COMS W4705: Natural Language Processing (sec:002) - Homework #1

Name: Geraldi Dzakwan (gd2551)

October 2, 2019

Problem 1

The data:

- Email1 (spam): buy car Nigeria profit
- Email2 (spam): money profit home bank
- Email3 (spam): Nigeria bank check wire
- Email4 (ham): money bank home car
- Email5 (ham): home Nigeria fly

(a) 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. $P(Class)$?

Answer:

If we don't know about the content, we would simply estimate the prior probability based on the class occurrence over the sample of the data. Mathematically, that would be:

$$P(Class = spam) = \frac{count(Class = spam)}{number\ of\ data} = \frac{3}{5} = 0.6$$

$$P(Class = ham) = \frac{count(Class = ham)}{number\ of\ data} = \frac{2}{5} = 0.4$$

(b) 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 distributions in a table.

Answer:

Probability of a word given a class can be described as: Out of every word that appears in all emails of a certain class (spam/ham), how many of them are that word? Suppose we create a vocabulary of word from the data, the equivalent mathematical notation would be:

$$P(Word|Class) = \frac{count(Word, Class)}{\sum_{w \in Vocabulary} count(w, Class)}$$

As an example, the word "buy" only appears once in spam emails (the first email). Meanwhile, for all three spam emails, there are 12 words in total (including "buy" that appears once). That gives us:

$$P(buy|spam) = \frac{count(buy, spam)}{\sum_{w \in Vocabulary} count(w, spam)} = \frac{1}{12} = 0.0833$$

The full probability distributions of $P(Word|Class)$ is given in the Table 1 below.

word	$P(word\|spam)$	$P(word\|ham)$
buy	$1/12 = 0.0833$	$0/7 = 0$
car	$1/12 = 0.0833$	$1/7 = 0.1429$
Nigeria	$2/12 = 0.1667$	$1/7 = 0.1429$
profit	$2/12 = 0.1667$	$0/7 = 0$
money	$1/12 = 0.0833$	$1/7 = 0.1429$
home	$1/12 = 0.0833$	$2/7 = 0.2857$
bank	$2/12 = 0.1667$	$1/7 = 0.1429$
check	$1/12 = 0.0833$	$0/7 = 0$
wire	$1/12 = 0.0833$	$0/7 = 0$
fly	$0/12 = 0$	$1/7 = 0.1429$

Table 1: Prior Probability Table

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

Answer :

To determine the label given a bag of words, we need to follow these steps:

- Calculate the probability of a class given a bag of words. Suppose there are n words in the bag, it can be mathematically written as:

$$P(Bag\ of\ Words|Class) = P(w_1, w_2, ..., w_n|Class)$$

Bag of words assumes that word position doesn't matter. Moreover, Naive Bayes has a conditional independence assumption that says for every word, $P(Word|Class)$ is independent of each other. Hence, we can simply multiply $P(Word|Class)$ for every word to get $P(Bag\, of\, Words|Class)$.

$$P(w_1, w_2, ..., w_n|Class) = \prod_{i=1}^{n} P(w_i|Class)$$

- Using Bayes theorem, probability of a class given a bag of words is:

$$P(Class|Bag\, of\, Words) = \frac{P(Bag\, of\, Words|Class) * P(class)}{P(Bag\, of\, Words)}$$

$$P(Class|w_1, w_2, ..., w_n) = \frac{(\prod_{i=1}^{n} P(w_i|Class)) * P(class)}{P(Bag\, of\, Words)}$$

We can use Table 1 to get $P(w_i|Class)$ for every i and use the result from problem 1a to get $P(Class)$.

- Calculate $P(Class = spam|w_1, w_2, ..., w_n)$ and $P(Class = ham|w_1, w_2, ..., w_n)$. Take the bigger probability and finally output its class. Note that P(Bag of Words) is the same for both probabilities so we don't need to compute it.

