Statistical Machine Translation
Probability Theory II

Jakub Waszczuk

Heinrich Heine Universität Düsseldorf

Winter Semester 2018/19

Recap

Probability space

A probability space is a triple $(\Omega, \mathfrak{A}, P)$, where:

- Ω is the sample space (the set of possible outcomes)
- $\mathfrak{A} \subseteq \wp(\Omega)$ is the algebra of events
- $P : \mathfrak{A} \to [0, 1]$ is the function assigning probabilities to events

Sum rule

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{1}$$

Conditional probability

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{2}$$

Product rule

$$P(A \cap B) = P(A|B) \cdot P(B) \tag{3}$$

Today

- Random variables
- Independence
- Bayes' theorem

Partition

Special case

$$P(A) = P(A \cap B) + P(A \cap \overline{B}) \tag{4}$$

In general

Let $B = B_1, \ldots, B_n$ be a sequence of mutually disjoint events such that $\bigcup_{i=1}^{n} B_i = \Omega$. We call it a **partition**. Then:

$$P(A) = \sum_{i=1}^{n} P(A \cap B_i) = \sum_{i=1}^{n} P(A|B_i)P(B_i) \tag{5}$$

Throwing coins

Proposition

Suppose you throw a coin n times and that the probability of getting *heads* is h. Then, the probability of throwing heads k times is:

$$p(k) = \binom{n}{k} \times h^k \times (1-h)^{(n-k)} \tag{6}$$

Interpretation

- The probability of a sequence with k heads and $n-k$ tails is $h^k \times (1-h)^{(n-k)}$
- $\binom{n}{k}$ is the number of distinct sequences with k heads and $n-k$ tails

Random Variables

What we assume

Let $(\Omega, \mathfrak{A}, P)$ be a probability space. For simplicity, we assume that it is *discrete*:

- $\mathfrak{A} = \wp(\Omega)$ (all events and outcomes are possible)
- Function $p : \Omega \to [0, 1]$, $p(x) := P(\{x\})$

Definition

A *random variable* is a function $X : \Omega \to \mathbb{R}$ which assignes a real value to every possible outcome.

Example

You roll a die in a casino. If you roll 6, you win 60$. Otherwise, you lose 10$. Is it worth it? Let's formalize this:

- $\Omega = \{1, 2, 3, 4, 5, 6\}$
- Random variable:

$$X(\omega) = \begin{cases} 60 & \text{if } \omega = 6 \\ -10 & \text{otherwise.} \end{cases}$$

Random Variables

Notation

A notation we will see frequently is $P(X = x)$, for a given variable X and its possible value x. What does it mean?

$(X = x)$ denotes the set of outcomes (event) for which the value of the variable X is x:

$$\{\omega : \omega \in \Omega, X(\omega) = x\} \tag{7}$$

Therefore:

$$P(X = x) = P(\{\omega : \omega \in \Omega, X(\omega) = x\}) = \sum_{\omega \in \Omega : X(\omega) = x} p(\omega) = \sum_{\omega \in \Omega} p(\omega)[X(\omega) = x] \tag{8}$$

Expected Value

Definition

The *expected value* (or *expectation*) of X is defined as:

$$\mathbb{E}(X) = \sum_{x \in \mathrm{Img}(X)} x \cdot P(X = x) \tag{9}$$

Equivalently:

$$\mathbb{E}(X) = \sum_{\omega \in \Omega} X(\omega) \cdot p(\omega) \tag{10}$$

Example

Getting back to our casino example; assuming that the die is fair:

$$\mathbb{E}(X) = 60 \cdot \frac{1}{6} + (-10) \cdot \frac{5}{6} = 1\,\frac{2}{3}$$

But if, for example, $p(6) = \frac{1}{8}$:

$$\mathbb{E}(X) = 60 \cdot \frac{1}{8} + (-10) \cdot \frac{7}{8} = -1.25$$

Random Variables

The Law of Large Numbers

Suppose we have a probability space and a corresponding random variable X. Suppose also that we randomly draw a given number of outcomes $\omega_i \in \Omega$ from our space and store the values $X(\omega_i)$ as results.

