COMS W4705: Natural Language Processing (Fall 2018)
Problem Set #1

Wenbo Gao - `wg2313@columbia.edu`

September 20, 2018

Problem 1

Consider the following training corpus of emails with the class labels **ham** and **spam**. The content of each email has already been processed and is provided as a bag of words.
Email1 (spam): buy car Nigeria profit
Email2 (ham): money profit home bank
Email3 (spam): Nigeria bank check wire
Email4 (ham): money bank car
Email5 (ham): home Nigeria fly

(a)

Problem. Based on this data, estimate the prior probability for a random email to be spam or ham if we don't know anything about its content, i.e. $P(Class)$?

Solution. A random variable $Class$ has two possible outcomes, ham and $spam$.
In this dataset, we have $\{spam, ham, spam, ham, ham\}$. Thus,

$$P[Class = ham] = \frac{3}{5}$$
$$P[Class = spam] = \frac{2}{5}$$

(b)

Problem. Based on this data, estimate the conditional probability distributions for each word given the class, i.e. $P(Word|Class)$. You can write down these distribution in a table.

Solution. $Word \in \{bank, buy, car, check, fly, home, money, Nigeria, profit, wire\}$

Word	bank	buy	car	check	fly	home	money	Nigeria	profit	wire
$P[Word\|Class = ham]$	$\frac{2}{3}$	0	$\frac{1}{3}$	0	$\frac{1}{3}$	$\frac{2}{3}$	$\frac{2}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0
$P[Word\|Class = spam]$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	0	0	0	1	$\frac{1}{2}$	$\frac{1}{2}$

(c)

Problem. Using the Naive Bayes' approach and your probability estimates, what is the predicted class label for each of the following emails? Show your calculation.

Solution.

- Nigeria

$$\begin{aligned}
&P[Class = ham|Sentence = \text{Nigeria}] \cdot P[Sentence = \text{Nigeria}] \\
&= P[Sentence = \text{Nigeria}|Class = ham] \cdot P[Class = ham] \\
&= P[Word = Nigeria|Class = ham] \cdot P[Class = ham] \\
&= \frac{1}{3} \cdot \frac{3}{5} = \frac{1}{5}
\end{aligned}$$

$$\begin{aligned}
&P[Class = spam|Sentence = \text{Nigeria}] \cdot P[Sentence = \text{Nigeria}] \\
&= P[Sentense = \text{Nigeria}|Class = spam] \cdot P[Class = spam] \\
&= P[Word = Nigeria|Class = spam] \cdot P[Class = spam] \\
&= 1 \cdot \frac{2}{5} = \frac{2}{5}
\end{aligned}$$

Thus, "Nigeria" is more likely to be *spam*.

- Nigeria home

$$\begin{aligned}
&P[Class = ham|Sentence = \text{Nigeria home}] \cdot P[Sentence = \text{Nigeria home}] \\
&= P[Sentence = \text{Nigeria home}|Class = ham] \cdot P[Class = ham] \\
&= P[Word = Nigeria|Class = ham] \cdot P[Word = home|Class = ham] \cdot P[Class = ham] \\
&= \frac{1}{3} \cdot \frac{2}{3} \cdot \frac{3}{5} = \frac{2}{15}
\end{aligned}$$

$$\begin{aligned}
&P[Class = spam|Sentence = \text{Nigeria home}] \cdot P[Sentence = \text{Nigeria home}] \\
&= P[Sentence = \text{Nigeria home}|Class = spam] \cdot P[Class = spam] \\
&= P[Word = Nigeria|Class = spam] \cdot P[Word = home|Class = spam] \cdot P[Class = spam] \\
&= 1 \cdot 0 \cdot \frac{2}{5} = 0
\end{aligned}$$

Thus, "Nigeria home" is more likely to be *ham*.

- home bank money

$$
\begin{aligned}
&P[Class = ham|Sentence = \text{home bank money}] \cdot P[Sentence = \text{home bank money}] \\
&= P[Sentence = \text{home bank money}|Class = ham] \cdot P[Class = ham] \\
&= P[Word = home|Class = ham] \cdot P[Word = bank|Class = ham] \\
&\cdot P[Word = money|Class = ham] \cdot P[Class = ham] \\
&= \frac{2}{3} \cdot \frac{2}{3} \cdot \frac{2}{3} \cdot \frac{3}{5} = \frac{8}{45}
\end{aligned}
$$

$$
\begin{aligned}
&P[Class = spam|Sentence = \text{home bank money}] \cdot P[Sentence = \text{home bank money}] \\
&= P[Sentence = \text{home bank money}|Class = spam] \cdot P[Class = spam] \\
&= P[Word = home|Class = spam] \cdot P[Word = bank|Class = spam] \\
&\cdot P[Word = money|Class = spam] \cdot P[Class = spam] \\
&= 0 \cdot \frac{1}{2} \cdot 0 \cdot \frac{2}{5} = 0
\end{aligned}
$$

Thus, ”home bank money” is more likely to be *ham*.

Problem 2

Show that, if you sum up the probabilities of all sentences of length n under a bigram language model, this sum is exactly 1 (i.e. the model defines a proper probability distribution). Assume a vocabulary size of V.

$$\sum_{w_1,w_2,\ldots,w_n} P(w_1, w_2, \ldots, w_n) = \sum_{w_1,w_2,\ldots,w_n} P(w_1|\text{start}) \cdot P(w_2|w_1) \cdots P(w_n|w_{n-1}) = 1$$

Hint: Use induction over the sentence length. Comment: This property actually holds for any n-gram model, but you only have to show it for bigrams.

proof by induction.

- Base case:

  $$\sum_{w_1} P(w_1) = \sum_{w_1} P(w_1|\text{start}) = 1$$

  is true.

- Induction step (from $n = k$ to $n = k + 1$):
  Assume for $n = k$,

  $$\sum_{w_1,w_2,\ldots,w_k} P(w_1, w_2, \ldots, w_k) = \sum_{w_1,w_2,\ldots,w_k} P(w_1|\text{start}) \cdot P(w_2|w_1) \cdots P(w_k|w_{k-1}) = 1$$

  is true.
  For $n = k + 1$,

  $$\begin{aligned} &\sum_{w_1,w_2,\ldots,w_{k+1}} P(w_1, w_2, \ldots, w_{k+1}) \\ &= \sum_{w_1,w_2,\ldots,w_{k+1}} P(w_1|\text{start}) \cdot P(w_2|w_1) \cdots P(w_{k+1}|w_k) \\ &= \sum_{w_1,w_2,\ldots,w_k} (P(w_1|\text{start}) \cdot P(w_2|w_1) \cdots P(w_k|w_{k-1}) \cdot \underbrace{\sum_{w_{k+1}} P(w_{k+1}|w_k))}_{=1} \\ &= \sum_{w_1,w_2,\ldots,w_k} P(w_1|\text{start}) \cdot P(w_2|w_1) \cdots P(w_k|w_{k-1}) \\ &= 1 \end{aligned}$$

  is also true.

Therefore,

$$\sum_{w_1,w_2,\ldots,w_n} P(w_1, w_2, \ldots, w_n) = \sum_{w_1,w_2,\ldots,w_n} P(w_1|\text{start}) \cdot P(w_2|w_1) \cdots P(w_n|w_{n-1}) = 1$$

is true. $\square$