode

- $\chi(a) = 00$

- $\chi(b) = 01$
- $\chi(c) = 10$
- $\chi(d) = 11$

(der auch Präfix-frei ist), dann bekommen wir

(120) $\mathcal{E}(X_\chi) = 0.1 \cdot 2 + 0.2 \cdot 2 + 0.3 \cdot 2 + 0.4 \cdot 2 = (0.1 + 0.2 + 0.3 + 0.4)2 = 2$

Der Kompressionsfaktor beträgt also nur $\frac{1}{2}$.

Wie ist die Entropie für P?

(121)
$H(P) = -(0.1log_2(0.1)+0.2log_2(0.2)+0.3log_2(0.3)+0.4log_2(0.4)) = 1.846439$

Nehmen wir an, dagegen an dass

$$P'(a) = \ldots = P'(d) = 0.25$$

ändert sich die Lage: in diesem Fall ist natürlich χ die optimale Kodierung. Die Entropie ändert sich wie folgt:

(122) $H(P') = -(4 \cdot (0.25log_2(0.25)) = 2$

Die Entropie ist größer, daher wird auch die Kompressionsrate schlechter sein. Am Ende gilt:

> Sei P eine Wahrscheinlichkeitsverteilung über Σ, ϕ ein Kode über $\{0, 1\}$. Dann kann der Kompressionsfaktor von ϕ niemals grösser sein als $\frac{1}{H(P)}$, berechnet zur Basis 2.

11.3 Bedingte Entropie

Die bedingte Entropie von zwei Variablen (über demselben Wahrscheinlichkeitsraum) ist wie folgt definiert (hier bedeutet $y \in Y$ soviel wie: y ist ein Wert, den Y annehmen kann):

$$H(X|Y) = \sum_{y \in Y} P(Y = y) H(X|Y = y) \tag{123}$$

Wenn wir diese Definition auflösen, bekommen wir:

$$H(X|Y) = \sum_{x \in X, y \in Y} P(X^{-1}(x) \cap Y^{-1}(y)) log \left(\frac{P(X^{-1}(x) \cap Y^{-1}(y))}{P(Y^{-1}(y))} \right) \tag{124}$$

Die bedingte Entropie ist also ein Maß dafür, wie stark die Werte einer Zufallsvariable Y die Werte einer Zufallsvariable X festlegen. Wenn der Wert von X durch den Wert von Y – egal wie er ist – immer festgelegt ist, dann ist

$$H(X|Y) = 0 \tag{125}$$

insbesondere also:

$$H(X|X) = 0 \tag{126}$$

Umgekehrt, falls der Wert von Y keinerlei Einfluss hat auf die Wahrscheinlichkeitsverteilung des Wertes von X, dann haben wir

$$H(X|Y) = H(X) \tag{127}$$

Es ist klar dass das hier nur für diskrete Wahrscheinlichkeitsräume funktionieren kann; in kontinuierlichen Räumen funktionieren diese Dinge etwas anders.

Es gibt auch eine Kettenregel für bedingte Entropie:

$$H(X|Y) = H(\langle X, Y \rangle) - H(Y) \tag{128}$$

wobei $\langle X, Y \rangle$ eine neue Variable ist, mit

$$P(\langle X, Y \rangle = (x, y)) = P(X = x, Y = y) = P(X^{-1}(x) \cap Y^{-1}(y)) \tag{129}$$

Wir nehmen also die Entropie der **Verbundverteilung**, und ziehen die Entropie von $H(Y)$ ab.

11.4 Kullback-Leibler-Divergenz

Die KL-Divergenz ist eine andere Art zu messen, wie ähnlich sich zwei Wahrscheinlichkeitsverteilungen P und Q sind. Die Definition ist wie folgt:

$$D_{KL}(P\|Q) = \sum_{\omega\in\Omega} P(\omega) log \frac{P(\omega)}{Q(\omega)} \tag{130}$$

An dieser Definition kann man ablesen:

1. $D_{KL}(P\|Q) = 0$ gdw. für alle $\omega \in \Omega$ gilt: $P(\omega) = Q(\omega)$; denn $log(1) = 0$.
2. In allen anderen Fällen ist $D_{KL}(P\|Q) > 0$ (das ist nicht wirklich leicht zu sehen).
3. $D_{KL}(P\|Q) \neq D_{KL}(Q\|P)$, d.h. wir haben ein asymmetrisches Maß.
4. Man kann es jedoch symmetrisch machen auf folgende Art und Weise:

$$D_2(P\|Q) = D_{KL}(P\|Q) + D_{KL}(Q\|P) = D_2(Q\|P)$$

Sie gibt uns also ein Maß dafür, wie weit Q von P **entfernt** ist. Dadurch unterscheidet sie sich konzeptuell von $H(X|Y)$, dass bestimmt wie stark X von Y *determiniert* wird.

Man bezeichnet $D_{KL}(P\|Q)$ auch als den **Informationsgewinn**, den man mit P gegenüber Q erzielt. Wenn wir z.B. das obige Kodierungsbeispiel fortführen, dann sagt uns $D_{KL}(P\|Q)$, wieviel Platz wir (im Durchschnitt) verschwenden, wenn wir eine Kodierung auf Q basieren, während die zugrundeliegende Wahrscheinlichkeitsverteilung P ist.

Dementsprechen nutzt man $D_{KL}(P\|Q)$ oft im Kontext, wo P die tatsächliche Verteilung ist, Q unser Modell, dass wir geschätzt haben.

Aufgabe 7

Abgabe bis zum 6.6.2017 *vor dem Seminar*, egal ob digital/analog und auf welchem Weg.

Wir nehmen eine diskrete Wahrscheinlichkeitsverteilung P über das kartesische Produkt $\Omega = \{a, b\} \times \{1, 2, 3\}$, mit

$P(a,1) = 0.4$
$P(a,2) = 0.1$
$P(a,3) = 0.05$
$P(b,1) = 0.05$
$P(b,2) = 0.1$
$P(b,3) = 0.3$

Wir definieren 2 Zufallsvariablen $X_1(x,y) = x$, $X_2(x,y) = y$; so ist z.B. $P(X_2 = 1) = P(\{(a,1),(b,1)\})$ etc.

1. Berechnen Sie $H(X_1)$, $H(X_2)$.

2. Berechnen Sie $H(X_2|X_1)$. Was bedeutet das Erbegnis?

12 Wahrscheinlichkeiten schätzen

12.1 Die Likelihood-Funktion

Es gibt in der Stochastik/Statistik eine Unterscheidung zwischen Wahrscheinlichkeit (*probability*) und Likelihood, die man sich etwas schwierig klarmacht, da im Deutschen (und der englischen Umgangssprache) beide Begriffe zusammenfallen. In gewissem Sinne ist likelihood aber das Gegenteil (oder Gegenstück) zu Wahrscheinlichkeit. Intuitiv gesagt können wir von Wahrscheinlichkeit sprechen, wenn wir die zugrundeliegenden *Parameter* eines Experimentes kennen. Mit Parameter bezeichnet der Statistiker das, was der Stochastiker als Wahrscheinlichkeitsfunktion bezeichnet; die Parameter zu kennen bedeutet also: die zugrundeliegenden Wahrscheinlichkeiten zu kennen. Beispielsweise: wenn ich weiß dass eine Münze fair ist, als die Parameter des Experimentes kenne, kann ich fragen: was ist die *Wahrscheinlichkeit*, dass ich dreimal Zahl werfe? Wahrscheinlichkeit in diesem engeren Sinne bezeichnet also die Wahrscheinlichkeit eines Ereignisses, gegeben einen zugrundeliegenden, *bekannten* (oder als bekannt angenommenen) Wahrscheinlichkeitsraum. Wahrscheinlichkeit in diesem engeren Sinn haben wir ausführlich behandelt. Wenn wir das Ereigniss 3-mal Zahl als ω bezeichen, θ als die Wahrscheinlichkeit von Zahl im einfachen Bernoulli-Raum, P_θ als die Wahrscheinlichkeit im Produktraum, dann ist die Lösung $P_\theta = \theta^3$.

Wenn man das Beispiel verallgemeinert, dann ist die Wahrscheinlichkeit also eine Funktion, die jedem Ergebniss (jeder Beobachtung) einen Wert in $[0, 1]$ zuweist.

Likelihood bezeichnet dagegen die Plausibilität von zugrundeliegenden Wahrscheinlichkeitsräumen (sprich: Parametern), gegeben eine Reihe von Beobachtungen die wir gemacht haben. Beispielsweise: wir werfen eine Münzen 100mal, und werfen immer Zahl (nennen wir diese Beobachtung wieder ω. Was ist die Plausibilität dafür, dass der Würfel fair ist? Allgemeiner: was ist die Plausibilität für beliebige Parameter (sprich: zugrundeliegende Münzwahrscheinlichkeiten) gegeben ω? Wir haben auch ein solches Problem bereits einmal behandelt (siehe die 3. Sitzung). Dort haben wir versucht, zugrundeliegenden Parametern Wahrscheinlichkeiten zuzuweisen, und haben dabei gesehen, dass man das nicht ohne weitere Annahmen lösen kann: wir können zwar den Satz von Bayes benutzen um das Problem anzugehen, aber um es letztendlich zu lösen, brauchen wir einige zusätzliche Annahmen, und wir werden immer nur einen beschränkten Raum von Hypothesen

zulassen.

Eine andere Lösung ist die, dass man eben nicht Wahrscheinlichkeiten von Parametern sucht, sondern sich auf **Likelihood** beschränkt. Was wir nämlich machen können ist folgendes. Sei Θ die Menge aller möglichen Parameter für eine gegebene Beobachtung ω (also alle Wahrscheinlichkeitsräume, die zu dem Experiment passen). Θ ist also eine Menge von Wahrscheinlichkeitsräumen. Wir bekommen eine Funktion

$$L_\omega : \Theta \to [0, 1],$$

wobei

$$L_\omega(\theta) = P_\theta(\omega) \tag{131}$$

L_ω gibt uns also für jeden Parameter θ an, wie wahrscheinlich ω ist unter der Annahme dass der zugrundeliegende Parameter θ ist. L_ω ist die **Likelihoodfunktion**. Hier wird klar, warum wir hier nicht von Wahrscheinlichkeiten sprechen sollten: der Wert $L_\omega(\theta)$ gibt uns *nicht* die Wahrscheinlichkeit von θ; insbesondere gilt im allgemeinen Fall:

$$\sum_{\theta \in \Theta} L_\omega(\theta) \neq 1 \tag{132}$$

(*Aufgabe*: zeigen Sie dass mit einem Beispiel!) Wir können hier also nicht von Wahrscheinlichkeiten sprechen. Was uns $L_\omega(\theta)$ uns gibt ist Wahrscheinlichkeit von ω gegeben θ, also $P_\theta(\omega)$. Das ist eben qua Definition die *Likelihood* von θ gegeben ω, $L_\omega(\theta)$; wir können das mit Plausibilität übersetzen.

12.2 Maximum Likelihood Schätzung I

Warum ist Likelihood interessant, wenn sie uns am Ende nichts sagt, was wir nicht schon aus der Wahrscheinlichkeitsfunktion P erfahren? Der Grund ist folgender:

$$L_\omega(\theta_1) = P_{\theta_1}(\omega) \tag{133}$$

sagt uns nichts über die Wahrscheinlichkeit von θ_1. Aber: Nehmen wir an, wir haben zwei mögliche zugrundeliegende Parameter θ_1, θ_2. Wir können nun deren Likelihood berechnen. Falls wir nun haben

$$L_\omega(\theta_1) \leq L_\omega(\theta_2), \tag{134}$$

dann sagt uns das sehr wohl etwas:

> Das Ergebnis ω unter der Annahme der Parameter in θ_2 wahrscheinlicher ist als unter der Annahme der Parameter θ_1.

Und daraus folgt: gegeben ω ist θ_2 wahrscheinlicher als θ_1 – wenn beide Parameter *apriori* gleich wahrscheinlich sind. Und das ist im Prinzip alles was wir wissen möchten: normalerweise interessiert uns nicht die genaue Wahrscheinlichkeit eines Parameters (im Normalfall: einer wissenschaftlichen Hypothese), uns interessiert was die beste Hypothese ist. Warum können wir das sagen? Die Korrelation von Lkelihood eines Parameters und seiner Wahrscheinlichkeit lässt sich aus dem Satz von Bayes herleiten. Wir schreiben

$$P_\theta(\omega) := P(\omega|\theta) = L_\omega(\theta)$$

Nun sei also $P(\omega|\theta_1) \leq P(\omega|\theta_2)$. Nach Bayes Theorem gilt:

$$\begin{aligned} & P(\theta_i|\omega) = P(\omega|\theta_i) \cdot \frac{P(\theta_i)}{P(\omega)} \\ \Leftrightarrow\ & P(\theta_i|\omega)P(\omega) = P(\omega|\theta_i)P(\theta_i) \end{aligned} \tag{135}$$

Nachdem $P(\omega)$ immer gleich ist für θ_1, θ_2 etc, spielt das für uns keine Rolle, und wir können es getrost weglassen. Wir haben daher:

$$P(\theta_i|\omega) \sim P(\omega|\theta_i)P(\theta_i), \tag{136}$$

wobei wir mit $\sim$ eine lineare Korrelation meinen: je größer der eine Term, desto größer der andere. Jetzt kommen wir vorerst nicht weiter, denn wir müssen immer noch die *a priori* Wahrscheinlichkeit der Parameter $P(\theta_i)$ berücksichtigen. Wir müssen also die Annahme machen, dass alle Parameter *a priori* gleich wahrscheinlich sind, was in vielen, jedoch nicht in allen Kontexten sinnvoll ist. (Z.B. wenn wir eine Münze finden, die absolut normal aussieht werden wir es für viel wahrscheinlicher halten, dass sie fair ist, als das sie eine starke Tendenz hat.) Dann fällt also auch der Term $P(\theta_i)$ weg (da er für alle $i = 1, i = 2...$ gleich ist), und wir haben:

$$P(\theta_i|\omega) \sim P(\omega|\theta_i). \tag{137}$$

Das ist genau was wir zeigen wollten: je größer $P(\omega|\theta_i) = L_\omega(\theta)$, desto größer ist $P(\theta_i|\omega)$, die Wahrscheinlichkeit der Parameter gegeben unsere Beobachtungen. Insbesondere gilt also:

(138) $L_\omega(\theta_1) \leq L_\omega(\theta_2)$

daher

(139) $P(\theta_1|\omega) \leq P(\theta_2|\omega)$

– natürlich nur unter der Annahme, dass alle Parameter *a priori* gleich wahrscheinlich sind.

Das führt uns zu der wichtigen Methode der Maximum Likelihood Schätzung. Wenn wir den Hypothesenraum Θ betrachten, dann haben wir natürlich mehr als zwei Hypothesen darin; genauer gesagt, im Normalfall werden wir *kontinuierlich viele* Parameter haben. "Kontinuierlich" bedeutet: "so viele, wie es reelle Zahlen gibt", ebenso wie abzählbar bedeutet: soviele wie die natürlichen Zahlen. Wir können uns also unmöglich hinsetzen und alle möglichen Parameter prüfen. Wir können also mittels Likelihood die Plausibilität von Hypothesen prüfen. Das nächste Problem ist: es gibt viel zu viele Hypothesen, als das wir sie alle prüfen könnten

Um das nächste Problem zu lösen brauchen wir zunächst etwas Notation. Sei

$$f : M \to \mathbb{R}$$

eine Funktion von einer beliebigen Menge in die reellen Zahlen (eigentlich reicht es schon, wenn die Menge linear geordnet ist, aber für uns spielt das keine Rolle). Dann definieren wir

(140) $argmax_{m \in M} f := \{m : \forall m' \in M,\ f(m') \leq f(m)\}$

D.h. argmax(f) liefert uns die $m \in M$, für die $f(m)$ maximal ist. Z.B.

(141) $argmax_{x \in \mathbb{R}}(-(x^2)) = 0$

da $f(x) = -(x^2)$ für $x = 0$ seinen größten Wert annimmt. $argmax_{x \in \mathbb{R}}(x^2)$ ist nicht definiert, da es für $f(x) = x^2$ keinen maximalen Wert $x \in \mathbb{R}$ gibt. $argmax(f)$ ist also nur definiert, wenn f *nach oben beschränkt* ist.

Die Maximum Likelihood Schätzung ist nun einfach die Methode, für ω und Θ den Parameter

(142) $argmax_{\theta \in \Theta} L_\omega(\theta)$

zu finden. Wie löst man dieses Problem? Nehmen wir an, unsere Beobachtungen sind binär, d.h. wir haben zwei mögliche Ergebnisse, und unsere Beobachtung ist eine Sequenz dieser Ergebnisse; nach unserer Konvention schreiben wir $\omega \in \{0,1\}^n$. In diesem Fall ist Θ, unsere möglichen Parameter, eine Menge von Bernoulli-Räumen; und weil jeder Bernoulli-Raum ein möglicher Parameter für unsere Beobachtung ist, ist Θ (bis auf Isomorphie) die Menge *aller* Bernoulli-Räume (bis auf Isomorphie bedeutet: wir haben alle, wenn wir erlauben die beiden Elemente in $\Omega = \{0,1\}$ beliebig anders zu benennen). Jeder Bernoulli-Raum ist (wieder bis auf Isomorphie) eindeutig charakterisiert durch $p := P(1)$; sobald dieser Wert gegeben ist, stehen alle anderen Dinge fest.

Das wiederum bedeutet: wir können jedem $\theta \in \Theta$ eine Zahl $p_\theta \in [0,1]$ zuweisen. In diesem Fall können wir also Likelihood-Funktion

$$L_\omega : \Theta \to \mathbb{R}$$

auffassen als eine Funktion

$$L_\omega : \mathbb{R} \to \mathbb{R}.$$

Es lässt sich zeigen, dass diese Funktion stetig und differenzierbar ist (im Sinne der Analysis). Daraus wiederum folgt, dass wir die Maxima mit den klassischen Mitteln der Analysis bestimmen können (erste Ableitung gleich 0 setzen, prüfen ob es ein Extremwert ist). In diesem Fall lässt sich das Problem also lösen.

Was ist, wenn unsere Beobachtungen nicht einem Bernoulli-Raum entsprechen? Wenn wir beispielsweise einen Würfel 10mal werfen? Um in diesem Fall eine Maximum Likelihood Schätzung vornehmen zu können, müssen wir diesen Raum vereinfachen: für jedes der 6 möglichen Ergebnisse in Ω partitionieren wir die Menge der Ergebnisse in den zwei Ereignisse: für ω ein Ergebnis nehmen wir die Partition $\{\{\omega\}, \Omega - \omega\}$. So haben wir wiederum einen Bernoulli-Raum, wobei ein nicht-Bernoulli Experiment in eine Reihe von Bernoulli-Experimenten aufgeteilt wird.

12.3 Ein Beispiel

Wir werden nun ein einfaches Beispiel aus der statistischen Sprachverarbeitung betrachten. Nehmen wir an, wir betrachten ein Korpus mit 1.000.000 Wörtern, und finden darin 60mal das Wort `Hase`. Was uns interessiert ist die Wahrscheinlichkeit, mit der das Wort `Hase` in einem beliebigen Text auftritt. Wir möchten nun die MLS-Methode dafür anwenden. Wie machen wir das? Wir benennen

$$p = P(\texttt{Hase})$$

die Wahrscheinlichkeit des Wortes `Hase`; wir haben $q = 1 - p$, also einen Bernoulli-Raum. ω ist die Beobachtung die wir gemacht haben: dass nämlich in einem Text von 1.000.000 Wörtern 60 mal `Hase` vorkommt.

Was ist unsere Likelihood-funktion? Hier können wir unser wissen über Binomialverteilungen nutzen, und bekommen

$$L_\omega(\theta) = P_\theta(\omega) = L_\omega(p_\theta) = \binom{1.000.000}{60} p_\theta^{60} \cdot (1 - p_\theta)^{1.000.000-60} \tag{143}$$

Wenn uns nur das Maximum interessiert, können wir den Term $\binom{1.000.000}{60}$ außer Betracht lassen; wir suchen also, etwas allgemeiner ausgedrückt,

$$argmax_{p\in[0,1]}(p^n \cdot (1-p)^{m-n}), \text{ for } m \geq n \tag{144}$$

Das Ergebnis ist – wenig überraschend:

$$argmax_{p\in[0,1]}(p^n \cdot (1-p)^{m-n}) = \frac{n}{m} \tag{145}$$

In diesem einfachen Beispiel sagt uns also die MLS, dass die Wahrscheinlichkeitstheorie mit unseren Intuitionen über die Korrelation von Frequenz Wahrscheinlichkeit übereinstimmt. Das heißt natürlich nicht, dass $\frac{60}{1.000.000}$ die beste Schätzung der Wahrscheinlichkeit von `Hase` in einem beliebigen Text ist; aber es sagt uns dass es die plausibelste Schätzung ist gegeben die Beobachtung die wir gemacht haben.

12.4 Definitionen

Wir haben also folgende Definitionen:

Definition 11 *Sei Ω ein Bernoulli-Raum. Eine Schätzung ist eine Funktion $S_n : \Omega_n \to \Theta$, wobei Θ die Menge der möglichen Parameter ist.*

Sei $\Omega = \{0, 1\}$. Wir bezeichen, für $\vec{\omega} = \langle \omega_1, ..., \omega_n \rangle$, $f_1(\vec{\omega}) = \sum_{i=1}^{n} \omega_i$, und $f_0(\vec{\omega}) = n - \sum_{i=1}^{n} \omega_i$.

Definition 12 *Die Maximum-Likelihood Schätzung für $P(1)$ gegeben Ω^n ist die Funktion*

$$MLS_n(\omega) := \frac{f_1(\omega)}{n}$$

Der entscheidende Punkt ist der folgende, den wir bereits oben angedeutet, wenn auch nicht wirklich bewiesen haben:

Satz 13 *Für jeden Bernoulli-Raum Ω und alle zugrundeliegenden Wahrscheinlichkeiten θ gilt: für $\omega \in \Omega^n$, $MLS_n(\omega) = argmax_{\theta \in \Theta} P(\omega|\theta)$; anders gesagt, unter der Annahme, dass alle Parameter $\theta \in \Theta$ gleich wahrscheinlich sind, ist $P(MLS_n(\omega)|\omega)$ die wahrscheinlichste Hypothese.*

Das ist der Grund warum MLS_n die Maximum-Likelihood Schätzung genannt wird. Neben einer ganzen Reihe positiver Eigenschaften hat sie vor allen Dingen eine: sie ist sehr einfach zu berechnen.

13 Markov-Ketten

13.1 Vorgeplänkel

Markov-Ketten sind stochastische Prozesse, bei denen die Wahrscheinlichkeiten eines Ergebnisses von einer begrenzten Reihe von vorherigen Ergebnissen abhängen. Man spricht auch von Markov-Prozessen, wobei dieser Begriff eher für kontinuierliche Prozesse verwendet wird, der Begriff Markov-Kette eher für diskrete Prozesse. Wir werden uns hier ausschlielich mit **diskreten Prozessen** beschäftigen. Markov-Prozesse sind diskret, wenn sie über eine diskrete Kette von Ereignissen definiert sind. Eine Kette ist eine lineare Ordnung $(M, <)$ mit $<\subseteq M \times M$, wie etwas die natürlichen Zahlen, eine beliebige Teilmenge davon, die rationalen, reellen Zahlen etc., geordnet nach "ist größer als".

Definition 14 *Eine Kette ist **diskret**, falls es für jedes $m \in M$, für dass es ein n gibt, so dass $n < m$, es auch ein n' gibt so dass für alle $o < m$ gilt: $o < n'$ oder $o = n'$.*

n' ist dann der unmittelbar Vorgänger von m. $(\mathbb{N}, <)$ ist diskret, wobei der unmittelbare Vorgänger von m, für $m \geq 2$, $m-1$ ist. Jede endliche Kette ist diskret. Die Ketten $(\mathbb{Q}, <)$ und $(\mathbb{R}, <)$ sind nicht diskret: denn was ist die größte rationale (reelle) Zahl, die echt kleiner als 2 ist?

Nehmen wir wieder einmal die Münze. Wie ist die Wahrscheinlichkeit dafür, mit 10 Würfen mindestens dreimal Zahl zu werfen? Diese Wahrscheinlichkeit kennen wir (unter der Annahme dass die Münze fair ist). Wir werden jetzt aber noch zusätzliche Informationen berücksichtigen: wie ist diese Wahrscheinlichkeit, gegeben dass wir mit den ersten 9 Würfen nur *einmal* Zahl geworfen haben? Offensichtlich 0, ganz unabhängig von der Münze! Umgekehrt, wie ist die Wahrscheinlichkeit, gegeben dass wir mit den ersten 9 Würfen bereits 6mal Zahl geworfen haben? Offensichtlich 1, ebenfalls unabhängig von den zugrundeliegenden Wahrscheinlichkeiten. Ein dritter Fall ist die Wahrscheinlichkeit unter der Annahme dass wir mit 9 Würfen 2mal Zahl geworfen haben - hier hängt die Wahrscheinlichkeit von der Münze selber ab, und ist 0.5 im Fall einer fairen Münze.

Wir verallgemeinern das Beispiel. Sei $\mathcal{P}$ ein Bernoulli-Raum mit $p = P(1)$, X_n eine (Reihe von) Zufallsvariable(n) mit

$$X_n(\langle \omega_1, ..., \omega_n \rangle) = \sum_{i=1}^{n} \omega_i,$$

für beliebige $n \in \mathbb{N}$. Wir bezeichnen mit S_n ein Ereignis von n Würfen (mit irgendwelchen Ergebnissen), S_{n-1} das Ereignis von n_1 Würfen etc. Uns interessiert die Wahrscheinlichkeit von $P(X_n = r)$, also r-mal Zahl von n Würfen. Diese Wahrscheinlichkeit von $X_n(S_n) = r$ hängt nun offensichtlich ab von $X_{n-1}(S_{n-1})$, wie wir oben gesehen haben; falls $X_{n-1}(S_{n-1}) = r - 1$, dann ist $P(X_n = r) = p$, falls $X_{n-1}(S_{n-1}) = r$, dann ist $P(X_n = r) = 1 - p$, und in allen anderen Fällen ist $P(X_n = r) = 0$.

NB: was hier wichtig ist $X_{n-1}(S_{n-1})$; alle vorigen Ergebnisse, also $X_{n-j}(S_{n-j})$ für $1 < j < n$ sind vollkommen unerheblich. Hier handelt es sich um ein typisches Beispiel von einer Markov Kette *erster Ordnung* - die Verteilung für S_n hängt ausschließlich an S_{n-1} und p; die Zukunft und die fernere Vergangenheit spielen überhaupt keine Rolle.

13.2 Markov-Ketten: Definition

Wir haben nun das wichtigste Merkmal von Markov-Ketten beschrieben: Ereignisse beeinflussen die Wahrscheinlichkeiten von Nachfolgeereignissen, aber nur einem begrenzten Abstand. Wir werden nun eine formale Definition liefern.

Definition 15 *Sei $S_n : n \in \mathbb{N}$ eine (endliche oder unendliche) Reihe von Ergebnissen, $X_n : n \in \mathbb{N}$ Reihe von Zufallsvariablen auf S_n. $X_n : n \in \mathbb{N}$ ist eine Markov-Kette $m - ter$ Ordnung, falls $P(X_{t+1} = x_{t+1} | X_t = x_t, X_{t-1} = x_{t-1}, ..., X_1 = x_1) = P(X_{t+1} = x_{t+1} | X_t = x_t, X_{t-1} = x_{t-1}, ..., X_{t-m} = x_{t-m})$*

Das bedeutet soviel wie: nur die letzten m Ergebnisse beeinflussen die Wahrscheinlichkeit von $P(X_t = x_t)$, alle anderen sind irrelevant. Beachten Sie, dass Ketten von Ereignissen der Form: der n-te Wurf ist Zahl, der $n+1$-te Wurf ist Kopf etc., wo alle Ereignisse unabhängig sind, Markov Ketten 0-ter Ordnung sind.

Unser obiges Beispiel war in gewissem Sinne irreführend für die Anwendung von Markov-Prozessen, denn in diesem Beispiel sind alle Ereignisse voneinander unabhängig. Markov-Ketten werden hingegen gerade dann benutzt, wenn diese Prämisse *nicht* gegeben ist, d.h. wenn wir nicht annehmen können dass die vorherigen Ergebnisse die nachfolgenden nicht beeinflussen. Ein besseres Beispiel wäre ein Spiel wie "Schiffe versenken": hier können Sie die von der Wahrscheinlichkeit sprechen, dass ein Spieler ein gewisses Feld wählt; aber natürlich nicht unabhängig von seinen bieherigen Entscheidungen: denn er wird gewisse strategisch interessante Felder wählen.

Auf der anderen Seite, wenn Sie einen Zustand betrachten als die Spielsituation nach einem bestimmten Zug, dann ist es allein der letzte Zustand, der die Wahrscheinlichkeit beeinflusst (d.h. die Spielsitation nach dem letzten Zug). Die Reihe der vorherigen Informationen hingegen wird völlig irrelevant, denn alle relevante Information steckt ja bereits im letzten Zustand (hier lassen wir natürlich Faktoren wie eine bestimmte Strategie des Spielers oder Gewohnheit außen vor).

Wir haben von Zuständen geredet, wie es bei Markov-Ketten üblich ist. Als Zustand betrachten wir Objekte der Form $X_n(S_n)$, also Bilder der Zufallsvariablen. Beachten Sie dass wir hier Zufallsvariablen in einem weiteren Sinne benutzen als gewöhnlich; insbesondere sind, in unserem letzten Beispiel, die Zustände *nicht* die Züge, sondern die Spielsituationen, die daraus resultieren! Andernfalls haben wir sicher keine Markov-Kette erster Ordnung, und im allgemeineren Fall nicht einmal eine Markov-Kette! Hieraus wird hoffentlich deutlich, warum wir von Zuständen sprechen.

13.3 (Teile der) Sprache als Markov-Prozess

Wir haben bereits in einigen Beispielen Texte behandelt, in denen gewisse Buchstaben eine gewisse Wahrscheinlichkeit des Auftretens haben. Wenn wir natürliche Sprachen wie Deutsch betrachten, dann macht es wenig Sinn mit solchen Wahrscheinlichkeiten zu rechnen: denn die entscheidende Voraussetzung für die Methode, mit der wir Wahrscheinlichkeiten von Worten berechnet haben, ist das Buchstaben in einem (deutschen) Text zufällig verteilt sind. Diese Annahme ist natürlich abwegig, wie wir bereits auf der Ebene der sog. Phonotaktik feststellen: `kle` ist eine mögliche Buchstabenfolge, während eine Folge wie `klp` nicht möglich ist als deutsche Buchstabenfolge. Diese einfache Regelmässigkeit ist eine von vielen, und wir können sie ganz einfach wie folgt erfassen:

(146) $P(\texttt{p}|\texttt{kl}) = 0$

Hier ist `x` eine Kurzschreibweise für das Ereigniss: "der n-te Buchstabe im Text ist x; wir können das notieren als $n = \texttt{x}$; und

(147) $P(\texttt{p}|\texttt{kl})$

ist eine Kurzform für:

(148) $P(n = \texttt{p}|n-1 = \texttt{l}, n-2 = \texttt{k})$

Wir können also phonotaktische Regeln als Markov-Kette kodieren.

Die große Frage ist jedoch: ist die Verteilung von Buchstaben in einem Text tatsächlich ein Markov Prozess? Diese Frage lässt sich natürlich, wie alle empirischen Fragen über Wahrscheinlichkeiten, nur näherungsweise betrachten. Wenn wir zunächst phonotaktische Beschränkungen betrachten, dann stellen wir fest dass es solche Beschränkungen nur im Rahmen einer Silbe gibt. (Das gilt wohlgemerkt nicht für alle Sprachen; es gibt phonotaktische Phänomene wie Vokalharmonie die über die Silbengrenze hinweg gelten.) Die mögliche Größe von Silben ist beschränkt: die (meines Wissens) Längste deutsche Silbe ist das Wort `springst` mit 8 Buchstaben. Und eigentlich können wir diese Zahl noch weiter verringern, denn die Buchstaben in Silbenauftakt (*onset*) und Coda haben keinen Einfluss aufeinander; aber darauf soll es uns nicht ankommen.

Wir können also aus diesem Grund behaupten, dass die Verteilung von Buchstaben in deutschen Texten mit einer Markov-Kette modelliert werden kann; und zumindest wird uns jeder Recht geben, dass dieses Modell besser ist als das krude Zufallsmodell. Wenn wir allerdings annehmen, dass auch Faktoren Syntax und Semantik eine Rolle spielen für die Verteilung von Buchstaben, dann ist unser Modell natürlich völlig inadequat.

13.4 Likelihood und Parameter-Schätzung bei für Markov-Ketten

Wir haben gesehen, wie wir effektiv Parameter aus Daten schätzen können mit der Maximum-Likelihood Schätzung. Wenn wir also annehmen, dass Buchstaben in Texten zufällig verteilt sind, dann können wir, gegeben einen Text der groß genug ist um einigermaßen zuverlässig zu sein, effektiv schätzen was die zugrundeliegenden Parameter sind. Können wir diese Methode effektiv erweitern für Markov-Ketten?

Sei $\mathfrak{T}$ ein Text. Wir bezeichnen mit $\mathtt{a}(\mathfrak{T})$ die Anzahl von `a`s in $\mathfrak{T}$ etc., mit $|\mathfrak{T}|$ bezeichnen wir die Anzahl der Zeichen in $\mathfrak{T}$. Wir haben gesehen dass wir die Maximum Likelihood für $P(\mathtt{a})$ effktiv schätzen können mit $\frac{\mathtt{a}(\mathfrak{T})}{|\mathfrak{T}|}$. Wir erweitern unsere Schätzung nun für Markov-Ketten. Einfachheit halber nehmen wir zunächst eine Markov-Kette erster Ordnung als Modell, obwohl das natürlich inadäquat ist, wie wir gesehen haben. Zunächst machen wir folgende Annahme: wir erweitern unser Alphabet Σ, das bereits das Leerzeichen enthält, um die Zeichen $\#_a, \#_e \notin \Sigma$. Wir nehmen an, dass

$\#_a$ (nur) am Anfang jedes Textes steht, $\#_e$ nur am Ende. Uns interessiert natürlich nicht die Wahrscheinlichkeit von # selbst, sondern die Wahrscheinlichkeit, dass ein Buchstabe am Anfang eines Textes steht. Diesen Fall müssen wir natürlich gesondert betrachten, denn in diesem Fall haben wir keine Vorgängerzustände, auf die wir uns berufen können. Wir müssen dabei beachten, dass wir für verlässliche Schätzungen für $P(\mathtt{a}|\#_e)$ eine Vielzahl von Texten betrachten müssen, da wir pro Text nur eine solche Folge haben.

Wir möchten zunächst die Wahrscheinlichkeit von $P(\mathtt{a}|\mathtt{x})$ für alle $\mathtt{x}\in\Sigma$ berechnen. Wir tun das auf eine denkbar einfache Art und Weise: wir erweitern unsere Notation $\mathtt{a}(\mathfrak{T})$ auf Worte, so dass $\mathtt{abc}...(\mathfrak{T})$ die Anzahl aller Vorkommen von $\mathtt{abc...}$ in $\mathfrak{T}$ ist. Wir sagen nun:

$$\hat{P}(\mathtt{a}|\mathtt{b}) := \frac{\mathtt{ba}(\mathfrak{T})}{\mathtt{b}(\mathfrak{T})}, \tag{149}$$

wobei $\hat{P}$ die von uns geschätzte Wahrscheinlichkeit bezeichnet. Diese Schätzung erlaubt es uns, für alle $\mathtt{a},\mathtt{b}\in\Sigma$ die Wahrscheinlichkeit $\hat{P}(\mathtt{a}|\mathtt{b})$ zu schätzen.