Then, according to the law of the large numbers, the mean of the obtained results is less likely to deviate from the expected value $\mathbb{E}(X)$ as the number of iterations get larger.

Corollary

In our casino example, if the dice is fair, the player will win, in the long run, $1\frac{2}{3}$\$ per roll.

Or, if $p(6) = \frac{1}{8}$, loose 1.25\$ per roll.

Variance and Standard Deviation

Definition

The *variance* of X measures the extent to which the actual values of the variable differ from the expected one:

$$\mathbb{V}(X) = \mathbb{E}((X - \mathbb{E}(X))^2) = \sum\nolimits_{x \in \mathrm{Img}(X)} (x - \mathbb{E}(X))^2 \cdot P(X = x) \tag{11}$$

The *standard deviation* is defined as:

$$\sigma(X) = \sqrt{\mathbb{V}(X)} \tag{12}$$

Example

Getting back again to the casino and assuming that the die is fair:

$$\mathbb{V}(X) = (60 - 1\frac{2}{3})^2 \cdot \frac{1}{6} + (-10 - 1\frac{2}{3})^2 \cdot \frac{5}{6} \approx 680$$

$$\sigma(X) = \sqrt{680.5555555555} \approx 26$$

Intuitively, the expected gain of the player is therefore equal to $1\frac{2}{3} \pm 26$.

Example

Setup

Let's change the rules of the game:

- As before: if you roll 6, you win 60$; otherwise, you lose 10$.
- The change: you have to roll the die 5 times, no more, no less.

Questions

- What is the expected gain of the player in this new version of the game?
- What is the variance and standard deviation of the gain?

Random Variables

Operators

- $X + Y$
$$(X + Y)(\omega) := X(\omega) + Y(\omega) \tag{13}$$
- XY
$$(XY)(\omega) := X(\omega) \cdot Y(\omega) \tag{14}$$

Calculation toolkit

- if α is a constant, then $\mathbb{E}(\alpha) = \alpha$
- if α is a constant, then $\mathbb{V}(\alpha) = 0$
- $\mathbb{E}(\mathbb{E}(X)) = \mathbb{E}(X)$
- $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$
- if α is a constant, $\mathbb{E}(\alpha \cdot X) = \alpha \cdot \mathbb{E}(X)$
- $\mathbb{E}(XY) = \mathbb{E}(X) \cdot \mathbb{E}(Y)$, but **only if X and Y are independent.** (TRICKY TO PROVE)
- $\mathbb{E}(X - \mathbb{E}X) = 0$
- $\mathbb{V}(X) = \mathbb{E}(X^2) - (E(X))^2$
- $\mathbb{V}(X + Y) = \mathbb{V}(Y) + \mathbb{V}(Y)$, **provided that X and Y are independent.**

Random Variables

Marginalization

- Suppose we have two random variables X, Y over the same probability space.
- Suppose that we also know the **joint** probability distribution of X and Y, that is, we know $P(X = x, Y = y)$ for any two values x, y of the random variables X, Y.
- **Question**: How can we determine $P(X = x)$?

$$P(X = x) = \sum\nolimits_{y \in \mathrm{Img}(Y)} P(X = x, Y = y) \tag{15}$$

This process is called **marginalization**.

Proof (intuition)

- $\{Y = y : y \in \mathrm{Img}(Y)\}$ is a partition of Ω
- Let $A \equiv (X = x)$ and $B_y \equiv (Y = y)$. Then, Eq. 15 follows directly from Eq. 5.

Independence

Definition

We say that two events $A, B \in \mathfrak{A}$ are **independent** if the following holds:

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

This implies that (if $P(B) \neq 0$):

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

Intuitively, knowing B does not tell us anything about A and vice versa.

Definition

Let X, Y be two random variables. We say that X and Y are independent if for each possible value x of X and each possible value y of Y it holds that:

$$P(X = x \cap Y = y) = P(X = x) \cdot P(Y = y) \tag{18}$$

Conditional Independence

Definition

Let $A, B, C \in \mathfrak{A}$ be three events. We say that A and B are **(conditionally) independent** given C if:

$$P(A \cap B|C) = P(A|C) \cdot P(B|C) \tag{19}$$

This implies that (if $P(B|C) \neq 0$):

$$P(A|B \cap C) = P(A|C) \tag{20}$$

Intuitively, knowing B does not tell us anything about A if we already know C. And vice versa, knowing A does not tell us anything about B if we already know C.

