COMS W4705: Natural Language Processing
Written Homework 1

Qianrui Zhao (qz2338)

September 29, 2018

Problem 1

(a)

$$P(Spam) = \frac{3}{5}$$
$$P(Ham) = \frac{2}{5}$$

(b) $P(Word|Class)$ in table

$P(Word\|Class)$	Spam	Ham
"buy"	$\frac{1}{12}$	0
"car"	$\frac{1}{12}$	$\frac{1}{7}$
"Nigeria"	$\frac{1}{6}$	$\frac{1}{7}$
"profit"	$\frac{1}{6}$	0
"money"	$\frac{1}{12}$	$\frac{1}{7}$
"home"	$\frac{1}{12}$	$\frac{1}{7}$
"bank"	$\frac{1}{6}$	$\frac{2}{7}$
"check"	$\frac{1}{12}$	0
"wire"	$\frac{1}{12}$	0
"fly"	0	$\frac{1}{7}$

(c) Predict class label

(a) Predict class of Nigeria

$$P(Spam|Nigeria) = \frac{P(Spam) \cdot P(Nigeria|Spam)}{P(Nigeria)} = \frac{\frac{3}{5} \cdot \frac{1}{6}}{P(Nigeria)} = \frac{\frac{1}{10}}{P(Nigeria)}$$

$$P(Ham|Nigeria) = \frac{P(Ham) \cdot P(Nigeria|Ham)}{P(Nigeria)} = \frac{\frac{2}{5} \cdot \frac{1}{7}}{P(Nigeria)} = \frac{\frac{2}{35}}{P(Nigeria)}$$

Because $P(Nigeria)$ is a positive number and $P(Spam|Nigeria) > P(Ham|Nigeria)$

Therefore, Predicted class for Nigeria is Spam.

(b) Predict class of Nigeria, Home

$$P(Spam|Nigeria, home) = \frac{P(Spam) \cdot P(Nigeria|Spam) \cdot P(Home|Spam)}{P(Nigeria) \cdot P(Home)}$$

$$= \frac{\frac{3}{5} \cdot \frac{1}{6} \cdot \frac{1}{12}}{P(Nigeria) \cdot P(Ham)} = \frac{\frac{1}{120}}{P(Nigeria) \cdot P(Home)}$$

$$P(Ham|Nigeria, Home) = \frac{P(Ham) \cdot P(Nigeria|Ham) \cdot P(Home|Ham)}{P(Nigeria) \cdot P(Home)}$$

$$= \frac{\frac{2}{5} \cdot \frac{1}{7} \cdot \frac{1}{7}}{P(Nigeria) \cdot P(Ham)} = \frac{\frac{2}{245}}{P(Nigeria) \cdot P(Home)}$$

Because $P(Nigeria) \cdot P(Home)$ is a positive number and $P(spam|Nigeria, Home) > P(Ham|Nigeria, Home)$

Therefore, predicted class for Nigeria, Home is Spam.

(c) Predict class of Home, Bank, Money

$$P(Spam|Home, Bank, Money) = \frac{P(Spam) \cdot P(Home|Spam) \cdot P(Bank|Spam) \cdot P(Money|Spam)}{P(Home) \cdot P(Bank) \cdot P(Money)}$$

$$= \frac{\frac{3}{5} \cdot \frac{1}{12} \cdot \frac{1}{6} \cdot \frac{1}{12}}{P(Home) \cdot P(Bank) \cdot P(Money)} = \frac{\frac{1}{1440}}{P(Home) \cdot P(Bank) \cdot P(Money)}$$

$$P(Ham|Home, Bank, Money) = \frac{P(Ham) \cdot P(Home|Ham) \cdot P(Bank|Ham) \cdot P(Money|Ham)}{P(Home) \cdot P(Bank) \cdot P(Money)}$$

$$= \frac{\frac{2}{5} \cdot \frac{1}{7} \cdot \frac{2}{7} \cdot \frac{1}{7}}{P(Home) \cdot P(Bank) \cdot P(Money)} = \frac{\frac{4}{1715}}{P(Home) \cdot P(Bank) \cdot P(Money)}$$

Because $P(Home) \cdot P(Bank) \cdot P(Money)$ is a positive number

and $P(Spam|Home, Bank, Money) < P(Ham|Home, Bank, Money)$

Therefore, predicted class of Home, Bank, Money is Ham.

Problem 2

Prove that sum of prob of all sentence with length == n is 1:

Assume vocabulary size is V, words $W = \{w_1, w_2, \dots, w_V\}, w^n$ means nth word in the sentence.

Lemma 01: Two words w^n and w^{n+1},

$$\sum_{w^{n+1} \in W} P(w^{n+1}, w^n) = (P(w_1, w^n) + P(w_2, w^n) + \cdots + P(w_V, w^n)) \cdot P(w^n) = P(w^n)$$

Lemma 02: When sentence length == 1,

$$\sum_{w_i \in W} P(w^1|\text{START}) = P(w_1|START) + P(w_2|START) \dots p(w_n|START) = 1$$

Induction formula can be built as:

$$\begin{aligned}
&\sum_{w^1, w_2, \dots, w^{n+1} \in W} P(w^1|\text{START}) P(w^2|w^1) \dots P(w^{n+1}|P(w_n)) \\
&= \sum_{w^1, w_2, \dots, w^n \in W} P(w^1|START) P(w^2|w^1) \dots P(w^n|P(w^{n-1})) (\sum_{w^{n+1} \in W} P(w^{n+1}|w^n)) \\
&= \sum_{w^1, w_2, \dots, w^n \in W} \frac{P(w^1|START) P(w^2|w^1) \dots P(w^n|P(w^{n-1}))}{P(w^n)} (\sum_{w^{n+1} \in W} P(w^{n+1}, w^n))
\end{aligned}$$

According to Lemma 01, where $\sum_{w^{n+1} \in W} P(w^{n+1}, w^n) = P(w^n)$

$$\begin{aligned}
&= \sum_{w^1, w_2, \dots, w^n \in W} \frac{P(w^1|START) P(w^2|w^1) \dots P(w^n|P(w^{n-1}))}{P(w^n)} \cdot P(w^n) \\
&= \sum_{w^1, w_2, \dots, w^n \in W} P(w^1|START) P(w^2|w^1) \dots P(w^n|w^{n-1})
\end{aligned}$$

Therefore, $\sum_{w^1, w_2, \dots, w^{n+1} \in W} P(w^1|START) P(w^2|w^1) \dots P(w^{n+1}|w^n) = \sum_{w^1 \in W} P(w^1|START)$

According to Lemma 2, where $\sum_{w^1 \in W} P(w^1|START) = 1$

$$\sum_{w^1, w_2, \dots, w^{n+1} \in W} P(w^1|START) P(w^2|w^1) \dots P(w^{n+1}|w^n) = 1$$