Diese Methode lässt sich leicht auf beliebige Markov-Ketten n-ter Ordnung verallgemeinern: sei $\vec{\mathtt{w}}$ ein Wort mit $|\vec{\mathtt{w}}| = n$; dann ist

$$\hat{P}(\mathtt{a}|\vec{\mathtt{w}}) := \frac{\vec{\mathtt{w}}\mathtt{a}(\mathfrak{T})}{\mathtt{a}(\mathfrak{T})}. \tag{150}$$

Mit diesen bedingten Wahrscheinlichkeiten können wir nicht ohne weiteres zu den unbedingten Wahrscheinlichkeiten zurückkommen: wir haben zwar die bekannten Regeln zur bedingten Wahrscheinlichkeit und Partitionen, und bekommen:

$$\hat{P}(\mathtt{a}) := \sum_{|\vec{\mathtt{w}}|=n} \hat{P}(\mathtt{a}|\vec{\mathtt{w}})\hat{P}(\vec{\mathtt{w}}), \tag{151}$$

Allgemeiner ausgedrückt, für eine Markov-Kette n-ter Ordnung, $|\vec{\mathtt{w}}| \leq n$, haben wir

$$\hat{P}(\mathtt{a}|\vec{\mathtt{w}}) := \sum_{|\mathtt{x}\vec{\mathtt{w}}|=n} \hat{P}(\mathtt{a}|\mathtt{x}\vec{\mathtt{w}})\hat{P}(\vec{\mathtt{x}}), \tag{152}$$

Aber um Wahrscheinlichkeiten zu berechnen, brauchen wir Wahrscheinlichkeit von Wörtern $\hat{P}(\vec{\mathtt{w}})$! Wahrscheinlichkeit von Wörtern berechnet sich wie folgt: sei $\vec{\mathtt{w}} = \mathtt{a_1 a_2 ... a_i}$. Dann ist

$$\begin{aligned} \hat{P}(\mathtt{a_1 a_2 a_3 ... a_i}) &= \hat{P}(\mathtt{a_1})\hat{P}(\mathtt{a_2}|\mathtt{a_1})\hat{P}(\mathtt{a_3}|\mathtt{a_1 a_2})...\hat{P}(\mathtt{a_n}|\mathtt{a_1 ... a_{n-1}}) \\ &= \prod_{i=1}^{n} \hat{P}(\mathtt{a_i}|\mathtt{a_1 ... a_{i-1}}) \end{aligned} \tag{153}$$

Wir müssen also, für eine Markov-Kette n-ter Ordnung, alle Wahrscheinlichkeiten $\hat{P}(\mathtt{a}|\vec{\mathtt{w}}) : 0 \leq |\vec{\mathtt{w}}| \leq n$ schätzen. Mit diesem Wissen und einiger Mühe lässt sich natürlich zeigen:

$$\hat{P}(\mathtt{a}) := \sum_{|\vec{\mathtt{w}}|=n} \hat{P}(\mathtt{a}|\vec{\mathtt{w}})\hat{P}(\vec{\mathtt{w}}) = \frac{\mathtt{a}(\mathfrak{T})}{|\mathfrak{T}|}, \tag{154}$$

wie wir das erwarten.

Aufgabe 8

Abgabe bis zum 6.6.2017 *vor dem Seminar*, egal ob digital/analog und auf welchem Weg.

Wir nehmen ein Markov Modell, das wie folgt spezifiziert ist (wir benutzen hier die eingeführten Kurzformen; $\ltimes$ steht für den Wortanfang, $\rtimes$ für das Ende):

- $P(a|\ltimes) = 0.4$
- $P(b|\ltimes) = 0.6$
- $P(\rtimes|\ltimes) = 0$
- $P(a|a) = 0.8$
- $P(b|a) = 0.1$
- $P(\rtimes|a) = 0.1$
- $P(a|b) = 0.2$
- $P(b|b) = 0.7$
- $P(\rtimes|b) = 0.1$.

Berechnen Sie die Wahrscheinlichkeit der folgenden Ereignisse in der dazugehörigen probabilistischen Sprache:

1. Der dritte Buchstabe in einem beliebigen Wort ist ein a.
2. Ein Wort hat n Buchstaben (bitte eine Formel über n!)

14 Parameter glätten – Smoothing 1 (add one)

Wie wir gesehen haben ist die ML-Schätzung problematisch, wenn unsere Daten sehr dünn sind – was z.B. insbesondere bei Markov-Ketten höherer Ordnung unvermeidlich ist: wenn unser Lexikon 10.000 Worte enthält, dann gibt es

$$10.000^5 = (10^4)^5 = 10^{20} \tag{155}$$

5-grame – das sind enorm viele, und es ist fast ausgeschlossen dass wir einen repräsentativen Einblick in die Verteilung für seltene 5-grame bekommen. Wenn aber ein Ergebnis nicht beobachtet wurde, bekommt es nach ML-Schätzung die Wahrscheinlichhkeit 0 – ein sehr extremer Wert, den man oft in dieser Form nicht will, denn er absorbiert alle anderen Werte.

Deswegen benutzt man verschiedene Verfahren um Parameter zu **glätten**, d.h. solche extremen Werte zu vermeiden. Das einfachste Verfahren ist das sog. **add-one smoothing**, das darauf basiert dass wir einfach so tun, als hätten wir jedes Ergebnis *mindestens einmal* beobachtet. Die Schätzung (bleiben wir beim Beispiel der Markov-Kette) sieht dann wie folgt aus:

$$\hat{P}_{\textit{add-one}}(a|w) = \frac{|D|_{wa} + 1}{|D|_w + |\Sigma|} \tag{156}$$

Wir nehmen also an dass w noch $|\Sigma|$-oft vorkommt, jedes mal gefolgt von einem anderen $a \in \Sigma$.

Diese Methode ist tatsächlich die einfachste um 0-Schätzungen zu vermeiden; sie ist allerdings oft kritisiert worden, aus folgendem Grund: *add-one-smoothing* sei so wie denen, die wenig haben, etwas wegzunehmen, um es denen zu geben, die gar nichts haben. Das ist natürlich bildlich gesprochen und bedeutet: durch diese Art von smoothing verschiebt sich viel Wahrscheinlichkeitsmasse von den selten gezählten zu den gar nicht gezählten. Nehmen wir beispielsweise an,

$$|D|_{wa} = 1,\ |D|_w = c|\Sigma|,\ |D|_{wb} = 0$$

Dann ist

$$\hat{P}_{\textit{add-one}}(a|w) = \frac{2}{(c+1)|\Sigma|} \tag{157}$$

und

$$\hat{P}_{add\text{-}one}(b|w) = \frac{1}{(c+1)|\Sigma|} \quad (158)$$

Damit ist klar dass je größer c ist, desto kleiner ist die Differenz der beiden. Weiterhin können wir folgendes sehen: falls

$$|D|_{vb} = n,\ |D|_v = m|\Sigma|$$

und

$$\frac{n}{m} < \frac{1}{c} \quad (159)$$

dann ist

$$\hat{P}_{add\text{-}one}(b|w) > \hat{P}_{add\text{-}one}(b|v) \quad (160)$$

obwohl ersteres nie beobachtet wurde, letzteres möglicherweise durchaus häufig!

Ein weiteres Problem ist folgendes: nehmen wir wieder das obige Beispiel, wir haben ein Korpus mit 1.000.000 Worten (*token*) und 10.000 *types*, und wir möchten damit 5-grame schätzen. Aus einer Division ergibt sich, dass wir selbst im allerbesten Fall

$$\frac{10^{20}}{10^6} = 10^{14} \quad (161)$$

5-grame nicht beobachten können. Jedes dieser 5-grame wird dann eine Wahrscheinlichkeit bekommen, die wir natürlich nicht genau bestimmen können; allerdings zeigt schon eine kurze Überlegung, dass der **allergrößte Teil** der gesamten Wahrscheinlichkeitsmasse auf diese 5-grame entfällt, obwohl wir sie noch nie beobachtet haben. Das ist natürlich ein großes Problem, denn es macht unsere gesamten Schätzungen ziemlich wertlos, da die allermeiste Information völlig uninformiert vergeben wird. Es gibt also eine Reihe Probleme mit dieser Schätzung.

15 Parameter glätten – Good-Turing smoothing (vereinfacht)

Good-Turing smoothing wurde von Alan Turings Assistenten Good entwickelt, um den deutschen Enigma-Kode zu entschlüsseln. Die Grundlage dieser Methode ist folgende: anstatt zu fragen:

> Wie ist die Wahrscheinlichkeit, ein Objekt der Art x zu treffen?

Fragen wir nun:

> Wie ist die Wahrscheinlichkeit, ein Objekt der Häufigkeit n zu treffen?

Z.B.: wie wahrscheinlich ist es, ein Objekt zu sehen, das die Häufigkeit 210 hat? Der Sinn dahinter ist folgender: uns interessiert natürlich insbesondere die Frage:

> Wie ist die Wahrscheinlichkeit, ein Objekt zu treffen, das wir noch nie zuvor beobachtet haben?

Denn diese Wahrscheinlichkeit ist genau diejenige, die wir den Objekten zuweisen wollen, die wir in unseren Daten nicht beobachtet haben. Und das schöne ist: das lässt sich sehr einfach schätzen: den jedesmal, wenn wir ein Objekt nur einmal beobachtet haben, haben wir eine neue Beobachtung gemacht; wir nehmen hierfür also einfach die (nach Maximum likelihood) geschätzte Wahrscheinlichkeit, ein Objekt nur einmal zu beobachten. Das Problem ist allerdings, eine konsistente Wahrscheinlichkeitsverteilung daraus zu bekommen.

Nehmen wir also folgende Definitionen:

- $\mathfrak{T}$ ist unser Datensatz.
- $G = |\mathfrak{T}|$ die Gesamtzahl der Beobachtungen.
- L ist die Menge der Beobachtungen, die wir gemacht haben (also die Menge der beobachteten types!).
- $D : L \to \mathbb{N}$ ist eine Menge von Paaren, die jeder Beobachtung ihre Häufigkeit zuweist.

- $N : \mathbb{N} \to \mathbb{N}$ ist eine Funktion ("Häufigkeiten der Häufigkeiten"), die jeder Häufigkeit einer Beobachtung ihre Häufigkeit zuweist; formal und besser verständlich:

$$N(n) = |\{x \in L : D(x) = n\}|$$

Die einfache Good-Turing Schätzfunktion funktioniert nun wie folgt: wir schätzen die Wahrscheinlichkeit eines nicht-beobachteten Ereignisses als

$$\hat{P}(neu) = \frac{N(1)}{G} \quad (162)$$

Dem liegt die Überlegung zugrunde, dass eine neue Beobachtung immer darin resultieren würde, diese Beobachtung einmal gemacht zu haben, also nehmen wir einfach die geschätzte Wahrscheinlichkeit hiervon. Wohlgemerkt: hier handelt es sich nicht um die Wahrscheinlichkeit *eines* unbeobachteten n-grams, sondern um die Wahrscheinlichkeit *aller*, die wir nicht beobachtet haben. Nehmen wir also an, es gibt c n-grame, die wir nicht beobachtet haben. Dann wäre also, für x ein solches n-gram,

$$\hat{P}(x) = \frac{\hat{P}(neu)}{c} = \frac{N(1)}{G \cdot c} \quad (163)$$

Denn in Ermangelung von weiterem Wissen sollten wir alle diese n-grame für gleich wahrscheinlich halten. Nun ist die Frage: wir schätzen wir die übrigen Wahrscheinlichkeiten? Hier lautet die Antwort: wir setzen

$$\hat{P}(neu) := \hat{P}(0) \quad (164)$$

– wir schätzen also die Wahrscheinlichkeiten von Häufigkeiten von Häufigkeiten, und verallgemeinern:

$$\hat{P}(n) = (n+1) \cdot \frac{N(n+1)}{G} \quad (165)$$

Wir schätzen also allgemein die Wahrscheinlichkeit von n mittels der Häufigkeit von $n+1$. Die Motivation hierfür ist folgende: wenn wir eine Beobachtung machen aus der Klasse von Objekten, die wir n mal beobachtet haben, dann haben wir sie $n+1$-Mal beobachtet – also nehmen wir diese Wahrscheinlichkeit. Der Term $n+1$ steht hier für die Tatsache, dass die Zahl $N(n+1)$

mit $n+1$ multipiziert werden muss, um die wahren Häufigkeiten abzubilden, wir brauchen das also für die Konsistenz. Für eine bestimmte Klasse von Beobachtungen $x \in L$ bekommen wir dann:

$$\hat{P}(x) = (n+1) \cdot \frac{N(n+1)}{G \cdot N(n)} \quad (166)$$

denn wir müssen die Wahrscheinlichkeiten wieder gleich verteilen.

Soweit, so gut – diese Schätzfunktion wird uns eine konsistente Wahrscheinlichkeitsfunktion liefern. Es gibt aber ein großes Problem hierbei: die Dünne der Daten (English: *sparseness*. Das ist ein großes Problem, wenn wir z.B. Zipf-verteilte Daten haben, dann kann das wie folgt aussehen:

n	$N(n)$
1	427
2	285
3	157
4	103
...	...
401	1
402	0
403	1

Hier haben wir ein Problem: nehmen wir an, a ist die Klasse so dass $N(a) = 401$). Nun gilt:

$$\hat{P}(a) = 402 \cdot \frac{N(402)}{G \cdot N(401)} = 402 \cdot \frac{0}{G} = 0 \quad (167)$$

Das ist natürlich Unsinn und völlig daneben. Das zugrundeliegende Problem ist: die Wahrscheinlichkeit von a hängt ab von der Wahrscheinlichkeit, mit der Objekte mit Häufigkeit $D(a)+1$ auftreten. Dadurch dass es keine solchen Objekte gibt, bekommen wir 0 als Ergebnis. Und es ist erst an dieser Stelle, dass Good-Turing smoothing kompliziert wird. Es gibt für dieses Problem 2 Lösungen:

1. Anstatt der Funktion N, die auf den Daten basiert und evtl. Lücken hat nutzen wir $S(N)$, was eine (lineare) Approximation von N darstellt.

$S(N)$ ist also eine lineare Funktion (allgemeiner: ein Polynom), das sich ähnlich wie N verhält, aber eben keine Lücken hat. In diesem Fall bekommen wir die Schätzfunktion:

$$(168)\quad hatP(n) = (n+1) \cdot \frac{S(N)(n+1)}{G \cdot NS(N)(n)}$$

Die Funktion N kann man normalerweise gut und effektiv approximieren mittels **linearer Regression**, die wir später behandeln werden.

2. Wir konstruieren eine Mischung aus einfachem Good-Turing und der approximierten Funktion. Das ist besser, aber deutlich komplizierter, denn wir müssen entscheiden, a) wann wir welche Methode wählen, und b) die Wahrscheinlichkeiten addieren nicht auf 1, wir müssen also **normalisieren**.

Das gute ist: es gibt fertige Pakete, die verschiedene Methoden von Good-Turing Schätzung fertig implementiert haben; das ganze von 0 an zu machen ist relativ kompliziert.

16 Parameter schätzen – Bayesianisch

16.1 Uniformes Apriori

Im Allgemeinen gilt, wie wir gesehen haben, für eine Hypothese H bezüglich der zugrunde liegenden Wahrscheinlichkeiten, D eine Reihe von Beobachtungen wir gemacht haben,

$$(169)\quad P(H|D) = P(D|H)\frac{P(H)}{P(D)} \propto P(D|H)P(H)$$

erstmal wegen Bayes, und zweitens weil der Term $P(D)$ unabhängig ist von H, also für die Suche eines Maximums über H (und viele andere Operationen) keine Rolle spielt.

Wie wir gesehen haben, basiert die "orthodoxe" Schätzung auf der Likelihood, die wiederum darauf basiert, die *a priori*-Wahrscheinlichkeit $P(H)$ zu unterdrücken: wenn wir annehmen, dass wir keine Informationen über $P(H)$ haben, können wir den Term auch weglassen:

$$(170)\quad P(D|H)P(H) \propto P(D|H) = L_P(H|D)$$

So kommen wir von der Wahrscheinlichkeit zur Likelihood von H, nämlich $P(D|H)$. In der bayesianischen Auffasseung gibt es aber praktisch immer eine *a priori* Information, die wir O nennen. Wir haben also:

$$(171)\quad P(H|DO) = P(D|HO)\frac{P(H|O)}{P(D|O)}$$

Bayesianische Parameterschätzung basiert nicht auf der

$$\text{Likelihood } P(D|H)$$

sondern auf der

$$\textit{aposteriori}\text{-Wahrscheinlichkeit } P(H|DO).$$

Information geht für uns niemals verloren, auch nicht durch unsere Beobachtung D, dementsprechend müssen wir über O Rechnung ablegen. Information kann zwar irrelevant werden – aber im Allgemeinen gibt es keinen Grund dafür! In unserem Fall besteht O darin, dass wir keine weitere spezielle Information bezüglich der Wahrscheinlichkeit von Ereignissen; die Frage ist,

wie wir das in formale Wahrscheinlichkeitsverteilung transformulieren. Der einfachste Fall ist (wie immer) die **uniforme Verteilung**, auch wenn man in manchen Fällen besser davon abweicht. Wie schätzen wir also die *aposteriori*-Wahrscheinlichkeiten?

Nehmen wir das Beispiel eines Textes mit Worten $a, b, c, ...$ Unser Text D ist die Beobachtung, die wir machen. Wir haben bereits gesehen, dass die Maximum-Likelihood von a errechnet wird durch

$$\frac{|D|_a}{|D|} \tag{172}$$

Wie geht die Schätzung bayesianisch? Zunächst folgende Konvention: wir nennen θ_a das, was wir vorher $\hat{P}(a)$ genannt haben, also

$$\theta_a \triangleq \hat{P}(a)$$

Der Vorteil hiervon ist: wir können nun θ_a als Zufallsvariable auffassen, die Werte in $[0, 1]$ annimmt mit einer unterschiedlichen Wahrscheinlichkeit.

$$P(\theta_a = x|DO) = P(D|\theta_a = x, O)\frac{P(\theta_a = x|O)}{P(D|O)} \tag{173}$$

$P(D|\theta_a = x, O)$ ist für uns die Wahrscheinlichkeit der Daten gegeben unsere Wahrscheinlichkeit von a ist x. Wohlgemerkt:

$$P(\theta_a = x|O) \neq \frac{1}{|\Sigma|} \tag{174}$$

die uniforme Verteilung verlangt vielmehr dass

$$\int_0^1 P(\theta_a = x|O)d(x) = 1 \tag{175}$$

und

$$\text{für alle } x, y \in [0, 1],\ P(\theta_a = x|O) = P(\theta_a = y|O) \tag{176}$$

$P(D|O)$ ist ein Term, der unabhängig ist von θ_a (und allen anderen Parametern). Wir haben also

$$P(\theta_a = x|D, O) = P(D|\theta_a = x, O)\frac{1}{C} \tag{177}$$

wobei C eine Normalisierungs-Konstante ist die unabhängig ist von allen Parametern. Wohlgemerkt ist

$$P(\theta_a|DO)$$

nicht die *aposteriori*-Wahrscheinlichkeit von a, sondern vielmehr eine Wahrscheinlichkeitsverteilung über mögliche Werte von θ_a.

Das heißt die eigentliche Schätzung steht natürlich noch aus, Hier hat man wieder dieselben Möglichkeiten wie vorher, und in unserem Fall läuft auf die ML-Methode hinaus.

16.2 Kein uniformes Apriori

Man kann sich fragen, worin die Bedeutung unseres *a priori* liegt, wenn er darin besteht, dass wir keine relevante Information haben. Die Antwort ist folgende: wie wir bereits gesagt haben, erwarten wir dass unsere Beobachtungen normalerweise *extremer* sind, als die zugrundeliegende Verteilung, insbesondere dort, wo wir wenige Beobachtungen haben. Während wir in der orthodoxen Likelihood eben *ad-hoc* eine Lösung dafür finden müssen, ist die Lösung in der bayesianischen Methode bereits eingebaut, nämlich mittels eine gut gewählten *apriori*-Verteilung.

Unsere *apriori* sagt uns bereits, welche Häufigkeiten wir erwarten, und jede Beobachtung, die davon abweicht, wird dadurch gemildert. Insbesondere wird – wenn die *apriori*-Wahrscheinlicheit > 0 ist, die *aposteriori*-Wahrscheinlichkeit ebenso > 0 sein; somit können wir sehr extreme Ergebnisse ausschließen.

Bisher hatte sich, durch unser uniformes Apriori, nichts geändert an der Schätzung, da das *apriori* nur als Konstante eingeflossen ist. Das ist anders wenn wir eine apriori-Verteilung wählen, die nicht uniform ist, und die beispielsweise berücksichtigt, dass extreme Werte unwahrscheinlicher sind als "mittlere" Werte. Ein Beispiel hierfür ist:

(178) $P(\theta_a = x|O_1) = C(x \cdot (1-x))$

wobei C eine **Normalisierungs-Konstante** ist, und O_1 das entsprechende *apriori* ist. Dabei entspricht O_1 dem Wissen, dass wir "gemäßigte" θ_a bevorzugen. Die Verteilung ist symmetrisch, mit einem Maximum an

(179) $P(\theta_a = 0.5|O_1) = 0.25$

und, was sehr wichtig ist,

(180) $P(\theta_a = 0|O_1) = P(\theta_a = 1|O_1) = 0$

d.h. unser apriori schließt aus, das die Wahrscheinlichkeit θ_a je 0 wird. Wichtig ist: die Regeln der Wahrscheinlichkeitstheorie sind so, dass wenn etwas kategorisch ausgeschlossen ist (wie $\theta_a = 0$) sich dass durch keine Beobachtung ändert! Das kann sinnvoll sein, wenn wir eine Münze werfen oder ähnliches: denn es ist *apriori* viel wahrscheinlicher, das der korrekte Parameter irgendwo in der Mitte liegt, während es an den Rändern immer unplausibler wird den korrekten Parameter zu finden.

Um zu sehen, dass $P(\theta_a|O_1)$ das eine Wahrscheinlichkeitsverteilung ist, müssten wir noch zeigen dass

$$\int_0^1 P(\theta_a = x|O_1)\ d(x) = \int_0^1 C \cdot x \cdot (1-x)d(x) = 1 \tag{181}$$

(wie ging das nochmal?)

$$\int x \cdot (1-x)d(x) = \int -x^2 + x\ d(x) = -\frac{x^3}{3} + \frac{x^2}{2} \tag{182}$$

Dementsprechend:

$$\int_0^1 x \cdot (1-x)d(x) = (-\frac{1^3}{3} + \frac{1^2}{2}) - (-\frac{0^3}{3} + \frac{0^2}{2}) = \frac{1}{2} - \frac{1}{3} = \frac{1}{6} \tag{183}$$

Das bedeutet, dass das Integral unserer Wahrscheinlichkeit nur 1/6 beträgt, wir brauchen also

$$C = 6 \tag{184}$$

Das bedeutet aber:

$$P(\theta = 0.5|O_1) = 1.5 \tag{185}$$

d.h. es ist keine echte Wahrscheinlichkeit mehr. Das ist aber kein Problem, wir müssen uns nur erinnern: wir haben jetzt kontinuierliche Verteilungen, die Wahrscheinlichkeit an einem **Punkt** ist immer 0; was interessant ist, ist die Masse in einem Integral.

Nun nehmen wir einmal an, wir haben a in unseren Daten D kein einziges Mal beobachtet. Wir haben nun zwei relevante Faktoren:

$$lim_{\theta_a \to 0} P(D|\theta_a, O_1) = 1 \tag{186}$$

d.h. für $\theta_a \to 0$ geht die Wahrscheinlichkeit von D gegen 1; aber:

(187) $lim_{\theta_a \to 0} P(\theta_a|O_1) = 0$

Das bedeutet: wenn wir

(188) $P(\theta_a|DO_1)$

maximieren möchten, dann reicht es nicht, $P(D|\theta_a, O_1)$ zu maximieren! Insbesondere werden wir niemals ein Maximum bei $\theta_a = 0$ haben.

Wo der genaue Wert landet, hängt also von der Interaktion der Verteilungen ab (anders als bei ML): insbesondere hängt es davon ab, wie oft wir das Experiment wiederholen. Das ist intuitiv klar: je mehr Beobachtungen machen, desto unwichtiger wird unser apriori. Gehen wir das an einem ganz konkreten Beispiel durch: Sei $|D_1| = 5$, $|D_1|_a = 0$. Dann haben wir

(189) $P(\theta_a = x|D_1, O) = ((1-x)^5)(x(1-x))C$

C ist eine Normalisierungskonstante (ergibt sich aus $P(D|O_1)$) die unabhängig von θ_a gleich bleibt. Wir brauchen also:

(190) $argmax_{0 \leq x \leq 1}((1-x)^5)(x(1-x)) = argmax_{0 \leq x \leq 1}((1-x)^6)x$

Wir haben

(191) $\frac{d}{d(x)}((1-x)^6)x = (x-1)^5(7x-1)$

und für

(192) $(x-1)^5(7x-1) = 0$

gibt es die Lösungen $x = 1$ und $x = \frac{1}{7}$, wobei ersteres natürlich kein Maximum sein kann. Also ergibt die ML-Schätzung in diesem Fall

(193) $\hat{P}_{ML}(\theta_a) = \frac{1}{7}$

Nun nehmen wir an, wir haben in D_2 20 Beobachtungen gemacht, und a war immer noch nicht darunter. In diesem Fall ist alles wie gehabt, nur ein Parameter ändert sich:

(194) $P(\theta_a = x|D_2, O_1) = (1-x)^{20}(x(1-x))C$

Uns interessiert also

$$(195)\quad argmax_{0\leq x\leq 1}(1-x)^{21}x$$

Also:

$$(196)\quad \frac{d}{d(x)}(1-x)^{21})x = (x-1)^{20}(22x-1)$$

dann setzen wir

$$(197)\quad (x-1)^{20}(22x-1) = 0 \Leftrightarrow x = \frac{1}{22}$$

Das Muster ist leicht zu erkennen: für n Beobachtungen, von denen keine a ist, schätzen wir mit ML Schätzung und unserem konvexen apriori

$$(198)\quad \hat{P}_{ML}(\theta_a) = \frac{1}{n+2}$$

– also bei $n = 0$, $\theta_a = \frac{1}{2}$, wie es sein sollte. Wir sehen also, dass unser konvexes Apriori das Smoothing vollkommen unnötig macht.

Umgekehrt, nehmen wir an wir bekommen ein Ergebnis wie

$$(199)\quad \frac{|D|_a}{|D|} = \frac{1}{3}$$

d.h. die (orthodoxe) ML-Schätzung liegt relativ nahe am apriori-wahrscheinlichsten Wert. Wie ist nun der aposteriori maximale Wert? Das hängt wiederum von $|D_3|$ ab; sagen wir $|D_3| = 21$. Dann haben wir

$$(200)\quad P(\theta_a = x|D_3, O_1) = \binom{21}{7}(1-x)^{14}x^7(1-x)x\frac{1}{C} = \binom{21}{7}(1-\theta_a)^{15}\theta_a^8\frac{1}{C}$$

Uns interessiert also

$$(201)\quad argmax_{0\leq x\leq 1}\binom{21}{7}(1-x)^{15}x^8$$

Das errechnet sich aus

$$(202)\quad \frac{d}{d(x)}\binom{21}{7}(1-x)^{15}x^8 = 0$$
$$\Leftrightarrow x = \frac{8}{23}$$

Wir weichen also nur um

$$\frac{8}{23} - \frac{1}{3} = \frac{24}{69} - \frac{23}{69} = \frac{1}{69} \tag{203}$$

von der "orthodoxen" ML-Schätzung ab. Allgemeiner gesagt

$$\begin{aligned} \frac{d}{d(x)}\binom{n}{k}(1-x)^{n-k}x^k &= 0 \\ \Leftrightarrow x &= \frac{k+1}{n+2} \end{aligned} \tag{204}$$

D.h. unser konvexes Apriori gibt uns eine Schätzung

$$\hat{P}(\theta_a|DO_1) = \frac{|D|_a + 1}{|D| + 2} \tag{205}$$

Also in auch in diesem Fall ist die Berechnung sehr einfach und hat den Vorteil, dass wir keinerlei weitere Methoden brauchen, um sehr extreme Ergebnisse abzumildern.

17 Numerische Parameter und Alternativen zu ML

ML für Erwartungswerte Nehmen wir einmal an, wir schätzen einen stetigen Parameter, also anstellen von θ_a (für $a \in \Sigma$) oder θ_x (für $x \in \{0,1\}$) schätzen wir θ_x (für $x \in [0,10]$). Das macht erstmal nicht so viel Sinn – wir müssten ja unendlich viele Parameter schätzen. Das macht aber durchaus Sinn wenn wir einen Erwartungswert schätzen: nehmen wir an, wir haben eine Zufallsvariable X deren Erwartungswert wir schätzen möchten. Das wäre z.B.: wir treffen Menschen und fragen Sie nach Ihrem Alter. Wir möchten den Altersschnitt schätzen, suchen also den Erwartungswert von X.

Wenn wir die zugrundeliegenden Wahrscheinlichkeiten kennen würden, dann müssten wir einfach nur $\mathcal{E}(X)$ berechnen; allerdings können wir uns nur auf eine Stichprobe berufen. Natürlich können wir einfach folgendes machen: sei D unser Datensatz, der wie folgt aussieht:

Alter	Anzahl
1	3
2	2
3	-
4	4
....	
91	1
92	1

D besteht also aus Paaren (n,m); ausserdem sei G die Gesamtgröße unserer Stichprobe, also

$$G = \sum_{(n,m)\in D} n \tag{206}$$

Der naheliegendste Ansatz wäre also:

$$\langle X\rangle_{ML} = \sum_{(n,m)\in D} \frac{n}{G} m = \frac{1}{G} \sum_{(n,m)\in D} nm \tag{207}$$

Diese Art den Erwartungswert zu berechnen entspricht der ML-Schätzung; das sieht man wie folgt: erinnern wir uns dass

$$P(X = n) = P(X^{-1}(n)) \tag{208}$$

Weiterhin ist

$$(209)\ \ \mathcal{E}(X) = \sum_m m \cdot P(X = n)$$

Wir sehen dass nach ML-Schätzung

$$(210)\ \ \hat{P}_{ML}(X^{-1}(n)) = \frac{m}{G} :\ (n, m) \in D$$

und dementsprechend ist (207) nichts anderes als der Erwartungswert von $\mathcal{E}(X)$ mit der unterliegenden Wahrscheinlichkeit $\hat{P}_{ML}$ nach ML geschätzt.

Least squared error ML ist durchaus sinnvoll im Szenario mit Alter. Allerdings haben wir bereits besprochen das diese Form der Schätzung Schwächen aufweist; insbesondere ist ihr vollkommen gleich, ob der Wert $\langle X \rangle_{ML}$ tatsächlich auftritt (in unserem Fall wahrscheinlich nicht – wir messen Alter in ganzen Zahlen, $\langle X \rangle_{ML}$ wird aber aller Voraussicht nach keine ganze Zahl sein). In unserem Fall lässt sich das durch Rundung beheben; im allgemeinen Fall ist das schwierig aufzulösen. Nehmen wir einmal folgendes an:

> wir sollen $\langle X \rangle_2$ auf eine gewisse Art schätzen, und jedesmal wenn ein neu gemessener Wert von unserer Schätzung abweicht, kostet uns das Geld (und zwar in Form einer Funktion über den Grad der Abweichung).

Wir haben also ein Interesse daran die Abweichung so gering als möglich zu halten. An dieser Stelle kommen wir zurück auf den Begriff der Varianz:

$$(211)\ \ V(X) = \mathcal{E}((X - \mathcal{E}(X))^2)$$

Die Varianz misst, wieviel wir im Quadrat erwarten abzuweichen von unserem Erwartungswert. Die Methode des **kleinsten quadratischen Fehlers** schätzt $\langle X \rangle_{LSE}$ so, dass die quadratische Abweichung der Daten von $\langle X \rangle_{LSE}$ minimal ist:

$$(212)\ \ \langle X \rangle_{LSE} = argmin_{x \in \mathbb{R}} \sum_{(n,m) \in D} ((x - n) \cdot m)^2$$

Das bedeutet: wir möchten die Abweichungen (im Quadrat) von unserem Wert minimieren. Das Quadrat kommt natürlich erstmal daher, dass wir positive Werte möchten. Im Allgemeinen hat die LSE-Schätzung einen bedeutenden Vorteil vor der ML-Schätzung:

- Die LSE-Schätzung ist sensibel für Abweichungen vom geschätzten Wert und versucht sie zu vermeiden;
- der ML-Schätzung sind Abweichungen egal, solange sie sich "ausgleichen".

Das Quadrat hat aber noch eine weitere Auswirkung: extreme Abweichungen werden stärker bestraft als geringe, d.h. "Ausreißer" werden überproportional gewichtet. Nehmen wir einmal an,

(213) $\langle X \rangle_{ML} = 38,$

allerdings gibt es in D einen Ausreißer $(120, 1)$ – wir haben also einen 120-jährigen getroffen! Das würde uns also für die LSE-Schätzung eine Abweichung von

(214) $(38 - 120)^2 = 6724$

liefern – und damit mehr ins Gewicht fallen als 6 70-jährige!

(215) $6 \cdot (38 - 70)^2 = 6 \cdot 1024 < 6724$

Das kann in manchen Fällen gewünscht sein, in anderen ist es das nicht. Wir haben also folgende Nachteile:

- Die LSE-Schätzung ist sehr anfällig gegenüber Ausreißern – sie mißt ihnen großes Gewicht bei;
- und sie ist deutlich komplizierter zu berechnen (auch wenn das heutzutage kein Problem mehr sein sollte.

Mediane Nehmen wir jetzt nun folgendes Szenario an: anstelle des Alters der Personen, die wir treffen, haben wir ihr Einkommen. Was ändert das? Nun, wie wir wissen sind die Einkommen grundsätzlich anders verteilt als die Altersstruktur: insbesondere haben wir eine normalerweise eine **Zipf-Verteilung**, d.h. sehr wenige sehr reiche Leute, und eine große Anzahl wenig wohlhabender Leute. Daraus folgt dass sowohl für ML- als auch für LSE-Schätzung der Wert stark nach oben gezogen wird: weder

$$\langle X \rangle_{ML}$$

noch

$$\langle X \rangle_{LSE}$$

liefern uns einen vernünftigen Wert für das Einkommen einer Person, der wir zufällig auf der Straße begegnen – die extremen Werte werden einfach zu stark berücksichtigt. Hier kann die **Median-Schätzung** hilfreich sein: der Median von D ist folgender Wert (D eine Menge von Paaren (n, m) mit (n =Einkommen,m =Anzahl der Verdiener, $G = |D|$),

$$\begin{aligned} & med(D) = m \Leftrightarrow \\ \text{(216)} \quad & \text{es gibt } \frac{G-1}{2} \text{ Datenpunkte } n' : n' \leq n \& \\ & \frac{G-1}{2} \text{ Datenpunkte } n' : n' \geq m \end{aligned}$$

Das liefert zumindest in diesem Fall eine vernünftige Schätzung: denn das Einkommen einiger weniger Superreichen hat ja tatsächlich keinen Einfluß darauf, welches Einkommen wir einem Menschen, den wir zufällig begegnen, zumessen!

Wenn wir also eine Zipf-Verteilung haben, dann wird uns die Media-Schätzung mit ziemlicher Sicherheit einen Wert am unteren Ende der Verteilung liefern.

18 Maximum Entropie Methoden

18.1 Definition

Maximum Entropie (ME) Methoden sind sehr allgemein und mächtig, und wir werden uns nur einen besonderen Fall anschauen. In der Praxis haben wir manchmal (oft) den Fall, dass wir eine Wahrscheinlichkeitsfunktion schätzen möchten, aber die Verteilung mehr Parameter hat, als durch die Daten vorgegeben werden. Das kann verschiedene Gründe haben:

1. Unser Weltwissen leitet uns zu der Annahme, dass es relevante Parameter gibt, die wir nicht direkt beobachten können (z.B. syntaktische Kategorien, Wortarten, Bedeutungen in der Sprachverarbeitung).

2. Unser Weltwissen leitet uns zu der Annahme, dass Parameter, die auf den ersten Blick relevant erscheinen, eigentlich irrelevant sind und nicht zur Schätzung hinzugezogen werden sollten (z.B. wenn wir annehmen wir haben die Markov Eigenschaft erster Ordnung).

Theoretisch sieht normalerweise das wie folgt aus: wir haben eine Reihe von Zufallsvariablen $X_1, ..., X_i$, die jeweils eine bedingte Verteilung haben für Zufallsvariablen $Y_1, ..., Y_i$ (sie hängen also von Y ab). $Y_1, ..., Y_i$ sind also **Prädiktoren** für $X_1, ..., X_i$; sie bestimmen deren Verteilung. Die dünne Datenlage erlaubt aber keine vollständige Schätzung nach ML (oder anders); alles was wir haben ist der Erwartungswert (und vielleicht Varianz, Standardabweichung etc.). Das äußert sich dann in einer Reihe gewisser Bedingungen, die unsere Funktionen erfüllen müssen, ohne dass sie dadurch vollständig determiniert wären; wir haben also eine Reihe von Beschränkungen der Form

$$\begin{aligned} &\sum_{x \in X_1} \hat{P}(X_1 = x | Y_1 = y) \cdot x_1 = \alpha_1 \\ &\qquad\qquad \vdots \\ &\sum_{x \in X_i} \hat{P}(X_i = x | Y_i = y) \cdot x_i = \alpha_i \end{aligned} \tag{217}$$

Diese Gleichungen liefern uns Bedingungen, die $\hat{P}$ erfüllen muss. Wenn wir abstrakt von so einer Liste von Gleichungen ausgehen, dann gibt es keine

Garantie, dass es tatsächlich ein $\hat{P}$ gibt, dass alle Gleichungen erfüllt. Da wir aber alle Gleichungen von *denselben Daten* schätzen, ist klar dass es mindestens eine Verteilung $\hat{P}$ gibt, die alle erfüllt (nämlich die volle ML-Schätzung für alle Parameter). Das Problem ist eher das umgekehrte: es gibt normalerweise viele, genauer gesagt unendlich viele Verteilungen, die diese Gleichungen erfüllen. Unsere Frage ist: welche sollen wir wählen? Und hier beginnt der Ansatz der Maximum Entropie (ME) Methode. Kurz gesagt besteht er darin, dass wir die Verteilung $\hat{P}$ wählen, die die obigen Gleichungen erfüllt *und* die maximale Entropie hat. Intuitiv bedeutet das: da Entropie ein Maß für Information bzw. Unsicherheit ist, wir möchten dass unsere Verteilung alle relevante Information beinhaltet, aber keine weitere Information darüber hinaus.

