Machine Learning
Preparation of the final exam

Laura Kallmeyer

Summer 2016, Heinrich-Heine-Universität Düsseldorf

Exercise 1 *Consider the following toy example from the LM homework exercises.*

Training data:

```
<s> I am Sam </s>
<s> Sam I am </s>
<s> Sam I like </s>
<s> Sam I do like </s>
<s> do I like Sam </s>
```

Bigram probabilities:

$P(\texttt{Sam}\|\texttt{<s>}) = \frac{3}{5}$	$P(\texttt{I}\|\texttt{<s>}) = \frac{1}{5}$	
$P(\texttt{I}\|\texttt{Sam}) = \frac{3}{5}$	$P(\texttt{</s>}\|\texttt{Sam}) = \frac{2}{5}$	
$P(\texttt{Sam}\|\texttt{am}) = \frac{1}{2}$	$P(\texttt{</s>}\|\texttt{am}) = \frac{1}{2}$	
$P(\texttt{am}\|\texttt{I}) = \frac{2}{5}$	$P(\texttt{like}\|\texttt{I}) = \frac{2}{5}$	$P(\texttt{do}\|\texttt{I}) = \frac{1}{5}$
$P(\texttt{Sam}\|\texttt{like}) = \frac{1}{3}$	$P(\texttt{</s>}\|\texttt{like}) = \frac{2}{3}$	
$P(\texttt{like}\|\texttt{do}) = \frac{1}{2}$	$P(\texttt{I}\|\texttt{do}) = \frac{1}{2}$	

1. *What are the probabilities of the following sentences?*

   (1) `<s>` *I like Sam* `</s>`

   (2) `<s>` *Sam I am Sam I like Sam* `</s>`

2. *What are the perplexities of these sentences?*

3. *How do you explain that the difference in probability we see here is not reflected in a corresponding difference in perplexity? In other words, why does the lower probability in this case not correlate with a higher perplexity?*

Solution:

1. (1) $\frac{1}{5} \cdot \frac{2}{5} \cdot \frac{1}{3} \cdot \frac{2}{5} = 0.0107$

   (2) $\frac{3}{5} \cdot \frac{3}{5} \cdot \frac{2}{5} \cdot \frac{1}{2} \cdot \frac{3}{5} \cdot \frac{2}{5} \cdot \frac{1}{3} \cdot \frac{2}{5} = 0.002304$

2. (1) $\frac{1}{\sqrt[4]{0.0107}} = 3.11$

   (2) $\frac{1}{\sqrt[8]{0.002304}} = 2.12$

3. The fact that perplexity is higher when probability is lower only holds of sentences of equal length. Perplexity factors out sentence length while the probabilities get lower with increasing sentence length. This is the reason why the perplexity of the longer sentence is lower even though its probability is lower as well.

Exercise 2 *Assume that we have used a classifier, for instance naive Bayes, for classifying documents with respect to sentiment. Classes are Pos (positive), Neg (negative) and Neu (neutral).*

We test our classifier on 10 documents for which gold classes are given. The testing has the following results:

Documents	*gold class*	*system class*
d_1	*Pos*	*Pos*
d_2	*Pos*	*Pos*
d_3	*Pos*	*Pos*
d_4	*Pos*	*Neu*
d_5	*Neg*	*Neg*
d_6	*Neg*	*Neu*
d_7	*Neg*	*Neg*
d_8	*Neu*	*Pos*
d_9	*Neu*	*Neu*
d_{10}	*Neu*	*Neu*

1. *Compute precision, recall, accuracy and F_1 for these classification resuls for all three classes.*
2. *Give the pooled confusion matrix for the results on the test set. Give the overall precision and recall by*
   (a) *macroaveraging, and*
   (b) *microaveraging.*

Solution

1.

