EECS 349 Machine Learning

Spam Filtering System

Ding Xiang

Oct. 28, 2017

Problem 1

To create a dictionary, we need to determine the probabilities of observing a word in spams and in hams. But there's a problem that if some word doesn't appear in hams or spams, then it may lead to the 0 probability, which may further lead to the problem of "log(0)". So here I use pseudo code by adding 1 for all counting numbers (on both numerators and denominators). So the probabilities are as below.

$$P(a|spam) = \frac{(Counting\ Number\ of\ Documents\ in\ Spams\ with\ that\ word) + 1}{(Total\ Number\ of\ Documents\ in\ Spams) + 1},$$

$$P(a|ham) = \frac{(Counting\ Number\ of\ Documents\ in\ Hams\ with\ that\ word) + 1}{(Total\ Number\ of\ Documents\ in\ Hams) + 1}.$$

Problem 2

Since parse function already makes a UNIQUE words list of email, so here I just check if each word is in the dictionary and if so then update the v_{spam} and v_{ham} based on formulation (7). Otherwise keep v_{spam} and v_{ham} unchanged. At last compare whether v_{spam} is larger than

v_{ham}. If $v_{spam} > v_{ham}$, then return True, else return Flase.

Problem 3

Yes, equation (6) and (7) always return the same results, if there were no issues with underflow or precision. The proof is shown below.

Since log function is monotonically increasing in $(0,\infty)$, which means if $0<x_i<x_j$, then $\log(x_i)< \log(x_j)$. So argmax expression of (6) is equivalent to argmax log(expression) of (6).

$$\begin{aligned} v_{NB} &= \arg\max_{v_j \in V} P(v_j)\prod_i P(a_i \mid v_j) \\ &= \arg\max_{v_j \in V} \log\left(P(v_j)\prod_i P(a_i \mid v_j)\right) \\ &= \arg\max_{v_j \in V} \left(\log\left(P(v_j)\right) + \sum_i \log\left(P(a_i \mid v_j)\right)\right) \end{aligned}$$

So (6) and (7) are equivalent if there were no issues with underflow or precision.

But if there's underflow or precision problem then (6) may not be equivalent to (7), since for an email with a large amount of unique words, the term $\prod_i P(a_i \mid v_j)$ in (6) will go to extremely small (extremely close to 0) which may not be stored or distinguished by computer, so that error generates in this case. But by using equation (7) we could largely get rid of this problem since first using log will make multiplication become summation so the results could be stored; second log will enlarge the subtle distinction of different values that are close to 0

(Because the gradient of log function goes to $+\infty$ as variable goes to 0). So log can capture this subtle distinction pretty well when variable is close to 0 and the log value can be stored well when we use summation in formula (7).

Problem 4

A) In my experiment, the data sets (including training and testing data) I use are "20021010_easy_ham.tar" and "20030228_spam.tar". In specific, I did 5 fold cross validations to see how this Naïve Bayesian Classifier performs. So first, I will divide 2551 emails in "20021010_easy_ham.tar" into 5 groups where each group has 510 ham emails (last group has 511 ham emails), and divide 501 emails in "20030228_spam.tar" into 5 groups where each group has 100 spam emails (the last group has 101 spam emails). Then each round choose 4 groups of "20021010_easy_ham.tar" and 4 groups of "20030228_spam.tar" as training data to build dictionary and calculate the prior probability, the rest 1 group of "20021010_easy_ham.tar" and 1 group of "20030228_spam.tar" will become testing data. Rotate this process for all 5 groups. Compare the average error rate between Naïve Bayesian Classifier and prior probability classifier.

So in each round training data and testing data are different. The reason I chose those as data sets, which are all from the given two folders, is that this will be more FAIR to compare whether Naïve Bayesian Classifier is better than prior probability classifier or not. If I use a totally different folder with new statistics and make them as test data, then the comparison result will be largely decided by how I chose that folder and how the prior probability of spam emails in that folder. For example, if I train the classifier by using “20021010_easy_ham.tar” and “20030228_spam.tar” but test the results on “20050311_spam_2.tar.bz2”, since training data has prior probability of spam 501/(2551+501)=16.4% < 50%, so all the prediction will become ham, which means the prior probability classifier will label all the emails in “20050311_spam_2.tar.bz2” as ham emails. Then the error rate of prior probability classifier becomes 100%, and then no matter how Naïve Bayesian Classifier performs, it will never be worse than prior probability classifier. To get avoid of the similar unfair comparisons, I chose the data sets in the way I mentioned earlier.