Using those steps, let's predict the label for these examples:

1. Nigeria

$$P(spam|Nigeria) = \frac{P(Nigeria|spam)P(spam)}{P(Nigeria)} = \frac{(2/12)(3/5)}{P(Nigeria)} = \frac{0.1}{P(Nigeria)}$$

$$P(ham|Nigeria) = \frac{P(Nigeria|ham)P(ham)}{P(Nigeria)} = \frac{(1/7)(2/5)}{P(Nigeria)} = \frac{0.057142857}{P(Nigeria)}$$

Because $P(spam|Nigeria) > P(ham|Nigeria)$, then predicted class label would be ***spam***.

2. Nigeria home

$$P(spam|Nigeria, home) = \frac{P(Nigeria|spam)P(home|spam)P(spam)}{P(Nigeria, home)}$$

$$P(spam|Nigeria, home) = \frac{(2/12)(1/12)(3/5)}{P(Nigeria, home)} = \frac{0.00833}{P(Nigeria, home)}$$

$$P(ham|Nigeria, home) = \frac{P(Nigeria|ham)P(home|ham)P(ham)}{P(Nigeria, home)}$$

$$P(ham|Nigeria, home) = \frac{(1/7)(2/7)(2/5)}{P(Nigeria, home)} = \frac{0.01632653}{P(Nigeria, home)}$$

Because $P(ham|Nigeria, home) > P(spam|Nigeria, home)$, then predicted class label would be ***ham***.

3. home bank money

$$P(spam|home, bank, money) = \frac{P(home|spam)P(bank|spam)P(money|spam)P(spam)}{P(home, bank, money)}$$

$$P(spam|home, bank, money) = \frac{(1/12)(2/12)(1/12)(3/5)}{P(home, bank, money)} = \frac{0.0006944}{P(home, bank, money)}$$

$$P(ham|home, bank, money) = \frac{P(home|ham)P(bank|ham)P(money|ham)P(ham)}{P(home, bank, money)}$$

$$P(ham|home, bank, money) = \frac{(2/7)(1/7)(1/7)(2/5)}{P(home, bank, money)} = \frac{0.00233236}{P(home, bank, money)}$$

Because $P(ham|home, bank, money) > P(spam|home, bank, money)$, then predicted class label would be ***ham***.

Problem 2

***Answer* :**

Let S(n) denotes what is described in the question:

$$S(n) = \sum_{w_1,w_2,...,w_n} P(w_1|start) * P(w_2|w_1) * ... * P(w_n|w_{n-1})$$

To proof that $S(n)$ equals to 1 for any $n \geq 1$, I am going to use induction. Thus, the proof will be divided into two parts:

1. Proving the base case.

   As a base, let's take $n = 1$. This means that there is only one word for every sentence. We can then restate our sum of the probabilities of all sentence formula by only considering w_1:

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

   Notice that since w_1 is the first word, the only possible word that can precede w_1 is the *start* token. This means that $P(w_1|start)$ is equal to $P(w_1)$. However, this will break if there are more than one words in the sentence, which in this case it won't break as we observe only $n = 1$.

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

   The sum of probabilities for every possible outcomes, given they are mutually exclusive, equals to 1. In this case, the possible values for w_1 are all the words in the vocabulary and they are all unique (mutually exclusive).

   $$S(1) = \sum_{w_1} P(w_1) = 1$$

   Thus, it is proved that $S(1) = 1$. Another way to look at it is to convert the form of our conditional probability. Basically, the probability of w_1 occurs given *start* can be calculated by the probability of them occurring together → $P(w_1, start)$ divided by the probability of *start* occurs → $P(start)$.

   $$S(1) = \sum_{w_1} P(w_1|start) = \frac{\sum_{w_1} P(start, w_1)}{P(start)}$$

   Finally, $\sum_{w_1} P(start, w_1)$ can be seen as the probability of *start* followed by any w_1 possible, thus making its just the same as the probability of *start*.

   $$S(1) = \frac{\sum_{w_1} P(start, w_1)}{P(start)} = \frac{P(start)}{P(start)} = 1$$

2. Proving the step case (inductive step).

This step is to prove that if for $n = k$, $S(n = k)$ is equal to 1, then for $k+1$, $S(n = k+1)$ is also equal to 1. In this case, $k \geq 1$ because of our base case.