Pos	gold yes	gold no
system yes	3	1
system no	1	5

Neg	gold yes	gold no
system yes	2	0
system no	1	7

Neu	gold yes	gold no
system yes	2	2
system no	1	5

Pos: $P = 0.75,\ R = 0.75,\ A = 0.8,\ F_1 = \frac{2PR}{P+R} = \frac{2\cdot 0.75\cdot 0.75}{1.5} = 0.75$

Neg: $P = 1,\ R = \frac{2}{3},\ A = 0.9,\ F_1 = \frac{2\cdot 1\cdot \frac{2}{3}}{\frac{5}{3}} = 0.8$

Neu: $P = 0.5,\ R = \frac{2}{3},\ A = 0.7,\ F_1 = \frac{2\cdot 0.5\cdot \frac{2}{3}}{\frac{7}{6}} = \frac{4}{7} = 0.57$

2. Pooled table:

	gold yes	gold no
system yes	7	3
system no	3	17

Macroaveraging: $P = \frac{0.75+1+0.5}{3} = \frac{2.25}{3} = 0.75,\ R = \frac{0.75+\frac{2}{3}+\frac{2}{3}}{3} = 0.69$
Microaveraging: $P = \frac{7}{10} = 0.7,\ R = \frac{7}{10} = 0.7$

Exercise 3 *Consider the following training data for text classification.*

document	*class*	*document*	*class*
aaa	*A*	*bb*	*B*
ab	*A*	*abb*	*B*

1. *Compute $P(A), P(a|A), P(b|A)$ and also $P(B), P(b|B), P(a|B)$ as done in a naive Bayes classifier.*
2. *Based on this, compute the probabilities $P(A|d)$ and $P(B|d)$ for some new text $d = aa$.*
   Note that with Bayes $P(A|d) = \frac{P(d|A)P(A)}{P(d)}$, similarly for B and, furthermore, $P(A|d)+P(B|d) = 1$. Consequently, $P(d) = P(d|A)P(A) + P(d|B)P(B)$.

Solution

1. $P(A) = \frac{1}{2}$, $P(a|A) = \frac{4}{5}$, $P(b|A) = \frac{1}{5}$
   $P(B) = \frac{1}{2}$, $P(a|B) = \frac{1}{5}$, $P(b|B) = \frac{4}{5}$

2. $P(A|d) = \frac{P(d|A)P(A)}{P(d)} = \frac{0.8^2 \cdot 0.5}{0.5 \cdot 0.8^2 + 0.5 \cdot 0.2^2} = 0.94$
   $P(B|d) = \frac{P(d|B)P(B)}{P(d)} = \frac{0.2^2 \cdot 0.5}{0.5 \cdot 0.8^2 + 0.5(0.2)^2} = 0.06$

Exercise 4 *Now consider the same training data for kNN classification with terms a, b and classes A, B.*

1. *Give the corresponding term-document matrix.*
2. *Compute the probabilities $P(A|d)$, $P(B|d)$ for the same d as before if we assume $k = 3$ and if we use the probability definition based on the cosine scores, as defined on slide 27.*

Solution

1.

	aaa	ab	bb	abb
a	3	1	0	1
b	0	1	2	2

2. The new document aa has a vector $[2, 0]$. Its cosine similarity to the four training instances are

	aaa	ab	bb	abb
aa	$\frac{6}{\sqrt{9}\sqrt{4}} = 1$	$\frac{2}{\sqrt{2}\sqrt{4}} = 0.71$	$\frac{0}{\sqrt{4}\sqrt{4}} = 0$	$\frac{2}{\sqrt{5}\sqrt{4}} = 0.45$

   For $k = 3$, we have to consider all training instances except *bb*.

   $P(A|aa) = \frac{1+0.71}{1+0.71+0.45} = 0.79$

   $P(B|aa) = \frac{0.45}{1+0.71+0.45} = 0.21$

Exercise 5 *Now consider again classifying sequences over $\{a, b\}$ into classes A or B (one class per sequence). We use logistic regression, i.e., MacEnt with the following indicator features:*

function	*weight*
$f_1(c, x) = \begin{cases} 1 \text{ if } a \text{ is the first element in } x \text{ and } c = A \\ 0 \text{ otherwise} \end{cases}$	$w_1 = 1.9$
$f_2(c, x) = \begin{cases} 1 \text{ if } b \text{ is the first element in } x \text{ and } c = B \\ 0 \text{ otherwise} \end{cases}$	$w_2 = 2.7$
$f_3(c, x) = \begin{cases} 1 \text{ if } b \text{ is in } x \text{ and } c = A \\ 0 \text{ otherwise} \end{cases}$	$w_3 = 0.3$
$f_4(c, x) = \begin{cases} 1 \text{ if } a \text{ is the last element in } x \text{ and } c = A \\ 0 \text{ otherwise} \end{cases}$	$w_4 = 0.5$
$f_5(c, x) = \begin{cases} 1 \text{ if } a \text{ is in } x \text{ and } c = B \\ 0 \text{ otherwise} \end{cases}$	$w_5 = 0.2$