Das ist leicht gesagt; es ist auch leicht in eine Formel geschrieben; sei $\mathcal{C}$ die Menge aller Verteilungen $\hat{P}$, die den obigen Gleichungen genüge tun. Was wir suchen ist

(218) $argmax_{\hat{P}\in\mathcal{C}} H(\hat{P})$

Das Problem ist: diese Formel zu finden. Und hier werden die Dinge spannend.

18.2 Ein einfaches Beispiel

Nehmen Sie an, sie reden mit einem Kollegen über seinen Arbeitsweg. Er sagt Ihnen:

- Mit den öffentlichen Verkehrsmitteln brauche ich durchschnittlich 45min;
- mit dem Auto durchschnittlich 40;
- mit dem Fahrrad durchschnittlich 35.

Natürlich hängt die Wahl des Verkehrsmittels von einer Menge Faktoren ab (das Wetter, der Verkehr etc.); außerdem haben wir nur die Erwartungswerte (also den Durchschnitt), keinesfalls die Wahrscheinlichkeitsverteilung: z.B. kann es sein dass beim Auto (wg. Verkehr) die Streuung sehr hoch ist, bei den öffentlichen eher gering. Davon wissen wir aber nichts!

Was wir aber wissen ist folgendes:

- Unser Kollege braucht durchschnittlich 41min für seinen Arbeitsweg.

Die Aufgabe ist nun: wir sollen $\hat{P}(\ddot{O}), \hat{P}(A), \hat{P}(F)$ schätzen. Alles was wir wissen sind die obigen Erwartungswerte, sowie

$$\hat{P}(\ddot{O}) + \hat{P}(A) + \hat{P}(F) = 1 \tag{219}$$

Um das ganze in eine einheitliche Form zu bringen, führen wir eine Zufallsvariable X ein, mit

$$\begin{aligned} P(X = 45) &= \hat{P}(\ddot{O}) \\ P(X = 40) &= \hat{P}(A) \\ P(X = 35) &= \hat{P}(F) \end{aligned} \tag{220}$$

Unsere Aufgabe ist nun, eine Verteilung P zu finden so dass gilt:

$$\sum_{x \in \{35,40,45\}} P(X = x) = 1 \qquad \mathcal{E}(X) = \sum_{x \in \{35,40,45\}} P(X = x) \cdot x = 41 \tag{221}$$

Intuitiv ist klar, dass das auf viele Arten und Weisen geschehen kann: es kann z.B. sein dass Ihr Kollege praktisch nie mit dem Auto, oft mit Öffis und noch öfter mit dem Fahrrad fährt; es kann aber genauso gut sein, dass er praktisch immer mit dem Auto und nur ausnahmsweise mit Öffis fährt. Wir haben aber keinerlei Wissen über diese Sachen, und beide Annahmen sind gleichermaßen ungerechtfertigt gegeben unser Wissensstand.

Was wir daher anwenden ist dass ME-Prinzip, dass nichts weiter ist als eine Generalisierung des Prinzips der Indifferenz, das besagt:

> Falls wir kein relevantes Vorwissen haben, nehmen wir die uniforme Verteilung an.

Wir haben nun aber durchaus relevantes Vorwissen. Unser ME-Prinzip sagt daher (da Entropie ein Maß der Unsicherheit ist):

> Von allen Verteilungen, die mit unserem Vorwissen konform sind, nehmen wir immer diejenige an, die die Maximale Entropie hat.

Auf diese Weise sind wir sicher, dass wir nicht mehr in die Verteilung hineinstecken, als wir wirklich wissen; wir bleiben uns also unserer Unsicherheit

bewußt. Das Problem ist: wie errechnen wir diese Verteilung? Zunächst das analytische Problem: wir suchen

(222)
$$\underset{P}{argmax}\, H(P) : \sum_{x\in\{35,40,45\}} P(X=x) = 1 \quad \& \quad \sum_{x\in\{35,40,45\}} P(X=x)\cdot x = 41$$

wobei gilt:

(223) $$H(P) = \sum_{x\in\{35,40,45\}} P(X=x) log_2(P(X=x))$$

Wie berechnen wir das? In unserem Fall geht das mit elementaren Methoden (die wir aus der Schule kennen); denn wir haben:

(224) $$35 = 35P(X=35) + 35P(X=40) + 35P(X=45)$$

und

(225) $$41 = 35P(X=35) + 40P(X=40) + 45P(X=45)$$

Nun subtrahieren wir die beiden Terme voneinander, damit bekommen wir:

(226) $$6 = 5P(X=40) + 10P(X=45)$$

also

(227) $$P(X=40) = \frac{6 - 10P(X=45)}{5} = \frac{6}{5} - 2P(X=45)$$

Wir können also $P(X = 40)$ loswerden; dasselbe gilt natürlich auch für $P(X = 35)$, indem wir (227) einsetzen in (219):

(228)
$$1 = P(X=35) + \frac{6}{5} - 2P(X=45) + P(X=45)$$
$$P(X=35) = P(X=45) - \frac{1}{5}$$

Wir können also beide Wahrscheinlichkeiten ausdrücken können als Formeln mit der einzigen Variable $P(X = 45)$. Daraus wiederum folgt:

(229) $$\sum_{x\in\{35,40,45\}} P(X=x) log_2(P(X=x))$$

lässt sich schreiben also Funktion mit *einer* einzigen Variable, für die wir also nur den Maximalwert suchen müssen:

$$\begin{aligned} & \underset{P(X=45)\in[0,1]}{argmax} \\ & f_1(P(X=45)) \cdot log_2(f_1(P(X=45))) \\ & +f_2(P(X=45)) \cdot log_2(f_2(P(X=45))) \\ & +P(X=45) \cdot log_2(P(X=45)) \end{aligned} \tag{230}$$

wobei sich f_1 und f_2 jeweils aus (228) und (227) ergeben. Das lässt sich mit den gewöhnlichen analytischen Methoden (Nullstelle der Ableitung) leicht ausrechnen.

18.3 Der allgemeinere Fall

In unserem Beispiel ließ sich der Wert gut berechnen, da er im Prinzip nur eine nicht-triviale Gleichung erfüllen musste. Gibt es eine ganze Reihe von Gleichungen, sind die Berechnungen kompliziert und man braucht fortgeschrittene Methoden (Lagrange-Multiplikatoren). Mittlerweile macht solche Sachen aber der Computer. Wichtig ist es zu verstehen worum es geht: die Schätzung verborgener Parameter, die durch unsere Daten nicht ausreichend determiniert sind.

In NLP Anwendungen kommt es oft dazu, dass wir gewisse Merkmale benutzen, uns aber über deren Bedeutung nicht ganz im Klaren sind. Ein typisches Beispiel ist die Übersetzungswahrscheinlichkeit eines Wortes, wobei als zusätzliches Merkmal das nachfolgende Wort gewählt wird. Das ersetzt aber nicht die einfache, unbedingte Übersetzungswahrscheinlichkeit, daher lässt sich die Wahrscheinlichkeit nicht eindeutig schätzen. Daher kommt auch die enorme Bedeutung von ME-Methoden.

19 Parameter für offene Skalen schätzen

19.1 Einleitung

Wenn wir – kontinuierliche oder diskrete – Skalen haben, die nach oben offen sind, gibt es einige besondere Dinge zu beachten. Das sieht man beispielsweise an folgendem Rätsel: Sie sitzen/liegen im Nachtzug, schlafen, wachen irgendwann auf, schauen auf dem Fenster. Sie sehen eine Straße mit Häusern, Sie sind also in einer Stadt, haben aber keine Anhaltspunkte für die Größe der Stadt. Sie sehen auch ein Taxi mit der Nummer 32 (nehmen wir an Taxis einer Stadt sind durchnummeriert). Sie sollen nun schätzen wie viele Taxis es gibt. Was schätzen Sie?

- ML sagt Ihnen: 32, denn das maximiert natürlich die Likelihood ihrer Beobachtung. Aber etwas an dieser Schätzung widerspricht unserer Intuition: ist es nicht unwahrscheinlich dass wir genau das "letzte" Taxi sehen?

- Intuitiv plausibler ist 64. Aber warum? Dazu müssen wir den Erwartungswert berücksichtigen: wenn wir 64 Taxis haben, alle gleich wahrscheinlich zu beobachten, dann liegt der Erwartungswert unserer Beobachtungen bei 32:

$$\sum_{i=1}^{64} i \cdot \frac{1}{64} = 32 \tag{231}$$

Warum haben wir in diesem Fall auf einmal Intuitionen, die so stark gegen ML sprechen? Der Grund liegt in der Natur des Parameter und Beobachtungen: dadurch dass wir wissen, dass es sich um eine Skala handelt, deren Parameter in eine Richtung offen sind, wissen wir auch, dass ML automatisch denjenigen Wert schätzt, der die Skala möglichst klein hält. Das impliziert dass unsere Beobachtung(en) am (in diesem Fall oberen) Rand der Skala liegen. Das ist aber nicht plausibel – viel plausibler ist es, dass sie sich um den Erwartungswert befinden.

19.2 Apriori Verteilungen über diskrete offene Skalen

Wir haben über *apriori*-Verteilungen gesprochen, und darüber, dass für endliche Räume die uniforme Verteilung die maximale Entropie hat. Nun

nehmen wir aber das obige Beispiel: wir haben eine abzählbar unendliche Menge von möglichen Parametern: es gibt

$$P(\text{es gibt } n \text{ Taxis}) : n \in \mathbb{N}$$

Wenn wir die Wahrscheinlichkeitsmasse uniform darüber verteilen, dann bekommt jedes n aber eine *apriori* Wahrscheinlichkeit von 0, wir haben also keine diskrete Verteilung mehr! Wenn wir also diskrete Verteilungen über abzählbar unendliche Mengen suchen, steht die uniforme Verteilung nicht mehr zur Verfügung. Was ist also die **neutralste Verteilung** über $\mathbb{N}$? Hier gibt es keine eindeutige Antwort, sondern eine ganze Familie von Funktionen.

Wir nennen Funktionen $P : \mathbb{M} \to [0, 1]$, die die Form haben

$$P_r(n) = (1 - r)r^{n-1} \tag{232}$$

mit $r \in [0, 1)$, **geometrische Verteilungen**. Jede geometrische Verteilung P_r hat den Erwartungswert $r/(1 - r)$, denn es gilt unabhängig von r dass

$$\sum_{i=1}^{\infty}(1 - r)r^{n-1} = r/(1 - r) \tag{233}$$

Ein Spezialfall hiervon ist die Funktion

$$P_{0.5}(n) = \frac{1}{2^n} \tag{234}$$

die wir bereit kennengelernt haben, und die nach der letzten Gleichung also den Erwartungswert 1 besitzt. Die Wichtigkeit der geometrischen Verteilungen wird durch folgendes Ergebnis belegt:

Lemma 16 *Für jeden Wert $r/(1 - r)$ ist die die geometrische Verteilung P_r die eindeutige Wahrscheinlichkeitsverteilung über $\mathbb{N}$ mit 1. diesem Erwartungswert und 2. der maximalen Entropie.*

Wir haben also eine Familie von Funktionen, die für ihren jeweiligen Erwartungswert die maximale Entropie haben. Man beachte auch folgendes: die Funktion

$$f(x) = \frac{x}{1 - x} \tag{235}$$

ist stetig und nimmt für $x \in [0,1)$ jeden Wert in $\mathbb{R}$ an; es gibt also für jeden Erwartungswert $x \in \mathbb{R}$ eine Verteilung P_r mit genau diesem Erwartungswert.

Es gibt jedoch noch das **Problem der Permutation**: für die Entropie spielt die Natur eines Ereignisses keine Rolle, sondern einzig dessen Wahrscheinlichkeit. Dementsprechend ändert sich die Entropie von P nicht unter Permutationen. Eine Permutation π ist eine Abbildung

$$\pi : \mathbb{N} \to \mathbb{N},$$

so dass

$$\pi[\mathbb{N}] = \mathbb{N} = \pi^{-1}[\mathbb{N}],$$

also eine Bijektion, für die außerdem jede Zahl ein Urbild hat ($f(n) = n + 1$ ist z.B. eine Bijektion, aber keine Permutation – die 1 hat kein Urbild). Es ist nun leicht zu sehen, dass für jede Permutation π gilt:

(236) $H(P_r) = H(P_r \circ \pi)$

Denn Addition ist kommutativ und kümmert sich also nicht um die Reihenfolge der Elemente. Allerdings ist

(237) $\mathcal{E}(P_r) < \mathcal{E}(P_r \circ \pi)$

(es folgt aus der Natur der geometrischen Verteilung dass jede Permutation den Erwartungswert nach oben schiebt). Weiterhin gibt es eine geometrische Verteilung $P_{r'}$ so dass

(238) $\mathcal{E}(P_{r'}) = \mathcal{E}(P_r \circ \pi)$

wobei dann natürlich $r' > r$ (je größer r in der geometrischen Verteilung, desto weiter nach rechts verschiebt sich der Erwartungswert. Daraus folgt natürlich wiederum, dass

(239) $H(P_{r'}) > H(P_r \circ \pi) = H(P_r)$,

also: je größer r, desto größer die Entropie. Da aber $r \in [0,1)$ liegt, gibt keinen Maximalwert.

19.3 Schätzen von kontinuierlichen Skalenparametern

(Nach Jaynes, Probability Theory, p190ff.) Nehmen wir einmal an, wir möchten schätzen, wie weit eine reellwertige Skala reicht, wobei wir eine Menge von Beobachtungen $D = \{x_1, ..., x_i\} \subseteq \mathbb{R}$ haben. Wir suchen $\alpha \in \mathbb{R}$, die Obergrenze der Skala, und unser *apriori* Wissen sagt uns, dass

$$(240)\quad P(x|\alpha, I) = \begin{cases} \frac{1}{\alpha}, \text{ if } 0 \leq x \leq \alpha \\ 0 \text{ andernfalls.} \end{cases}$$

Es ist leicht zu sehen, dass wir hier im Prinzip das Taxi-Problem aufgreifen, nur eben mit reellwertigen Parametern und der entsprechenden kontinuierlichen Wahrscheinlichkeitsfunktion. Eine Verteilung wie in (240) nennt man auch **rechteckig**; wem der Grund unklar ist, der zeichne sich den Graphen. Die Wahrscheinlichkeit unserer Daten, gegeben einen Parameter α und $0 \leq x_1, \ldots, x_i \leq \alpha$ lässt sich leicht berechnen als

$$(241)\quad P(D|\alpha, I) = \prod_{n=1}^{i} P(x_n|\alpha, I) = \frac{1}{\alpha^i}$$

Wenn wir die aposteriori-Verteilung möchten, brauchen wir einfach den Satz von Bayes der uns sagt:

$$(242)\quad P(\alpha|D, I) = P(D|\alpha, I)\frac{P(\alpha|I)}{P(D|I)}$$

$P(D|I)$ ist natürlich erstmal uninteressant (aber später wichtig); was jedoch wichtig ist, ist die apriori Wahrscheinlichkeit $P(\alpha|I)$. Wir legen einmal folgendes apriori fest:

$$(243)\quad P(\alpha|I) = \begin{cases} \alpha_1 - \alpha_0, \text{ falls } \alpha_0 \leq \alpha \leq \alpha_1 \\ 0 \text{ andernfalls.} \end{cases}$$

für feststehende α_0, α_1. Das setzt natürlich voraus, dass $x_1, ..., x_i \leq \alpha_1$, ansonsten haben wir eine logische Inkonsistenz.

19.4 Jeffreys Apriori-Verteilung

Harold Jeffreys hat als erster bemerkt, dass eine ebene Verteilung für einen kontinuierlichen, offenen Parameter nicht wirklich optimal ist um völlige Ignoranz zu modellieren. Stattdessen sollte die Verteilung uniform über den

Logarithmus des Parameters sein, d.h. es gibt eine konstante c so dass gilt:

(244) $P(log(\alpha)|I)c \propto \frac{1}{\alpha}$

(was heit das erste, wie kommt eins zum anderen???)

20 Induktives Lernen

20.1 Der Rahmen

Klassifikation ist ein erstes Beispiel für induktive Inferenz. Was dabei induziert wird ist eine

$$\text{Funktion } f, \text{ (z.B. } F : M \to N)$$

und zwar eine diskrete Funktion, d.h. eine Funktion die nur endlich viele verschiedene Eingaben nimmt und damit nur endliche viele Ausgaben liefert. Das Klassifikationsproblem ist also folgendes:

> Gegeben eine endliche Teilmenge von Instanzen von f, liefere eine Funktion h die f *approximiert*.

Wir sagen, dass $(m, n) \in M \times N$ eine **Instanz** von f ist, falls $f(m) = n$.

- Falls f eine stetige Funktion ist (z.B. $f : \mathbb{R} \to \mathbb{R}$), dann spricht man von *Regression*,
- falls f nur endlich viele Eingaben (und damit Ausgaben) hat, spricht man von **Klassifikation**.

Wir nennen wir unsere **Hypothese**.

$$h, \text{ wobei } h : M \to N$$

Das Grundproblem ist dass wir f normalerweise nicht kennen, d.h. wir können nie wissen, ob unsere Induktion erfolgreich war oder nicht. Alles was wir wissen können ist ob h übereinstimmt mit f auf dem endlichen Datensatz, den wir zur Verfügung haben. Eine entscheidende Rolle spielt dabei der sogenannte

Hypothesenraum H,

d.i. eine Menge von möglichen Funktionen, aus der wir h auswählen. Der Raum **H** ist durchaus nicht vorgegeben im Rahmen des Induktionsproblems, und die Wahl ist oft alles andere als einfach.

Das sieht man sehr schön am Beispiel einer *Regression*. Nehmen wir an, wir möchten eine Funktion

$$f : \mathbb{R} \to \mathbb{R}$$

induzieren, z.B. um die Korrelation von Tagestemparatur und Straftaten an einem gewissen Ort zu bestimmen (letztere gemittelt über einen längeren Zeitraum).

Was wir also gegeben haben ist eine Menge von Zahlenpaaren der Form

$$(10, 25.3),$$
$$(14, 27.8)$$
$$\text{etc.}$$

Nennen wir diese Menge D, unseren Datensatz. Da D eine Teilmenge des **Funktionsgraphen** von f ist, ist unsere Aufgabe ist wie folgt umrissen:

Finde eine Funktion $h \in H$, so dass für alle $(x, y) \in D$, $h(x) \approx y$.

Man nennt das auch die Konsistenzbedingung: die Hypothese soll konsistent mit den Daten sein. Und jetzt die Frage: was ist H? Hier gibt es folgende Überlegungen:

Je einfacher h ist, desto überzeugender würden wir es finden.

Z.B.: sei

$$h_1(x) = \frac{x}{2} + k \tag{245}$$

Das wäre sehr schön, und wir könnten sagen: ein Anstieg von 2° Celsius bedeutet eine zusätzliche Straftat. Wir könnten auch sagen: wenn es im August im Schnitt 25° wärmer ist als im Dezember, dann haben wir im Schnitt 12.5 Straftaten mehr pro Tag. Das wäre also eine sehr interessante Entdeckung!

Andererseits, es ist sehr unwahrscheinlich dass ein so komplexer, mittelbarer Zusammenhang so einfach ist, und so ist es sehr unwahrscheinlich, dass für alle $(x, y) \in D$ wir tatsächlich $h_1(x) = y$ haben. Es gibt sicherlich eine Funktion h_2, die in dieser Hinsicht wesentlich besser ist, z.B.

$$h_2(x) = x^5 + 6x^4 - 14x^3 + 15x^2 - 8x \tag{246}$$

Nehmen wir an, h_2 ist genauer auf D als h_1. Würden Sie sagen, dass h_2 plausibler ist? Eher nicht: wir würden sagen, dass die Komplexität von h_2 ein Anzeichen dafür ist, dass sie "maßgeschneidert" ist auf D und

schlecht generalisiert.

Das liegt v.a. daran, dass h_2 extrem komplex ist im Vergleich zu h_1. Wir treffen hier auf ein sehr grundlegendes Prinzip, nämlich das sog. **Rasiermesser von Ockham** (*Ockham's razor*), das besagt:

> *Die beste Hypothese aus einer Anzahl von Hypothesen die konsistent sind mit den Daten ist die einfachste.*

Allerdings sieht man bereits an unserem Beispiel, dass das eine sehr weiche Bedingung ist: denn h_2 passt besser als h_1, und es hängt nun alles davon ab, wie wir Konsistenz definieren. Es handelt sich also um eine weiche Richtlinie (die nichtsdestotrotz von grundlegender Bedeutung ist).

Wir können dieses Problem evtl. vermeiden, indem wir unseren Hypothesenraum *a priori* beschränken. Z.B. können wir sagen: uns interessieren nur die Polynome 2ten Grades, also Funktionen der Form

(247) $x^2 + ax + b$

Dabei gibt es folgendes zu beachten:

- Je kleiner der Hypothesenraum H, desto einfacher ist es, zwischen den konsistenten Hypothesen einen Kandidaten auszuwählen.
- Aber: je kleiner die Hypothesenraum, desto größer ist auch die Wahrscheinlichkeit, dass die korrekte Funktion gar nicht darin enthalten ist, also $f \notin H$.

Es gibt also Gründe die dafür und dagegen sprechen, H zu verkleinern. Wenn z.B. $f \notin H$, dann haben wir natürlich keine Möglichkeit, die korrekte Funktion zu induzieren. Da wir f nicht kennen, gibt es keine Möglichkeit, dass auszuschließen.

Nehmen wir z.B. an, die korrekte Korrelation (die wir natürlich nicht kennen) wäre

(248) $f(x) = ax + b + c\sin(x)$

das bedeutet: wir haben eine wachsende Wellenfunktion: Kriminalität erlebt bei steigenden Temperaturen immer wieder Scheitelpunkte.

- Solange wir also annehmen dass H aus Polynomialen besteht, werden wir niemals die richtige Funktion finden, sondern immer unmöglichere Polynomialfunktionen suchen müssen, solange wir mit neuen Daten konfrontiert werden!

Wir sehen also wie wichtig der richtige Hypothesenraum ist!

21 Klassifikation

21.1 (Boolesche) Entscheidungsfunktionen

Klassifikation ist ein erstes Beispiel für induktive Inferenz. Was dabei induziert wird ist eine Funktion f, und zwar eine diskrete Funktion, d.h. eine Funktion die nur endlich viele verschiedene Eingaben nimmt und damit nur endliche viele Ausgaben liefert. Wir werden uns hauptsächlich einen Spezialfall der Klassifikation anschauen, nämlich die **Boolesche Klassifikation**. Boolesche Klassifikation ist deswegen speziell, weil wir eine Boolesche (Wahrheits-)Funktion lernen. Wir suche eine Funktion,

- die für eine Eingabe x entweder "ja" oder "nein" liefert;
- wir fassen "ja" als 1, nein als 0 auf;
- weiterhin basiert eine solche Funktion auf einer Menge von **Attributen**, die auch entweder den Wert 0 oder 1 haben (das werden wir lockern), also erfüllt sind oder nicht.

Wir haben also eine Funktion

(249) $f : \{0,1\}^n \to \{0,1\}$

Um mit dem Konzept vertraut zu werden, erstmal folgendes Beispiel (aus Russel & Norvig): es geht um die Entscheidung, ob wir in einem Restaurant warten, bis wir einen Tisch zugewiesen bekommen, oder weitergehen; also eine binäre Entscheidung. NB: wir suchen also unsere eigene Entscheidungsfunktion, möchten also eine Funktion die uns für jedes Restaurant sagt, ob wir warten würden!

Die Attribute sind hier nicht alle binär, aber das tut erstmal nichts zur Sache. Als erstes stellen wir die Liste der Merkmale zusammen, die für unsere Entscheidungsfunktion relevant sind (schöner wäre es natürlich, wenn wir diese Attribute automatisch erstellen könnten, dazu später mehr). Unsere Merkmale sind:

1. Alternativen: gibt es passende Alternativen in der Nähe?
2. Theke: können wir uns an die Theke setzen und schonmal ein Bier trinken?
3. Fr/Sa: ist es Freitag oder Samstag?
4. Betrieb: wie viel Betrieb ist im Lokal? (Werte: leer, einige Leute, voll)
5. Regen: regnet es draußen?
6. Reservierung: haben wir reserviert?
7. Typ: was für eine Art Restaurant haben wir (französisch, italienisch, deutsch)
8. Geschätzte Wartezeit (von uns geschätzt): 0-10,10-30,30-60,>60

Das sind also die Faktoren, die bestimmen, ob wir auf einen freien Tisch warten. Nicht alle Attribute sind binär; wie können sie aber leicht darauf reduzieren; z.B. Attribut 4. kann aufgespalten werden in 2 Attribute: Leer: ja/nein und Voll: ja/nein. Unser Hypothesenraum besteht also aus allen Funktionen

(250) $h : \{0,1\}^3 \times \{0,1,2\} \times \{0,1\}^2 \times \{0,1,2\} \times \{0,1,2,3\} \to \{0,1\}$

Wie viele solche Funktionen gibt es? Nehmen wir einfachheitshalber mal an, **H** wäre die Menge aller Funktionen

(251) $h' : \{0,1\}^8 \to \{0,1\}$

Wie groß ist unser Hypothesenraum? Man könnte meinen er wäre nicht übermäßig groß; aber der Eindruck täuscht:

$$\text{es gibt } 2^{2^8} \text{ solche Funktionen, also } 2^{64}$$

– eine wahnsinnig große Zahl. Unser Hypothesenraum ist also riesig! Unser Ziel muss es sein, eine möglichst einfache Funktion aus diesem Raum zu wählen, die (nach unseren Begriffen) gut verallgemeinert. Hierbei greift man auf die sogenannten **Entscheidungsbäume** zurück.

Bsp.	Alt	Theke	Fr/Sa	Bet	Reg	Res	Typ	Wart	Warten?
d1	1	0	0	halb	0	1	fr	0-10	1
d2	1	0	0	voll	0	0	it	30-60	0
d3	0	1	0	halb	0	0	de	0-10	1
d4	1	0	1	voll	1	0	it	10-30	1
d5	1	0	1	voll	0	1	fr	>60	0
...									

Table 1: Ein Ausschnitt aus unserem Datensatz

21.2 Entscheidungsbäume

Boolesche Funktionen lassen sich einfach als Tabellen auffassen; wir nehmen nun wieder unser Beispiel, um das darzustellen: Tabelle 1 ist nur ein kleiner Ausschnitt unserer Funktion; wir können auch annehmen, es handelt sich um unseren Datensatz D. Ein **Entscheidungsbaum** ist einfach ein Baum,

1. in dem jeder Knoten ein Merkmal repräsentiert,
2. jedes Blatt einen Wert, den die Funktion annimmt;
3. auf jedem Pfad von der Wurzel zu einem Blatt kommt dabei jedes Merkmal höchstens einmal vor.

Jede Boolesche Funktion lässt sich als als Entscheidungsbaum darstellen: wir können einfach den Baum nehmen, in dem jede *Zeile* unserer Tabelle einem *Pfad* entspricht. Es gibt gewisse Boolesche Funktionen, die lassen sich nicht oder nur sehr schwer kompakt repräsentieren, z.B.

> die **Paritätsfunktion** (f nimmt den Wert 1 an, wenn eine gerade Zahl von Argumenten den Wert 1 annimmt), oder
>
> die **Majoritätsfunktion** (f nimmt den Wert 1 an, falls mindestens die Hälfte seiner Argumente den Wert 1 annimmt).

Allerdings gibt es auch Entscheidungsbäume, die eine wesentlich kompaktere Darstellung erlauben. Wenn wir das obige Beispiel betrachten, dann fällt uns z.B. auf dass wann immer die geschätzte Wartezeit >60 Minuten beträgt, dann warten wir niemals darauf dass ein Tisch frei wird. Wenn sich dieses Muster durch alle unsere Beobachtungen zieht, dann können wir also

dieses Merkmal an die Wurzel unseres Baumes setzen, und dann können wir in einigen Fällen den Baum an dieser Stelle schon mit dem Blatt 0 beenden. Algorithmen zur Induktion von Entscheidungsbäumen beruhen genau auf dieser Beobachtung:

> Wir können die Komplexität von Booleschen Funktionen messen nach der Komplexität der Entscheidungsbäume.

Das wiederum passt zu unseren obigen Beobachtung, dass einfache Funktionen eher sinnvolle, interessante Generalisierungen liefern als komplexe. Wir bekommen also folgendes:

> Gegeben eine Menge D von Daten, finde den enfachsten Entscheidungsbaum, der mit D konsistent ist; die zugehörige Boolesche Funktion ist unsere Hypothese h.

Wie finden wir? Man benutzt hier das sog. **Splitting**: wir nehmen das Merkmal, dass für unsere Unterscheidung **am informativsten** ist, und setzen es an die Wurzel des Entscheidungsbaumes. Dann nehmen wir das nächst-informativste Merkmal, setzen es als nächsten Knoten etc. Wie macht man das? Hier nutzen wir wieder einmal das Konzept der **Entropie**. Dafür müssen wir zunächst etwas arbeiten:

- Unser zugrundeliegende Raum ist eine Menge von Funktionen $X : M_1 \times M_2 \times \ldots \times M_i \to \{0, 1\}$.
- Wenn wir nun ein $n : 1 \leq n \leq i$ wählen, dann haben wir eine Funktion $X_n : M_n \mapsto (M_1 \times \ldots \times M_{n-1} \times M_{n+1} \times \ldots \times M_i \to \{0, 1\})$
- Das bedeutet: für jedes Merkmal, dass einen gewissen Wert annimmt, bekommen wir eine neue Funktion über die verbliebenen Merkmale.
- Wir möchten das Merkmal finden, das uns am besten die Menge der verbliebenen Funktionen aufteilt; insbesondere sollten die Teilmengen disjunkt sein!

Wir suchen also erstmal Merkmale M_n, für die gilt:

> Falls $m, m' \in M_n$, $m \neq m'$, dann ist $X_n(m) \cap X_n(m') = \emptyset$.

Das ist aber ein Kriterium, das gleichzeitig zu schwach (viele Merkmale können es erfüllen) und zu stark ist (in manchen Fällen wird es kein Merkmal geben, dass dieses Kriterium erfüllt.

Wir müssen also mal wieder Zuflucht zu Wahrscheinlichkeiten nehmen. Wir bauen daher den Wahrscheinlichkeitsraum $\mathfrak{A}$, wobei gilt:

1. $\Omega = M_1 \times M_2 \times \ldots \times M_i \to \{0, 1\}$ (die Menge der Ereignisse),
2. und für jedes $d \in \Omega$ gilt:

$$P(d) = \frac{1}{|D|} \text{ falls } f \in D, \text{ wobei } D \text{ unser Datensatz ist.}$$

Auf diesem Raum können wir nun eine Reihe von Zufallsvariablen $X_n : n \leq i$ definieren (wir fassen hier den Begriff etwas allgemeiner):

$$\text{Für } d = (m_1, \ldots, m_n, \ldots, m_i, x) \ (x \in \{0, 1\}),$$

gilt:

$$X_n(d) = m_n.$$

Man beachte, dass der **Zielwert** x (0 oder 1) hier nur ein weiteres Merkmal unter vielen ist! Nun hat jede dieser Zufallsvariablen eine Entropie, die sich errechnet als

$$H_P(X_n) = \sum_{m \in M_n} P(X_n = m) \cdot log(P(X_n = m)) \tag{252}$$

Damit bemessen wir, wie informativ eine Variable ist, und da die Variablen einem Merkmal entsprechen, bemessen wir also indirekt, wie informativ ein Merkmal ist. Das wäre aber zu allgemein: wir möchten ja nicht irgendein Merkmal vorhersagen, sondern ein ganz bestimmtes, unser **Zielmerkmal**. Hierzu brauchen wir das Konzept der **bedingten Entropie**:

$$\begin{aligned} H(X|Y) &= \sum_{y \in Y} H(X|Y = y) \\ &= \sum_{x \in X, y \in Y} P(X^{-1}(x) \cap Y^{-1}(y)) log \left(\frac{P(X^{-1}(x) \cap Y^{-1}(y))}{P(Y^{-1}(y))} \right) \end{aligned} \tag{253}$$

Insbesondere interessiert uns die Entropie der Variable X_{Ziel}, also des Zielwertes, gegeben dass wir den Wert eines Merkmals kennen:

$$H_P(X_{Ziel}|X_n) \tag{254}$$

Was jedoch wichtiger ist als dieser Wert (der ja auch sehr extrem sein kann, auch wenn M_n keinen Einfluss darauf hat) ist der **Informationsgewinn**; der ist wie folgt definiert:

(255) $IG_P(X_{Ziel}|X_n) = H_P(X_{Ziel}) - H_P(X_{Ziel}|X_n)$

Je geringer die bedingte Entropie im Vergleich zur unbedingten ist, desto größer ist der Informationsgewinn. Falls

(256) $H_P(X_{Ziel}|X_n) = 0$

also der Wert von X_{Ziel} vollständig von X_n bestimmt wird, dann ist

(257) $IG_P(X_{Ziel}|X_n) = H_P(X_{Ziel})$

Das bedeutet: wir gewinnen sämtliche Information, die in X_{Ziel} enthalten ist. Was wir damit also suchen ist:

(258) $\underset{1\leq n\leq i}{argmax}\ IG_P(X_{Ziel}|X_n)$

Das liefert uns das Merkmal, welches wir ganz oben in unseren Entscheidungsbaum stellen. Danach iterieren wir das mit den verbliebenen Variablen/Merkmalen: als nächstes interessiert uns

(259) $\underset{1\leq n\leq i}{argmax}\ H_P(X_{Ziel}|X_{max}) - H_P(X_{Ziel}|X_{max}, X_n)$

und so weiter, so dass wir also $i!$ Schritte benötigen (ein Schritt ist hier die Berechnung der bedingten Entropie). Das ist ein gutes Ergebnis, da die Anzahl der Merkmale normalerweise überschaubar ist!

21.3 Overfitting I

Vorher haben wir die Tatsache benutzt, das gewisse Merkmale informativer sind als andere. Es gibt hierbei aber ein mögliches Problem: dass ein Merkmal *zu informativ* ist, nämlich keine Generalisierung enthält. Das passiert insbesondere, wenn das Merkmal viele Werte annehmen kann, schlimmstenfalls mehr als unser Datensatz an Punkten enthält. Ein Beispiel hierfür wäre, wenn wir ein Merkmal **Datum** hinzunehmen. Unter der Annahme, dass wir an jedem Tag nur einmal essen gehen, ist klar dass wir damit einen perfekten

Bsp.	Alt	Theke	Fr/Sa	Bet	Reg	Res	Typ	Wart	Tag	Warten?
d1	1	0	0	halb	0	1	fr	0-10	5	1
d2	1	0	0	voll	0	0	it	30-60	18	0
d3	0	1	0	halb	0	0	de	0-10	9	1
d4	1	0	1	voll	1	0	it	10-30	26	1
d5	1	0	1	voll	0	1	fr	>60	17	0
...										

Table 2: Die Daten mit dem Tag des Monats

Prädiktor für unseren Datensatz haben: das Datum gibt uns eindeutig die richtige Klassifizierung. Das Problem ist: es gibt dabei keine Generalisierung! Das bleibt bestehen wenn wir ein Merkmal haben **Tag des Monats** – auch das mag bei einem relativ kleinen Datensatz ein guter Prädiktor sein, hat aber vermutlich keine Relevanz.

Das zugrundeliegende Problem ist also, dass

$$\frac{|M|}{|D|},$$

der relativ groß ist, im schlimmsten Fall > 1. Wie gehen wir mit diesem Merkmal um? Wir können ja nicht davon ausgehen, dass die Irrelevanz eines Merkmals derart offen zutage liegt. Hier können wir die klassische statistische Analyse nutzen: die **Nullhypothese** ist, dass das Merkmal keinen Einfluss hatte auf unsere jeweilige Entscheidung. Wir können nun versuchen, diese Hypothese zu widerlegen: wir müssen belegen, dass es wahrscheinlich ist, dass die Verteilung des Merkmals M rein zufällig ist.

