COMS W4705: Natural Language Processing
Written Homework 1

Sample Solutions

October 2, 2018

Problem 1

Prior Probabilities

P(SPAM) = $\frac{\text{Number of emails categorized as spam}}{\text{Total number of emails}} = \frac{3}{5}$

Similarly, P(HAM) = $\frac{2}{5}$

Conditional Probabilities

To calculate $P(Word|Class = Spam)$, we count the occurrences of the word in class Spam and divide it by the total number of words in class Spam. Similarly for the class of Ham emails.
Total number of words in class Spam = 12
Total number of words in class Ham = 7

SPAM	HAM	Words
1/12	0/7	Buy
1/12	1/7	Car
2/12	1/7	Nigeria
2/12	0/7	Profit
1/12	1/7	Money
1/12	2/7	Home
2/12	1/7	Bank
1/12	0/7	Check
1/12	0/7	Wire
0/12	1/7	Fly

Predicted Class Labels

Predicted Label: SPAM
P(HAM | Nigeria) = $\alpha\cdot$ P(HAM).P(Nigeria | HAM) = $\alpha\cdot \frac{2}{5} \times \frac{1}{7} = \alpha\cdot \frac{2}{35}$
P(SPAM | Nigeria) = $\alpha\cdot$ P(SPAM).P(Nigeria | SPAM) = $\alpha\cdot \frac{3}{5} \times \frac{2}{12} = \alpha\cdot \frac{1}{10}$
As P(SPAM | Nigeria) is greater, predicted class is SPAM.
Predicted Label: HAM
P(HAM | Nigeria home) = $\alpha\cdot$ P(HAM).P(Nigeria home | HAM) = $\alpha\cdot \frac{2}{5} \times \frac{1}{7} \times \frac{2}{7} = \alpha\cdot \frac{0.8}{49}$
P(SPAM | Nigeria home) = $\alpha\cdot$ P(SPAM).P(Nigeria home | SPAM) = $\alpha\cdot \frac{3}{5} \times \frac{2}{12} \times \frac{1}{12} = \alpha\cdot \frac{1}{120}$
Since P(HAM | Nigeria Home) is greater, the predicted class is HAM.

Predicted Label: HAM
P(HAM | home bank money) = $\alpha\cdot$ P(HAM).P(home bank money | HAM) =$\alpha\cdot$ $\frac{2}{5}\times\frac{2}{7}\times\frac{1}{7}\times\frac{1}{7}=\alpha\cdot$ 0.00233
P(SPAM | home bank money) = $\alpha\cdot$ P(SPAM).P(home bank money | SPAM) = $\alpha\cdot$ $\frac{3}{5}\times\frac{1}{12}\times\frac{2}{12}\times\frac{1}{12}$ =$\alpha\cdot$ 0.00069
Since P(HAM | home bank money) is greater, predicted class is HAM.

Problem 2

Base case: For a 1-word sentence, $\sum_{w_1} P(w_1|START) = 1$ by definition.

Inductive hypothesis: Assume that the sum of the probability of all sentences of length $n-1$ is 1:

$$\sum_{w_1,\ldots,w_{n-1}} P(w_1,\ldots w_{n-1}) = 1.$$

Inductive step: Let $P(w_{n-1}=q) = \sum_{w_1,\ldots w_{n-2}} P(w_1,\ldots,w_{n-2},w_{n-1}=q)$ be the probability of all sentence of length $n-1$ that ends in q. Because of the inductive hypothesis,

$$\sum_q P(w_{n-1}=q) = 1.$$

We can then write the sum for length n sentences as

$$\sum_r \sum_q P(w_{n-1}=q, w_n=r).$$

$$= \sum_r \sum_q P(w_{n-1}=q)\cdot P(w_n=r|w_{n-1}=q)$$

Because $\sum_q P(w_{n-1}=q) = 1$ and there are V different words q, we can rewrite this as:

$$= \sum_q \sum_r \frac{1}{V} P(w_n=r|w_{n-1}=q)$$

$$= \frac{1}{V}\sum_q \sum_r P(w_n=r|w_{n-1}=q)$$

For a specific word q we know that $\sum_r P(w_n=r|w_{n-1}=q) = 1$. Since there V words, $\sum_q \sum_r P(w_n = r|w_{n-1}=q) = V$. So the total sum becomes

$$\frac{1}{V}V = 1 \quad \square$$

Problem 3

The expression for regular add-one smoothing is

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

This approach teakes care of the case in which w_3 was not seen in this context. It also handles the case of unseen bi-grams, in the sense that there is no division by 0 error. However: it suffers another flaw. If $count(w_1, w_2) = 0$, then the probability deteriorates to a uniform distribution of $\frac{1}{V}$ for each word.
The easiest solution to this problem is to simply use the smoothed bi-gram probability as a back-off in the case that $count(w_1, w_2) = 0$.

$$P(w_3|w_1, w_2) = \begin{cases} \frac{count(w_2,w_3)+1}{count(w2)+V} & \text{if } count(w_1, w_2) = 0 \\ \\ \frac{count(w_1,w_2,w_3)+1}{count(w_1,w_2)+V} & \text{otherwise} \end{cases}$$

The problem description was a little unclear. We accepted a wide range of solutions, as long as they addressed the following points:

1. Using add-one smoothing for trigram estimates.
2. Providing a solution that does not deteriorate to a uniform distribution for trigram probability if bigram context is unseen.
3. Ensuring the final trigram probability expression is a valid probability distribution.