So totally 2551+501=3052 files in the data. The percentage of spam in data is 16.4%. I don’t think they are representative of the spam for today since first, the statistic was done in 2002 and 2003 which is 15 years ago, the email content of spams and hams may changed dramatically; Second based on the statistic of 2017 Q2 from SECURELIST (https://securelist.com/spam-and-phishing-in-q2-2017/81537/), the spam rate is about 57.47%, which is much higher than 16.6% in 2002 and 2003.

B) I will do the experiment as I mentioned. I will do 5 fold cross validation. So first, I will divide 2551 emails in “20021010_easy_ham.tar” into 5 groups where each group has 510 ham emails (last group has 511 ham emails), and divide 501 emails in “20030228_spam.tar” into 5 groups where each group has 100 spam emails (the last group has 101 spam emails). Then each round choose 4 groups of “20021010_easy_ham.tar” and 4 groups of “20030228_spam.tar” as training data to build dictionary and calculate the prior probability, the rest 1 group of “20021010_easy_ham.tar” and 1 group of “20030228_spam.tar” will become testing data. Rotate this process for all 5 groups. Compare the average error rate between Naïve Bayesian Classifier and prior probability classifier.

In each round, the error rate is defined as below.

$$error_i = \frac{Number\ of\ Misclassified\ Emails}{Number\ of\ Total\ Testing\ Emails}, i = 1,2, \ldots ,5$$

The average error rate is

$$AverageError = \frac{1}{5}\sum_{i=1}^{5} error_i$$

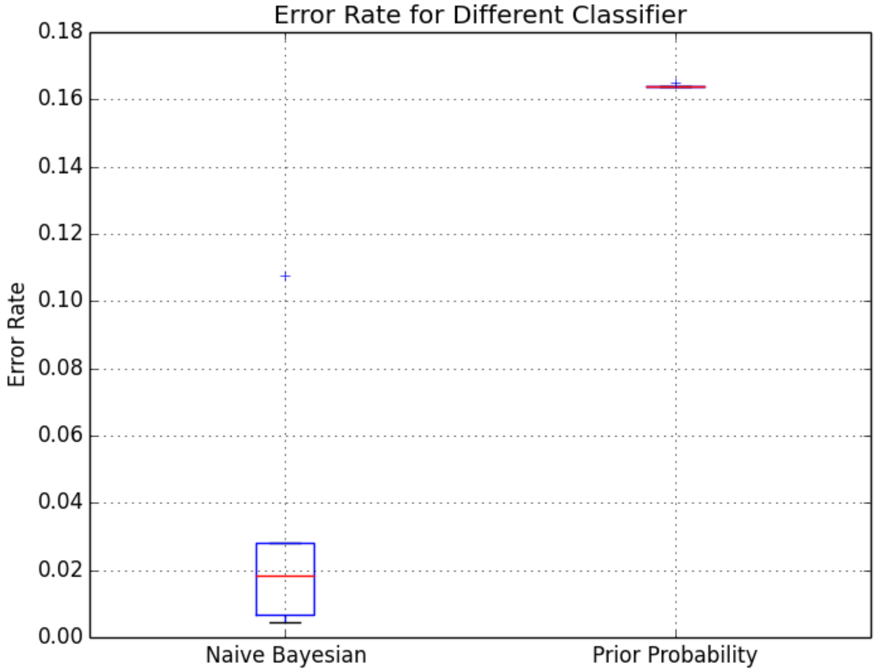
C) Yes. The Naïve Bayesian Classifier is better than the classifier just using prior probability of spam. The experiment results are shown below.

Table 1. Error Rates for Different Classifiers

Error Rate	Average	Group 1	Group 2	Group 3	Group 4	Group 5
NB*	3.31%	10.77%	1.83%	0.67%	2.83%	0.46%
PP*	16.41%	16.39%	16.39%	16.39%	16.39%	16.5%

*where NB denotes Naïve Bayesian Classifier; PP denotes Prior Probability Classifier.

The boxplot of two different classifier error rate is shown in the following figure.



From the table and picture above we can see that Naïve Bayesian Classifier performs better than Prior Probability Classifier in each group comparison and also performs much better in average.