Dafür überlegen wir zunächst:

- Wie viele Werte kann das Merkmal M annehmen? Wir nennen diese Zahl $|M|$.
- Wie würde es aussehen, wenn diese Merkmale rein zufällig über die anderen verteilt würden? Es würde zunächst gleichmäßig gestreut sein, d.h. keine besondere Ko-Okkurenz mit anderen Merkmalen haben.

Den zweiten Punkt kann man wie folgt verdeutlichen: da $|M|$ in kritischen Fall relativ groß ist. muss man ein Merkmal M' nehmen mit möglichst kleinem $|M'|$. Ein besonderes Beispiel hierfür wäre das "Zielmerkmal" $\{0, 1\}$, das wir eigentlich vorhersagen möchten. In diesem Fall ist die **Nullhypothese** klar numerisch formulierbar; wir benutzen unsere Zufallsvariablen X_n, setzen fest (qua Definition):

$$M = M_j \qquad M' = M_k$$

Nun sollte laut Nullhypothese gelten:

$$\text{für alle } m \in M_j, m' \in M_k,\ d \in D: \\ P(X_j = m | X_k = m') \approx P(X_j = m)$$

Da $P(X_j = m)$ aber naturgemäss (qua Annahme dass $\frac{|M|}{|D|}$ relativ groß ist) eine Zahl ist, die für uns schwierig von 0 zu unterscheiden ist, ist das noch problematisch; wir können aber folgendes machen: nehmen wir Einfachheit halber an, alle anderen Merkmale außer M sind binär. Dann können wir eine neue Zufallsvariable Y annehmen, die eine Summe von Werten denotiert:

$$Y(M) = \sum_{j \neq k} P(X_j = m | X_k = m') \tag{260}$$

(wobei i die Gesamtanzahl der Merkmale ist) als das Ergebnis eines $i-1$-Fach wiederholten Zufallsexperimentes lesen, wobei jeweils mit einem sehr großen Würfel geworfen wurde. Dementsprechend haben wir also eine Multinomialverteilung mit einem Erwartungswert

$$\mathcal{E}(Y) = \frac{i-1}{|M|} \tag{261}$$

mit einer entsprechenden symmetrischen Verteilung, Varianz und Standardabweichung. Das bedeutet: wir können die üblichen Methoden der Vertrauensgrenzen etc. ohne weiteres anwenden.

21.4 Overfitting II

Wir können auch im Rahmen unserer Methodik der Informationstheorie bleiben, und den Begriff der **bedingten Entropie** nutzen. Hier nochmals die Definition:

$$H(X|Y) = \sum_{y \in Y} H(X|Y = y) \\ = \sum_{x \in X, y \in Y} P(X^{-1}(x) \cap Y^{-1}(y)) log \left(\frac{P(X^{-1}(x), Y^{-1}(y))}{P(Y^{-1}(y))} \right) \tag{262}$$

Nach unseren Annahmen für $M = M_j$ hat die Zufallsvariable X_j sicher eine hohe/maximale Entropie. Es ist also genau die Eigenschaft, die sie eigentlich

positive hervorheben, die sie auch problematisch macht! Hier sehen wir die zwei Seiten derselben Medaille: je größer $|M_j|$, desto größer die Entropie von $H_P(X_j)$; aber je größer $|M_j|$, desto größer die Gefahr, dass das Merkmal eigentlich keine relevante Information enthält. Wir können uns nun mit der bedingten Entropie helfen: sei X_{Ziel} die Zufallsvariable, die das Zielmerkmal unserer Daten liefert. Wir können nun z.B.

(263) $H_P(X_j|X_{Ziel})$

berechnent. Falls nun gilt:

(264) $H_P(X_j|X_{Ziel}) \approx H_P(X_j)$

dann wissen wir, dass das Ergebnis einen geringen Einfluss auf X_j hat (den Tag des Monats). Im Umkehrschluss bedeutet dass, dass auch andersrum wenig Information fließt; wir haben zwar diesen Eindruck, aber das ist nur der Größe $|M_j|$ geschuldet.

Bsp.	Alt	Theke	Fr/Sa	Bet	Reg	Res	Typ	Wart	Warten?
d1	1	0	0	halb	0	1	fr	0-10	1
d2	1	0	0	voll	0	0	it	30-60	0
d3	0	1	0	halb	0	0	de	0-10	1
d4	1	0	1	voll	1	0	it	10-30	1
d5	1	0	1	voll	0	1	fr	>60	0
d6	1	0	1	voll	1	0	it	10-30	0
d7	0	1	0	halb	1	0	de	10-30	1
...									

Table 3: Ein Datensatz, der unsere Entscheidung nicht funktional bestimmt.

22 Probabilistische Graphische Modelle I - Bayesianische Netze

22.1 Einleitung

Nehmen wir einmal an, unser Datensatz ist so gestrickt, dass er keine Funktion mehr ist: das Zielmerkmal ist nicht mehr eindeutig durch die übrigen Merkmale determiniert. In Tabelle 3 etwa unterscheiden sich d4 und d6 nur durch den Zielwert, alle anderen Merkmale sind gleich!

Es ist klar, dass wir in diesem Fall keinen Entscheidungsbaum induzieren können: die Entscheidungen sind ja durch keinen Baum eindeutig bestimmt! Eine häufige Ursache für derartige Konstellationen ist, dass unsere gelisteten Faktoren nicht die einzig relevanten sind. Z.B. unsere Laune, Hunger, Begleitung etc. mag ebenfalls eine Rolle spielen, nur dass das Faktoren sind, über die wir keine Information haben. Das kann verschiedene Gründe haben:

- Die Information ist nicht oder nur schwer beobachtbar (z.B. unsere Laune)
- Der Indikator hat zu viele Werte, um sinnvoll benutzt zu werden (z.B. Begleitung)
- Es gibt einfach Nicht-determinismus!

Das bedeutet aber natürlich nicht, dass wir keine wertvolle Information aus den Merkmalen bekommen für unser Zielmerkmal: wir können z.B. leicht

sehen dass uns die Wartezeit immer noch eine ziemlich relevante Information liefert: indem sie nämlich die Wahrscheinlichkeitsverteilung ändert:

$$(265)\quad P(\text{Warten} = 1) = \frac{4}{7}$$

– d.h. wir haben wenig Information; aber wenn wir nach der obigen Methode die bedingte Wahrscheinlichkeit schätzen, dann bekommen wir:

$$(266)\quad P(\text{Warten} = 1|\text{Wartezeit} < 30) = \frac{4}{5}$$

Wir können insbesondere leicht sehen, dass die vorher angewandte Methodik der bedingten Wahrscheinlichkeit, bedingten Entropie nach wie vor problemlos angewendet werden kann. Aber was machen wir mit dieser Information? Die Frage ist also, welches Modell wir nutzen sollen; unsere allgemeinen Erwägungen bringen uns zu der Auffassung, das wir die in den Daten enthaltene Information nutzen sollten, um das Modell *möglichst einfach* zu gestalten. Wir werden hier ein besonders interessantes Modell betrachten, die sog. Bayesianischen Netze.

22.2 Definitionen

Ein **Graph** ist eine Struktur (V, E), wobei V eine Mengeo von Knoten ist (*vertices*), $E \subseteq V \times V$ die Kanten (*edges*). In Graphen gilt normalerweise: falls $(v_1, v_2) \in E$, dann $(v_2, v_1) \in E$, d.h. die Kanten haben keine Verbindung, sie repräsentieren nur Verbindungen. Ein **gerichteter** Graph ist ein Graph, in dem diese Bedingung fallengelassen wird: Kanten sind gerichtet. Ein gerichteter azyklischer Graph ist ein gerichteter Graph, in dem folgendes gilt: ein **Zyklus** ist eine Sequenz von Kanten

$$(v_1, v_2), (v_2, v_3), ..., (v_{i-1}, v_i), \text{ bei der } v_i = v_1$$

gilt. Wir folgen also den (gerichteten) Kanten und kommen zum Ausgangspunkt zurück. Ein **gerichteter azyklischer Graph** ist nun einfach ein gerichteter Graph, der keine Zyklen enthält. Wir werden hier normalerweise endliche Graphen betrachten.

Wir haben bereits Markov-Ketten kennengelernt; was wir nun machen ist folgendes: wir betrachten Ketten als Spezialfälle von gerichteten azyklichen Graphen mit der **Markov Eigenschaft**, und wollen nun zum allgemeineren Fall. Zunächst müssen wir die Definition betrachten: eine Markov-Kette hatte die Form

$$X_1, X_2, X_3, ..., X_n, ...$$

wobei jedes X_i eine Zufallsvariable war. Weiterhin gilt:

Fall $i < j < k$, dann ist $P(X_k = y_k | X_j = y_j, X_i = y_i) = P(X_k = y_k | X_j = y_j)$

D.h. bedeutet der Informationsfluss entlang der Kette wird dadurch, dass wir den Wert eines Zwischengliedes kennen, *blockiert*. (Erinnern Sie sich, dass viele andere Dinge, die man auf den ersten Blick meinen würde, nicht gelten!) Anders gesagt, wenn wir den Wert eines Gliedes kennen, sind die Werte aller vorigen Glieder irrelevant.

Wie verallgemeinern wir das? Zunächst folgende Begriffe: ein GAG entspricht einer **partiellen Ordnung** $\leq$, welche die transitive Hülle der Kanten E ist. Das heißt sie erfüllt folgende Axiome:

1. Reflexivität: $x \leq x$
2. Antisymmetrie: $x \leq y \ \& \ y \leq x \Rightarrow x = y$
3. Transitivität: $x \leq y \ \& \ y \leq z \Rightarrow x \leq z$

(Das entspricht der natürlichen Ordnung der Zahlen oben). $<$ ist die irreflexive Variante von $\leq$. Ein weiterer Begriff ist der des unmittelbaren Vorgängers. Wir definieren:

$$eltern(v) = \{v' : v' < v, \text{ es gibt kein } v'' : v' < v'' < v\}$$

Bayesianische Netze basieren darauf, dass wir einen GAG (V, E) haben, und jedem $v_k \in V$ eindeutig eine Zufallsvariable X_k zugewiesen wird. Wir können bereits einen der wichtigsten Begriffe der Bayesianischen Netze formulieren: wir sagen X_k **blockiert** einen Pfad von X_i nach X_j, falls v_k auf diesem Pfad von v_i nach v_j liegt.

Nun kommt die Definition von Bayesianischen Netzen: ein solches Netz ist eine Struktur $(V, E, \mathbf{X})$, wobei $\mathbf{X}$ eine Menge von Zufallsvariablen ist die eindeutig Knoten in V zugeordnget werden, und die die Markov-Eigenschaft im Hinblick auf (V, E) erfüllen. Aber was genau heißt das? Tatsächlich ist das keine leichte Frage, und die Antwort hält einige Überraschungen bereit.

Definition 17 *Eine Menge von Verteilungen $\boldsymbol{X}$ erfüllt die Markov Eigenschaft im Hinblick auf einen GAG (V, E), falls für alle $X_k \in \boldsymbol{X}$, $\boldsymbol{Y} \subseteq X$ gilt: falls $Y < X$ für alle $Y \in \boldsymbol{Y}$, dann ist*

$$P(X|\,\boldsymbol{Y}, eltern(X)) = P(X|eltern(X))$$

Das bedeutet: um die genaue Verteilung einer Variable gegeben eine Teilmenge ihrer Vorgänger zu kennen, reicht es aus, die Werte der Eltern zu kennen. Die Beschränkung auf Vorgänger ist sehr wichtig, wie wir später sehen werden! Für uns ist zunächst das wichtigste: ein Bayesianisches Netz wird induziert durch eine Reihe bedingter Wahrscheinlichkeitsverteilungen

$$P(X_v|X_{v_1}, ..., X_{v_k}), \text{ wobei } \{v_1, ..., v_k\} = eltern(v).$$

Das ganze sieht normalerweise wie folgt aus (praktisch): wir nehmen an, die Zufallsvariablen nehmen nur endlich viele Werte an (aber das ist nur eine pädagogische Vereinfachung). Das bedeutet: wir spezifizieren

$$P(X_v = x|X_{v_1} = y_1, ..., X_{v_k} = y_k),$$

für alle

$$x \in X_v,\ y_1 \in X_{v_1}, ..., y_k = X_{v_k}),$$

also jeden Wert, den die Variablen annehmen können. Damit, und mit den Regeln der Wahrscheinlichkeitstheorie, ist die resultierende Wahrscheinlichkeitsverteilung vollkommen determiniert (und sie ist die Verteilung eines Bayesianischen Netzes!).

Bsp.	Jahreszeit	Temperatur	Regen	Schnee/Eis	Unfälle
d1	Sommer	mittel	1	0	40
d2	Winter	niedrig	0	1	90
d3	Winter	niedrig	0	0	31
d4	Sommer	hoch	0	0	30
d5	Winter	niedrig	0	0	45
...					

Table 4: Abhängigkeiten und bedingte Unabhängigkeiten

22.3 Die Intuition

Die Intuition hinter diesen Strukturen ist folgende: es mag durchaus sein, dass eine ganze Reihe von Faktoren Einfluss hat auf die Wahrscheinlichkeit eines Ereignisses; allerdings ist die Wahrscheinlichkeit bereits von einer Teilmenge der Ereignisse bestimmt – wenn wir gewisse Dinge wissen, dann spielen andere keine Rolle, die sonst allerdings eine Rolle spielen würden. Wir können z.B. den Datensatz in Tabelle 4 betrachten:

Hier entspricht jeder Datenpunkt einem bestimmten Tag. Es ist klar, dass jeder einzelne Faktor einen Einfluss auf die Zahl der Unfälle hat, und dementsprechend auf die Unfallwahrscheinlichkeit (wir können das errechnen als Unfälle/Bevölkerung, als einfachste Lösung).

Des weiteren ist aber auch folgendes klar: wenn wir wissen, dass Schnee/Eis positiv ist, dann spielen Temperatur und Jahreszeit keine Rolle mehr. Sommer/Winter mögen relevant sein, aber nur insofern, als sie die Häufigkeit von Regen beeinflussen; ebenso weil Winter die Wahrscheinlichkeit von Schnee/Eis erhöht, wodurch die Unfallzahlen am stärksten steigen. Natürlich spielen noch eine Reihe anderer Faktoren eine wichtige Rolle; insbesondere z.B. das *Verkehrsaufkommen* – darüber haben wir aber keine gesicherten Information, daher müssen wir das als einen reinen **Störfaktor** auffassen, also einen Faktor, der sich auf unberechenbare Art und Weise in unseren Wahrscheinlichkeiten widerspiegelt. Z.B. d5 ließe sich auf diese Art und Weise erklären. Allerdings gibt es hier zu beachten: es kann natürlich auch sein, dass dieser Faktor *systematisch* wirkt, nämlich dadurch, dass winters ein höheres Verkehrsaufkommen ist als im Sommer.

Dadurch, dass es im Normalfall derartige Störfaktoren gibt, wird man auch mit aus solchen Datensätzen geschätzten Wahrscheinlichkeiten *praktisch niemals* eine echte konditionale Unabhängigkeit finden – das wäre sogar

dann unwahrscheinlich, wenn die Daten von einem echten Bayesianischen Netz generiert würden, rein aus Zufall. Daher ist es nicht immer einfach, den korrekten unterliegenden Graphen zu finden, und man wendet oft eine Mischung aus *common sense* und Datenanalyse an, um zu einem befriedigenden Ergebnis zu kommen. Dazu später mehr!

22.4 Rechnen mit BNs

Für Markov Ketten haben wir bereits folgende Beobachtung gemacht:

> Information läuft immer in beide Richtungen der Kette, (Un-) Abhängigkeit ist immer symmetrisch.

Dasselbe gilt für BNs; es gibt allerdings auch einen sehr wichtigen Unterschied: für Markov Ketten gibt es eine **Symmetrie**: die Verteilungen selber geben uns niemals Aufschluss über Direktionalität; wir wählen sie immer willkürlich. Für BNs gilt das **nicht**; das ist eine der wichtigsten Beobachtungen in diesem Kontext. Um das zu sehen, müssen wir zunächst lernen, wie wir Wahrscheinlichkeiten in einem BN effektiv berechnen (gegeben $(V, E, \mathbf{X})$) und entsprechende bedingte Wahrscheinlichkeitsverteilungen

$$P(X|eltern(X)), \text{ für alle } X \in \mathbf{X}$$

Zunächst berechnen wir einfache *unbedingte Wahrscheinlichkeiten* der Form $P(X_v = x)$. Das ergibt sich aus folgender Gleichung:

(267)
$$P(X_v = x) = \sum_{y_1 \in X_1, ..., y_i \in Y_i} P(X_v = x | X_1 = y_1, ..., X_i = y_i) P(X_1 = y_1, ..., X_i = y_i),$$

wobei $\{X_1, ..., X_i\} = eltern(X)$. 267 ist nichts weiter als die bekannte Regel der **Marginalisierung**. Das bedeutet aber: um die unbedingte Wahrscheinlichkeit eines *einzelnen* Ereignisses $X_v = x$ auszurechnen, muss man bereits die gesamten unbedingten Vorgängerwahrscheinlichkeiten berechnen – für alle Werte, die sie annehmen können!

> Es ist leicht zu sehen dass damit der Rechenaufwand um $P(X_v = x)$ zu berechnen *exponentiell* ist in der Anzahl von Vorgängern von X_v.

Das ist natürlich ein Problem, dass sich allerdings nur stellt, wenn uns diese Frage wirklich *ad-hoc* interessiert (was normalerweise eher selten der Fall ist). Angesichts dessen ist auch klar, warum wir die Anzahl der Knoten/Kanten immer möglichst klein halten sollten: abgesehen von allgemeinen Erwägungen (Ockhams Rasiermesser) sprechen auch ganz konkrete Berechenbarkeitsüberlegungen dafür!

Als nächstes interessiert uns die Frage, wie wir *beliebige bedingte Wahrscheinlichkeiten* brerechnen. Das geht nach folgender Gleichung:

(268)
$$P(X_v = x|\mathbf{Y}) = \sum_{y_1 \in X_1, ..., y_i \in Y_i} P(X_v = x|\mathbf{Y}, X_1 = y_1, ..., X_i = y_i) P(X_1 = y_1, ..., X_i = y_i|\mathbf{Y}),$$

wobei wiederum $\{X_1, ..., X_i\} = eltern(X) - \mathbf{Y}$, also die Menge der Eltern-Variablen ist, die nicht in $\mathbf{Y}$ enthalten ist. (268) liefert einen allgemeinen Fall; die Berechnung kann relativ einfach sein, falls

$$\mathbf{Y} \cap eltern(X)$$

verhältnismäßig groß ist, oder aber $\mathbf{Y}$ aus Vorgängern von X_v besteht. Falls $\mathbf{Y}$ aber *Nachfolger* von X_v enthält, wird die Berechnung nochmals aufwändiger. Die Berechnung einer Verteilung

$$P(X_1 = x_1, ..., X_i = x_i|Y_1 = y_1, ..., Y_i = y_i)$$

Lässt sich wiederum sehr einfach auflösen nach der Produktregel.

22.5 Konditionale (Un-)Abhängigkeit

Einer der wichtigsten Begriffe ist der der konditionalen Unabhängigkeit in BNs. Sie ist wie folgt definiert:

$(X \| Y | \mathbf{Z})$ gdw. für alle $x \in X$, $y \in Y$, $P(X = x | Y = y, \mathbf{Z}) = P(X = x | \mathbf{Z})$

In Worten bedeutet das soviel wie: wenn wir die Werte der Variablen $\mathbf{Z}$ kennen, dann spielt der Wert von Y keine Rolle für die Verteilung von X. Insofern ist die Markov Bedingung in BNs nur ein Fall von konditionaler Unabhängigkeit, der erfüllt sein muss. Es gilt aber noch viel mehr:

(269) $(X \| Y | Y)$

gilt immer, ebenso

(270) $(X \| Y | \mathbf{Y})$, falls $Y \in \mathbf{Y}$

Weiterhin gilt:

(271) $(X \| Y | \mathbf{Z}) \Leftrightarrow (Y \| X | \mathbf{Z})$

Konditionale Unabhängigkeit ist also eine symmetrische Eigenschaft (wie wir das erwarten würden: Information fließt immer in beide Richtungen.

Interessanterweise lässt sich die konditionelle Unabhängigkeit zweier Variablen auf rein strukturelle Eigenschaften des unterliegenden Graphen reduzieren, nämlich durch das Konzept der *d*-**Separation**. Ein wichtiger Begriff ist der eines **Pfades**, der etwas unintuitiv definiert wird; ein Pfad in einem Netz ist einfach eine Sequenz von Kanten

$$(v_1, v_2), (v_2, v_3), ..., (v_{i-1}, v_i)$$

in dem

1. adjazente Knoten jeweils identisch sind, und
2. für alle (v_l, v_{l+1}) des Pfades gilt: entweder $(v_l, v_{l+1}) \in E$ oder $(v_{l+1}, v_l) \in E$.

Wir sagen v liegt auf dem Pfad P mit der offensichtlichen Bedeutung dass für

$$P = (v_1, v_2), (v_2, v_3), ..., (v_{i-1}, v_i)$$

wir folgendes haben:

$$v = v_k \text{ für ein } k \text{ so dass } 1 \leq k \leq i$$

Teilpfade sind zusammenhängende Teilsequenzen eines Pfades. Wir unterscheiden zwei Arten von (Teil-)Pfaden, nämlich

1. eine *Kette* hat die Form $(v_1, v_2), (v_2, v_3)$, wobei $(v_1, v_2) \in E$ und $(v_2, v_3) \in E$; (also $v_1 \rightarrow v_2 \rightarrow v_3$)

2. eine *Gabel* hat die Form $(v_1, v_2), (v_2, v_3)$, wobei $(v_2, v_1) \in E$ und $(v_2, v_3) \in E$; (also $v_1 \leftarrow v_2 \rightarrow v_3$)

3. ein V hat die Form $(v_1, v_2), (v_2, v_3)$, wobei $(v_1, v_2) \in E$ und $(v_3, v_2) \in E$; (also $v_1 \rightarrow v_2 \leftarrow v_3$)

Die entscheidende Definition ist folgende:

Definition 18 *Ein Pfad P in (V, E) ist d-separiert von einer Menge von Knoten $M \subseteq E$, falls eines der folgenden gilt:*

1. *P enthält einen Teilpfad $(v_1, v_2), (v_2, v_3)$, die entweder eine Kette oder eine Gabel ist, und $v_2 \in M$; oder*

2. *P enthält einen Teilpfad $(v_1, v_2), (v_2, v_3)$, der ein V ist, und $v_2 \notin M$.*

Zwei Knoten v, v' in (V, E) sind d-separiert von einer Menge $M \subseteq V$, genau dann wenn alle Pfade in (V, E) von v nach v' d-separiert sind von M.

Wir haben Knoten im Graphen eindeutig Zufallsvariablen zugeordnet. Dementsprechend können wir sagen dass in einem BN $(V, E, \mathbf{X})$ eine Variable Z zwei Variablen X, Y d-separiert. Das entscheidende Ergebnis ist folgendes:

Theorem 19 *Für jedes BN $(V, E, \boldsymbol{X})$, $X, Y \in \boldsymbol{X}$, $\boldsymbol{Z} \subseteq \boldsymbol{X}$ gilt: $(X \| Y | \boldsymbol{Z})$ genau dann wenn X, Y d-separiert sind von $\boldsymbol{Z}$ in $(V, E, \boldsymbol{X})$.*

Das bedeutet: um das Kriterium der konditionellen Unabhängigkeit zu verifizieren reicht es, einen Blick auf den Graphen des BN zu werfen!

22.6 Minimalität und Direktionalität

Der Normalfall ist allerdings der, dass wir das BN nicht als gegeben bekommen; stattdessen haben wir eine Wahrscheinlichkeitsverteilung über die Werte der Variablen; also eine Funktion

$$(\#)\ P : M_1 \times M_2 \times ... \times M_i \to [0, 1],$$

die die üblichen Bedingungen eines Wahrscheinlichkeitsraumes erfüllt (man nennt das eine **multivariate Verteilung**. Wichtig ist aber: es handelt sich hier nicht um einen Produktraum im engeren Sinne, d.h. die Wahrscheinlichkeiten der Komponenten sind *nicht* voneinander unabhängig. Das Ziel ist nun folgendes: wir suchen ein graphisches Modell (d.h. für uns ein BN), dass **minimal** ist, d.h. eine minimale Anzahl von Abhängigkeiten hat. Die Motivation hierfür ist folgende: je weniger Anhängigkeiten ein Modell hat, desto *einfacher* ist, und je einfacher es ist, desto *besser*.

Es gibt hier aber noch eine wichtige Eigenschaft: wir haben bereits gesagt, dass Abhängigkeiten in BNs direktional sind, d.h. sie gehen von A nach B, nicht umgekehrt. Diese Direktionalität kann man verknüpfen mit dem Begriff der **Kausalität**: Kausalität ist – im Gegensatz zu konditionaler Abhängigkeit – ein asymetrischer Begriff. Dadurch wird klar, dass konditionale Abhängigkeit kein guter Indikator für Kausalität sein kann. Wir haben bereits gesehen, dass es oftmals unklar ist, in welche Richtung Kausalität fließt: erinnern wir uns an das Beispiel von Regen/Temperatur: Regen beeinflusst die Temperatur (kühlt ab); aber die Temperatur hat auch Einfluß auf den Regen (bei großer Kälte kein Regen, bei Wärme öfter Gewitter). Hier kommt uns nun unser Modell zur Hilfe:

> Wenn es für jede Verteilung P wie in (#) ein eindeutiges, minimales BN gäbe, dass P erzeugt, dann könnten wir an der Direktionalität seiner Kanten die Richtung der kausalen Wirkung ablesen.

Leider gilt das nur bedingt: wir wissen bereits, dass für ein Netz der Form

$$X_1 \to X_2 \to X_3$$

wir ein äquivalentes Netz haben der Form:

$$Y_1 \leftarrow Y_2 \leftarrow Y_3$$

Wir können das transformieren mittels:

$$(272)\quad P(Y_1|Y_2) = P(X_2|X_1)\frac{P(X_1)}{P(X_2)}$$

und

$$(273)\quad P(Y_2|Y_3) = P(X_3|X_2)\frac{P(X_2)}{P(X_3)}$$

wobei man $P(X_3)$ etc. nach den gewohnten Regeln ausrechnet. Ebenso für ein Netz der Form

$$X_1 \leftarrow X_2 \rightarrow X_3$$

Denn wir können auch hier die Richtungen umdrehen, z.B. zu

$$(+)\ Y_1 \rightarrow Y_2 \rightarrow Y_3$$

(Ein ähnliches Argument wie oben) Das gilt allerdings nicht mehr, wenn wir sog. V-Strukturen haben:

$$(*)\ X_1 \rightarrow X_2 \leftarrow X_3$$

Warum ist das so? Die Antwort hat mit dem Stanford-Musiker Paradoxon zu tun. Zunächst führen wir das Symbol $\dot{\sim}$ mit der Bedeutung der *konditionalen Abhängigkeit* ein, also

$$(274)\quad X_1 \dot{\sim} X_2|X_3 \Longleftrightarrow \neg(X_1\|X_2|X_3)$$

Zunächst haben wir in (*) eine konditionale Abhängigkeit von X_2 und X_1, egal ob X_3 gegeben ist oder nicht; ebenso umgekehrt für X_2, X_3 gegeben X_1; also

$$X_1 \dot{\sim} X_2|X_3 \quad X_3 \dot{\sim} X_2|X_1$$

Das allein ist aber noch kein Argument, denn dasselbe gilt für (+). Allerdings haben wir auch noch folgendes:

$$X_1 \dot{\sim} X_3|X_2$$

Das gilt nun weder für Ketten noch für Gabeln, wie wir oben gesehen haben! Daraus können wir folgern: Für eine Verteilung über drei Variablen, die folgende 3 Bedingungen erfüllt:

$$X_1 \not\perp\!\!\!\perp X_2 | X_3 \quad X_3 \not\perp\!\!\!\perp X_2 | X_1 \tag{275}$$

$$X_1 \not\perp\!\!\!\perp X_3 | X_2 \tag{276}$$

$$X_1 \| X_3 | \emptyset \tag{277}$$

In diesem Fall ist eine V-Struktur das eindeutige minimale BN. Das wiederum erlaubt es uns, die Direktionalität der Pfeile zu inferieren! Selbstverständlich gilt das auch *a fortiori* für größere Graphen/Verteilungen mit mehr Variablen, in denen diese Teilverteilungen auftreten.

> Gegeben eine Verteilung mit n Variablen können wir tatsächlich – bis auf Direktionalität in Gabeln und Ketten – einen eindeutiges minimales BN finden.

Das ist noch etwas ungenau; genauer wird es durch folgenden Definitionen und Ergebnisse (die auch noch etwas salopp formuliert sind):

Definition 20 *Zwei GAGs G_1, G_2 sind **probabilistisch äquivalent**, wenn für jedes BN, dass auf G_1 basiert, es ein auf G_2 basiertes BN gibt, dass dieselbe Wahrscheinlichkeitsverteilung beschreibt.*

Ein wichtiges Ergebnis ist folgendes:

Theorem 21 *Zwei GAGs G_1, G_2 sind probabilistisch äquivalent, genau dann wenn das gleiche Skelett (ungerichtete Kanten) und die gleiche Menge an V-Strukturen darüber haben.*

22.7 Von der Verteilung zum Graphen

Nun die Frage: wie komme ich von einer **multivariaten Verteilung** zu den bedingten Wahrscheinlichkeiten, die ich für mein BN brauche? Hier können wir einmal mehr unsereursprüngliche Definition der bedingten Wahrscheinlichhkeit ausgraben. Sei eine multivariate Verteilung

$$P : M_1 \times M_2 \times ... \times M_i \rightarrow [0, 1]$$

gegeben. Dann berechnen wir die bedingte Wahrscheinlichkeit $P(m_1|m_2, ..., m_i)$ wie folgt:

$$(278)\quad P(m_1|m_2, ..., m_i) = \frac{P(m_1, m_2, ..., m_i)}{P(m_2, ..., m_i)}$$

dasselbe für andere bedingten Wahrscheinlichkeiten. $P(m_2, ..., m_i)$, $P(m_2)$ etc. sind wiederum marginale Wahrscheinlichkeiten, die man nach den gewohnten Regeln berechnet:

$$(279)\quad P(m_1) = \sum_{m_2 \in M_2, ..., m_i \in M_i} P(m_1, m_2, ..., m_i)$$

Mit diesen beiden Regeln können wir also sämtliche bedingten und unbedingten (marginalen) Wahrscheinlichkeiten ausrechnen, und dementsprechend auch ein minimales BN konstruieren. Es sollte aber klar sein, dass das mit einem erheblichen Aufwand verbunden ist, v.a. wenn es eine große Zahl von Variablen gibt.

Aufgabe 9

Abgabe bis zum 4.7.2017 vor dem Seminar. Nehmen Sie folgende Graphen:
$G1 = (\{1, 2, 3, 4\}, \{(1, 3), (2, 3), (2, 4), (3, 4)\}$
$G2 = (\{1, 2, 3, 4, 5\}, \{(1, 2), (1, 3), (4, 3), (4, 5)\}$
$G3 = (\{1, 2, 3, 4\}, \{(1, 2), (1, 3), (2, 3), (4, 3)\}$
Nehmen Sie, jeder dieser Graphen beschreibt ein bayesianisches Netz mit Variablen $\{X_1, ..., X_5\}$, die den Knoten wie offensichtlich zugeordnet sind. Sagen Sie, ob folgendes gilt, mit Begründung:

1. in $G1$: $X_1 \| X_4 | X_3$
2. in $G2$: $X_2 \| X_5 | X_3$
3. in $G3$: $X_1 \| X_4 | X_3$

Aufgabe 10

Abgabe bis zum 4.7.2017 vor dem Seminar. Nehmen sie folgende multivariate Wahrscheinlichkeitsverteilung

$$P : M_1 \times M_2 \times M_3 \to [0, 1],$$

wobei

$$M_1 = M_2 = M_3 = \{0, 1\}$$

Und die Wahrscheinlichkeitsverteilung wie folgt ist (es gibt $2^3 = 8$ Ereignisse):

Ereignis	Wahrscheinlichkeit
(0,0,0)	0.252
(0,0,1)	0.096
(0,1,0)	0.252
(0,1,1)	0.054
(1,0,0)	0.028
(1,0,1)	0.024
(1,1,0)	0.168
(1,1,1)	0.126

Liefern Sie die ein minimales BN mit Variablen X_1, X_2, X_3, das diese Verteilung generiert!

23 PAC-Lernen

23.1 Einleitung

Es gibt eine Vielzahl von formalen und computationellen Lerntheorien; die einzige, die (meines Wissens) wirklich in der Praxis relevant geworden ist, ist das PAC-Lernen, weil man darin auch nach endlich vielen Schritten starke Aussagen über den Lernerfolg machen kann.

Das Problem bei der Induktion von einer gewissen Menge von Beobachtungen ist, dass immer eine Ungewissheit bleibt: ist unsere Generalisierung richtig? Hier sorgen viele Faktoren für Ungewissheit:

- Vielleicht ist die korrekte Hypothese gar nicht im Hypothesenraum;
- vielleicht ist sie darin, aber wir (unser Algorithmus) hat nicht die plausibelste Hypothese (gegeben die Datenlage) ausgewählt, weil er unsere Herangehensweise nicht optimal ist;
- oder aber: wir haben alles bestmöglich gemacht, aber wir hatten einfach Pech mit unseren Beobachtungen: anstatt normaler, repräsentativer Ereignisse haben wir unwahrscheinliche, irreführende Beobachtungen gemacht.

PAC-Lernen konzentriert sich insbesondere auf den letzten Punkt. Das entscheidende ist: natürlich kann es immer sein, dass unsere Beobachtungen nicht repräsentativ sind, aber mit zunehmender Größe unseres Datensatzes wird das immer unwahrscheinlicher.

PAC steht für *probably approximately correct*, und intuitiv gesagt bedeutet PAC-Lernen: wir lernen auf eine Art und Weise, dass es immer unwahrscheinlicher wird, dass wir unsere Hypothese mehr als eine beliebig kleine Distanz von der korrekten Hypothese entfernt ist. Das bedeutet umgekehrt: eine Hypothese, die ernsthaft falsch ist, wird fast mit Sicherheit als falsch erkannt; wenn wir eine Hypothese für richtig halten, dann ist sie mit großer Wahrscheinlichkeit sehr nahe an der korrekten Zielhypothese. Um so etwas sagen zu können, brauchen wir allerdings die passenden Rahmenbedingungen.

23.2 Definitionen

Zunächst müssen wir eine Reihe von Annahmen. Die erste ist die Annahme der **Stationarität**:

> Alle relevanten Beobachtungen, die wir machen, werden von derselben Wahrscheinlichhkeitsverteilung generiert.

Das ist eine sehr wichtige Annahme, und in gewissem Sinn die Voraussetzung induktiven Lernens: wenn die Verteilung im Laufe der Zeit sich (beliebig) ändert, dann erlauben uns die Beobachtungen, die wir gemacht haben, keinerlei Rückschlüsse auf zukünftige Beobachtungen, und jede Form von Induktion ist unmöglich.

Weiterhin haben wir folgendes;

- M ist die Menge aller möglichen Beobachtungen (korrekte Klassifikationen etc., üblicherweise bekannt)
- P ist eine Wahrscheinlichkeitsverteilung über M, die uns sagt wie wahrscheinlich eine Beobachtung ist (üblicherweise unbekannt)
- f ist die Zielfunktion, die wir lernen möchten (unbekannt; wir nehmen wieder an, wir lernen eine Funktion)
- H ist die Menge der Hypothesen, die uns zur Verfügung stehen (bekannt)
- $N = |D|$ ist die Anzahl der Beobachtungen, anhand derer wir unsere Hypothese $h \in H$ auswählen (bekannt bzw. variabel)

Wenn nun f die Zielfunktion ist, h eine Hypothese, dann können wir die Fehlerhaftigkeit von h genau quantifizieren (zumindest abstrakt; konkret kennen wir natürlich die Zahlen nicht):

(280) $error(h) = P(h(x) \neq f(x)|x \in M)$

das bedeutet, etwas genauer,

(281) $error(h) = P(X) : X = \{x \in M : h(x) \neq f(x))$

Das setzt natürlich voraus, dass P eine *diskrete* Wahrscheinlichkeitsfunktion über M ist, sonst können wir nicht garantieren, dass die Wahrscheinlichkeit definiert ist.

Wir sagen dass h **annähernd korrekt** ist, falls $error(h) \leq \epsilon$, wobei ϵ eine beliebig kleine Konstante ist.