1. *Give the weighted feature sums for aa and class A and for aa and class B.*
2. *Based on this, compute the probabilities $P(A|aa)$ and $P(B|aa)$.*
3. *How do we have to change the weights of the feature f_3 if we want to exclude class A for all documents containing any b?*

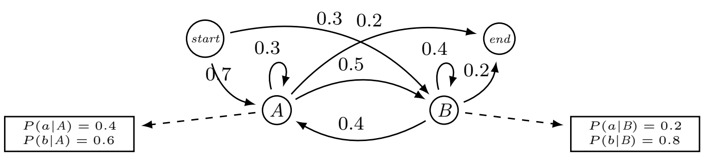
Solution

1. A: $1.9 + 0.5 = 2.6$
   B: 0.2

2. $P(A|aa) = \frac{e^{2.6}}{e^{2.6}+e^{0.2}} = 0.92$
   $P(B|aa) = \frac{e^{0.2}}{e^{2.6}+e^{0.2}} = 0.08$

3. We have to give it a highly negative weight. I.e., a weight towards $-\infty$ excludes a feature for a specific class.

Exercise 6 *Consider the following HMM for tagging sequences over $\{a, b\}$ with a sequence of classes over $\{A, B\}$.*



Calculate the forward and backward matrices for an input ba. What is the probability of ba according to this HMM?

Solution

α:

A	0.42	$8.88 \cdot 10^{-2}$
B	0.24	$6.12 \cdot 10^{-2}$
t	1	2

β:

A	$4.4 \cdot 10^{-2}$	0.2
B	$4.8 \cdot 10^{-2}$	0.2
t	1	2

$P(ba) = 3 \cdot 10^{-2}$

Exercise 7 *Take the same HMM and assume that we have training data consisting of a single sequence, aab. The forward and backward matrices for this sequence are*

α:

A	0.28	$4.32 \cdot 10^{-2}$	$1.56 \cdot 10^{-2}$
B	$6 \cdot 10^{-2}$	$3.28 \cdot 10^{-2}$	$2.77 \cdot 10^{-2}$
t	*1*	*2*	*3*

β:

A	$2.51 \cdot 10^{-2}$	0.12	0.2
B	$2.75 \cdot 10^{-2}$	0.11	0.2
t	*1*	*2*	*3*

Assume that we want to do a first iteration of EM parameter estimation, starting from the weights that are given and based on our toy training corpus of aab.

1. *What is the probability of the training data, given the current parameters?*

2. *Calculate the new transition probability $\hat{a}_{A,A}$ that we obtain in the next step.*

Solution

1. $P(aab) = 8.65 \cdot 10^{-3}$

2. $\xi_1(A, A) = \frac{\alpha_{1,A} \cdot a_{A,A} \cdot b_A(a) \cdot \beta_{2,A}}{P(aab)} = 0.45$

   $\xi_2(A, A) = \frac{\alpha_{2,A} \cdot a_{A,A} \cdot b_A(b) \cdot \beta_{3,A}}{P(aab)} = 0.18$

   $\xi_1(A, B) = \frac{\alpha_{1,A} \cdot a_{A,B} \cdot b_B(a) \cdot \beta_{2,B}}{P(aab)} = 0.36$

   $\xi_2(A, B) = \frac{\alpha_{2,A} \cdot a_{A,B} \cdot b_B(b) \cdot \beta_{3,B}}{P(aab)} = 0.4$

   $\hat{a}_{A,A} = \frac{\xi_1(A,A)+\xi_2(A,A)}{\xi_1(A,A)+\xi_2(A,A)+\xi_1(A,B)+\xi_2(A,B)+\frac{\alpha_{3,A} \cdot a_{A,end}}{P(aab)}} = 0.36$