COMS 4705 Natural Language Processing (2020 Spring) Homework 1

Shusen Xu - sx2261@columbia.edu

February 18, 2020

Problem 1 Text Classification with Naive Bayes

1. **Prior Probability of a random email**

$$P_{mle}(Class = spam) = \frac{3}{5}$$

$$P_{mle}(Class = ham) = \frac{2}{5}$$

2. **Estimate the conditional probability distribution**
Bag of Words:

Indexes	Word	Count
0	buy	1
1	car	2
2	Nigeria	3
3	profit	2
4	money	2
5	home	3
6	bank	3
7	check	1
8	wire	1
9	fly	1

Conditional Probability distribution:

Word	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	$\frac{1}{12}$	$\frac{2}{7}$
bank	$\frac{1}{6}$	$\frac{1}{7}$
check	$\frac{1}{12}$	0
wire	$\frac{1}{12}$	0
fly	0	$\frac{1}{7}$

3. **Naive Bayes Classifier**

Since $y* = \arg\max_y P(y) \prod_i P(x_i|y)$ (Naive Bayes classifier)

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

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

Since $\frac{1}{10} > \frac{2}{35}$, the class of "Nigeria" is spam.

(b) Nigeria home
$P(spam) \cdot P(Nigeria|spam) \cdot P(home|spam) = \frac{3}{5} \cdot \frac{1}{6} \cdot \frac{1}{12} = \frac{1}{120}$

$P(ham) \cdot P(Nigeria|ham) \cdot P(home|ham) = \frac{2}{5} \cdot \frac{1}{7} \cdot \frac{2}{7} = \frac{4}{245}$

Since $\frac{1}{120} < \frac{4}{245}$, the class of "Nigeria home" is ham.

(c) home bank money
$P(spam) \cdot P(home|spam) \cdot P(bank|spam) \cdot P(money|spam) = \frac{3}{5} \cdot \frac{1}{12} \cdot \frac{1}{6} \cdot \frac{1}{12} = \frac{1}{1440}$

$P(ham) \cdot P(home|ham) \cdot P(bank|ham) \cdot P(money|ham) = \frac{2}{5} \cdot \frac{2}{7} \cdot \frac{1}{7} \cdot \frac{1}{7} = \frac{4}{1715}$

Since $\frac{1}{1440} < \frac{4}{1715}$, the class of "home bank money" is ham.

Problem 2 Bigram Models

Goal: Prove that, if you sum up the probabilities of all sentence of length n under a bigram language model, this sum is exactly 1 by induction.

$$\sum_{w_1,w2,\ldots w_n} P\left(w_1, w_2, \ldots, w_n\right) = \sum_{w_1,w2,\ldots w_n} P\left(w_1|start\right) \cdot P\left(w_2|w1\right) \cdots P\left(w_n|w_{n-1}\right) = 1$$

1. Initialization
   when the length of sentence is 1:
   since $\sum_{w_1} p\left(w_1\right)$ sum over all possible w_1, the value of it obviously is 1. So, combining with the definition of bigram model, we can easily get the following equation.

   $$\sum_{w_1} p\left(w_1\right) = \sum_{w_1} p\left(w_1|start\right) = 1$$

2. Maintenance:

   (a) First, assume when the length of sentence is equal to n, the following equation holds true.

   $$\sum_{w_1,w2,\ldots w_n} P\left(w_1, w_2, \ldots, w_n\right) = \sum_{w_1,w2,\ldots w_n} P\left(w_1|start\right) \cdot P\left(w_2|w1\right) \cdots P\left(w_n|w_{n-1}\right) = 1$$

   (b) we need to proved when the length of sentence is equal to n+1, the following equation holds true.

   $$\sum_{w_1,w2,\ldots w_{n+1}} P\left(w_1, w_2, \ldots, w_n, w_{n+1}\right) = \sum_{w_1,w2,\ldots w_n} P\left(w_1|start\right) \cdot P\left(w_2|w1\right) \cdots P\left(w_{n+1}|w_n\right) = 1$$

   Firstly, according to the bigram model definition which based on Markov assumption, we can get the following equation:

   $$\sum_{w_1,w2,\ldots w_{n+1}} P\left(w_1, w_2, \ldots, w_n, w_{n+1}\right) = \sum_{w_1,w2,\ldots w_n} P\left(w_1|start\right) \cdot P\left(w_2|w1\right) \cdots P\left(w_{n+1}|w_n\right)$$

   Secondly, let's denote sentence $(w_1, w2, \ldots w_n)$ as event A. Notes: since $\sum_{w_{n+1}} P(w_{n+1}|A)$ summer over all possible w_{n+1} under the condition that the previous words sequence is A, so obviously the equation is equal to 1. so we can get:

   $\sum_{w_1,w2,\ldots w_{n+1}} P\left(w_1, w_2, \ldots, w_n, w_{n+1}\right) = \sum_{A,w_{n+1}} P\left(A, w_{n+1}\right)$

   $= \sum_{A,w_{n+1}} P\left(A\right) \cdot P\left(w_{n+1}|A\right)$

   $= \sum_{A} P\left(A\right) \cdot \left(\sum_{w_{n+1}} P(w_{n+1}|A)\right)$

$= \sum_{A} P(A) \cdot 1 = \sum_{w_1, w2, \ldots w_n} P(w_1, w_2, \ldots, w_n) = 1$

3. we proved that: the equation is true when length of sentence is equal to n+1. By induction, we can conclude that for any length of sentence, this equation always holds true.