(ϵ müssen wir natürlich festlegen). Beachten Sie aber dass hierbei 2 unbekannte auftauchen:

1. f, die Zielhypothese die wir nicht kennen, und
2. P, die Verteilung über den Daten, die wir nicht kennen.

Das eigentlich geniale am PAC-Lernen ist dass wir so arbeiten, dass sich die unbekannten "rauskürzen".

Zunächst unterscheiden wir zwei Arten von Hypothesen, nämlich solche, die **ernsthaft falsch**, und solche, die **annähernd korrekt** sind, auf in

$$H_{\downarrow} = \{h : error(h) > \epsilon\} \tag{282}$$

$$H_{\uparrow} = \{error(h) \leq \epsilon\} \tag{283}$$

Wir können uns $H_{\uparrow}$ vorstellen als eine Kugel, die einen gewissen Radius um die korrekte Hypothese hat.

Nun nehmen wir eine Hypothese h, die wir erstellt haben. Nach unserer Konstruktion gilt: h ist konsistent mit den N Beobachtungen, die wir gemacht haben. Uns interessiert die Wahrscheinlichkeit

$$P(h \in H_{\downarrow}),$$

also die Wahrscheinlichkeit, dass unsere Hypothese "ernsthaft falsch" ist. Nun können wir sagen, dass die Wahrscheinlichkeit, dass unsere Hypothese falsch ist, und dennoch ein Beispiel richtig klassifiziert, allerhöchstens $1 - \epsilon$ ist – denn wir haben eine Wahrscheinlichkeitsmasse von $\geq \epsilon$ auf die falsch klassifizierten Beispiele gesetzt:

$$P(h(x) = f(x)|h(x) \in H_{\downarrow}) \leq 1 - \epsilon \tag{284}$$

Diese Tatsache allein scheint nicht sonderlich interessant, denn ϵ ist üblicherweise ziemlich klein. Wir haben aber

$$\epsilon > 0,$$

und deswegen gilt: für alle $\delta > 0$ gibt es ein $n \in \mathbb{N}$, so dass

(285) $(1-\epsilon)^n < \delta$

Diese Beobachtung ist entscheidend, denn wir haben:

(286) $P(h(x) = f(x) : \forall x \in D | h \in H_{\downarrow}) \leq (1-\epsilon)^N$

Denn wir gehen davon aus, dass alle unsere Beispiele in D korrekt sind; es gibt also keine Störungen in unseren Daten. Dementsprechend sinkt die Wahrscheinlichkeit, dass wir eine Hypothese in $H_{\downarrow}$ haben, die konsistent mit unseren Daten ist, mit der Zahl der Beobachtungen die wir machen. Als nächstes sehen wir, dass wir die Wahrscheinlichkeit beachten müssen, dass *irgendeine* Hypothese in $H_{\downarrow}$ konsistent ist mit unseren Daten. Das ist natürlich

(287) $P(\exists h \in H_{\downarrow}.\forall x \in D : h(x) = f(x)) \leq |H_{\downarrow}|(1-\epsilon)^N \leq |H|(1-\epsilon)^N$

Das setzt natürlich voraus, dass $|H|$ endlich ist, sonst ist der Term undefiniert. Wenn wir also

- ein beliebes ϵ auswählen, dass eine Abweichung als "ernsthaft falsch" definiert,
- ein beliebiges δ, dass die maximale Wahrscheinlichkeit festlegt, dass die Hypothese ernsthaft falsch ist,
- dann müssen wir nur ein N finden so dass

(288) $|H|(1-\epsilon)^N \leq \delta$

Um diesen Term nach N aufzulösen, muss man etwas tricksen. Man kann zeigen dass

(289) $1 - \epsilon \leq e^{-\epsilon} = \frac{1}{e^{\epsilon}}$

(denn je kleiner ϵ, desto kleiner e^{ϵ}, desto größer $1/e^{\epsilon}$). Also reicht es, N so zu wählen dass

$$
\begin{aligned}
|H|(e^{-\epsilon})^N &\leq \delta \\
\Leftrightarrow ln(|H|(e^{-\epsilon})^N) &\leq ln(\delta) \\
\Leftrightarrow ln(|H|) + ln(e^{-\epsilon}) \cdot N &\leq ln(\delta) \\
\Leftrightarrow ln(|H|) - \epsilon \cdot N &\leq ln(\delta) \\
\Leftrightarrow -ln(|H|) + \epsilon \cdot N &\geq ln(\frac{1}{\delta}) \\
\Leftrightarrow \epsilon \cdot N &\geq ln(\frac{1}{\delta}) + ln(|H|) \\
\Leftrightarrow N &\geq \frac{1}{\epsilon} \cdot (ln(\frac{1}{\delta}) + ln(|H|))
\end{aligned}
\tag{290}
$$

Das bedeutet: wenn wir N entsprechend wählen, dann gilt für eine Hypothese h, die mit N Beispielen konsistent ist:

> Mit einer Wahrscheinlichkeit von mindestens $1 - \delta$ hat h eine Fehlerrate von höchstens ϵ.

Mit anderen Worten: sie ist wahrscheinlich annähernd korrekt. Diese Nummer N – gegeben ϵ und δ – nennt man die Stichprobenkomplexität des Hypothesenraumes H (denn sie hängt natürlich von H ab).

Betrachten wir das Kriterium

$$N \geq \frac{1}{\epsilon} \cdot (ln(\frac{1}{\delta}) + ln(|H|)) \tag{291}$$

Dann fällt uns auf:

- δ (unsere verbleibende Unsicherheit) spielt logarithmisch eine Rolle (also mit schrumpfenden δ wächst N eher langsam);
- $|H|$ – unser Hypothesenraum – spielt ebenfalls logarithmisch eine Rolle (also mit wachsendem $|H|$ wächst N eher langsam);
- ϵ ist ein linearer Faktor, also mit schrumpfenden ϵ wächst N proportional.

PAC-Lernen ist insofern sehr vorteilhaft, als dass wir nach einer endlichen Anzahl von Datenpunkten starke Aussagen über die Qualität unserer Hypothesen machen können. Man beachte insbesondere, dass P hierbei keine Rolle spielt, PAC-Lernen ist also unabhängig von der zugrundeliegenden Verteilung! Auf der anderen Seite haben wir aber eine Vielzahl von Voraussetzungen, die oft nicht erfüllt sind.

24 EM-Algorithmen: Parameter schätzen von unvollständigen Daten

24.1 Einleitung

EM-Algorithmen (**E**xpectation-**M**aximization) gehören eigentlich zum Thema der Parameter-Schätzung. Es handelt sich aber nicht um eine alternative Methode der Schätzung (EM-Algorithmen basieren meist auf ML-Schätzung), sondern um eine Methode, wie wir Parameter schätzen können von Daten, die **unvollständig** sind in Hinblick auf relevante Parameter. Etwas formaler können wir das wieder mit dem Begriff der Zufallsvariable fassen. Sei X eine Zufallsvariable im allgemeinen Sinn dass

$$X : \Omega \to \Omega',$$

d.h. X bildet Ergebnisse auf Ergebnisse ab. Es kann dabei gut sein dass X relevante Information "verschluckt", wie etwa die Würfelvariable

$$X_W : \{1, ..., 6\} \times \{1, ..., 6\} \to \{1, ..., 12\}$$

Wie wir gesehen haben, verbirgt diese Variable die Information, wie wir unser Ergebnis (z.B. 11 Augen) gewürfelt haben.

EM-Algorithmen brauchen wir, wenn wir annehmen, dass unsere Daten die Form haben

$$X(d) : d \in D,$$

wobei eigentlich D von einer Wahrscheinlichkeitsfunktion generiert wird, deren Parameter wir schätzen wollen. Das bedeutet wir können eigentlich nicht *unmittelbar von den Daten* unsere Parameter schätzen, sondern müssen zuerst unsere passenden Daten **rekonstruieren**.

Wie machen wir das? Zunächst einige grundsätzliche Bemerkungen:

- Wir können in diesem Fall nicht garantieren, dass wir die optimale Lösung finden;
- das ist aber ein mathematisches Problem, kein grundsätzliches: es gibt (meistens) eine theoretisch beste Lösung!
- Ziel muss also sein, dieser Lösung möglichst nahe zu kommen!

Das mathematische Problem liegt darin, dass wir **zwei** unbekannte haben statt einer:

1. Die optimalen Parameter (Wahrscheinlichkeiten), und
2. die optimale Struktur der Ereignisse, über die die Parameter verteilt wurden.

Struktur des EM-Algorithmus Das ist das Ziel von EM-Algorithmen. Grob lässt sich ihre Funktionsweise wie folgt charakterisieren:

- Zunächst müssen wir Wahrscheinlichkeiten **initialisieren**. Das kann beliebig sein, geschieht oft aber nach einer im Spezialfall sinnvollen Methode.
- **Expectation**-Schritt: gegeben unsere Wahrscheinlichkeiten und Daten d nehmen wir dasjenige d', dass 1. **maximal Wahrscheinlich** ist, und $X(d') = d$ erfüllt.
- **Maximization**-Schritt: nun nehmen wir d' als gegeben; da d' alle relevanten Strukturen enthält, können wir – nach gewohnter Methode – Parameter schätzen.
- Wir nehmen diese Wahrscheinlichkeiten als gegeben und gehen damit zurück zu Schritt 2 (Möglichkeit 1) oder wir sind zufrieden uns nehmen die Wahrscheinlichkeiten als gegeben und beenden (Möglichkeit 2).

Der EM-Algorithmus ist also ein Algorithmus, der beliebig iteriert werden kann.

24.2 Ein Beispielproblem

Greifen wir das obige Beispiel (leicht manipuliert) auf: wir betrachten Fußballspiele im Hinblick auf Fehlentscheidungen des Schiedsrichters. Uns interessiert insbesondere: wieviele Fehlentscheidungen gibt es zugunsten der Heim, wieviele zugunsten der Auswärtsmannschaft.

- Wir bekommen einen Datensatz D, der besteht aus Zahlenpaaren $(0, n_1), ..., (12, n_{12})$, die uns jeweils sagen, wie oft wir eine gewisse Zahl Fehlentscheidungen in einem Spiel haben.

	$x_2 = 1$	$x_2 = 2$	$x_2 = 3$	$x_2 = 4$	$x_2 = 5$	$x_2 = 6$	...
$x_1 = 1$	$n(1,1)$	$n(1,2)$	$n(1,3)$	$n(1,4)$	$n(1,5)$	$n(1,6)$	...
$x_1 = 2$	$n(2,1)$	$n(2,2)$	$n(2,3)$	$n(2,4)$	$n(2,5)$	$n(2,6)$	...
$x_1 = 3$	$n(3,1)$	$n(3,2)$	$n(3,3)$	$n(3,4)$	$n(3,5)$	$n(3,6)$	...
$x_1 = 4$	$n(4,1)$	$n(4,2)$	$n(4,3)$	$n(4,4)$	$n(4,5)$	$n(4,6)$	...
$x_1 = 5$	$n(5,1)$	$n(5,2)$	$n(5,3)$	$n(5,4)$	$n(5,5)$	$n(5,6)$	...
$x_1 = 6$	$n(6,1)$	$n(6,2)$	$n(6,3)$	$n(6,4)$	$n(6,5)$	$n(6,6)$	...
....							

Table 5: Ein Datensatz D', wie wir ihn brauchen

- Nehmen an, dass die Entscheidungen pro/contra Heimmannschaft voneinander unabhängig sind, d.h. wir schätzen deren Wahrscheinlichkeiten unabhängig, mit der Verbund-Wahrscheinlichkeit als einem Produkt.
- Wir kennen aber **nicht** die Zahl der Entscheidungen pro/contra Heimmannschaft, sondern nur deren Summe!

Als ein kleines Intermezzo überlegen wir kurz, wie wir diese Wahrscheinlichkeiten schätzen würden, wenn wir die vollen Daten gegeben hätten:

Nehmen wir an, wir haben den Datensatz D' mit $|D|$ definiert als:

$$|D| = \sum_{i=1}^{k} \sum_{j=1}^{6} n(i,j) \tag{292}$$

mit k dem festgelegten Maximum an Fehlentscheidungen. Dann liefert uns die ML-Methode einfach:

$$\hat{P}(i,j) = \frac{n(i,j)}{|D|} \tag{293}$$

Diese Schätzung wäre aber wahrscheinlich inkonsistent mit unserem Wissen: wir definieren die marginalen Wahrscheinlichkeiten wie üblich durch

$$\hat{P}(x_1 = i) = \sum_{j=1}^{k} \hat{P}(i,j) \tag{294}$$

und dann kann es gut sein dass

$$\hat{P}(1,2) \neq \hat{P}(x_1 = 1) \cdot \hat{P}(x_2 = 2) \tag{295}$$

was unserem *apriori*-Wissen widerspricht. Wir müssen also, um eine konsistente Verteilung mit unserem Vorwissen zu bekommen, die Wahrscheinlichkeiten

$$\hat{P}(x_1 = 1), \hat{P}(x_1 = 2), ..., \hat{P}(x_2 = k)$$

unabhängig voneinander schätzen, und die Verbundwahrscheinlichkeiten entsprechend definieren:

(296) $\hat{P}(1,2) := \hat{P}(x_1 = 1) \cdot \hat{P}(x_2 = 2)$

(das ist also qua Definition richtig). Wir schätzen diese Wahrscheinlichkeiten ganz einfach durch

$$\hat{P}(x_1 = i) = \frac{\sum_{j=1}^{k} n(i,j)}{|D|} : 1 \leq i \leq k \tag{297}$$

$$\hat{P}(x_2 = i) = \frac{\sum_{j=1}^{k} n(j,i)}{|D|} : 1 \leq i \leq k \tag{298}$$

Das liefert eine konsistente Wahrscheinlichkeitsverteilung, in der die beiden Teilereignisse unabhängig sind; darüber hinaus liefert es diejenige Verteilung, die gegeben unsere Daten, von allen Verteilungen, in denen die beiden Ergebnisse unabhängig sind, die maximale Likelihood hat.

Das eigentliche Problem ist aber folgendes: wir haben nicht den Datensatz D', sondern nur den Datensatz D, der wie folgt aussieht:

Fehlentscheidungen	Anzahl
2	n(2)
3	n(3)
4	n(4)
....	
11	n(11)
12	n(12)

Was machen wir also?

24.3 Der EM-Algorithmus auf unserem Beispiel

Initialisierung Was können wir also tun? Zunächst müssen wir die Wahrscheinlichkeiten **initialisieren**. Wir tun das wie oben beschrieben, indem wir zunächst die Wahrscheinlichkeiten initialisieren. Hierzu ist es wohl besser, zunächst konkrete Zahlen anzuschauen:

Fehlentscheidungen	Anzahl
1	1054
2	1232
3	1584
4	1784
5	2013
6	2494
7	2801
8	2693
9	2456
11	1704
12	1453

Wir haben übrigens $|D| = 21268$; und wir können D auch als Funktion

$$d : \{2, ..., 12\} \to \mathbb{N}$$

auffassen. Wie initialisieren wir $P(x_1 = 1)$ etc.? Wir sehen, dass die Wahrscheinlichkeiten leicht asymmetrisch sind (nach oben verschoben); wir setzen sie also mehr oder weniger willkürlich (wir könnten aber auch Maximum-Entropie-Methoden benutzen)

Fehlentscheidungenx	$\hat{P}_0$ Heim	$\hat{P}_0$ Auswärts
1	0.14	0.15
2	0.14	0.16
3	0.16	0.16
4	0.17	0.18
5	0.19	0.19
6	0.2	0.16

Hiermit können wir nun jeweils die Wahrscheinlichkeit

$$\hat{P}_0(1,1) = \hat{P}_0(x_1 = 1) \cdot \hat{P}_0(x_2 = 1) \tag{299}$$

berechnen, und dementsprechend auch die Wahrscheinlichkeit

$$\hat{P'}_0(4) = \hat{P}_0(1,3) + \hat{P}_0(3,1) + \hat{P}_0(2,2) \tag{300}$$

oder allgemeiner:

$$\hat{P'}_0(n) = \sum_{n_1+n_2=n} \hat{P}_0(n_1, n_2) \tag{301}$$

$d'(x_1, x_2)$	$x_2 = 1$	$x_2 = 2$	$x_2 = 3$	$x_2 = 4$	$x_2 = 5$	$x_2 = 6$
$x_1 = 1$	$d'(1,1)$	$d'(1,2)$	$d'(1,3)$	$d'(1,4)$	$d'(1,5)$	$d'(1,6)$
$x_1 = 2$	$d'(2,1)$	$d'(2,2)$	$d'(2,3)$	$d'(2,4)$	$d'(2,5)$	$d'(2,6)$
$x_1 = 3$	$d'(3,1)$	$d'(3,2)$	$d'(3,3)$	$d'(3,4)$	$d'(3,5)$	$d'(3,6)$
$x_1 = 4$	$d'(4,1)$	$d'(4,2)$	$d'(4,3)$	$d'(4,4)$	$d'(4,5)$	$d'(4,6)$
$x_1 = 5$	$d'(5,1)$	$d'(5,2)$	$d'(5,3)$	$d'(5,4)$	$d'(5,5)$	$d'(5,6)$
$x_1 = 6$	$d'(6,1)$	$d'(6,2)$	$d'(6,3)$	$d'(6,4)$	$d'(6,5)$	$d'(6,6)$

Table 6: Der konstruierte Datensatz

Expectation Nun kommen wir zum entscheidenden Schritt: wir wenden unsere Schätzungen auf unsere Daten an, und konstruieren aus D – dem unvollständigen Datensatz – den vollständigen Datensatz D' (mit zugehöriger Funktion d')

wobei gilt (dies ist der entscheidende Schritt:

$$d'(i,j) = d(i+j) \cdot \frac{\hat{P}_0(i,j)}{\hat{P}'(i+j)} \tag{302}$$

Z.B. haben wir:

$$d'(1,3) = d(4) \cdot \frac{\hat{P}_0(1,3)}{\hat{P}'(4)} = 1784 \cdot \frac{0.0224}{0.0224 + 0.024 + 0.0224} = 401.1163 \tag{303}$$

Hingegen:

$$d'(3,1) = d(4) \cdot \frac{\hat{P}_0(3,1)}{\hat{P}'(4)} = 1232 \cdot \frac{0.024}{0.0224 + 0.024 + 0.0224} = 429.7674 \tag{304}$$

Das war der **Expectation-Schritt**: wir haben jetzt eine Tabelle, in der wir wissen, welches die plausibelste Verteilung an Würfen war, um unsere Daten zu generieren. Das das keine ganzen Zahlen sind und somit eigentlich nicht sein kann, soll uns nicht stören, wir können ja später noch runden.

Maximization Nun machen wir den nächsten Schritt: wir nehmen die neu-gewonnenen Häufigkeiten, um damit die Wahrscheinlichkeiten neu zu

schätzen. Das geht ganz einfach nach dem gewohnten ML-Rezept:

$$\hat{P}_1(x_1 = i) = \frac{\sum_{j=1}^{6} d'(i,j)}{|D'|} : 1 \leq i \leq 6 \tag{305}$$

$$\hat{P}_1(x_2 = i) = \frac{\sum_{j=1}^{6} d'(j,i)}{|D'|} : 1 \leq i \leq 6 \tag{306}$$

Wir haben nun neue Wahrscheinlichkeiten gewonnen. An dieser Stelle gibt es nun zwei Möglichkeiten:

1. Wir sind mit dem gewonnenen zufrieden und behalten die geschätzten Wahrscheinlichkeiten. Dafür scheint es aber noch etwas früh. Daher:

2. Wir nutzen die neuen Wahrscheinlichkeiten als Ausgangpunkt, um den Expectation-Schritt zu wiederholen; wir konstruieren also D'' und $\hat{P}_2$ etc.

24.4 Der Algorithmus (allgemeine Form)

Die Prozedur des EM-Algorithmus ist im Allgemeinen wie folgt:

(1) Initialisiere P_0, D, d.
(2) für jedes $1, 2, 3, ..., n$, mache folgendes:
(3) **E-Schritt**: berechne die Funktion d_{i+1} mittels $d_{i+1}(x) = d_i(x) \cdot P_i(x|X(x))$
// X ist die "vergessliche Funktion"
(4) **M-Schritt**: berechne die ML-Schätzung $\hat{P}$ für unser Modell über d_{i+1}
(5) setze $\hat{P} = P_{i+1}$
(6) gebe P_{i+1} aus
(7) Ende

Das bedeutet: wir haben einen Datensatz D und Reihe von Datensätzen $D = D_0, D_1, D_2, ...$, die immer besser werden (hoffentlich), wir haben eine vergessliche Funktion X, so dass

$$X : D_0 \mapsto D, X : D_1 \mapsto D, X : D_2 \mapsto D, ...$$

Weiterhin haben wir eine Reihe von Wahrscheinlichkeitsfunktionen $P_0, P_1, P_2, ...$, die immer feiner werden. Aber wer garantiert uns, dass die Wahrscheinlichkeiten tatsächlich immer besser werden?

Theorem 22 *Die Ausgabe des EM-Algorithmus ist eine Folge von Wahrscheinlichkeitsfunktionen*

$$P_0, P_1, P_2, ...,$$

so dass für die vergessliche Funktion X, den Ausgangsdatensatz D gilt:

$$P_0 \circ X^{-1}(D) \leq P_1 \circ X^{-1}(D) \leq P_2 \circ X^{-1}(D) \leq P_3 \circ X^{-1}(D) \leq ...$$

Das bedeutet, dass die Likelihood unserer Daten mit jeder neuen Wahrscheinlichkeitsfunktion größer wird. Das ist gut, aber garantiert bei weitem nicht, dass wir eine optimale Lösung finden: wir können immer in lokalen Maxima hängenbleiben, und diese können sogar relativ schlecht sein.

25 Der EM-Algorithmus in der maschinellen Übersetzung

25.1 Grundbegriffe der maschinellen Übersetzung

Es gibt verschiedene Möglichkeiten, probabilistische Sprachen auf probabilistische Relationen zu verallgemeinern. Die wichtigste Konzeption für uns ist folgende: eine **probabilistische Relation** über M, N ist eine Funktion $R : M \times N \to [0, 1]$, so dass gilt:

$$(307) \quad \sum_{n \in N} R(m, n) = 1,$$

für alle $m \in M$. Ein gutes Beispiel für eine solche Relation ist für uns die ein probabilistisches Lexikon: M ist eine Menge von deutschen Wörtern, N ist eine Menge von englischen Wörtern, und R sagt uns: eine deutsches Wort m wird mit Wahrscheinlichkeit x in ein englisches Wort n übersetzt. Die Wahrscheinlichkeiten summieren sich wie immer zu eins, d.h. es ist sicher für jedes deutsche Wort, dass es als *irgendein* englisches Wort übersetzt wird.

Eine solche Relation liefert uns keine echte Wahrscheinlichkeitsverteilung über $M \times N$, sondern nur über N gegeben ein m. Man schreibt sie daher auch gerne als eine bedingte Wahrscheinlichkeit: $P_R(n|m) := R(m, n)$; wir fassen also diese Wahrscheinlichkeit einer Übersetzung auf als die bedingte Wahrscheinlichkeit eines englischen Wortes gegeben ein deutsches Wort. Die Bedingung in (1) sichert die Konsistenz der Verteilung. Diese Konvention ist weit verbreitet in der Literatur; positiv ist dass wir die Relation soz. direkt in den Wahrscheinlichkeitskalkül eingebettet haben. Negativ ist dass wir damit bereits gewisse Ressourcen des Kalküls verbraucht haben; wir können also beispielsweise nicht mehr von bedingten Wahrscheinlichkeiten von Übersetzungen sprechen, da wir schreiben müssten: $P(m_1|n_1||m_2|n_2)$, was außerhalb unseres Kalküls liegt. Weiterhin ist etwas unklar was eigentlich der zugrundeliegende Wahrscheinlichkeitsraum sein soll; aber solche ontologischen Fragen werden wir uns in Zukunft nicht mehr stellen.

Unser nächstes Ziel ist folgendes: nehmen wir an, wir haben unser probabilistisches Lexikon (notiert als bedingte Wahrscheinlichkeit). Was uns natürlich interessiert ist nicht die Wahrscheinlichkeit von Wortübersetzungen, sondern von Übersetzungen von Sätzen. Wir können wir unser Modell erweitern? Die naheliegendste Lösung wäre: gegeben ein deutscher Satz $d_1 d_2 ... d_i$,

ein englischer Satz $e_1e_2...e_i$ (die d_j denotieren deutsche Wörter, e_j englische), definieren wir:

(308) $P_R(e_1e_2...e_i|d_1d_2...d_i) := P_R(e_1|d_1) \cdot P_R(e_2|d_2) \cdot ... \cdot P_R(e_i|d_i)$

Das sieht einfach aus und garantiert Konsistenz. Das Problem ist nur: es ist zu einfach. Wer sagt uns, dass deutsche Sätze immer gleich lang sind wie ihre englischen Übersetzungen? Außerdem: wer sagt dass das i-te Wort im deutschen Satz dem i-ten Wort im englischen Satz entspricht? Nehmen wir nur einmal die Sätze ICH KENNE IHN NICHT und I DO NOT KNOW HIM. Hier sieht man leicht dass unser Modell vollkommen inadäquat ist.

Wir lösen dieses Problem wie folgt. Gegeben ein deutscher Satz der Länge i, ein englischer Satz der Länge j, definieren wir eine *Alinierungsfunktion* wie folgt:

(309) $a_{ji} : \{1, 2, ..., j\} \to \{1, 2, ..., i, 0\}$

Eine Alinierungsfunktion weist jedem englischen Satz höchstens ein deutsches Wort zu; die 0 bedeutet: das englische Wort hat keine deutsche Entsprechung. Das eröffnet eine Menge von Möglichkeiten; da wir oft alle Möglichkeiten berücksichtigen müssen, denotieren wir die Menge aller Alinierungsfunktionen a_{ji} mit $A(j, i)$. Dieser *Funktionenraum* wächst exponentiell, mit $|A(j, i)| = (i+1)^j$. Wir haben aber immer noch intrinsische Beschränkungen: es ist möglich dass beliebig viele englische Wörter als die Übersetzung eines deutschen Wortes sind; aber es ist nicht möglich dass zwei deutsche Wörter als ein englisches übersetzt werden! Wir werden das später berücksichtigen. Was wir nun ausrechnen können ist:

(310) $P_R(e_1e_2...e_i|a, d_1d_2...d_j) := \prod_{k=1}^{i} P_R(e_k|d_{a(k)})$

Einfachheit halber und um Indizes zu sparen schreiben wir in Zukunft für deutsche Sätze $\vec{d}$, für englische $\vec{e}$; mit $|\vec{d}|, |\vec{e}|$ bezeichnen wir die Länge der Sätze. Beachten Sie dass für die Wahrscheinlichkeit einer Übersetzung die deutschen Worte, die kein Urbild im englischen haben, keine Rolle spielen! Aus Gleichung (310) können wir mit unseren Regeln ableiten:

(311) $P_R(\vec{e}, a|\vec{d}) := P_R(\vec{e}|a, \vec{d}) \cdot P(a|\vec{d})$

Den ersten Term haben wir bereits; über den zweiten Term haben wir uns allerdings noch keine Gedanken gemacht; was ist die Wahrscheinlichkeit einer Alinierung? Intuitiv sollte aber klar sein: für Sprachen wie Deutsch und Englisch, (oder besser noch: Englisch und Französisch), die eine relativ ähnliche Wortstellung haben, ist eine Alinierung, die keine großen Positionswechsel macht, wahrscheinlicher als eine Alinierung die die Wortfolge komplett umdreht. Wenn wir keinerlei Informationen dieser Art haben und für uns also alle Alinierungen gleich wahrscheinlich sind, dann haben wir

$$P(a_{ji}) = \frac{1}{|A(j,i)|} \tag{312}$$

Wir sehen also, dass wir den Alinierungen nur im Bezug auf eine deutsche und englische Satzlänge eine Wahrscheinlichkeit zuweisen können; in der obigen Formel haben wir das implizit gelassen. Der Grund, warum wir die Alinierungen auf die linke Seite des | haben wollen ist folgender: wir können sie nun "ausmarginalisieren", d.h. durch eine Summe über alle möglichen Alinierungen die Wahrscheinlichkeit $P(\vec{e}|\vec{d})$ berechnen:

(313)

$$P(\vec{e}|\vec{d}) = \sum_{a \in A(|\vec{e}|,|\vec{d}|)} P_R(\vec{e}|a,\vec{d}) \cdot P(a|\vec{d}) = \sum_{a \in A(|\vec{e}|,|\vec{d}|)} \left(\left(\prod_{i=1}^{|\vec{e}|} P_R(e_i|d_{a(i)}) \right) P(a|\vec{d}) \right)$$

Diese Formel involviert also eine exponentiell wachsense Summe von Produkten; daher können wir sie praktisch nicht ausrechnen. Glücklicherweise kann man diese Formel wesentlich vereinfachen, unter der Annahme dass alle Alinierungen gleich wahrscheinlich sind:

$$\begin{aligned} &\sum_{a \in A(|\vec{e}|,|\vec{d}|)} \left(\left(\prod_{i=1}^{|\vec{e}|} P_R(e_i|d_{a(i)}) \right) \frac{1}{(|\vec{d}|+1)^{|\vec{e}|}} \right) \\ &= \frac{1}{(|\vec{d}|+1)^{|\vec{e}|}} \sum_{a \in A(|\vec{e}|,|\vec{d}|)} \prod_{i=1}^{|\vec{e}|} P_R(e_i|d_{a(i)}) \qquad (314) \\ &= \frac{1}{(|\vec{d}|+1)^{|\vec{e}|}} \prod_{i=1}^{|\vec{e}|} \sum_{j=0}^{|\vec{d}|} P_R(e_i|d_j) \end{aligned}$$

Wenn Ihnen das rätselhaft vorkommt sind Sie nicht allein. Die Umformung wird erreicht durch wiederholtes Ausklammmern von Termen; genaueres erfahren Sie in der Literatur. Die Wichtigkeit dieser Umformung ist kaum zu überschätzen: anstatt exponentiell viele Multiplikationen müssen wir nur noch linear viele Multiplikationen ausführen; und die Anzahl der Additionen ist damit ebenfalls linear beschränkt. Wir sind also von praktisch unberechenbar zu problemlos berechenbar gegangen. Diese Umformung funktioniert allerdings nur, wenn alle Alinierungen gleich wahrscheinlich sind!

Wir haben oben den Term $P(a|\vec{d})$. Was ist die Bedeutung von $\vec{d}$ für $P(a)$? Dieser Term spielt tatsächlich nur eine Rolle wegen der Länge von $\vec{d}$: je größer $\vec{d}$ ist, desto mehr Funktionen gibt es, und deswegen ändert sich auch die Wahrscheinlichkeitsverteilung. Anders verhält es sich aber, wenn wir sowohl $\vec{e}$ als auch $\vec{d}$ als gegeben annehmen: beide zusammen beeinflussen die Wahrscheinlichkeit so stark, dass wir die unabhängigen Wahrscheinlichkeiten nicht mehr brauchen, denn wir haben:

$$P(a_{|\vec{e}||\vec{d}|}|\vec{e},\vec{d}) = \frac{P(\vec{e},a|\vec{d})}{P(\vec{e}|\vec{d})} \tag{315}$$

Das ist eine direkte Anwendung der Definition der bedingten Wahrscheinlichkeit. Das bedeutet zum Beispiel: angenommen dass alle Alinierungen apriori gleich wahrscheinlich sind, gilt dasselbe *nicht* für die bedingten Wahrscheinlichkeiten der Alinierungen: sofern nicht auch die lexikalischen Übersetzungswahrscheinlichkeiten gleich verteilt sind, macht ein Satzpaar gewisse Alinierungen wahrscheinlicher als andere, weil eben umgekehrt auch gewisse Alinierungen die Übersetzung wahrscheinlicher machen als andere. Diese Tatsache macht sich der EM-Algorithmus zunutze.

25.2 Wahrscheinlichkeiten schätzen

Wenn wir die Wahrscheinlichkeiten von (Wort-)Übersetzungen und Alinierungen bereits kennen, dann gibt es für uns eigentlich nichts mehr zu tun. Das Problem ist dass wir sie normalerweise nicht kennen, sondern erst *schätzen* müssen. Wenn wir ein zweisprachiges Korpus haben, in dem jedes einzelne Wort mit seiner Übersetzung aliniert ist (so dass es mit unseren intrinsischen Beschränkungen konform geht), dann ist es eine leichte Übung, die Wahrscheinlichkeiten nach *maximum likelihood* Methode zu schätzen.

(Wenn Sie diese Übung nicht leicht finden, dann ist das ein Grund mehr sie zu machen!)

Aber wir brauchen noch weniger. Nehmen wir an, wir haben nur die lexikalischen Übersetzungwahrscheinlichkeiten gegeben. Dann können wir für jeden Satz in unserem Korpus, nach Formel (315), die Wahrscheinlichkeit der Alinierungen ausrechnen. Wenn wir dann die unbedingten Wahrscheinlichkeiten einer Alinierung a_{ij} ausrechnen wollen, dann summieren wir die bedingten Wahrscheinlichkeiten für alle Satzpaare die relevant sind ($|\vec{e}| = i, |\vec{d}| = j$), und Teilen die Anzahl durch dieser Satzpaare in unserem Korpus.

Umgekehrt, nehmen wir an wir haben keine lexikalischen Wahrscheinlichkeiten, aber dafür die Wahrscheinlichkeiten der Alinierung. Dann können wir folgendes machen. Gegeben ein Satzpaar $\vec{e}, \vec{d}$, mit $i \leq |\vec{e}|$, $j \leq |\vec{d}|$, können wir ausrechnen wie wahrscheinlich es ist, dass das i-te Wort von $\vec{e}$, e_i, mit dem j-ten Wort von $\vec{d}$, d_j, aliniert ist. Wir schreiben dafür: $P_a(j|i, |\vec{e}|, |\vec{e}|)$, also die Wahrscheinlichkeit dass $a_{|\vec{e}||\vec{d}|}(i) = j$. Das errechnet sich wie folgt:

$$P_a(j|i, |\vec{e}|, |\vec{d}|) = \sum_{a \in A(|\vec{e}|,|\vec{d}|), a(i)=j} P(a) \tag{316}$$

Wir können nun die Übersetzungswahrscheinlichkeit $P_R(e|d)$ (wobei die Indizes nur noch die Worte identifizieren sollen; die Position spielt keine Rolle mehr) berechnen: wir multiplizieren jedes Vorkommen, dass e mit d in unserem Korpus aliniert ist, mit der Wahrscheinlichkeit der Alinierung (letzte Gleichung), und addieren die so gewichtete Anzahl der Vorkommen zusammen. Das bedeutet, eine wahrscheinliche Alinierung zählt mehr, eine unwahrscheinliche weniger. Die resultierende Zahl ist noch keine Wahrscheinlichkeit; sie kann leicht grßer als 1 sein. Um zu *normalisieren* (damit bezeichnet man: eine Gewichtung in eine Wahrscheinlichkeit umwandeln), müssen wir noch durch einen geeigneten Term teilen. Dieser Term sind die nach Wahrscheinlichkeit der Alinierung gewichteten Häufigkeiten *irgendeines* englischen Wortes, das mit d aliniert ist. Wenn wir das in eine Formel bringen, sieht das in etwa wie folgt aus. Wir benutzen das sog. *Kronecker-*δ, wobei $\delta(x, y) = 1$ falls $x = y$, und andernfalls $\delta(x, y) = 0$.

$$\hat{P}(e|d) = \frac{\sum_{i \leq |\vec{e}|} \sum_{j \leq |\vec{d}|} \delta(e, e_i)\delta(d, d_j)(P_a(j|i, |\vec{e}|, |\vec{d}|))}{\sum_{i \leq |\vec{e}|} \sum_{j \leq |\vec{d}|} \delta(d, d_j)(P_a(j|i, |\vec{e}|, |\vec{d}|))} \tag{317}$$

Diese Formel liefert uns die Wahrscheinlichkeit nur für ein einzelnes Satzpaar $\vec{e}, \vec{d}$, wobei $\vec{e} = e_1...e_{|\vec{e}|}$, $\vec{d} = d_1...d_{|\vec{d}|}$. Das ist natürlich zuwenig, wir müssen über das ganze Korpus zählen. Da wir momentan keine Indizes für das Korpus haben, belassen wir es bei der einfachen Formel; um sie wirklich adäquat zu machen müssten wir das ganze Korpus indizieren, und falls i, j in verschiedenen Sätzen stehen, dann ist $P_a(j|i) = 0$.

