COMS W4705: Natural Language Processing
Written Homework 1

Minghao Li (ml4025)

February 13, 2018

Problem 1

(a) Based on the number of each class(3 for spam and 2 for ham), the prior probability is simply the number of files in each class divided by the total file number. So the results are:

$$P(spam) = \frac{3}{5}$$
$$P(ham) = \frac{2}{5}$$

(b) Given each class, we could calculate the probability of each word in that class by the formula:

$$P(word|Class) = \frac{count(word\_in\_Class)}{count(total\_words)}$$

For example, when calculating $P(buy|spam)$:

$$P(buy|spam) = 1/12$$

in which "1" is the times that "buy" appears in the Class spam, while "12" is the total number of all words in Class spam.

By operating the same process on the texts, we could get the following result table.

Table 1: Conditional Probability Table

	buy	car	Nigeria	profit	money	home	bank	check	wire	fly
spam	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{2}{12}$	$\frac{2}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{2}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	0
ham	0	$\frac{1}{7}$	$\frac{1}{7}$	0	$\frac{1}{7}$	$\frac{2}{7}$	$\frac{1}{7}$	0	0	$\frac{1}{7}$

(c) This part is based on the Naive Bayes Classification method, which is described by the following formula:

$$P(label|w_1, w_2, ..., w_n) = \frac{P(w_1, w_2, ..., w_n|label)P(label)}{P(w_1, w_2, ..., w_n)}$$

where n is the number of words in the sentence. For the denominator part, given a specific input sentence, the Probability is a constant since we only have one tested data once. So the formula can be simplified as:

$$P(label|w_1, w_2, ..., w_n) = \alpha\, P(w_1, w_2, ..., w_n|label)P(label)$$

Because each word w_i is independent distributed, the result can be furthered simplified below:

$$P(label|w_1, w_2, ..., w_n) = \alpha \, P(w_1|label)P(w_2|label)...P(w_n|label)P(label)$$

Firstly, given the sentence "Nigeria":

$$P(spam|sentence) = \alpha \, P(Nigeria|spam)P(spam) = \frac{\alpha}{10}$$

$$P(ham|sentence) = \alpha \, P(Nigeria|ham)P(ham) = \frac{2\alpha}{35}$$

It shows $P(spam|sentence) > P(ham|sentence)$, so it should be a *spam*.

Secondly, given the sentence "Nigeria home": using the formula and same process above, we could get the following probability of each class:

$$P(spam|sentence) = \frac{\alpha}{120} < P(ham|sentence) = \frac{4\alpha}{125}$$

So the second sentence should be classified as *ham*.

Thirdly, given the sentence "home bank money", using the same formula and process:

$$P(spam|sentence) = \frac{6\alpha}{8460} < P(ham|sentence) = \frac{4\alpha}{1715}$$

So the third sentence should be classified as *ham*.

Problem 2

I will use mathematic induction to prove this sum equal to 1.

Firstly, when the sentence length is only 1, there's V different sentences, each of which has a appearance possibility $\frac{1}{V}$, so the sum of all these possibilities are $V * \frac{1}{V} = 1$

Secondly, suppose that for a sentence with $n-1$ length, the sum of possibilities of all sentences are 1. That is :

$$\sum_q P(w_1, w_2, w_3, ..., w_{n-1} = q) = \sum_q P(W_{n-1} = q) = 1$$

in which the q refers to each possible word in *vocabulary*.

Then for the sentence with n length, which means we add one more word to the end of the above sentence, we have the following possibility:

$$\sum_{w_1, w_2, ..., w_{n-1}, w_n} P(w_1, w_2, ...w_{n-1}, w_n) = \sum_q \sum_v P(w_{n-1} = q, w_n = v)$$

Some interpretation concerning the above relation: we only want to concentrate on the last two words because of the bi-gram model. Since the first part which I ignore is sum to 1. Then it's reasonable to use this equation. Plus, q and v refers to the $n-1$th and nth word in the sentence, which is selected randomly from vocabulary.

$$\begin{aligned}
\sum_{w_1, w_2, ..., w_{n-1}, w_n} P(w_1, w_2, ...w_{n-1}, w_n) &= \sum_q \sum_v P(w_{n-1} = q) P(w_n = v | w_{n-1} = q) \\
&= \frac{1}{v} \sum_q \sum_v P(w_n = v | w_{n-1} = q) \\
&= \frac{1}{V} \sum_q 1 \\
&= \frac{1}{V} * V \\
&= 1
\end{aligned}$$

Finally, we can prove that from n=1 to n, the conclusion works well.

Problem 3

(a) The naive method based on add-one smoothing definition:

We have already known that for estimating $P(w_2|w_1)$, we have formula:

$$P(w_2|w_1) = \frac{count(w_1, w_2) + 1}{count(w_1) + V}$$

Similarly, when it comes to $P(w_3|w_1, w_2)$, we have

$$P(w_3|w_1, w_2) = \frac{count(w_1, w_2, w_3) + 1}{count(w_1, w_2) + V}$$

Considering when $count(w_1, w_2) = 0$, it roughly should be larger than V since the denominator is counting w_1 and w_2. So it should roughly be $2V$.

Besides, for introducing $count(w_1)$ into the result, we should also replace $count(w_1, w_2)$ with something else.

$$count(w_1, w_2) = count(w_1) * P(w_2|w_1) = \frac{count(w_1, w_2) + 1}{count(w_1) + V} * count(w_1)$$

By applying it to the $P(w_3|w_1, w_2)$, I got the final result below:

$$P(w_3|w_1, w_2) = \frac{count(w_1, w_2, w_3) + 1}{\frac{count(w_1,w_2)+1}{count(w_1)+V} count(w_1) + 2V}$$

(b) Using Backoff method for trigram and bigram.

The core thought in this method is:

If we have no trigrams like: (w_{n-2}, w_{n-1}, w_n) to calculate $P(w_n|w_{n-2}w_{n-1})$, then we can simply use $P(w_n|w_{n-1})$ instead. The same replacement for $P(w_n|w_{n-1})$ using $P(w_n)$

describe the function as below:

case 1: if count$(w_1, w_2, w_3) > 0$:

$$P(w_3|w_2, w_1) = \frac{count(w_1, w_2, w_3) + 1}{count(w_1, w_2) + V}$$

case 2: elseif count$(w_3, w_2) > 0$:

$$P(w_3|w_2) = \alpha_1 \frac{count(w_3, w_2) + 1}{count(w_3, w_2) + V}$$

where α_1 is the parameter to make the sum of new probability equal to 1. case 3: else :

$$P(w_3) = \alpha_2 \frac{count(w_3) + 1}{N + V}$$

where α_1 is the parameter to make the sum of new probability equal to 1. N is the number of all tokens in the corpus.