Assignment 2 - Solutions

Exercise 1 (IIR 13) [2 P.]

Based on the data below, estimate a multinomial Naive Bayes classifier (the type of NB classifier we introduced in class) and apply the classifier to the test document. Calculate the probability that the classifier assigns the test document to $c = \textit{Japan}$ or $\overline{c}$.

	docID	words in document	in c = Japan?
training set	1	Kyoto Tokyo Taiwan	yes
	2	Japan Kyoto	yes
	3	Taipei Taiwan	no
	4	Macao Taiwan Beijing	no
test set	5	Taiwan Taiwan Kyoto	?

Solution

We can compute the probability of a document d being in a class c with the following formula:

$$P(c|d) \quad \propto \hat{P}(c) \prod_{1 \leq k \leq n_d} \hat{P}(t_k|c) \tag{1}$$

Thus, for d_5 and $c = \textit{Japan}$, we need ...

i) The prior probabilities $\hat{P}(c)$ and $\hat{P}(\overline{c})$:

$$\hat{P}(c) = \hat{P}(\overline{c}) = \frac{N_c}{N} = \frac{N_{\overline{c}}}{N} = \frac{2}{4} = \frac{1}{2}$$

ii) The conditional probabilities $\hat{P}(Taiwan|c)$, $\hat{P}(Taiwan|\overline{c})$, $\hat{P}(Kyoto|c)$ and $\hat{P}(Kyoto|\overline{c})$. We compute them by means of the following formula:

$$\hat{P}(t|c) = \frac{C(t,c)+1}{\sum_{t' \in V} C(t',c) + |V|}$$

The vocabulary has 7 terms: Kyoto, Tokyo, Taiwan, Japan, Taipei, Macao, Beijing. There are 5 tokens in the concatenation of all c documents. There are 5 tokens in the concatenation of all $\overline{c}$ documents. Thus, the denominators have the form (5+7). The conditional probabilities for both classes are then as follows:

$$\begin{aligned}
\hat{P}(Taiwan|c) &= (1+1)/(5+7) = 2/12 \\
\hat{P}(Taiwan|\overline{c}) &= (2+1)/(5+7) = 3/12 \\
\hat{P}(Kyoto|c) &= (2+1)/(5+7) = 3/12 \\
\hat{P}(Kyoto|\overline{c}) &= (0+1)/(5+7) = 1/12
\end{aligned}$$

Now we can put it all together and compute the class to which the test document will be assigned using the formula (1):

$$\hat{P}(c|d) \propto 1/2 \cdot (2/12)^2 \cdot 3/12 = 1/2 \cdot 12/(12 \cdot 12 \cdot 12) = 1/2 \cdot 12/1728 = 1/288$$
$$\hat{P}(\overline{c}|d) \propto 1/2 \cdot (3/12)^2 \cdot (1/12) = 1/2 \cdot 9/(12 \cdot 12 \cdot 12) = 1/2 \cdot 9/1728 = 1/384$$

Thus, the classifier assigns the test document to $c = \textit{Japan}$.

Exercise 2 (IIR 3) [1 P.]

If you wanted to search for `s*ng` in a permuterm wildcard index, what key(s) would one do the lookup on?

Solution

We would perform the lookup on the key: `ng$s*`.

Exercise 3 (IIR 3) [3 P.]

Compute the Levenshtein matrix for the distance between the strings "obama" (input) and "romney" (output). Use this format (as introduced in class):

		c		a		t		c		a		t	
	0	**1**	**1**	**2**	**2**	**3**	**3**	**4**	**4**	**5**	**5**	**6**	**6**
c	**1**	**0**	2	2	3	3	4	**3**	5	5	6	6	7
	1	2	**0**	**1**	**1**	**2**	**2**	**3**	**3**	**4**	**4**	**5**	**5**
a	**2**	2	**1**	**0**	2	2	3	3	4	**3**	5	5	6
	2	3	**1**	2	**0**	**1**	**1**	**2**	**2**	**3**	**3**	**4**	**4**
t	**3**	3	**2**	2	**1**	**0**	2	2	3	3	4	**3**	5
	3	4	**2**	3	**1**	2	**0**	**1**	**1**	**2**	**2**	**3**	**3**

After you have calculated the distance between the two strings: Trace the editing operations for one possible editing path as demonstrated in class:

cost	operation	input	output
1	insert	*	c
1	insert	*	a
1	insert	*	t
0	(copy)	c	c
0	(copy)	a	a
0	(copy)	t	t

Solution

Levenshtein matrix:

		r		o		m		n		e		y	
	0	**1**	**1**	2	2	3	3	4	4	5	5	6	6
o	1	1	2	**1**	3	3	4	4	5	5	6	6	7
	1	2	1	2	**1**	2	2	3	3	4	4	5	5
b	2	2	2	2	2	**2**	3	3	4	4	5	5	6
	2	3	2	3	2	3	**2**	3	3	4	4	5	5
a	3	3	3	3	3	3	3	**3**	4	4	5	5	6
	3	4	3	4	3	4	3	4	**3**	4	4	5	5
m	4	4	4	4	4	3	4	4	4	**4**	5	5	6
	4	5	4	5	4	5	3	4	4	5	**4**	5	5
a	5	5	5	5	5	5	4	4	5	5	5	**5**	6
	5	6	5	6	5	6	4	5	4	5	5	6	**5**

Possible editing path:

cost	operation	input	output
1	insert	*	r
0	(copy)	o	o
1	replace	b	m
1	replace	a	n
1	replace	m	e
1	replace	a	y

Exercise 4 (IIR 13) [2 P.]