Question

Let's $A, B, C \in \mathfrak{A}$ be three events. Suppose we don't know anything about them. Which of the following two assumptions is stronger?

- A and B are independent
- A and B are independent given C

Conditional Independence

Example

Let's consider three events, on any particular day, all occurring in Düsseldorf:

- R – it is raining
- C – somebody has a car accident
- U – Hans takes an umbrella on his way to work

Questions

Which of the following can be simplified/reduced and how?

- $P(C, U)$
- $P(C, R|U)$
- $P(C, U|R)$

You should adopt certain rational assumptions:

- Hans does not take umbrella on his way to work every day
- Hans does not use his umbrella to break the headlights of the cars passing by
- ...

Bayes' theorem

$$P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{P(A \cap B)}{P(A)} = \frac{P(A \cap B)}{P(B)} \cdot \frac{P(B)}{P(A)} = \frac{P(A|B) \cdot P(B)}{P(A)} \tag{21}$$

Bayes' theorem: example

Example

Suppose we know that we have a biased coin, with $h = 0.4$. We throw the coin and get the following sequence:

$$H, T, T, T, H, H, T, H, T, H \tag{22}$$

Thus, instead of getting H the expected 4 times, we got it 5 times.

We can calculate the probability of such an event happening:

$$p(5) = \binom{10}{5} \times 0.4^5 \times 0.6^5 = 0.201$$

Suppose, however, that the coin is not biased. Then we get:

$$p(5) = \binom{10}{5} \times 0.5^5 \times 0.5^5 = 0.236$$

Question

Let's assume that we know that the coin is either biased with $h = 0.4$ or not biased at all ($h = 0.5$). What is the probability of the coin being biased if we throw 5 heads out of 10?

Bayes' theorem: example

Events

- B – the coin is biased with $h = 0.4$
- N – the coin is not biased ($h = 0.5$)
- E – we get heads 5 times in 10 trials

Calculations

Let $\alpha := P(B)$:

$$P(B|E) = P(E|B) \cdot \frac{P(B)}{P(E)} = 0.201 \cdot \frac{\alpha}{P(E)}$$

$$P(E) = P(E \cap B) + P(E \cap N) = P(E|B) \cdot P(B) + P(E|N) \cdot P(N) =$$

$$0.201 \cdot \alpha + 0.236 \cdot (1 - \alpha) = 0.236 - 0.035\alpha$$

Bayes' theorem: example

Events

- B – the coin is biased with $h = 0.4$
- N – the coin is not biased ($h = 0.5$)
- E – we get heads 5 time

Result

$$P(B|E) = 0.201 \cdot \frac{\alpha}{0.236 - 0.035\alpha}$$

Prior

$P(B) = \alpha$ can be seen as a **parameter** representing our **prior** knowlege about the coin.

- if $\alpha = 0.5$, then $P(B|E) = 0.46$
- if $\alpha = 0.6$, then $P(B|E) = 0.56$
- if $\alpha = 0.0$, then $P(B|E) = 0.0$
- if $\alpha = 1.0$, then $P(B|E) = 1.0$

Bayes' theorem

General interpretation

Let α represent model parameters and D the observed event (data!). Then:

$$P(\alpha|D) = \frac{P(D|\alpha) \cdot P(\alpha)}{P(D)} \tag{23}$$

where:

- $P(D|\alpha)$ – the so-called *likelihood*
- $P(D)$ – the probability of D regardless of parameters (we can often ignore it!)
- $P(\alpha)$ – the *prior*

Estimation

- *Maximum likelihood esimates* (MLE):

$$\arg\max_\alpha P(D|\alpha) \tag{24}$$

- *Maximum a-posteriori esimates* (MAP):

$$\arg\max_\alpha P(\alpha|D) \tag{25}$$