Die letzte Formel ist sehr unschön: wir können zwar jetzt mithilfe der Umformung (314) die Wahrscheinlichkeiten $P(\vec{e}|\vec{d})$ recht effizient berechnen; um allerdings die Funktion P_a zu berechnen müssen wir die Berechnung aber dennoch für alle Alignments ausführen (zumindest für alle $a : a(i) = j$), und auch diese Zahl wächst exponentiell mit der Länge der Sätze. D.h. wir haben trotz allem exponentiell viele Rechenschritte.

25.3 Der EM-Algorithmus: Vorgeplänkel

Wir kommen also von Alinierungswahrscheinlichkeiten zu Übersetzungswahrscheinlichkeiten, und von Übersetzungswahrscheinlichkeiten zu Alinierungswahrscheinlichkeiten. Das Problem ist: normalerweise haben wir keine von beiden. Was also ist zu tun? Hier hilft der EM-Algorithmus (EM steht wahlweise für *estimation maximization* oder *expectation maximization*.)

Zunächst stehen wir also ratlos vor unserem Korpus, in dem nur Sätze aliniert sind. Als erstes machen wir, was wir immer machen wenn wir ratlos sind: wir nehmen an dass Übersetzungswahrscheinlichkeiten und Alinierungswahrscheinlichkeiten uniform sind.

Als nächstes machen wir da weiter, wo wir eben aufgehört haben: wir schätzen Übersetzungswahrscheinlichkeiten von (uniform) gewichteten Alinierungshäufigkeiten. Was gewinnt man dadurch? Nun, die Gewichte sind zwar uniform, aber die Häufigkeiten sind es nicht: unser zweisprachiges Korpus ist ja nach Sätzen aliniert, und daher kann es sein dass wir 3mal (dog, Hund) haben, aber nur 1mal (dog, Katze). Unser Korpus enthält also durchaus schon einige Information! Und wenn unser Korpus groß genug ist, dann reicht diese Information, um unsere Maschine in Gang zu bringen.

Wenn wir die ersten Häufigkeiten haben, dann sind unsere neu geschätzten Übersetzungswahrscheinlichkeiten (hoffentlich) nicht mehr uniform. Wir könnten jetzt so vorgehen: wir benutzen die neuen Wahrscheinlichkeiten, um die Wahrscheinlichkeiten der Alinierung zu berechnen. Allerdings ist genau das problematisch, da die Alinierungen zuviele sind. Wir benutzen wieder einen kleinen Trick. Was wir suchen ist der Zähler von Formel (317); wir schreiben

ihn aber auf eine etwas andere Art und Weise:

$$
\begin{aligned}
C(e|d,\vec{e},\vec{d})) &:= \sum_{a\in A(|\vec{e}|,|\vec{d}|)} \left(P(a|\vec{e},\vec{d}) \sum_{i=1}^{|\vec{e}|} \delta(e,e_i)\delta(d,d_{a(i)}) \right) \\
&= \sum_{a:a(i)=j} \left(P_a(j|i,\vec{e},\vec{d}) \sum_{i=1}^{|\vec{e}|} \delta(e,e_i)\delta(d,d_j) \right)
\end{aligned} \tag{318}
$$

NB: $P_a(j|i,\vec{e},\vec{d}) \neq P_a(j|i,|\vec{e}|,|\vec{d}|)$! Um Missverständnisse zu vermeiden schreiben wir statt $P_a(j|i,|\vec{e}|,|\vec{d}|) := P(a(i)=j|\ |\vec{e}|,|\vec{d}|)$.

Der Zähler aus (317) und Formel (318) sind gleich. Allerdings haben wir einen großen Vorteil gewonnen: wir brauchen nicht mehr sämtliche Alinierungen, sondern nur noch die Alinierungen gegeben $\vec{e},\vec{d}$. Erinnern Sie sich dass diese Wahrscheinlichkeit völlig bestimmt ist durch

$$\frac{P(\vec{e},a|\vec{d})}{P(\vec{e}|\vec{d})} \tag{319}$$

D.h. wir können sie "lokal" mit einem Satz berechnen, ohne auf andere Sätze zurückgreifen zu müssen. Aber auch diese Berechnung ist noch aufwändig, da wir immer noch eine Summe über eine exponentiell wachsende Zahl von Alinierungen haben. Wir müssen nun wieder einige algebraische Tricks anwenden, die starke Ähnlichkeit mit Formel (314) haben. Der zweite Teil von (318) ist dagegen ziemlich trivial zu berechnen; wir konzentrieren uns also auf die erste Hälfte:

$$P(a(i)=j|\vec{e},\vec{d}) = \frac{P(\vec{e},a(i)=j|\vec{d})}{P(\vec{e}|\vec{d})} \tag{320}$$

Das ist einfach die Definition der bedingten Wahrscheinlichkeit. Wir lösen

die Formel zunächst auf:

$$\frac{P(\vec{e}, a(i) = j|\vec{d})}{P(\vec{e}|\vec{d})}$$

$$= \frac{\sum_{a:a(i)=j} \prod_{k=1}^{|\vec{e}|} P(a|\vec{d}) P(e_k|d_{a(k)})}{\sum_{a \in A(|\vec{e}|,|\vec{d}|)} \prod_{k=1}^{|\vec{e}|} P(a|\vec{d}) P(e_k|d_{a(k)})} \tag{321}$$

$$= \frac{\sum_{a:a(i)=j} \prod_{k=1}^{|\vec{e}|} P(e_k|d_{a(k)})}{\sum_{a \in A(|\vec{e}|,|\vec{d}|)} \prod_{k=1}^{|\vec{e}|} P(e_k|d_{a(k)})}$$

Da wir annehmen, dass alle Alinierungen gleich wahrscheinlich sind, können wir den Term $P(a|\vec{d})$ ausklammern (Distributivgesetz) und rauskürzen. Wir schreiben nun die Summe, die über alle Alinierungen läuft, etwas expliziter auf:

$$= \frac{\sum_{a:a(i)=j} \prod_{k=1}^{|\vec{e}|} P(e_k|d_{a(k)})}{\sum_{a \in A(|\vec{e}|,|\vec{d}|)} \prod_{k=1}^{|\vec{e}|} P(e_k|d_{a(k)})}$$

$$= \frac{P(e_i|d_j) \sum_{a(1)=0}^{|\vec{d}|+1} \cdots \sum_{a(i-1)=0}^{|\vec{d}|+1} \sum_{a(i+1)=0}^{|\vec{d}|+1} \cdots \sum_{a(|\vec{e}|)=0}^{|\vec{d}|+1} \prod_{k=1}^{|\vec{e}|} P(e_k|d_{a(k)})}{\sum_{a(1)=0}^{|\vec{d}|+1} \cdots \sum_{a(|\vec{e}|)=0}^{|\vec{d}|+1} \prod_{k=1}^{|\vec{e}|} P(e_k|d_{a(k)})} \tag{322}$$

Wir haben hier nur die Alinierungen explizit ausgeschrieben. Als nächstes benutzen wir den Trick, den wir schon in Formel (314) benutzt haben; durch iteriertes anwenden des Distributivgesetzes können wir das Produkt über die Summen heben:

$$= \frac{P(e_i|d_j) \sum_{a(1)=0}^{|\vec{d}|+1} \cdots \sum_{a(i-1)=1}^{|\vec{d}|+1} \sum_{a(i+1)=0}^{|\vec{d}|+1} \cdots \sum_{a(|\vec{e}|)=0}^{|\vec{d}|+1} \prod_{k=1}^{|\vec{e}|} P(e_k|d_{a(k)})}{\sum_{a(1)=0}^{|\vec{d}|+1} \cdots \sum_{a(|\vec{e}|)=0}^{|\vec{d}|+1} \prod_{k=1}^{|\vec{e}|} P(e_k|d_{a(k)})}$$

$$= \frac{P(e_i|d_j) \prod_{k=1}^{|\vec{e}|} \sum_{a(1)=0}^{|\vec{d}|+1} \cdots \sum_{a(i-1)=1}^{|\vec{d}|+1} \sum_{a(i+1)=0}^{|\vec{d}|+1} \cdots \sum_{a(|\vec{e}|)=0}^{|\vec{d}|+1} P(e_k|d_{a(k)})}{\prod_{k=1}^{|\vec{e}|} \sum_{a(1)=0}^{|\vec{d}|+1} \cdots \sum_{a(|\vec{e}|)=0}^{|\vec{d}|+1} P(e_k|d_{a(k)})} \tag{323}$$

$$= \frac{P(e_i|d_j)}{\sum_{a(i)=0}^{|\vec{d}|} P(e_i|d_{a(i)})}$$

D.h. am Ende haben wir eine sehr einfache Formel dastehen; wir können die Wahrscheinlichkeit $P(a(i) = j|\vec{e}, \vec{d})$ in linear vielen Schritten über die Länge von $\vec{e}, \vec{d}$ berechnen.

25.4 Der eigentliche Algorithmus

Folgende Konventionen: mit $\mathfrak{K}$ bezeichnen wir unser Korpus; da die Reihenfolge der Satzpaare keine Rolle spielt, nehmen wir an dass $\mathfrak{K} := \{(\vec{e}_i, \vec{d}_i) : i \in I\}$ eine Menge von $|I|$ Satzpaaren ist. Mit lex(E) bzw. lex(D) bezeichnen wir das englische bzw. deutsche Lexikon.

Wir nehmen jetzt den Term und definieren ihn in seiner vereinfachten Form. Da wir die Wahrscheinlichkeiten nur noch lokal berechnen, fallen dabei einige Indizes weg; insbesondere brauchen die Wortpaare, deren Übersetzungswahrscheinlichkeit wir schätzen möchten, keinen Index mehr. Beachten Sie aber dass es sich bei der Umformung um die (verkürzte) Umformung in (323) handelt.

$$\begin{aligned} C(e|d, \vec{e}, \vec{d})) &:= \sum_{a:a(i)=j} [P_a(j|i, \vec{e}, \vec{d}) \sum_{i=1}^{|\vec{e}|} \delta(e, e_i)\delta(d, d_j)] \\ &= \frac{P(e|d)}{\sum_{j=0}^{|\vec{d}|} P(e|d_j)} \sum_{i=1}^{|\vec{e}|} \delta(e, e_i)\delta(d, d_{a(i)}) \end{aligned} \tag{324}$$

Wir ändern nun diese Formel, um sie induktiv anwenden zu können: nehmen Sie an, wir haben eine Sequenz von Wahrscheinlichkeitsfunktionen $P_n : n \in \mathbb{N}_0$. Dann definieren wir

$$C_n(e|d, \vec{e}, \vec{d})) := \frac{P_n(e|d)}{\sum_{j=0}^{|\vec{d}|} P_n(e|d_j)} \sum_{i=1}^{|\vec{e}|} \delta(e, e_i)\delta(d, d_{a(i)}) \tag{325}$$

Wir nehmen also in C_n Referenz auf die Wahrscheinlichkeitsfunktion die wir benutzen. Wir kommen nun zum eigentlichen Algorithmus. Wir setzen für alle $e \in lex(E), d \in lex(D)$,

$$P_0(e|d) := \frac{1}{|lex(E)|} \tag{326}$$

Wir setzen also die Übersetzungswahrscheinlichkeiten uniform. Die (unbedingten) Alinierungswahrscheinlichkeiten setzen wir ebenfalls uniform; daran wird sich auch im Laufe des Algorithmus nichts ändern.

Als nächstes definieren wir:

$$(327)\quad P_{n+1}(e|d) := \frac{\sum_{(\vec{e},\vec{d})\in\mathfrak{K}} C_n(e|d,\vec{e},\vec{d})}{\sum_{e'\in lex(E)}\sum_{(\vec{e},\vec{d})\in\mathfrak{K}} C_n(e'|d,\vec{e},\vec{d})}$$

Und damit sind wir auch schon fertig: wir haben P_0, und gegeben irgendein P_n können wir auch P_{n+1} ausrechnen mithilfe der Funktion C_n. Beachten sie dass Formel (327) genau dasselbe macht wie (317), nur dass

1. wir keine unbedingten Alinierungswahrscheinlichkeiten brauchen,
2. sauber über das Korpus quantifiziert haben, und
3. das Ganze wesentlich einfacher berechnen können!

Was wir also möchten ist $P_n(e|d)$ für ein ausreichend großes $n \in \mathbb{N}$. Zwei Dinge sind entscheidend:

1. Die Folge von Verteilung $P_n(e|d) : n \in \mathbb{N}$ konvergiert gegen eine Verteilung $P_\infty(e|d) : n \in \mathbb{N}$, und
2. $P_\infty(e|d) : n \in \mathbb{N}$ ist ein *lokales Maximum* der Likelihood Funktion von $P(e|d)$ gegeben $\mathfrak{K}$.

25.5 EM für IBM-Modell 1: Ein Beispiel

Nehmen wir an, wir haben die folgenden Daten (ein satzaliniertes, zweisprachiges Korpus):

Korpus:

1. (Hund bellte, dog barked)
2. (Hund, dog)

1. Initialisiere uniform:

$P_0(e\|d)$	dog	barked
Hund	$\frac{1}{2}$	$\frac{1}{2}$
bellte	$\frac{1}{2}$	$\frac{1}{2}$
NULL	$\frac{1}{2}$	$\frac{1}{2}$

2a. Anteilige Häufigkeiten auf Satzebene In Bezug auf Folie 8: Die schwarzen Brüche sind $P_0(e|d)$. Die Summe $\sum_{j=0}^{|\vec{d}|} P(e|d_j)$ wird in der letzten Zeile gebildet. In rot sehen Sie dann die eigentlichen anteiligen Häufigkeiten bzw. wie sie berechnet wurden.

$C_1(e\|d, \vec{e}_1, \vec{d}_1)$	dog	barked
Hund	$\frac{1}{2}\cdot\frac{2}{3} = \frac{1}{3}$	$\frac{1}{2}\cdot\frac{2}{3} = \frac{1}{3}$
bellte	$\frac{1}{2}\cdot\frac{2}{3} = \frac{1}{3}$	$\frac{1}{2}\cdot\frac{2}{3} = \frac{1}{3}$
NULL	$\frac{1}{2}\cdot\frac{2}{3} = \frac{1}{3}$	$\frac{1}{2}\cdot\frac{2}{3} = \frac{1}{3}$
$\sum \quad (\rightarrow Z)$	$\frac{3}{2}$	$\frac{3}{2}$

$C_1(e\|d, \vec{e}_2, \vec{d}_2)$	dog
Hund	$\frac{1}{2}\cdot 1 = \frac{1}{2}$
NULL	$\frac{1}{2}\cdot 1 = \frac{1}{2}$
$\sum \quad (\rightarrow Z)$	1

Die Zahlen $\frac{2}{3}$ ist der Kehrwert des Nenners in 327 (daher Multiplikation statt Division), der sich wiederum durch (325) und (324) berechnen läßt (so kommen wir auf $\frac{3}{2}$).

2b. Anteilige Häufigkeiten auf Korpusebene

$C(e\|d)$	dog	barked	$\sum \quad (\rightarrow C(d))$
Hund	$\frac{1}{3} + \frac{1}{2} = \frac{5}{6}$	$\frac{1}{3}$	$\frac{7}{6}$
bellte	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{2}{3}$
NULL	$\frac{1}{3} + \frac{1}{2} = \frac{5}{6}$	$\frac{1}{3}$	$\frac{7}{6}$

Wir sehen: hier zeigen unsere Daten einen deutlichen Effekt!

3. Neue Parameter

$P_1(e\|d)$	dog	barked
Hund	$\frac{5}{6}\cdot\frac{6}{7} = \frac{5}{7}$	$\frac{1}{3}\cdot\frac{6}{7} = \frac{2}{7}$
bellte	$\frac{1}{2}$	$\frac{1}{2}$
NULL	$\frac{5}{7}$	$\frac{2}{7}$

Iteriere 2a + 2b + 3 Bemerkung: Sowohl die Häufigkeiten als auch die Parameter $P_k(e|d)$ stimmen mit denen des allgemeinen Algorithmus (Folie 6+7) überein!

Aufgabe 12

Abgabe bis zum 4.7. vor dem Seminar.

Führen Sie das Beispiel fort, indem Sie zwei weitere Iterationen des EM-Algorithmus machen, und liefern Sie die resultierenden Übersetzungswahrscheinlichkeiten $P_3(\text{dog}|\text{Hund})$ etc.

26 Naive Bayes Klassifikatoren (aka *idiot Bayes)*

Die Unabhngigkeit der Merkmale, Maximierung von Likelihood x Apriori.

27 Lineare Regression

27.1 Der einfache lineare Fall

Ein sehr wichtiges und einfaches Verfahren des maschinellen Lernens ist die lineare Regression. Hier wird versucht, eine Funktion mittels einer linearen Funktion zu approximieren. Etwas allgemeiner verwendet man lineare Regression für die Approximation mittels Polynomialfunktionen beliebigen Grades. Nehmen wir erstmal eine einfache lineare Funktion $f : \mathbb{R} \to \mathbb{R}$, die die Form haben soll:

$$f(x) = ax + b \tag{328}$$

wobei $a, b \in \mathbb{R}$ die Parameter sind. In diesem Fall muss unser Datensatz $D \subseteq \mathbb{R} \times \mathbb{R}$ einfach eine Abhängigkeit zweier reellwertiger Parameter darstellen. Wir setzen als Konvention $D = \{(a_1, b_1), ..., (a_n, b_n)\}$, und für $x = a_i$ schreiben wir y_x für b_i, also der Wert den D x zuweist. Wir suchen nun diejenige lineare Funktion, die die Differenz minimiert, also

$$argmin_{a,b\in\mathbb{R}} \sum_{(x,y)\in D} (ax + b - y_x)^2 \tag{329}$$

Das Quadrat ist dafür da, dass Werte positiv werden – sonst würden sich negative und positive Abweichungen ausgleichen. Damit gewichten wir natürlich weitere Abweichungen stärker, was nicht unbedingt erwünscht ist; allerdings gibt es kaum andere Möglichkeiten: die Betragsfunktion $|-|$ ist nicht differenzierbar, wir brauchen allerdings die erste Ableitung der Funktion, wie wir unten sehen werden.

Wir machen nun einen Trick: eigentlich sind die Parameter a, b festgelegt, während x das variable Argument der Funktion ist. Weil wir aber nur an denjenigen x interessiert sind, die in unserem Datensatz auftauchen (d.h. endlich viele), während wir alle reellen Parameter berücksichten müssen. Daher ist die Funktion, die wir minimieren müssen, eigentlich folgende:

$$\sum_{(a,b)\in D} (ax + y - b)^2 = (a_1x + y - b_1)^2 + ... + (a_n + y - b_n)^2 \tag{330}$$

Hier haben wir einfach die Konstanten und Variablen vertauscht, und daraufhin eine arithmetische Umformung vorgenommen. Am Ende bekommen

wir die einfache Form (denn $a_1, ..., a_n, b_1, ...b_n$ sind einfache gegebene Konstanten)

$$f(x,y) = ax^2 + by^2 + cxy + dx + ey + f \tag{331}$$

Denn alle Summanden haben eine dieser Variablenformen als Koeffizient, und wir suchen einfach

$$argmin_{x,y\in\mathbb{R}} ax^2 + by^2 + cxy + dx + ey + f \tag{332}$$

Das berechnet man mit der gewohnten Methode: wir bilden (in diesem Fall partielle) Ableitungen und konstruieren damit den Gradienten. Das ist natürlich besonders einfach:

$$\nabla f(x,y) = ((2ax + cy + d), (2by + cx + e)) \tag{333}$$

Wir haben also 2 Gleichungen, die wir auf 0 setzen müssen:

$$2ax + cy + d = 0 \tag{334}$$
$$2by + cx + e = 0 \tag{335}$$

Wir haben 2 Gleichungen und 2 Variablen, also eine Lösung:

$$x = -\frac{cy + d}{2a} \tag{336}$$

also

$$2by - \frac{c^2y + cd}{2a} + e = 0 \leftrightarrow \tag{337}$$
$$y = \frac{cd - 2ae}{4ab - c^2} \tag{338}$$

Wir haben also die Nullstelle für x, y berechnet, und wissen somit, wie wir die Funktion minimieren können.

27.2 Der komplexe lineare Fall

Im komplexeren nehmen wir an, dass

$$f(x) = a_n x^n + a_{n-1}x^{n-1} + ... + a_0 \tag{339}$$

f ist also nun ein Polynom, keine lineare Funktion mehr. Warum ist das immer noch lineare Regression? Weil wir, bei der eigentlichen Regression, also der Suche nach den Parametern die die beste Funktion ausmachen, x als eine Konstante behandeln (wir setzen nämlich Datenpunkte ein, bekommen also einfach konstante Werte in $\mathbb{R}$), während die eigentlichen Variablen die Werte $a_0, ..., a_n$ sind. In diesen Werten ist die resultierende Funktion nach wie vor linear – wir haben also einen Fall der etwas komplexer ist als der vorhergehende, aber nach wie vor durch die Lösung linearer Gleichungssysteme lösbar ist.

$$f(x_0, ..., x_n) = \sum_{(a,b)\in D} (a^n x_n + a^{n-1} x_{n-1} + ... + x_0 - b)^2 \quad (340)$$

Hier sind a^n etc. und b fixe reelle Zahlen, während die Variablen nur linear auftreten. Wir müssen nun

$$\nabla f : \mathbb{R}^{n+1} \to \mathbb{R}^{n+1} \quad (341)$$

konstruieren, kriegen also ein lineares Gleichungssystem mit $n+1$ Variablen und $n+1$ Gleichungen, das wir entsprechend lösen können. Lineare Regression lässt sich also mit elementaren mathematischen Methoden lösen, und das ist der große Vorteil dabei.

Eine weitere Erweiterung, die ohne große Probleme funktioniert, ist folgende: anstatt einer Funktion $f : \mathbb{R} \to \mathbb{R}$ suchen wir eine Funktion $f : \mathbb{R}^n \to \mathbb{R}$, die die Form hat

$$f(x_1, ..., x_n) = a_1 x_1 + a_2 x_2 + ... + a_n x_n + b \quad (342)$$

Auch das geht ohne große Probleme mit den obigen elementaren Methoden, und auch die Erweiterung auf Polynomialfunktionen (auch wenn natürlich alles etwas schwieriger wird).

27.3 Lineare Regression in R

In R gibt es ein einfaches Kommando zur (einfachen) lineaeren Regression, nämlich `lm`. Erstmal brauchen wir Daten; dazu nehmen wir eine Menge $D \subseteq \mathbb{R}^2$. In R geht das einfacher, wenn man zwei Vektoren nimmt:

$$v_1 = (x_1, ..., x_n),\ v_2 = (y_1, ..., y_n), \text{ wobei } D = \{(x_1, y_1), ..., (x_n, y_n)\}$$

Denn hiermit kann man eine einfache Funktion nutzen:

> x < − c(-4,-2.8,-1.5,0,0.7,1.9,2.3)
> y < − c(7,4.3,5.2,3.8,3.6,2,0.8)
> plot(x,y,xlim = c(-5,5),ylim=(-7,7))

Hier werden also die Werte von `x` gegen `y` geplottet, wobei die Zuordnung anhand der Stelle im Vektor erfolgt. Nun möchten wir diese Funktion linear approximieren, also mittels eines Modelles

(343) $f(x) = ax + b$

Das geht in R ganz einfach mit dem Befehl:

> lm1 ¡- lm(y ∼ x)

Wir haben hier das Modell lm1, und das soll y als linear abhängige Variable von x vorhersagen.

> lm1
Call:
lm(formula = y x)

Coefficients:
(Intercept) x
3.4308 -0.7896

Wir bekommen also den *intercept*, eine konstante die uns sagt wo die Funktion die y-Achse kreuzt, und den *slope*, also die Steigung, der die lineare Anhängigkeit von x liefert; wir haben also:

(344) $lm1(x) = -0.7896x + 3.4308$

Das ist also die beste lineare Approximation unserer Daten. Wie gut sie ist kann man mit der Funktion `residuals` sehen, die jeweils liefert:

(345) $residuals(f) = (f(x_1) - y_1, ..., f(x_n) - y_n)$

In unserem Fall also:

> residuals(lm1)
1 ...
0.41079392 ...

Man kann die Regressionsfunktion auch schön visualisieren, mittels:

> abline(lm1, col = "red")

Um Modelle höherer Ordnung in R zu bekommen, gibt es meines Wissens nach keinen direkten Befehl. Man kann das aber recht einfach machen. Zunächst muss man wissen, wie man eine Variable x in Abhängigkeit mehrerer Variablen setzt. Das geht ganz einfach:

> lm2 <- lm(y ~ x+z)

Jetzt müssen wir nur noch den passenden Vektor z erstellen:

> z <- 1:7
> for(i in 1:7)z[i]=(x[i])$\hat{}$2
> lm2 <- lm(y+ ~ x + z)

In diesem Fall bekommen wir eine Polynom zweiter Ordnung für unsere Funktion; wir können die Ordnung weiter erhöhen, bis unsere Residuen bei 0 liegen werden. Die große Frage ist: wird unser Modell dadurch besser?

27.4 Meta-Parameter, Overfitting, Underfitting

Bei linearer Regression geht es darum, Parameter zu schätzen, wie immer beim maschinellen Lernen. Es gibt aber auch *Meta-Parameter*, die wir *apriori* festlegen, und die durch die Daten nicht beeinflusst werden. In diesem Fall ist das der Grad der Funktion, die wir für die Regression nutzen, also linear, quadratisch etc. Um hier den richtigen Wert zu finden, müssen wir die Daten richtign einschätzen: ist der Grad zu niedrig, finden wir nicht die passende Generalisierung (underfitting); ist der Grad zu hoch, übergeneralisieren wir. Um Meta-Parameter richtig zu schätzen – bzw. um Anhaltspunkte zu haben, wie sie am besten liegen in einer gewissen Klasse von Problemen – nimmt man üblicherweise eine Trennung vor von Trainingsdaten und Testdaten. Die Testdaten werden nicht genutzt, um das System zu optimieren, aber hinterher

wird das System auf den Testdaten *evaluiert*. Auf diese Weise kann man prüfen (bis zu einem gewissen Maß, ob wir die korrekten Generalisierungen getroffen haben.

Es gibt fr das ML zwei groe Probleme, oder besser gesagt: zwei Arten, wie unsere Ergebnisse danebenliegen knnen, die eine ist *overfitting*, die andere *underfitting*. Einfach gesagt handelt es sich um folgende Probleme: *overfitting* bedeutet, dass wir keine ausreichenden Generalisierungen treffen, weil wir zu sehr an den Daten bleiben; *underfitting* bedeutet, dass die Dateen nicht gengend bercksichtigen und daher die Muster nicht erfassen. *Overfitting* geschieht, wenn unsere Modellklasse zu mchtig ist, also zu viele Parameter hat. *Underfitting* bekommen wir, wenn unsere Modellklasse nicht mchtig genug ist, also die Variation der Daten gar nicht erfassen kann.

Man kann das gut anhand von einem Beispiel erklren: nehmen wir an, wir sehen eine Elster im Park ein Nest bauen. Daraus kann man verschiedene Generalisierungen ziehen:

(12) a. Elstern bauen im Park Nester.
b. Elstern bauen Nester.
c. Vögel bauen Nester.

In diesem Fall wre a. eine Form von Overfitting: wir haben einen Parameter zuviel, nmlich den Ort, der keine Rolle spielt, und daher entgeht uns die richtige Generalisierung b. Umgekehrt ist c. eine falsche Generalisierung, denn wir haben einen Parameter zuwenig, die Vogelart, die wir brauchen um eine korrekte Generalisierung zu treffen. Es geht also darum, die richtige Zahl von Parametern zu finden (in unserem Beispiel: Faktoren die relevant sind), um die richtigen Generalisierungen zu finden. Dafr gibt es allerdings kein Patentrezept, wie wir auch im obigen Beispiel sehen: oft hilft es einfach nur, wenn wir domnenspezifisches Wissen haben, dass wir anwenden (mehr dazu gleich).

Ein weiteres Problem hierbei ist, dass es oft Strparameter gibt, also Parameter, die nicht relevant sind fr das was wir suchen. das knnen typischerweise Mefehler sein, aber auch durchaus systematische Dinge: nehmen wir an, wir suchen den Erwartungswert einer Normalverteilung. Dann ist die Varianz dieser Verteilung ein Strparameter, da sie durchaus unsere Beobachtungen beeinflusst, aber keine Relevanz hat fr das was wir suchen. Da es oft diesen Strparameter gibt, ist es auch nicht immer wichtig, dass unsere Modelle die Daten exakt reproduzieren: das wre nmlich oft bereits eine Form von

overfitting. Viel wichtiger ist, dass die Abweichung nicht systematisch ist, und das wir eben nicht zuviele Parameter haben.

27.5 Parameter und Hyperparameter, Test und Trainingsdaten

Die Frage, ob wir over- oder underfitting haben, lt sich nicht beantworten, indem wir nur einen Datensatz betrachten. Stattdessen teilen wir den Datensatz in zwei Teile: nmlich die **Trainingsdaten** und die **Testdaten**. Wir nutzen die Trainingsdaten fr die Optimierung des Modells (das eigentliche Lernen). Wenn das abgeschlossen ist, dann schauen wir, wie gut das Modell auf den Testdaten funktioniert. Die Idee dahinter ist folgende: wenn wir unser Modell auf den Trainingsdaten overfitten, dann wird es auf den Testdaten schlechter abschneiden; dasselbe gilt frs underfitten. Wichtig ist hierbei: das Modell darf whrend des Trainings die Testdaten nicht sehen, und wir mssen streng verhindern, dass Informationen aus den Testdaten im Training verwendet werden. Wenn wir nun zuviele Parameter verwenden, dann bekommen wir gute Ergebnisse auf den Trainingsdaten, aber nicht auf den Testdaten, denn diese berflssigen Parameter konnten ja nicht in Hinblick auf diese Daten optimiert werden. Gleichzeitig gibt es natrlich keinen Grund, warum wir underfitten sollten, denn wenn ein Parameter relevant ist, wird er sowohl fr Trainings- als Testdaten relevant sein. Diese Methode ist also sehr geeignet um das oben beschriebene Problem zu umgehen.

Die Partition der Daten in Training und Test sollte rein zufllig geschehen; blicherweise nimmt man ein Verhltnis 4:1 an. Es kann evtl. Probleme geben, wenn wir nicht gengend haben, wir also keine Daten fr den Test "entbehren" knnen. In diesem Fall nutzt man die Methode der **cross-validation**: wir partionieren unsere Daten in Test- und Trainingssatz, wobei der letztere viel zu klein ausfllt. Das wird dann allerdings immer wieder iteriert (z.B. ist jeder Datenpunkt einmal Trainingsatz). Das liefert jedesmal ein (etwas) anderes Modell, aber am Ende knnen wir ber die Ergebnisse mitteln und so sehen, ob das Modell insgesamt stimmt.

Das bringt uns auf eine weitere wichtige Unterscheidung: bei cross-validation haben wir streng genommen jedesmal andere Parameter; was gleich bleibt sind die **Hyperparameter**. Hyperparameter sind Eigenschaften des Modells, die nicht von den Daten abhngen, z.B. dass das Modell eine lineare Funktion ist (ein Polynom). brigens ist auch die Anzahl und Art der Param-

eter, die unser Modell hat, ein Hyperparameter. Mit cross-validation zeigen wir also, dass wir korrekte Hyperparameter gewhlt haben.

Hier gibt es allerdings eine Kleinigkeit zu beachten: nehmen wir an, es gibt einen Datensatz D, mit dem wir wiederholt verschiedene Modelle ($\cong$ Hyperparameter) testen, indem wir ihn auf verschiedene Art und Weise in Test- und Trainingssatz spalten. Auf diese Weise schlieen wir overfitting für die Parameter aus. Eine andere Sache, die jedoch passieren kann, ist dass wir auf diese Art und Weise overfitting für die Hyperparameter bekommen; denn diese werden immer wieder auf demselben Datensatz trainiert. Das ist insbesondere ein Problem fr Gebiete, auf denen Daten nur schwer zu bekommen sind. Deswegen ist es fr alle Gebiete wichtig, dass wir beständig neue Daten haben.

28 Logistische Regression – Aktivierungsfunktionen

28.1 Definitionen

Logistische Regression ist im engeren Sinne keine Regression, sondern Klassifikation; es ist aber eine Klassifikation, die auf linearer Regression aufbaut. Logistische Regression funktioniert im einfachen Fall, wenn wir 2 Klassen haben, zwischen denen wir auswählen, z.B. A und B. Unser Problem ist also folgendes: wir möchten eine Eingabe in $\mathbb{R}$ nach A oder B klassifizieren. Wir machen das mittels der logistischen Sigmoid-Funktion

$$S(x) = \frac{1}{1 + e^{-x}} \tag{346}$$

Einige Beobachtungen: wir haben

- $lim_{x \to -\infty} S(x) = 0$
- $lim_{x \to \infty} S(x) = 1$
- $S(x) \in (0, 1)$ f.a. $x \in \mathbb{R}$
- $S(x)$ ist streng monoton steigend in x

Allgemeiner nennt man solche Funktionen **Sigmoidfunktionen**. Ein anderes Beispiel für eine solche Funktion ist

$$T(x) = \frac{x}{1 + |x|} \tag{347}$$

$$U(x) = \frac{x}{1 + x^2} \tag{348}$$

$$tanh(x) = \frac{sinh(x)}{cosh(x)} = \frac{e^x - e^{-x}}{e^x + e^{-x}} = 1 - \frac{2}{e^{2x} + 1} \tag{349}$$

Was allen diesen Funktionen gemeinsam ist, ist dass sie im Bereich um 0 relativ schnell von 0 Richtung 1 wechseln, sonst fr groe positive/negative Zahlen langsam Richtung 1/0 konvergieren. Es gibt also nur einen kleinen Bereich, in dem eine nicht extreme Änderung im Argument eine signifikante Änderung im Wert verursacht; ansonsten gibt es relativ schnell eine Sättigung. solche

Prozesse findet man oft in der Natur, z.B. bei Populationen. Solche logistischen Funktionen haben allerdings die Einschränkung dass sie nur $\mathbb{R} \to \mathbb{R}$ definiert sind. Aktivierungsfunktionen spielen in neuronalen Netzen eine große Rolle, und werden normalerweise auf die einzelnen Komponenten von Vektoren angewendet.

Wenn wir nun 2 Klassen haben, dann können wir einfach sagen: wir interpretieren den Wert als eine **Wahrscheinlichkeit**, also:

$$P(X = A|x) = S(f(x)) \tag{350}$$

wobei f eine lineare Funktion ist, die wir mittels linearer Regression induzieren. Wir transformieren also eine Funktion in die reellen Zahlen zu einer Funktion nach $[0, 1]$:

$$f : \mathbb{R} \to \mathbb{R} \Longrightarrow S \circ f : \mathbb{R} \to [0, 1]$$

Das Ergebnis können wir dann als Wahrscheinlichkeit interpretieren. Das gibt uns nun einen Klassifikator:

$$C(x) = \begin{cases} A, \text{ falls } P(X = A|x) > 0.5 \\ B \text{ andernfalls} \end{cases} \tag{351}$$

Das funktioniert natürlich nur, falls wir nur zwei Klassen (hier A, B) zur Verfügung haben. Intuitiv bilden wir hier die Eingabe auf eine Wahrscheinlichkeit ab, und Klassifizieren nach Maximum Likelihood. Natürlich kann auch das sehr leicht generalisiert werden auf beliebige Eingaben $\mathbf{x} \in \mathbb{R}^n$.

28.2 Bedeutung

Die Bedeutung der logistischen Regression ist erstmal folgende: wir suchen eine Klassifikationsfunktion

$$f : \mathbb{R}^n \to \{0, 1\}$$

Wir tun das aber mittels zweier Zwischenschritte:

$$\mathbb{R}^n \underset{f}{\Longrightarrow} \mathbb{R} \underset{S}{\Longrightarrow} \{0, 1\}$$

Die zugrundeliegende Annahme ist hierbei, dass es einen kritischen Bereich gibt, auf dem die Klassifikation von 0 zu 1 springt und umgekehrt ist; aber damit man das sieht, muss man zunächst eine **numerische Transformation** durchführen, nämlich die lineare Regression. Genauer gesagt, die lineare Funktion f kann die numerischen Werte so transformieren, dass gilt:

$$\text{Falls } f(x) > s, \text{ dann } C(x) = 1, \text{ und falls } f(x) \leq s, \text{ dann } C(x) = 0$$

Das bedeutet, wir haben eine *quasi-lineare Abhängigkeit* von unseren Eingabedaten und den Klassen. Das kann man noch genauer fassen, es gibt nämlich eine genaue Beschreibung wenn wir den **Hyperparameter** des **Grades der Funktion** mitberücksichtigen. Nehmen wir an, wir suchen eine Funktion $f : \mathbb{R} \to \{0, 1\}$. Wir wissen, dass

- Eine lineare Funktion ist hat eine Gerade als Graphen, kann also einen gewissen Wert nur einmal annehmen
- Ein Polynom zweiten Grades hat einen Wendepunkt, kann einen gewissen Wert höchstens zweimal annehmen.
- ...
- Ein Polynom $n$ten Grades hat n Wendepunkte, kann einen gewissen Wert höchstens n-mal annehmen.