Rank the documents in collection $\{d_1, d_2\}$ for query q using the language model approach to IR introduced in class with Jelinek-Mercer smoothing. Use the mixture coefficient $\lambda = 0.4$.

- d_1: The European Union Act 2011 prevents additional powers being passed to Brussels without a referendum
- d_2: EU will not ban Chanel 5 perfumes over allergy findings
- Query q: European Union

Solution

$$P(q|d) = \prod_{1 \leq k \leq |q|} [\lambda P(t_k|M_d) + (1-\lambda) P(t_k|M_c)]$$

$$\begin{aligned} P(q|d_1) &= [0.4 \cdot 1/15 + 0.6 \cdot 1/25] \cdot [0.4 \cdot 1/15 + 0.6 \cdot 1/25] \approx 0.051^2 \approx 0.0026 \\ P(q|d_2) &= [0.4 \cdot 0/10 + 0.6 \cdot 1/25] \cdot [0.4 \cdot 0/10 + 0.6 \cdot 1/25] = 0.024^2 = 0.000576 \end{aligned}$$

$\Rightarrow$ Ranking: $d_1 > d_2$

Exercise 5 (IIR 12/13) [3 P.]

Consider the following frequencies for the class *coffee* for four terms in the first 100,000 documents of Reuters-RCV1:

term	N_{00}	N_{01}	N_{10}	N_{11}
brazil	98,012	102	1835	51
council	96,322	133	3525	20
producers	98,524	119	1118	34
roasted	99,824	143	23	10

a) Which two terms will be selected in frequency-based feature selection and why?

b) Compute the MI values and order the terms according to MI. Which two terms will be selected in MI-based feature selection?

Solution

a) The terms *brazil* and *producers* will be selected in frequency-based feature selection because they are the most frequent terms in the class *coffee*.

b) *brazil*:

$$\begin{aligned} I(U;C) &= \frac{98012}{100000}\log_2\frac{100000\cdot 98012}{(102+98012)(1835+98012)} \\ &\quad +\frac{102}{100000}\log_2\frac{100000\cdot 102}{(102+98012)(51+102)} \\ &\quad +\frac{1835}{100000}\log_2\frac{100000\cdot 1835}{(51+1835)(1835+98012)} \\ &\quad +\frac{51}{100000}\log_2\frac{100000\cdot 51}{(51+1835)(51+102)} \\ &\approx 0.0015536892 \end{aligned}$$

council:

$$\begin{aligned} I(U;C) &= \frac{96322}{100000}\log_2\frac{100000\cdot 96322}{(133+96322)(3525+96322)} \\ &\quad +\frac{133}{100000}\log_2\frac{100000\cdot 133}{(133+96322)(20+133)} \\ &\quad +\frac{3525}{100000}\log_2\frac{100000\cdot 3525}{(20+3525)(3525+96322)} \\ &\quad +\frac{20}{100000}\log_2\frac{100000\cdot 20}{(20+3525)(20+133)} \\ &\approx 0.0001774273 \end{aligned}$$

producers:

$$
\begin{aligned}
I(U;C) &= \frac{98524}{99795}\log_2\frac{99795\cdot 98524}{(119+98524)(1118+98524)} \\
&\quad +\frac{119}{99795}\log_2\frac{99795\cdot 119}{(119+98524)(34+119)} \\
&\quad +\frac{1118}{99795}\log_2\frac{99795\cdot 1118}{(34+1118)(1118+98524)} \\
&\quad +\frac{34}{99795}\log_2\frac{99795\cdot 34}{(34+1118)(34+119)} \\
&\approx 0.0010479995
\end{aligned}
$$

roasted:

$$
\begin{aligned}
I(U;C) &= \frac{99824}{100000}\log_2\frac{100000\cdot 99824}{(143+99824)(23+99824)} \\
&\quad +\frac{143}{100000}\log_2\frac{100000\cdot 143}{(143+99824)(10+143)} \\
&\quad +\frac{23}{100000}\log_2\frac{100000\cdot 23}{(10+23)(23+99824)} \\
&\quad +\frac{10}{100000}\log_2\frac{100000\cdot 10}{(10+23)(10+143)} \\
&\approx 0.0006484759
\end{aligned}
$$

Terms ranked by MI:

(1) brazil
(2) producers
(3) roasted
(4) council

The terms *brazil* and *producers* will be selected in MI-based feature selection.