First, expand the form of $S(k)$:

$$S(k) = \sum_{w_1, w_2, ..., w_k} P(w_1|start) * P(w_2|w_1) * ... * P(w_k|w_{k-1})$$

$$S(k) = \sum_{w_1}(P(w_1|start)*\sum_{w_2}(P(w_2|w_1)*...*\sum_{w_{k-1}}(P(w_{k-1}|w_{k-2})*\sum_{w_k} P(w_k|w_{k-1}))))$$

Since these are bigram probabilities in which every probability depends only on two components w_i and w_{i-1} $(1 \leq i \leq k)$, we can simplify the form above:

$$S(k) = \sum_{w_1} P(w_1|start) * \sum_{w_2}\sum_{w_1} P(w_2|w_1) * ... * \sum_{w_k}\sum_{w_{k-1}} P(w_k|w_{k-1})$$

Next, expand for $n = k+1$:

$$S(k+1) = \sum_{w_1, w_2, ..., w_{k+1}} P(w_1|start) * P(w_2|w_1) * ... * P(w_k|w_{k-1}) * P(w_{k+1}|w_k)$$

$$S(k+1) = \sum_{w_1}(P(w_1|start)*\sum_{w_2}(P(w_2|w_1)*...*\sum_{w_k}(P(w_k|w_{k-1})*\sum_{w_{k+1}} P(w_{k+1}|w_k))))$$

Again, since these are bigram probabilities, we can simplify the form above:

$$\sum_{w_1} P(w_1|start)*\sum_{w_2}\sum_{w_1} P(w_2|w_1)*...*\sum_{w_k}\sum_{w_{k-1}} P(w_k|w_{k-1})*\sum_{w_{k+1}}\sum_{w_k} P(w_{k+1}|w_k)$$

The left parts (all elements except $\sum_{w_{k+1}}\sum_{w_k} P(w_{k+1}|w_k)$) is basically $S(n = k)$. So, we can substitute that and it gives us:

$$S(k+1) = S(k) * \sum_{w_{k+1}}\sum_{w_k} P(w_{k+1}|w_k)$$

We can apply our assumption that $S(k) = 1$ to our formula above:

$$S(k+1) = 1 * \sum_{w_{k+1}}\sum_{w_k} P(w_{k+1}|w_k) = \sum_{w_{k+1}}\sum_{w_k} P(w_{k+1}|w_k)$$

What's left is to determine if the value of $\sum_{w_{k+1}}\sum_{w_k} P(w_{k+1}|w_k)$ equals to 1. To do that, we can first convert the form of our conditional probability. Basically, the probability of w_{k+1} occurs given w_k can be calculated by the probability of them occurring together → $P(w_k, w_{k+1})$ divided by the probability of w_k occurs → $P(w_k)$.

$$S(k+1) = \sum_{w_{k+1}} \sum_{w_k} P(w_{k+1}|w_k) = \sum_{w_{k+1}} \sum_{w_k} \frac{P(w_k, w_{k+1})}{P(w_k)}$$

We can then evaluate the inner summation first ($\sum_{w_k}$):

$$S(k+1) = \sum_{w_{k+1}} \frac{\sum_{w_k} P(w_k, w_{k+1})}{\sum_{w_k} P(w_k)}$$

Notice that $\sum_{w_k} P(w_k, w_{k+1})$ can be seen as the probability of w_{k+1} occurs no matter what w_k is (all possibilities of w_k are included), thus making it's just the same as the probability of w_{k+1}. Meanwhile, for the denominator, it's clear that the sum of probability of all w_k is equal to 1.

$$S(k+1) = \sum_{w_{k+1}} \frac{P(w_{k+1})}{1} = \sum_{w_{k+1}} P(w_{k+1})$$

Again, using the notion that the sum of probability of all events, in this case all w_{k+1}, is equal to 1, we get:

$$S(k+1) = 1$$

Finally, it is proved that for every $k \geq 1$, $S(k) = S(k+1) = 1$.

Because $S(1) = 1$ (base case is proved) and for $k \geq 1$, $S(k+1) = S(k)$ (induction step is proved), then we can say that the statement below is true for $n \geq 1$.

$$S(n) = \sum_{w_1, w_2, ..., w_n} P(w_1|start) * P(w_2|w_1) * ... * P(w_n|w_{n-1}) = 1$$