Dann heißt das soviel wie:

- Falls wir eine lineare Funktion haben, dann gibt ein $\alpha \in \mathbb{R}$, so dass falls $x < \alpha$, dann $f(x) = 1$ (bzw. 0), ansonsten $f(x) = 0$ (bzw. 1). Wir spalten also den Eingaberaum $\mathbb{R}$ in zwei Teile.

- Falls wir ein Polynom zweiten Grades haben, dann gibt $\alpha_1, \alpha_2 \in \mathbb{R}$, so dass falls $\alpha_1 < x\alpha_2$, dann $f(x) = 1$ (bzw. 0), ansonsten $f(x) = 0$ (bzw. 1). Wir spalten also den Eingaberaum $\mathbb{R}$ in drei Teile.
- ...
- Falls wir ein Polynom n-ten Grades haben, haben wir $\alpha_1, ..., \alpha_n$, die $\mathbb{R}$ in $n + 1$ Teile spalten, und wir klassifizieren auf dieser Basis.

28.3 Lernen & Anwendung

Wir betrachten folgendes Beispiel: wir möchten, als abhängige, diskrete Variable vorhersagen, ob ein Student die Prüfung besteht (0 oder 1; abhängig bedeutet nur: wir möchten das vorhersagen). Als Prädiktor nutzen wir die Anzahl der Stunden, die der Student gelernt hat. Unsere Regressionsfunktion soll eine einfache lineare Funktion sein. Unser Datensatz sieht nun wie folgt aus (nennen wir den Datensatz $D1$):

Stunden gelernt	4	5	6	7	8	9	10
Bestanden	0	0	1	0	1	1	1

Eine lineare Regression liefert hier folgendes Ergebnis (danke R):

(352) $f(x) = 0.1786x - 0.6786$

Wir haben also eine positive Abhängigkeit von Lern-Stunden und Klausur-Bestehen (zum Glück); aber dieses Ergebnis ist noch unbefriedigend: wir können das nicht sinnvoll als Wahrscheinlichkeit interpretieren. Wir setzen aber nun diesen Term ein in unsere Aktivierungsfunktion, und defineren:

(353) $P(\text{bestehen}|n \text{ Stunden lernen}) = \frac{1}{1 + e^{-f(x)}} = \frac{1}{1 + e^{-(0.1786x - 0.6786)}}$

Das liefert uns nun ordentlich Werte:

Stunden	4	5	6	7	8	9	10
P(bestehen)	0.508	0.553	0.597	0.639	0.679	0.717	0.752

Das ist schon besser, aber nicht wirklich optimal, insbesondere stört die Asymmetrie (bei 4 haben wir bereits eine ziemlich hohe Wahrscheinlichkeit!).

Die logistische Funktion interagiert mit der linearen auf eine Art und Weise, die nicht optimal ist. Wir sollten also stattdessen folgendes suchen:

$$(354)\ \ argmin_{a,b\in\mathbb{R}} \sum_{(x,y)\in D} \frac{1}{1+e^{-(ax+b)}} - y$$

Wir nutzen also die Aktivierungsfunktion, um den Unterschied zwischen 0 und 1 sichtbar zu machen, da eine lineare Funktion hierzu nicht in der Lage ist. Um 354 auszurechnen gibt reichen keine elementaren Methoden, man muss also etwas komplexere numerische Optimierung anwenden (der Computer kann das). R macht das mit dem Kommando:

> glm(x ∼ y, family = binomial())

Nun definieren wir:

> $g <-$ function(x){1/(1+exp(-(1.251*x-8.113)))}

Das sollte nun stimmen; wir bekommen dementsprechend:

Stunden	4	5	6	7	8	9	10
P(bestehen)	0.0427	0.135	0.353	0.656	0.869	0.959	0.988

Tadaa! Wir bekommen also eine Wahrscheinlichkeit. Was aber interessanter ist, ist folgende Frage: gegeben diese Ergebnisse, wie ist die Wahrscheinlichkeit, dass die Abhängigkeit des Bestehens von der Lernzeit reiner Zufall ist? Das ist eine klassische statistische Frage, und im Zusammenhang mit logistischer Regression gibt es den sog. **Wald-test**:

$$(355)\ \ \frac{(\theta_{ML}-\theta)^2}{var(\theta_{ML})}$$

Er gibt die Wahrscheinlichkeit, dass die gegebene Verteilung zufällig ist, also dass es keine Abhängigkeit der beiden Variablen (Lernzeit / Bestehen) gibt. Das ist verwandt mit dem Test des Likelihood-Verhältnisses, was ein wenig einfacher zu verstehen ist. Seien

- ω unsere Beobachtungen, also die Daten in D1;
- H_0 die Nullhypothese, also Unabhängigkeit der Variablen: Lernzeit hat keinen Einfluss auf das Bestehen, Wahrscheinlichkeit von Bestehen oder Nicht-bestehen wird durch Maximum Likelihood geschätzt.

- H_1 die Alternativhypothese, also unsere durch logistische Regression berechnete bedingte Verteilung.

Dann haben wir den Likelihood-Quotienten:

$$R(\omega) := \frac{P(\omega|H_0)}{P(\omega|H_1)} \qquad \text{(Sonderfall für } P(\omega|H_1) = 0) \tag{356}$$

Hierauf kann man nun den Schwellentest S_t anwenden, $t > 0$, und üblicherweise $t = 0.05$. Zur Erinnerung: das ist der Test, der sich für H_0 entscheidet falls $R(\omega) > t$, und für H_1 andernfalls, also:

$$S_t(\omega) = \begin{cases} H_0, \text{ falls } R(\omega) > t \\ H_1 \text{ andernfalls.} \end{cases} \tag{357}$$

Das können wir/Sie nun ausrechnen:

Aufgabe 11

Abgabe bis zum 11.7.2017 vor dem Seminar.

1. Berechnen Sie für das Beispiel in diesem Abschnitt (die Daten $D1$, H_0 und H_1 wie oben angegeben) den Quotienten $R(\omega)$.

2. Ist das Ergebnis signifikant, das heißt, würde der Schwellentest H_0 zurückweisen (wir setzen $t = 0.05$)?

29 Nearest neighbour Regression

Nearest neighbour regression ist eine denkbar simple Methode: wir haben eine Funktion $f : V \to C$, wobei V ein Vektorraum ist, und C eine (endliche) Menge von Klassen; wir haben eine Menge $D \subseteq V \times C$ von Datenpunkten, mittels derer wir die Funktion f "lernen" sollen. Was wir hierfür erstmal brauchen ist der Begriff der **Norm**:

Eine Norm, definiert auf einem Vektorraum V ist das eine Funktion

$$\| - \| : V \to \mathbb{R}_0^+$$

es werden also beliebige Vektoren auf einen nicht-negativen Wert abgebildet. Zustzlich muss $\| - \|$ noch folgende Bedingungen erfllen f.a. $\vec{v} \in V$, $\lambda \in \mathbb{R}$.

1. $\|\vec{v}\| = 0 \Rightarrow \vec{v} = 0_V$ (die 0 des Vektorraums, neutral für Addition)
2. $\|\lambda \cdot \vec{v}\| = |\lambda| \cdot \|\vec{v}\|$, wobei $|\lambda|$ der Betrag ist (respektiert Skalarmultipikation)
3. $\|\vec{v} + \vec{w}\| \leq \|\vec{v}\| + \|\vec{w}\|$ (allgemeine Dreiecksungleichung)

Die intuitivste Norm ist die *euklidische*, die jedem Vektor seine **Lnge** zuweist (wenn wir einen Vektor als eine Linie vom Ursprung auf seine Koordinaten (im n-dimensionalen Raum) auffassen. Diese Norm basiert auf einer Verallgemeinerung des Satz des Pythagoras:

$$\|(v_1, ..., v_n)\| = \sqrt{v_1^2 + ... + v_n^2} \tag{358}$$

In dieser geometrischen Interpretation wird Bedingung 3 zur **Dreiecksungleichung**: in jedem rechwinkligen Dreieck ist die Lnge der Hypothenuse geringer als die Summe der Lnge der Katheten. Es gibt aber noch viele weitere Normen, z.B. die sog. p-Norm, wobei $p \geq 1$ eine reelle Zahl ist:

$$\|(v_1, ..., v_n)\|_p = \left(\sum_{i=1}^{n} |v_i|^p \right)^{\frac{1}{p}} \tag{359}$$

Für $p = 1$ vereinfacht sich das zu

$$\|(v_1, ..., v_n)\|_1 = \sum_{i=1}^{n} |v_i| \tag{360}$$

Das ist die sog. Manhatten-Norm, weil man immer rechtwinklig um die Blocks fahren muss – das gibt im 2-dimensionalen Fall also die krzeste Strecke in Manhattan an.

Darauf basiert der Begriff der Metrik; jede Norm induziert eine Metrik d mittels

(361) $d(\vec{v}, \vec{w}) = \|\vec{v} - \vec{w}\|$

wobei natürlich

(362) $\vec{v} - \vec{w} = (v_1 - w_1, ..., v_i - w_i)$

Es ist nicht schwer zu sehen dass gilt:

- $d(x, y) \geq 0$ (positiv)
- $d(x, y) = d(y, x)$ (symmetrisch)
- $d(x, y) \leq d(x, z) + d(z, y)$ (Dreiecksungleichung)

Die **euklidische Distanz** ist definiert durch die euklidische Norm, mit

(363) $d_2(\vec{v}, \vec{w}) = \|\vec{v} - \vec{w}\|_2$

Nun machen wir einfach folgendes: gegeben einen beliebigen Vektor $\vec{v}$, eine Norm $\| - \|$, definieren wir

(364) $nn(\vec{v}) = argmin_{(\vec{w},c)\in D} d(\vec{v}, \vec{w})$

Wir suchen uns also den **nächsten Nachbarn** nach unserer Metrik. Als nächstes machen wir das denkbar naheliegendste: wir machen genau das, was der nächste Nachbar macht (Einfachheit halber fassen wir unseren Datensatz D als partielle Funktion $D : V \to C$ auf)

(365) $NNR_D(\vec{v}) = D(nn(\vec{v}))$

In Wort: $NNR_D : V \to C$ ist die nearest neighbour regression über D, und diese (vollständige) Funktion funktioniert, indem sie für jeden Eingabevektor $\vec{w}$ zunächst den nächstliegenden Vektor $\vec{v}$ im Datensatz findet (ein endliches Suchproblem), um ihm dann dieselbe Klasse zuzuordnen, die auch $\vec{v}$ zugeordnet wird.

Ein einfaches Beispiel wäre z.B. eine Sprach- oder Bilderkennung: bei Bildern wäre das Pixelbild ein Vektor, in dem jeder Pixel eine Komponente ist, und der Wert ist die Farbe der Pixel. Bei Spracherkennung kann man die verschiedenen Frequenzbereiche F1-F3 auf einen Vektor verteilen, mit dem Wert als Frequenz; die Klassifikation wäre dann die Frage: welcher Gegenstand ist auf dem Bild abgebildet (aus einer endlichen Menge), bzw. welcher laut wird geformt? NNR ist einfach folgende Methode: finde den ähnlichsten Punkt in unserem Datensatz und klassifiziere entsprechend.

Was ganz interessant ist: NNR kann nicht durch eine lineare Funktion beliebiger Ordnung modelliert werden. Das sieht man sehr leicht an einem Beispiel: man nehme einen Datensatz

$$D = \{((0,0),a),((0,1),b),((0,2),a),((0,3),b),...\}$$

NNR wird hier beliebig oft wechseln können zwischen a und b; jedes lineare Modell wird nur eine konstante Anzahl von wechseln ermöglichen können.

30 Principle component analysis

Kovarianz und Kovarianz-Matrix

Hier sind die Daten Vektoren, und die Komponenten werden als Zufallsvariablen gedacht.

$cov(X, Y) = \mathcal{E}[(X - E[X])(Y - \mathcal{E}[Y])]$

(siehe Bengio, DL)

31 k-means clustering

(siehe Bengio, DL)

32 Zur Methodik des maschinellen Lernens

32.1 Abriß der Methode

Im Rahmen des maschinellen Lernens und der Wahrscheinlichkeitstheorie wird natürlich oft die Frage gestellt: ist es so dass ein Lern-Algorithmus bessere Ergebnisse erzielt als ein anderer? Tatsächlich wird das meistens empirisch überprüft, nach folgender Methode:

Wir haben einen Datensatz D gegeben.

Wir partitionieren den Datensatz D in zwei disjunkte Teilsätze $D_{training}, D_{test}$.

Wir nutzen nun $D_{training}$, um unseren Klassifikator zu **trainieren**, also z.B. um unseren Entscheidungsbaum festzulegen, unser BN zu induzieren etc.

Zuletzt nutzen wir D_{test}, um unseren Klassifikator zu **evaluieren**, d.h. wir prüfen, wie gut er auf diesen Daten abschneidet. Diese Evaluation wird normalerweise als Kennzeichen aufgefasst für die Qualität unseres Klassifikators.

Wir sehen also, dass diese Methodik in gewissem Sinne eine Alternative zur Methodik der klassischen/bayesianischen Statistik liefert: anstatt den gesamten Datensatz auf Regelmäßigkeiten zu prüfen, nutzen wir einen Teil, ziehen Generalisierungen, und prüfen dann, ob die Generalisierungen richtig sind. Das soll uns vor dem Problem des *Overfitting* schützen: jeder noch so große Datensatz erlaubt "falsche" Generalisierungen, die auf Artefakten beruhen. Ein einfaches Beispiel hierfür: jedes Korpus K hat einen längsten Satz S mit k Worten. Dementsprechend könnten wir immer den Schluss ziehen:

$$\text{jeder Satz hat Länge} \leq k.$$

Die Tatsache, dass wir unsere Generalisierungen prüfen auf einem Datensatz, den wir vorher nicht gesehen haben, soll das verhindern bzw. einschränken. Insbesondere glaubt man:

> Wenn eine Generalisierung über $D_{training}$ auch für D_{test} gilt, dann hat sie gute Chancen, allgemein korrekt zu sein.

Dafür gibt es natürlich keine Garantie, aber zumindest zeigt dass, das die Generalisierungen keine Artefakte der Daten sind.

32.2 Zwei Probleme

1. Es gibt bei dieser Methode allerdings 2 Dinge zu beachten. Das erste ist folgendes Problem: das Aufsplitten der Daten ergibt nur dann einen Sinn und einen echten Vorteil, wenn der Datensatz für den Lernalgorithmus **tatsächlich unsichtbar** ist. Das ist aber leichter gesagt als getan: man nehme z.B. folgenden Fall:

> Wir trainieren den Klassifikator C auf dem Datensatz $D_{training}$; dann testen wir ihn auf D_{test}. Dabei bemerken wir: C macht auf D_{test} einen sehr charakteristischen Fehler relativ häufig. Also ändern wir C dergestalt zu C', dass wir wissen, dass es diesen Fehler nicht mehr macht. Die Ergebnisse sind dementsprechend auch besser.

Ist nun aber C' besser als C? Das lässt sich schwer sagen, und zwar aus folgendem Grund: zwar wurde C' **optimiert**, aber: es wurde optimiert im Hinblick auf D_{test} – von daher ist es nicht verwunderlich, dass C' besser darauf abschneidet als C! Man sagt auch: es sind Informationen von D_{test} nach C' eingeflossen. Der gesamte Vorteil der obigen Methode bestand aber darin, dass das nicht geschah, damit wir eben prüfen konnten dass Generalisierungen nicht nur im Hinblick auf die Daten geschehen! Um also den Vorteil wirklich zu verifizieren, bräuchten wir einen neuen Testsatz D_{test2} – der selten zur Verfügung steht.

2. Das zweite Problem ist grundlegender: wir wissen nicht, ob unser Klassifikator C *nur zufällig* eine gute Performanz auf D_{test} hat. Denn in vielen Fällen ist die Menge der klassifizierten Objekte unendlich, und D ist nur ein kleiner Ausschnitt daraus (eine sog. Stichprobe). Und wir wissen nie, ob wir nicht in dieser Stichprobe eine systematische Tendenz haben (durch die Art, wie wir sie genommen haben), die auf der Population insgesamt nicht erfüllt ist!

> In unserem Beispiel der Entscheidungsbäume: es kann gut sein, dass in unseren Daten gewisse seltene Konstellationen niemals

vorkommen. Dementsprechend unterschätzen wir gewisse Faktoren, weil sie einfach in D (also sowohl $D_{training}$ als auch D_{test}) nie eine Rolle spielen.

Hiergegen kann man natürlich nichts machen. Insbesondere führt uns diese Überlegung auf die NFL-Theoreme.

32.3 Gibt nix umsonst – die *no-free-lunch* Theoreme I

Stellen wir uns folgendes vor: wir interessieren uns für das Problem der Erfüllbarkeit Boolescher Formeln. Eigentlich sind das 2 Probleme:

1. Gegeben eine Formel ϕ, ist ϕ erfüllbar (gibt es ein Modell, in dem ϕ wahr ist)?

2. Ist ϕ in **allen** Modellen wahr?

Diese Probleme sind eng miteinander verbunden, denn:

$$\phi \text{ ist nicht erfüllbar gdw. } \neg\phi \text{ in allen Modellen wahr ist.}$$

NB: beide Problem sind NP-Vollständig, das heißt: sie können gelöst werden von einer nicht-deterministischen Turing-Maschine in polinomieller Zeit, und – was wichtiger ist – jedes Problem, dass von eine solchen Maschine in polinomieller Zeit gelöst werden kann, kann darauf reduziert werden (dieser Begriff ist etwas technisch).

Nehmen wir einmal an, wir möchten nun einen **probabilistischen Algorithmus** entwickeln, der mittels Wahrscheinlichkeiten seine nächsten Züge auswählt. Wir haben also eine Art Weiterentwicklung der nicht-deterministischen TM, da wir Züge mit unterschiedlichen Wahrscheinlichkeiten ausführen.

Die Wahrscheinlichkeiten wollen wir natürlich nicht erfinden, sondern wir können, z.B., die Wahrscheinlichkeiten trainieren auf einem gegebenen Datensatz von Formeln und Lösungswegen (wo wir immer den besten Lösungsweg nachgehen).

Als Ergebnis haben wir einen Algorithmus, der verschiedene Lösungswege mit verschiedenen Wahrscheinlichkeiten ausführt. Als nächstes testen wir sie auf einem Testsatz. Das muss natürlich so aussehen, dass wir ihn mehrmals auf demselben Datensatz laufen lassen, und das Mittel über die Anzahl der

Lösungsschritte nehmen (denn wir haben ja Nicht-Determinismus, wollen also die durchschnittliche Anzahl von Schritten – sonst hätte es ja sein können, dass wir einfach Glück hatten! Nun finden wir, dass unser Algorithmus gute Ergebnisse liefert. Was sagt uns das für den allgemeinen Fall?

> Die Antwort hierauf geben die NFL-Theoreme, und sie lautet: **nichts**.

Wie ist das möglich? Kurz gesagt besagen die NFL-Theoreme: unser Algorithmus funktioniert nur dann besser auf einer gewissen Klasse von Formeln, wenn er auf einer anderen Klasse schlechter funktioniert.

Allgemeiner formuliert: auf einer gegebenen Klasse von Problemen operiert jeder probabilistische Algorithmus im Mittel gleich gut. Das bedeutet: wir haben nur einen Vorteil, wenn die Klasse von Problemen selbst probabilistischer Natur ist (gewisse Formeln sind wahrscheinlicher als andere), und unser Algorithmus *Wissen über die zugrunde liegende Wahrscheinlichkeitsverteilung inkorporiert.*

32.4 *NFL*-Theoreme und maschinelles Lernen

Die NFL-Theoreme werfen auch ein Schlaglicht auf das Grundproblem des maschinellen Lernens, nämlich die Induktion: wie können wir von einem endlichen Datensatz D zu einem (unendlichen) Satz der *möglichen* Daten T generalisieren, gegeben dass es (unendlich) viele T gibt, die mit D kompatibel sind?

33 Fuzzy Logik

Fuzzy logic is neither a poor man's logic, nor a poor man's probability.
Petr Hajek

33.1 Einleitung

Krause Logik beschreibt Logiken, die entstehen, wenn man nicht mehr davon ausgeht, dass Aussagen wahr (1) oder falsch (0) sind, sondern einen beliebigen Wahrheitswert in [0,1] annehmen können. Der Name kommt eben daher: unsere Wahrheitswerte sind nicht mehr hart und scharf, sondern eben unscharf oder "kraus". Man darf sich aber nicht täuschen: Krause Logik hat unscharfe, "weiche" Wahrheitsverhältnisse zum Gegenstand; die mathematische Theorie der krausen Logik ist aber scharf und präzise wie die Theorie jeder anderen Logik auch. Insbesondere werden wir sehen, dass sich krause Logiken in den größeren Zusammenhang der substrukturellen Logiken einordnen lassen. Das bedeutet, dass krause Logiken "ganz gewöhnliche" Logiken sind, die mit den gewöhnlichen Werkzeugen der Metalogik analysiert werden können (Beweiskalküle, algebraische Semantik etc). Es gibt mehrere Ausgangspunkte für die krause Logik; einer ist die Arbeit des polnischen Logikers Lukasiewicz über mehrwertige Logik aus den 1930er Jahren. Eine andere Wurzel ist die Arbeit über krause Mengen, die in die 1960er Jahre zurückgeht (Lotfi Zadeh). Das Standardwerk zum Thema, an das ich mich im Zweifelsfall halte, ist "Metamathematics of Fuzzy Logic" von Petr Hajek.

Krause Logiken haben mittlerweile eine sehr große Bandbreite von Anwendungen. Die Hauptanwendung ist aber nach wie vor die **Kontrolle von physischen Systemen** (damit ist gemeint: Systemen, die in der "echten Welt" funktionieren müssen), so etwa Klimatisierungen, U-Bahnen (erste große Anwendung in der Nanboku-Linie, Sendai, Jp.), Roboter, Fabriken etc. Diese Präsentation wird wie folgt vorgehen:

1. erst betrachten wir ein Beispiel;
2. dann fangen wir an mit den Grundlagen der krausen Mengenlehre,
3. gehen dann in die Semantik der krausen Logik, und betrachten zuletzt Beweiskalküle.

Diese Darstellung ist sicher nicht alleine seligmachend, entspricht aber am ehesten dem natürlichsten Zugang zu krauser Logik aus der Theorie der Wahrscheinlichkeit.

33.2 Ein Beispiel: Klimatisierung

Die Aufgabe einer Klimatisierung ist vor der Hand einfach:

- Sie geben eine Zieltemperatur ein;
- das System soll heizen/kühlen, um die Zieltemperatur zu erreichen.

Seltsamerweise ist die Sache dann doch wesentlich komplizierter. Das liegt unter anderem an folgenden Punkten:

1. Tagüber sollte es allgemein wärmer sein als nachts.
2. Dieselbe Raumtemperatur wird als wärmer empfunden, wenn die Sonne scheint.
3. Wenn jemand die Temperatur (deutlich) hinuntersetzt, möchte er einen Kühlungseffekt. Aber: üblicherweise wird die Senkung übertrieben, und später wieder korrigiert. In der Zwischenzeit wird aber relativ viel Energie verschwendet!
4. Wenn jemand die Temperatur minimal korrigiert, dann ist derjenige interessiert an einer exakten Temperatur, keiner schnellen Korrektur. Mit diesem Wissen lässt sich Energie sparen.
5. Wird die Temperatur häufig geändert, sollte sie sensibler reagieren als andernfalls.
6. Gibt es starke Variationen in der gemessenen Temperatur, deutet das auf häufige Nutzung hin – die Kontrolle sollte also sensibler reagieren als andernfalls!

NB: es handelt sich hier nicht um Kleinigkeiten: Klimatisierung ist ein wichtiger Faktor des globalen Energieverbrauchs. Die Klimatisierung eines großen Bürogebäudes ist ein bedeutender Kostenfaktor für jede Firma!

Wir haben nun folgende Situation: alle 6 Faktoren spielen eine Rolle, und lassen sich relativ leicht als Eingaben für das System nutzen. Aber:

keine der Eingaben ist binär; sie treffen alle mehr oder weniger zu. Außerdem *addieren* sich Effekte teilweise, teilweise heben sie einander auf. Das bedeutet: für unser Kontrollsystem ist jeweils wichtig, nicht nur ob, sondern auch **in welchem Maße** eine Bedingung erfüllt ist. Jede Handlung soll berücksichtigen, in welchem Maße alle Bedingungen erfüllt sind. Das ist aber gar nicht einfach, denn eine normale Logik kommt dafür nicht in Betracht!

33.3 Krause Mengenlehre

Krause Mengen sind eine natürliche Verallgemeinerung von (diskreten) Wahrscheinlichkeiten: man bildet eine Menge nach $[0, 1]$ ab, ohne sich weiter um Konsistenz zu kümmern. Da die Konsistenz keine Rolle spielt, spielen auch alle anderen Beschränkungen keine Rolle (Additivtät etc.). Also formal gesprochen: eine **krause Menge** M ist eine Funktion

$$M : X \to [0, 1]$$

Die große Frage ist: was ist X? Wir wollen ja den Definitionsbereich von M nicht von vornherein beschränken; wir müssen also verlangen, dass X die universelle Klasse ist (denn es gibt keine universelle Menge). Wir verlangen weiterhin noch dass $M^{-1}(0, 1]$ eine *Menge* im technischen Sinn ist. Damit möchten wir sichergehen, dass $\{x : M(x) > 0\}$ eine Menge ist. Man kann dieses Problem umgehen, indem man eine **Referenzmenge** festlegt, und krause Mengen nur für die Referenzmenge definiert.

Wir kommen nun zur ersten wichtigen Definition: krause Mengen sind intendiert als Generalisierung von normalen Mengen; in der krause Mengenlehre nennen wir normale Mengen **knackig**. Also gilt: eine krause Menge M ist knackig, wenn

$$M[X] = \{0, 1\};$$

d.h. jedes Objekt wird auf 0 oder 1 angebildet. Damit haben wir, was man eine **charakteristische Funktion** ist:

$$x \text{ ist ein Element von } M, \text{ gdw. } M(x) = 1.$$

Man kann knackige Mengen mit ihren charakteristischen Funktionen eindeutig charakterisieren (und umgekehrt); aus diesem Grund werden wir öfter von einer Charakterisierung zur anderen wechseln, ohne das explizit zu sagen.

Das bedeutet: krause Mengen generalisieren die Element-Relation zu verschiedenen Graden von Element-Sein . Die Intuition dahinter ist einfach: in der formalen Semantik sind Eigenschaften Mengen; wenn wir nun sagen etwas hat eine gewisse Eigenschaft (z.B. "groß"), dann kann das mehr oder weniger wahr sein, nicht nur wahr oder falsch (dass viele andere Probleme dabei ignoriert werden: "große Ameise", "kleiner Elefant", nehmen wir in Kauf). Die erste große Frage ist:

> was ist mit den klassischen Relationen und Operationen der Mengenlehre, wie können wir die benutzen?

Wir werden im folgenden die wichtigsten Konzepte der Reihe nach durchgehen.

Die krause **Teilmengenrelation** ist wie folgt definiert:

$$\text{wir sagen } A \subseteq B, \text{ falls f.a. } x \in X,\ A(x) \leq B(x).$$

Sei M eine krause Menge; die **krause Potenzmenge** von M ist die Menge aller krausen Teilmengen, geschrieben

$$\mathcal{F}(M) := \{A : A \subseteq M\}.$$

NB: die krause Potenzmenge selbst ist eine knackige Menge! Die **Kardinalität** einer krausen Menge errechnet man mit der einfachen Formel

$$|M| = \textstyle\sum_{x \in X} M(x).$$

Ein wichtiges Konzept, dass krause mit knackigen Mengen verbindet, ist das Konzept des α-Schnittes, wobei $\alpha \in [0, 1]$. Wir definieren den α-Schnitt von M als

(366) $${}^{\alpha}M := \{x : M(x) \geq \alpha\}$$

Es gibt auch eine schärfere Version

(367) $${}^{\alpha+}M := \{x : M(x) > \alpha\}$$

wir nennen das den scharfen Schnitt. Es ist klar dass

$${}^{\alpha_1}M \subseteq^{\alpha_2} M \text{ gdw. } \alpha_2 \leq \alpha_1.$$

Wir haben also, für alle $\alpha \in [0,1]$, eine knackige Menge, die immer kleiner wird indem α größer wird. Zwei Begriffe sind besonders hervorzuheben, nämlich zunächst der **Kern** von M, definiert als

(368) $K(M) := {}^{1}M$

und der **Umfang** von M, definiert als

(369) $U(M) := {}^{0+}M$

.Ein weiterer wichtiger Begriff ist die **Höhe** von M,

(370) $H(M) := max_{x \in X} M(x)$

und der Wertebereich von M,

$$\{\alpha \in [0,1] : \exists x \in X : M(x) = \alpha\}.$$

Diese Konzepte sind offensichtlich; was etwas weniger offensichtlich ist folgendes: jede Krause Menge wird eindeutig charakterisiert durch die Menge ihrer α-Schnitte, denn wir haben

(371) $M(x) = sup\{\alpha \cdot^{\alpha} M(x) : \alpha \in [0,1]\}$

Diese Formel sollte klar sein, wenn wir bedenken dass ${}^{\alpha}M(x) \in \{0,1\}$; das heißt wir wählen in der obigen Formel einfach das größte α so dass ${}^{\alpha}M(x) = 1$. Wir benutzen übrigens das Supremum, nicht das Maximum, und so bekommen wir

(372) $sup\{\alpha \cdot^{\alpha} M(x) : \alpha \in [0,1]\} = sup\{\alpha \cdot^{\alpha+} M(x) : \alpha \in [0,1]\}$

Wir können also eine krause Menge M eindeutig darstellen als eine knackige Menge von Paaren

$$\{\langle M_\alpha, \alpha\rangle\}, \text{ wobei } \alpha \in [0,1]$$

wobei auch jedes A_α knackig ist und eben den (scharfen) α-Schnitt von M darstellt.

Das bringt uns zu dem zentralen Konzept der **Schnittwürdigkeit**. Wir haben gesagt dass krause Mengen knackige Mengen generalisieren; wir stehen

also oft vor der Frage: gegeben eine Operation (wie Schnitt), Eigenschaft (wie Konvexität im Hinblick auf R) oder Relation (wie Teilmenge) auf knackigen Mengen, wie sollen wir diese Operation auf beliebige krause Mengen erweitern? Oft gibt es eine Vielzahl von Möglichkeiten hierfür, aber es gibt nur wenige (oft nur eine) davon, die Schnittwürdig ist.

Definition 23 *Sei F eine n-äre Operation, P eine Eigenschaft, R eine m-äre Relation auf knackigen Mengen, und F', P', R' jeweils ihre Erweiterung auf krause Mengen.*

1. *F' ist schnittwürdig, falls f.a. krause Mengen $M_1, ..., M_n$, $\alpha \in [0,1]$ gilt: ${}^{\alpha}F'(M_1, ..., M_n) = F({}^{\alpha}M_1, ..., {}^{\alpha}M_n)$.*

2. *P' ist schnittwürdig, falls f.a. krause Mengen M, $\alpha \in [0,1]$ gilt: M hat Eigenschaft P', genau dann wenn ${}^{\alpha}M$ Eigenschaft P hat.*

3. *R' ist schnittwürdig, falls f.a. krause Mengen $M_1, ..., M_m$, $\alpha \in [0,1]$ gilt: $R'(M_1, ..., M_m)$ gilt genau dann wenn $R({}^{\alpha}M_1, ..., {}^{\alpha}M_m)$.*

Die Schnittwürdigkeit ist der Ritterschlag für eine krause Operation/Eigenschaft/Relation. Z.B. die krause Teilmengenrelation, die wir oben beschrieben haben, ist schnittwürdig, wie man leicht sehen kann. Ein Beispiel für eine schnittwürdige Eigenschaft ist die krause Konvexität, die wir hier für krause Mengen von reellen Zahlen definieren: M ist konvex, wenn gilt:

falls $x \leq y \leq z$, dann liegt $M(y)$ zwischen (oder auf) $M(x)$ und $M(z)$.

Der folgende Satz ist fundamental, und ergibt sich aus GLeichung (366) und den dazgehörigen Erwägungen:

Theorem 24 *Jede Operation, Eigenschaft, Relation der klassischen Mengenlehre hat höchstens eine krause Erweiterung, die schnittwürdig ist.*

Das folgt, da die Eigenschaften in Definition (nach (1)) die krause Erweiterung eindeutig bestimmen. Wir können diese Erweiterung auch **kanonisch** nennen. Zwei Anmerkungen sind wichtig:

1. Im Lemma steht *höchstens*; der Grund dafür ist: es gibt nicht immer eine sinnvolle krause Erweiterung von krausen Relationen. Z.B. sehe ich nicht, wie man die Relation $\in$ sinnvoll auf krause Mengen erweitern soll.

2. Schnittwürdigkeit ist wichtig, aber nicht entscheidend: schnittwürdige Erweiterungen sind, wie man sieht, eindeutig durch klassische Mengenlehre bestimmt, und damit zwar kanonisch, aber auch manchmal etwas "langweilig": wenn wir nur mit schnittwürdigen Erweiterungen arbeiten würden, wir würden das beste verpassen; denn es sind die nicht schnittwürdigen Eigenschaften, in denen krause Mengen ihre genuinen Eigenschaften entfalten.

Gleichzeitig liefert uns Theorem 24 und Gleichung (366) eine Methode, um Operationen, Relationen etc. über knackigen Mengen effektiv auf krause Mengen zu erweitern. Wir sagen daher auch: wir kräuseln die Operation.

Eine wichtige Technik, die damit zusammenhängt, ist die **Kräuselung von Funktionen**. Seien M, N knackige Mengen, und $f : M \to N$ eine Funktion. Die Funktion wird gekräuselt zu einer Funktion

$$F : \mathcal{F}(M) \to \mathcal{F}(N),$$

also einer Funktion der krausen Potenzmengen. Die Funktion F wird bestimmt durch das **Erweiterungsprinzip**:

$$F(A) = B \iff \text{f.a. } y \in N,\ B(y) = max_{y=f(x)} A(x) \tag{373}$$

D.h. $B(y)$ wird bestimmt durch den maximalen Wert $A(x)$, vorausgesetzt dass $f(x) = y$. Dadurch ist die Funktion F *eindeutig* bestimmt. Die inverse Funktion F^{-1} ist bestimmt durch

$$(F^{-1}(N))(x) = N(y),\ \text{wobei } y = f(x) \tag{374}$$

Damit haben wir $F^{-1}(F(M)) \supseteq M$, und falls f eine Bijektion ist, haben wir $F^{-1}(F(M)) = M$.

33.4 Modifikatoren

Wir kommen nun zu den sog. **Modifikatoren**. Wenn eine krause Menge eine Eigenschaft wie "groß" beschreibt, dann beschreibt ein Modifikator ein Adverb wie "sehr" – und damit meine ich, dass die jetzige Analogie problematisch ist wie die vorige, sich aber (unter Nicht-Linguisten) eingebürgert hat. Ein Modifikator ist – nach üblicher Verwendung – eine monotone, stetige Funkton

$$m : [0,1] \to [0,1].$$

Monoton bedeutet dass falls $x \leq y$, dann $m(x) \leq y$, stetig bedeutet, dass minimale Veränderungen des Arguments minimale Veränderungen des Wertes implizieren. Sei m ein Modifikator; er modifiziert eine krause Menge M zu einer krause Menge $m(M)$ durch folgende Gleichung:

(375) $m(M)(x) = m(M(x))$

Der Modifikator interessiert sich also nicht für x selber, sondern nur für den Wert $M(x)$. Im Hinblick auf unsere späteren logischen Konzepte könnten wir sagen: Modifikatoren sind "Wahrheitsfunktional"; sie interessieren sich nur für die Wahrheitwerte, nicht für die Proposition selber. Das wird für die meisten krausen Operatoren gelten, denn es ist meist der einzige weg, sie allgemein zu definieren. Eine sehr offensichtliche Familie von Modifikatoren lässt sich wie folgt beschreiben:

(376) $m_\lambda(x) = x^\lambda$,

wobei $\lambda \in \mathbb{R}^+$ ein beliebiger Parameter aus den (streng) positiven reellen Zahlen ist. Es ist leicht zu prüfen dass für alle $\lambda \in \mathbb{R}^+$ m_λ ein Modifikator im obigen Sinne ist. Der Fall

(377) $\lambda = 0$

ist natürlich pathologisch denn wir haben $x^0 = 1$ f.a. $x \in \mathbb{R}$; deswegen kann er durchaus Probleme bereiten (wenn wir mit universellen Klassen arbeiten).

33.5 Komplemente

Knackige Komplemente sind definiert für Mengen i.H.a. eine Referenzmenge: sei X die Referenzmenge, M eine (knackige) Menge; dann ist das Komplement von A

(378) $\overline{A} := \{x : x \in X, x \notin A\}$

Krause Komplemente sind eine Verallgemeinerung davon, und sind *nicht eindeutig*, d.h. es gibt viele Arten, ein krauses Komplement zu bilden. Ein krauses Komplement ist ebenfalls eine Funktion

$$c : [0,1] \to [0,1],$$

die zusätzlich folgende Bedingungen erfüllt:

1. $c(1) = 0$,
2. $c(0) = 1$,
3. für $x, y \in [0,1]$, falls $x \leq y$, dann $c(y) \leq c(x)$,
4. (fakultativ) c ist eine stetige Funktion,
5. (fakultativ) f.a. $x \in [0,1]$ gilte: $c(c(x)) = x$

Das Komplement von M $c(M)$ wird wiederum definiert durch

$$c(M)(a) = c(M(a)).$$

Bedingungen 1 und 2 Stellen sicher dass c das klassische, knackige Komplement generalisiert; 3 verlangt dass es es die krause Teilmengenrelation umkehrt: wenn a stärker zu M gehört als b, dann gilt für $c(M)$ das Gegenteil. Bedingung 4 ist klar; Bedingung 5 ist die sogenannte *Involutivität* (doppelte Negation hebt sich auf), eine wichtige logische Eigenschaft, die aber normalerweise nicht notwendig ist.

Wie gesagt gibt es für krause Mengen viele Komplemente, da es viele Funktionen gibt, die Bedingung 1-5 erfüllen. Eine Familie von Komplementen, die besonders leicht zugänglich ist, sind die sog. *Yager*-Komplemente:

(379) $c_\lambda(x) = (1 - x^\lambda))^{1/\lambda}$,

wobei $\lambda \in \mathbb{R}^+$ (diesmal streng positiv). Hier bekommen wir für jedes λ ein Komplement. Ein besonderer Fall ist der Fall $\lambda = 1$; dann reduziert sich die Gleichung zu

(380) $c(x) = 1 - x$

Das liefert uns das sog. kanonische krause Komplement, und wir schreiben $c_1(M) \equiv \overline{M}$, wobei $\overline{M}(x) = 1 - M(x)$.

Lemma 25 *c_1 ist nicht schnittwürdig.*

Beweis. Nehmen wir $\alpha = 0.3$, $M(a) = 0.6$. Dann ist $\overline{M}(a) = 0.4$, also $a \in^{\alpha} \overline{M}$. Allerdings ist $a \in^{\alpha} M$, also $a \notin {}^{\alpha}\overline{M}$, also ${}^{\alpha}\overline{M} \neq {}^{\alpha}\overline{M}$. ⊣

Allgemeiner funktioniert diese Konstruktion immer wenn $M(a) > 0.5$ und $M(a) + \alpha \leq 1$. Also ist nicht einmal das kanonische krause Komplement schnittwürdig, und ich bin mir nicht sicher wie und ob man ein schnittwürdiges Komplement konstruieren kann.

33.6 Schnitt

Schnitt und Vereinigung werden genauso behandelt wie die anderen Operationen bisher; der einzige Unterschied ist, dass wir hier binäre Operationen haben an Stelle unärer. Wir haben deswegen Funktionen

$$s, v : [0,1] \times [0,1] \to [0,1],$$

und definieren dann

$$M \cup_v N(a) = v(M(a), N(a)) \tag{381}$$

$$M \cap_s N(a) = s(M(a), N(a)) \tag{382}$$

Wiederum gilt, dass diese Operationen nicht eindeutig sind, sie muüssen aber bestimmte Bedingungen erfüllen. Die Funktion s muss eine sog. **Dreiecksnorm** (t-Norm) sein, die Funktion v muss eine **Dreieckskonorm** (t-Konorm) sein. Die Bedingungen sind wie folgt:

Definition 26 *$s[0,1] \times [0,1] \to [0,1]$ ist eine t-Norm, falls sie folgende Bedingungen erfüllt:*

1. *$s(x,1) = x$; (1 neutral)*
2. *$s(x,y) = s(y,x)$; (kommutitativ)*
3. *aus $x \leq y$ folgt $s(z,x) \leq s(z,y)$ (monoton);*
4. *$s(x,s(y,z)) = s(s(x,y),z)$ (assozativ).*

Was man vielleicht vermisst (um sicherzustellen dass s den klassischen Schnitt generalisiert) ist $s(x,0) = 0$. Das lässt sich aber wie folgt ableiten: wir haben $s(0,1) = 0$, und da $x \leq 1$, folgt

$$s(x,0) = s(0,x) \leq s(0,1) = 0 \tag{383}$$

Ebenso gilt:

(384) $s(x,x) \leq s(x,1) = x$

Die umgekehrte Ungleichgung können wir aber nicht aus den obigen Prinzipien ableiten! Es gibt folgende bekannte t-Normen:

- Kanonischer krauser Schnitt: $s(x,y) = min(x,y)$ (Gödel-Norm);
- Produkt: $s(x,y) = x \cdot y$;
- beschränkte Differenz: $s(x,y) = max(0, x+y-1)$ (Lukasiewicz-Norm);
- drastischer Schnitt: $s_{min}(x,y) = \begin{cases} x, \text{ falls } y = 1 \\ y, \text{ falls } x = 1 \\ 0 \text{ andernfalls} \end{cases}$

Insbesondere gilt, f.a. t-Normen i, $x, y \in [0,1]$,

(385) $i_{min}(x,y) \leq i(x,y) \leq min(x,y)$

Wir haben also minimale und maximale t-Normen, und alle anderen liegen dazwischen. Eine Methode, t-Normen zu konstruieren, die den Zwischenraum ausfüllen, sind die sog. Yager-Schnitte. Das ist wiederum eine Familie von Schnitten i_λ, die definiert ist durch

(386) $i_\lambda(x,y) = 1 - min(1, [(1-x)^\lambda + (1-y)^\lambda]^{1-\lambda})$

wobei $\lambda \in \mathbb{R}^+$. Es ist nicht leicht zu sehen dass i_λ, für $\lambda \to \infty$, gegen min konvergiert. min ist nicht nur die maximale t-Norm, es ist auch die einzige krause Schnittmengenbildung, die schnittwürdig ist; um Mißverständnisse zu vermeiden, nennen wir diese Norm ks

Lemma 27 *ks ist schnittwürdig: f.a. $\alpha \in [0,1]$ gilt: ${}^\alpha(M \cap_{ks} M) =^\alpha M \cap^\alpha N$.*

Beweis. ${}^\alpha(M \cap_{ks} M) \subseteq^\alpha M \cap^\alpha N$: sei $a \in^\alpha (M \cap_{ks} M)$; dann ist $min(M(a), N(a)) \geq \alpha$, also $M(a) \geq \alpha, N(a) \geq \alpha$, also $a \in^\alpha M \cap^\alpha N$.

${}^\alpha M \cap^\alpha N \subseteq^\alpha (M \cap_{ks} M)$: sei $a \in^\alpha (M \cap_{ks} M)$, dann ist $a \in^\alpha M, a \in^\alpha N$, also $M(a) \geq \alpha, N(a) \geq \alpha$, also $M \cap_{ks} N(a) \geq \alpha$. ⊣

ks hat also eine besondere Bedeutung. ks hat noch eine weitere Eigenschaft hat: es ist die einzige t-Norm die **idempotent** ist, also

$$s(x,x) = x$$

erfüllt. Ich habe aber keinen Beweis dafür. t-Normen sind nicht nur wichtig für krause Mengen, sondern für krause Logiken: wir werden sie benutzen, um unsere Konjunktion zu interpretieren.

33.7 Vereinigung

Vereinigung ist eine duale Operation zu Schnitt; wir benutzen hier statt t-Normen sog. t-Konormen:

Definition 28 *$v[0,1] \times [0,1] \to [0,1]$ ist eine t-Konorm, falls sie folgende Bedingungen erfüllt:*

1. *$v(x,0) = x$; (0 neutral)*
2. *$v(x,y) = s(y,x)$; (kommutitativ)*
3. *aus $x \leq y$ folgt $v(z,x) \leq v(z,y)$ (monoton);*
4. *$v(x,v(y,z)) = v(v(x,y),z)$ (assozativ).*

Aus diesen Regeln können wir – nach demselben Muster wie oben – ableiten, dass $v(x,1) = 1$; umgekehrt haben wir $v(x,x) \geq x$, aber nicht die umgekehrte Ungleichung. Wichtige t-Konormen sind

1. kanonische Vereinigung: $kv(x,y) = max(x,y)$;
2. algebraische Summe: $v(x,y) = x + y - xy$ (vgl. Wahrscheinlichkeiten!)
3. begrenzte Summe: $v(x,y) = min(1, x+y)$;
4. drastische Vereinigung: $v_{max}(x,y) = \begin{cases} x, \text{ falls } y = 0 \\ y, \text{ falls } x = 0 \\ 1 \text{ andernfalls} \end{cases}$

Wiederum kann man recht leicht verifizieren dass f.a. t-Konormen v gilt:

(387) $kv(x,y) \leq v(x,y) \leq v_{max}(x,y)$;

wir haben also eine minimale und eine maximale Vereinigung. Widerum ist kv idempotent, $kv(x,x) = x$, und die einzige idempotente t-Konorm (nach dem was ich lese). Außerdem ist kv die (einzige) schnittürdige krause Vereinigung; der Beweis läuft wie oben ab. Eine Familie von t-Konormen, mit denen man die Lücke zwischen minimalen und maximalen Konormen füllen kann, sind die Yager-Vereinigungen, die wie folgt definiert sind:

(388) $v_\lambda(x,y) = min(1, x^\lambda y^\lambda)^{1/\lambda})$

wobei $\lambda \in \mathbb{R}^+$. Wiederum konvergiert diese Operation auf kv für $\lambda \to \infty$. Eine weitere Anmerkung sollte ich machen: es ist bekannt, dass klassischer Schnitt, Vereinigung und Komplement durch eine Reihe von Gesetzen miteinander verbunden sind (das wird oft unter Booleschen Algebren behandelt). Diese Gesetze gelten *nicht* für die krausen Erweiterungen, und zwar in keinem Fall. Man kann zwar unter gewissen Umständen manche Gesetze erfüllen, aber niemals alle.

33.8 Allgemeine Logik

Ich gebe zunächst einen Überblick über die Vorraussatzungen, die wir an klassischer Logik brauchen. Die ist weder vollständig noch selbsterklärend, kann aber in jeder Einführung in die Logik nachgelesen werden. Wir haben eine (induktiv definierte) Menge von Formeln, über eine abzählbare Menge von Variablen *Var* und die Konnektoren

$$\{\neg, \wedge, \vee, \to, \bot, \top\}.$$

Wir nennen die resultierende Menge *Form*(*Var*). Eine **Valuation** ist eine Funktion

$$v : Var \to \{0,1\}.$$

Jeder unserer Konnektoren der Stelligkeit n wird interpretiert als Funktion von

$$\{0,1\}^n \to \{0,1\},$$

und auf diese Art erweitern wir v zu einer Funktion

$$Form(\, Var) \to \{0,1\}.$$

Sei Γ eine Menge von Formeln, ϕ eine Formel. Wir schreiben

$$\Gamma \models \phi,$$

falls gilt: f.a. v, falls $v(\gamma) = 1$ f.a. $\gamma \in \Gamma$, dann ist $v(\phi) = 1$. Das ist die **semantische Konsequenz**.

Wir kommen nun zur syntaktischen Konsequenz, Ableitbarkeit und Beweistheorie. Wir präsentieren nur den sog. **Hilbert-Kalkül**. Hilbert Kalküle sind einfach zu präsentieren, aber sperrig zu benutzen und deswegen unbeliebt. Ich benutze sie auch nicht gern, aber die Krausen Logiken die ich kenne werden allesamt nur im Hilbert Stil präsentiert.

Ein Hilbert Kalkül besteht normalerweise aus einer Menge von Axiomen. Diese Menge ist normalerweise überschaubar und wird endlich präsentiert; die Axiome für klassische Logik sind

(c1) $\phi \to (\psi \to \phi)$
(c2) $(\phi \to (\psi \to \chi)) \to ((\phi \to \psi) \to (\phi \to \chi))$
(c3) $(\neg\phi \to \neg\psi) \to (\psi \to \phi)$

Das sind zwar nur drei, aber man muss im Kopf behalten dass damit soz. unendlich viele Formeln repräsentiert werden: die griechischen Buchstaben sind sog. Metavariablen, für die wir beliebige Formeln substitutieren können. Die Axiome bleiben gültig, sofern wir die Substitution *einheitlich* machen: gleiche Metavariablen werden gleich substituiert. Zusätzlich zu den Axiomen gibt es **Inferenzregeln**, im propositionalen Hilbert-Kalkül nur eine, nämlich Modus Ponens:

$$\text{(MP)}\ \frac{\phi \to \psi \quad \phi}{\psi}$$

Das war es auch schon; was wir noch brauchen ist der Begriff des Beweises: ein Hilberbeweis von einer Formel ϕ ist eine endliche Folge

$$\langle \psi_1, ..., \psi_i \rangle,$$

so dass

1. $\phi = \psi_i$,
2. f.a. $j \leq i$ gilt: ϕ_j ist entweder die Instanz eines Axioms, oder es gibt $j', j'' < j$, so dass ϕ_j abgeleitet werden kann aus $\phi_{j'}, \phi_{j''}$ mit Modus Ponens.

Wir schreiben $\vdash_H \phi$, falls es einen Hilbert Beweis von ϕ gibt. Die folgende Eigenschaft nennt man (schwache) **Vollständigkeit**:

Lemma 29 $\vdash_H \phi$ *gdw.* $v(\phi) = 1$ *f.a. Valuationen* v.

Wir können den Begriff erweitern: wir sagen $\langle \psi_1, ..., \psi_i \rangle$ ist ein Beweis von $\Gamma \vdash_H \phi$, falls gilt: jedes ϕ_j ist ein Axiom, ableitbar aus Vorgängern, oder in Γ. Wir nehmen uns also zusätzliche "Axiome". Das sind aber keine abstrakten Axiome über Metavariablen, sondern konkrete Instanzen im Normalfall. Das nächste Theorem, die starke Vollständigkeit, ist folgendes:

Theorem 30 $\Gamma \models \phi$ *gdw.* $\Gamma \vdash_H \phi$.

Das nennt man die Vollständigkeit des Beweiskalküls.
Ein weiterer wichtiger Satz ist das sog. Deduktionstheorem:

Theorem 31 $\Gamma \cup \{\phi\} \vdash_H \psi$ *gdw.* $\Gamma \vdash_h \phi \to \psi$.

Das sagt uns dass die metalogische Relation $\vdash_H$ genau dem logischen Konnektor $\to$ entspricht. Das müsste für den Anfang reichen.

33.9 Krause Logik - im engeren Sinn

Die Ausführungen über klassische Logik waren in erster Hinsicht dazu gedacht, die Zielsetzung für unsere Logik zu verdeutlichen. Die Ziele sind:

1. eine Semantik zu konstruieren
2. einen Beweiskalkül zu finden, der Vollständigkeit im Hinblick auf diese Semantik liefert, und
3. prüfen, inwieweit klassische metalogische Resultate noch Gültigkeit haben.

Krause Logik ist – wie klassische Logik – wahrheitsfunktional. Damit ist klar, wie sich Valuationen und Konnektoren verhalten: Valuationen bilden *Var* nach $[0, 1]$ ab, und n-äre Konnektoren werden interpretiert also Funktionen

$$[0, 1]^n \to [0, 1].$$

Die Frage ist nun: wie genau? Der normale Ansatz (bzw. der Ansatz von Hajek) ist durchaus willkürlich: wir beginnen mit der Konjunktion und definieren daraus alle anderen Konnektoren. Und zwar interpretieren wir die Konjunktion $\wedge$ als eine t-norm $*$ (welche t-Norm ist eine andere Frage). Wir haben also

$$v(\phi \wedge \psi) = v(\phi) * v(\psi).$$

Das ist natürlich zuwenig, um die anderen Konnektoren zu definieren. Wir können auch nicht einfach t-Konormen und Komplemente nehmen, da diese eine Reihe von wichtigen Bedingungen nicht erfüllen. Stattdessen gehen wir folgenden weg: wir benutzen die Konjunktion um die Implikation zu definieren, und definieren dann die üblichen Konnektoren.

Wie kommt man von der Konjunktion zur Implikation? Der entscheidende Begriff ist algebraischer Natur, nämlich der des **Residuums**. Residua generalisieren nicht nur klassische Implikaiton, sondern praktisch alle geläufigen Implikationen, bis zu dem Punkt dass ein Konnektor nur dann eine Implikation darstellt, wenn er eine Form von Residuum ist. Die Motivation ist folgende. Wir schreiben nun $\models$ in einem generalisierten Sinne:

$$\psi \models \phi \text{ gdw. f.a. } v \text{ gilt: } v(\psi) \leq v(\phi);$$

das bedeutet zunächst soviel wie: "ϕ ist wahrer als ψ"; es bedeutet aber darüberhinaus: wann imm ϕ einen Wahrheitsgrad hat, hat ψ einen größeren. Man kann es also auch lesen als: aus ϕ folgt ψ. Nimm nun an, wir haben

$$\phi \wedge \psi \models \chi.$$

Dann sollte gelten:

$$\phi \models \psi \rightarrow \chi.$$

Der Grund dafür ist folgender: wenn aus ϕ und ψ χ folgt, dann folgt aus ϕ dass wenn ψ gilt, dann gilt auch χ; das ist unsere elementare Intuition über die Implikation, die ja das innerlogische Gegenstück zur (meta)logischen Konsequenz ist.

Interessanterweise sollte auch das duale Gegenstück gelten: falls $\phi \models \psi \rightarrow \chi$, dann sollte auch gelten dass $\phi \wedge \psi \models \chi$: denn wenn aus ϕ folgt dass $\psi \rightarrow \chi$, dann sollte aus ϕ und ψ auch χ folgen; das ist unsere elementare Intuition über Transitivität logischer Schlüsse. Was wir also haben ist:

$$\text{(Res) } \phi \models \psi \to \chi \text{ gdw. } \phi \wedge \psi \models \chi.$$

Das ist (für kommutatives $\wedge$) das sog. Gesetz der Residua, dass sämtliche mir bekannten Implikationen erfüllen. Da wir das haben, können wir nun anfangen, die semantischen Konsequenzen dieses Gesetzes zu betrachten. Wir lesen

- '$\wedge'$ als '$*'$,
- ' $\models'$ als ' $\leq'$;
- die semantische Übersetzung von ' $\to'$ nennen wir ' $\Rightarrow'$, ohne zu wissen was genau sie bedeutet – aber sie wird definiert über (Res).

Dann bekommen wir:

$$v(\phi) \leq (v(\psi) \Rightarrow v(\chi)) \text{ gdw. } v(\phi) * v(\psi) \leq v(\chi).$$

Da es sich hier um Funktionen auf Werten in $[0,1]$ handelt, ist klar, dass

$$\Rightarrow\colon [0,1]^2 \to [0,1]$$

auch eine solche Funktion sein muss. Was viell. weniger klar ist, ist dass $\Rightarrow$ durch das Gesetz der Residua bereits eindeutig bestimmt ist durch $*$:

$$x \Rightarrow y = max\{z : x * z \leq y\}.$$

Damit dieses Maximum existiert, muss $*$ stetig sein; damit lässt sich eindeutige Existenz einfach zeigen. Wir haben also nun die Implikation und Konjunktion. Das reicht aber immer noch nicht, um alle Konnektoren zu definieren (im klassischen Sinn). Was wir uns noch dazu nehmen ist die Konstante 0, mit

$$v(0) = 0 \text{ (alternativ: } 0 : [0,1]^0 \to \{0\}).$$

Wir definieren

(389) $\neg\phi := \phi \to 0$

Das bedeutet – zur Illustration:

(390) $v(\neg\phi) = v(\phi \to 0) = max\{x : v(\phi) * x \leq 0\}$

Für min, die maximale t-Norm, folgt also

$$v(\neg(\phi)) = 0 \text{ gdw. } v(\phi) < 1.$$

Wenn wir z.B. den drastischen Schnitt, die minimale t-Norm, nehmen, folgt

(391) $v(\neg\phi) = max\{x : v(\phi) * x \leq 0\} = max\{x : x < 1\}$,

was natürlich nicht existiert. Hier sehen wir, dass wir stetige t-Normen brauchen. Da die drastische t-Norm wegfällt, haben wir noch 3 Interpretationen von $*$:

1. $* = min$,
2. $* = \cdot$, und
3. $* = max(0, x + y - 1)$.

Es ist leicht zu sehen dass alle diese Funktionen stetig sind. Wir schauen uns nun an, wie die Implikationen hierfür definiert sind. Wichtig ist zu beachten: wenn wir max schreiben, dann meinen wir den maximalen Wert in $[0, 1]$, denn andere Werte stehen uns gar nicht zur Verfügung.

1. $x \Rightarrow y := max\{z : min(x, z) \leq y\} = \begin{cases} 1, \text{ falls } x \leq y; \\ y \text{ andernfalls.} \end{cases}$
2. $x \Rightarrow y := max\{z : x \cdot z \leq y\} = \begin{cases} 1, \text{ falls } x \leq y; \\ \frac{y}{x} \text{ andernfalls.} \end{cases}$
3. $x \Rightarrow y := max\{0, z : x + z - 1 \leq y\} = \begin{cases} 1, \text{ falls } x \leq y; \\ (1 - x) + y \text{ andernfalls.} \end{cases}$

Wir nennen die Implikation 1. die Gödel, 3. die Lukasiewicz Implikation. Lukasiewicz Implikation muss man kurz erklären: falls $x \leq y$, dann ist $x + 1 - 1 \leq y$; andernfalls haben wir $x + z - 1 = y$ gdw. $z = 1 + y - x$. In der Produktnorm haben wir

$$v(\neg x) = x \Rightarrow 0 = \frac{0}{x} = 0 \text{ falls } x > 0;$$

ansonsten definieren wir

(392) $0 \Rightarrow 0 := 1$

In der Lukasiewicz-Norm haben wir

(393) $v(\neg x) = x \Rightarrow 0 = 1 - x$,

also das kanonische Komplement. Wichtig ist folgendes: wir haben für alle Implikationen

$$\text{falls } x \leq y, \text{ dann } x \Rightarrow y = 1.$$

Das ist kein Zufall: $\leq$ entspricht $\models$ entspricht logischer Konsequenz; also heißt das:

$$x \leq y \text{ gdw. “aus } x \text{ folgt } y\text{” wahr ist gdw. } x \Rightarrow y = 1.$$

Natürlich kann $v(\phi) \leq v(\chi)$ auch rein zufällig sein für ein gewisses v; uns interessiert aber immer was gilt *für alle* Valuationen v.

Was wir nun haben ist eine Semantik für krause Logiken. Diese Semantik wird also eindeutig festgelegt durch die Auswahl einer bestimmten t-Norm. Wir haben also theoretisch genausoviele krause Logiken wie stetige t-Normen – fast: denn isomorphe t-Normen liefern identische Semantiken!

Definition 32 *Eine t-Norm $*_1$ ist isomorph zu $*_2$, falls es eine bijektive Funktion $i : [0,1] \to [0,1]$, so dass $i(x *_1 y) = i(x) *_2 i(y)$.*

Es gibt also genau soviele krause Logiken, wie stetige t-Normen bis auf Isomorphie – nämlich 3:

Theorem 33 *Für jede stetige t-Norm t gilt: t ist isomorph entweder zur Gödel t-Norm, Lukasiewicz t-Norm, oder Produkt t-Norm.*

Der Beweis ist durchaus jenseits dessen, was ich hier behandeln will, wird aber ausführlich dargestellt in Hajeks Buch.

Es gibt also 4 krause Logiken:

1. Gödels Logik, mit $* := min$;
2. Produkt Logik, mit $* := \cdot$;
3. Lukasiewicz's Logik, mit $* := max(0, a + b - 1)$
4. Hajeks Logik (*basic logic*, *BL*), mit beliebigen t-Normen.

Jede dieser Logiken ist vollständig für die oben beschriebene Semantik und der entsprechenden Definition von $*$; Hajeks Logik ist vollständig für *jede* dieser Definitionen. D.h. jede der drei erstgenannten Logiken ist eine Erweiterung von Hajeks Logik. Wir werden daher mit *BL* anfangen.

33.10 Hajeks Logik

Eine Konvention in der abstrakten Logik ist folgende: wir identifizieren eine Logik mit der Menge ihrer Theoreme (oder Tautologien, falls wir semantisch denken). Logiken sind also schlichtweg Mengen. In diesem Sinne können wir sagen: Logik $\mathcal{L}_1$ ist *größer* als Logik $\mathcal{L}_2$; das bedeutet: jedes Theorem von $\mathcal{L}_2$ ist ein Theorem von $\mathcal{L}_1$. Das setzt natürlich voraus, das wir entweder dieselben Konnektoren haben in $\mathcal{L}_1, \mathcal{L}_2$, oder das wir zusätzlich eine **Einbettung** angeben, mit der jeder Konnektor von $\mathcal{L}_2$ in eine $\mathcal{L}_1$-Formel übersetzt wird (diese Einbettung muss natürlich injektiv sein). Z.B. ist klassische Logik *maximal*, d.h. es gibt keine größere Logik mit denselben Konnektoren, die nicht trivial ist. Von allen krausen Logiken ist Hajeks Logik *minimal*, denn egal an welcher t-Norm wir sie interpretieren, unsere Theorem sind immer Tautologien; und umgekehrt ist jede Formel, die unter jeder Evaluation in einer t-Norm wahr ist, ein Theorem von Hajeks Logik. Deswegen heißt sie auch **basic logic**, oder kurz BL.

33.11 Syntax und Semantik von BL

BL hat drei "primitive" Konnektoren; das bedeutet, aus diesen dreien werden alle anderen definiert, und nur für diese drei brauchen wir eine Semantik. Die Konnektoren sind

$$\&, \to, \bot;$$

sie werden interpretiert als

- $*$ (eine beliebige stetige t-Norm),
- $\Rightarrow$ (das dazugehörige Residuum),
- und 0, die arithetmetische 0.

Wir haben

- $Var = \{p_0, p_1, ...\}$,
- und falls ϕ, χ wohlgeformt sind, sind es auch $\phi\&\chi, \phi \to \chi, \bot$.

Es gibt aber eine Reihe anderer Konnektoren, die durch diese definiert werden können:

– $\neg\phi \equiv \phi \to \bot$

– $\phi \wedge \chi \equiv \phi \& (\phi \to \chi)$

– $\phi \vee \chi \equiv ((\phi \to \chi) \to \chi) \wedge ((\chi \to \phi) \to \phi)$

– $\phi \leftrightarrow \chi \equiv (\phi \to \chi) \& (\chi \to \phi)$

Was wir haben auf der semantischen Seite ist:

$$v : Var \to [0,1];$$

wir erweitern das zu $\overline{v}$ wie folgt:

– $\overline{v}(\phi \& \chi) = \overline{v}(\phi) * \overline{v}(\chi)$

– $\overline{v}(\phi \to \chi) = \overline{v}(\phi) \Rightarrow \overline{v}(\chi)$.

– $\overline{v}(\bot) = 0$.

Ein erstes schönes Ergebnis ist folgendes:

Lemma 34 *1.* $\overline{v}(\phi \wedge \chi) = min(\overline{v}(\phi), \overline{v}(\chi))$*;*

2. $\overline{v}(\phi \vee \chi) = max(\overline{v}(\phi), \overline{v}(\chi))$.

Wir beweisen nur die erste Behauptung.
Beweis. 1. Wir zeigen dass $x * (x \Rightarrow y) = min(x, y)$. Fall i: $x \leq y$; dann ist $x \Leftarrow y = 1$; also ist $x * (x \Rightarrow y) = x = min(x, y)$. Fall ii: falls $y < x$. Dann gibt es ein $z \in [0, 1]$, s.d. $z * x = y$, denn für $z = 0$ haben wir $z * x = 0$, und für $z = 1$ haben wir $z * x = x$, und $y \in [0, x)$ und $*$ ist stetig. Für das maximale $z : z * x = y$ haben wir $z = x \to y$, also ist $x * (x \to y) = y = min(x, y)$. $\dashv$

Definition 35 *Wir sagen ϕ ist eine BL-Tautologie, falls f.a. $v : Var \to [0,1]$ gilt: $\overline{v}(\phi) = 1$.*

Wir axiomatisieren BL wie folgt:

(A1) $(\phi \to \psi) \to ((\psi \to \chi) \to (\phi \to \chi)$

(A2) $(\phi\&\chi) \to \phi$

(A3) $(\phi\&\chi) \to (\chi\&\phi)$

(A4) $\phi\&(\phi \to \chi) \to (\chi\&(\chi \to \phi))$

(A5) $(\phi \to (\psi \to \chi) \leftrightarrow ((\phi\&\psi) \to \chi)$

(A6) $((\phi \to \psi) \to \chi) \to (((\psi \to \phi) \to \chi) \to \chi)$

(A7) $\bot \to \phi$

Natürlich stehen diese 7 wieder für unendlich viele Formeln. NB: die Axiome referieren nur auf die primitiven Konnektoren, denn alle anderen sind ja über sie definiert. Die einzige Ausnahme ist (A5); allerdings ist hier der Doppelpfeil nur eine Abkürzung für zwei Implikationen. Einige Worte zur Erklärung: (A1) ist eine Form des klassischen Dreierschlusses: falls ψ aus ϕ folgt, dann folgt alles, was ψ folgt, auch aus ϕ. (A2) ist klar, (A3) garantiert die Kommutativität von &. (A4),(A5) definieren die Interaktion von $\&, \to$. (A6) ist eine Form des Beweises über Fälle: falls χ aus $\psi \to \phi$ folgt und aus $\phi \to \psi$, dann gilt χ – denn da Formeln in $[0, 1]$ interpretiert werden, muss eines der beiden gelten (lineare Ordnung!).

Die einzige Deduktionsregel ist *modus ponens* (MP); Beweise sind wie für klassische Logik definiert.

Folgendes Ergebnis ist eine fundamentale Voraussetzung dafür, dass unser Kalkül korrekt ist:

Lemma 36 *Alle BL-Axiome sind krause Tautologien, d.h. für jede Interpretation $v : var \to [0, 1]$ und jede stetige t-Norm $*$ haben wir $\overline{v}(\phi) = 1$*

Das ist offensichtlich für (A2),(A3),(A7); für die anderen Axiome muss man etwas arbeiten. Wir lassen den Beweis aus. Eine weitere Eigenschaft ist:

Lemma 37 *Falls $\phi \to \psi$ eine Tautologie ist, ϕ eine Tautologie ist, dann ist ψ eine Tautologie.*

Beweis. Wir haben oben gesehen dass f.a. t-Normen, $x \Rightarrow y = 1$ gdw. $x \leq y$. Da ϕ eine Tautologie ist, ist $\overline{v}(\phi) = 1$; es muss aber $\overline{v}(\phi) \leq \overline{v}(\psi)$ sein, also $\overline{v}(\psi) = 1$. $\dashv$

Was das zeigt ist das unser Kalkül **korrekt** ist: alles, was wir damit beweisen, ist eine Tautologie. Man sagt auch: die Menge der BL-Tautologien sind abgeschlossen unter *modus ponens*.

Das Gegenstück hierzu ist die **Vollständigkeit**: jede Tautologie soll in unserem Kalkül beweisbar sein. Man zeigt das wie folgt: zunächst gibt man BL eine sog. **algebraische Semantik**, die BL-Algebren. Hierfür lässt sich leicht Vollständigkeit beweisen. Dann zeigt man, dass sich jede BL-Algebra als ein Produkt von t-Normen darstellen lässt. Der Beweis ist also eher algebraisch; ich werde ihn hier nicht darstellen.

33.12 Theorien und ihre Anwendung

BL liefert uns zunächst nur Sätze, die allgemein gültig sind (in jeder stetigen t-Norm). Die sind natürlich für die Anwendungen nicht sehr interessant; was aber interessant ist ist die Axiomatik im Rahmen dieser Logik. Wir führen nun den Begriff der **Theorie** ein: eine Theorie T ist eine (endliche) Menge von Formeln. Wir können auch den Begriff des Beweises erweitern: wenn wir

$$\vdash_{BL} \alpha$$

schreiben für: α ist beweisbar in BL, dann meinen wir mit

$$T \vdash_{BL} \alpha$$

dass α in BL *und* den zusätzlichen Annahmen in T beweisbar ist. Z.B.: wenn wir das $p\&q$ in unserer Theorie haben, dann ist in BL p beweisbar, denn

$$\{p\&q\} \vdash_{BL} p.$$

Bevor wir Theorien anwenden können, müssen wir uns noch kurz über propositionale Semantik Gedanken machen. Nehmen wir an, p bedeutet so etwas wie: "es ist kalt". Klassische gesprochen ist das wahr oder falsch; für uns sind die Dinge anders: die Bedeutung von p ist eine Funktion

$$p : \mathbb{R} \to [0, 1].$$

Das ist so zu verstehen: wir messen die Temperatur, und abhängig davon ändert sich der Wahrheitswert von p. Jetzt können wir ein Axiom hinzufügen:

$$p \to q,$$

wobei q die Bedeutung hat: "Heizung läuft". Auch das ist ein numerischer Parameter, dessen Wahrheit in $[0, 1]$ liegt, wobei die Skala natürlich zu definieren ist. Nun bedeutet

$$v \models p \to q$$

nicht, dass sobald $v(p) = 1$, dann $v(q) = 1$, sondern es bedeutet:

$$v(p) \leq v(q)$$

Die Formel ist also erfüllt, wenn wir "mehr heizen, als es kalt ist".

Nun füttern wir unser Kontrollsystem mit Daten zu unseren Variablen, und wann immer $T \vdash_{BL} \alpha$ gilt, soll unser System sicherstellen dass α gilt. Wir müssen natürlich sicherstellen, dass es in seiner Macht liegt.

Was wir dabei natürlich eigentlich benutzen (theoretisch) ist die Relation $\models_{BL}$: wir können unsere Meßdaten als eine Belegung v auffassen; die erweitern wir auf eine Art und Weise so dass f.a. nichtlogischen Axiome ϕ gilt dass

$$\overline{v}(\phi) = 1.$$

Wir möchten also, falls gilt:

$$\overline{v} \models_{BL} \alpha \text{ impliziert } \overline{v} \models_{BL} \beta$$

dann soll unser System Sorge tragen dass β gilt. Der Punkt ist: mit unserem Vollständigkeitsergebnis fallen $\vdash_{BL}$ und $\models_{BL}$ zusammen! Hier gibt es einige Dinge zu beachten (Konsistenz, Kontrolle).

Wenn man noch weiter geht, könnte man verlangen: β soll so manipuliert werden, dass $\overline{v}(\beta)$ den (mindesten) Wert annimmt, den es logisch Annehmen muss. Leider ist das nicht so ohne weiteres möglich, unser Kalkül ist nicht so stark, dass es das leisten könnte: nimm an, wir haben v so dass

(394) $min(\overline{v}(\phi), \overline{v}(\phi \to \chi)) = x$

Es kann dennoch sein dass

(395) $\overline{v}(\chi) < x$

Dazu müssen wir annehmen, dass

(396) $\overline{v}(\phi) < 1$,

außerdem

(397) $\overline{v}(\chi) < \overline{v}(\phi)$

Dann ist

(398) $\overline{v}(\phi) \Rightarrow \overline{v}(\chi) = max\{z : \overline{v}(\phi) * z \leq \overline{v}(\chi)\}$

Falls nun $* = \cdot$, dann ist

(399) $\overline{v}(\chi) < min(\overline{v}(\phi) \Rightarrow \overline{v}(\chi), \overline{v}(\phi))$

Der einzige Fall, wo das nicht gilt, ist tatsächlich, falls $* = min$, also Gödel Logik.

Dann stellt sich die Frage: warum brauchen wir BL, wenn wir ohnehin nur mit den diskreten Werten richtig arbeiten können? Wir können auch mit BL krause Sachverhalte erfassen: insbesondere $\rightarrow$ erlaubt es uns, beliebige Größenrelationen zu beschreiben, denn es gilt

$$\phi \rightarrow \chi \text{ genau dann wenn } \phi \leq \chi.$$

Wir können also folgendes machen: wir legen für $v(p)$ einen bestimmten Wert fest, und nehmen das Axiom $\phi \rightarrow p$. Wann immer

(400) $\overline{v}(\phi) > \overline{v}(p)$,

dann haben wir

(401) $\overline{v}(\phi \rightarrow p) = 0$

und mit diesem Sachverhalt können wir weiter räsonnieren (über Negation etc.).