Artificial Intelligence - CSE 537

Final Report Project 05: Machine Learning

Submitted By:

Tarun Jindal (112687340)

Anuroop Katiyar (112609023)

Spam Filter

***Approach**:*

For training our classifier we have taken the train data set and created 2 separate dictionaries having ham words and spam words. Each dictionary stores the unique words given in the train data set along with the value as the frequency of occurrence of that words. While parsing the mails in the train data set, we have also counted other parameters like no. of spam/ham mails, no. of unique words in ham/spam mails and total no. of words in spam/ham mails.

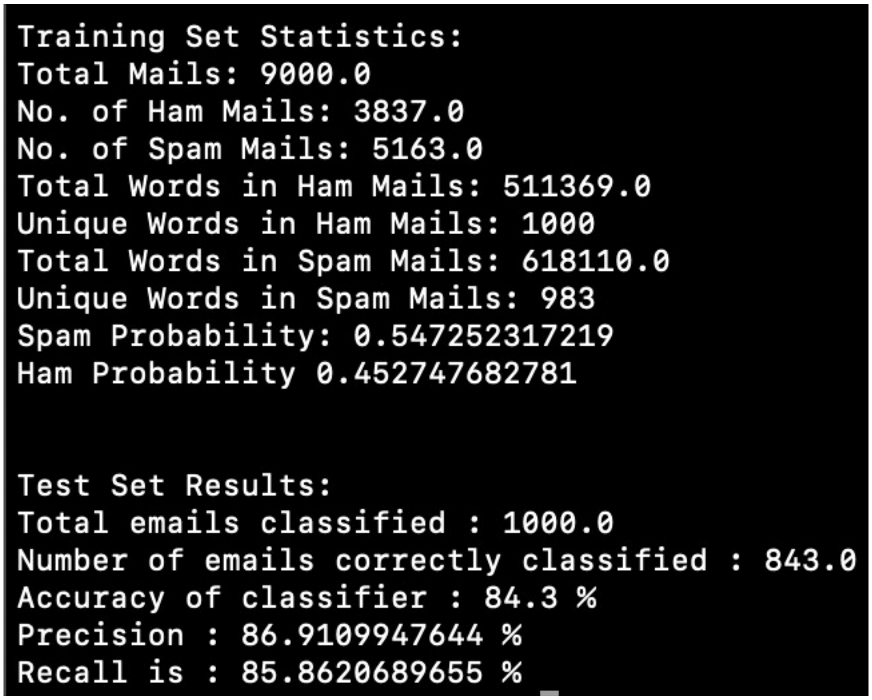
After parsing all the data, we calculate the prior probabilities of spam or ham mail like P(spam|D) and P(ham|D). Now, we scan each mail in the test data set and for each word in the test mail, we calculate spam/ham probability i.e. P(word|spam) and P(word|ham). We classify the mail as spam or ham on the basis of one that has higher probability.

Optimization:

To improve our accuracy, we have done the below 2 changes:

1. Multiplying lots of probabilities, which are between 0 and 1 by definition, can result in floating point underflow. Since log(xy) = log(x) + log(y), we performed all computations by summing logs of probabilities rather than multiplying probabilities.
2. In cases of words which our not present in our dictionary, we use Laplace Smoothing (alpha=1) to assign a small probability instead of zero.

Results:



```
Training Set Statistics:
Total Mails: 9000.0
No. of Ham Mails: 3837.0
No. of Spam Mails: 5163.0
Total Words in Ham Mails: 511369.0
Unique Words in Ham Mails: 1000
Total Words in Spam Mails: 618110.0
Unique Words in Spam Mails: 983
Spam Probability: 0.547252317219
Ham Probability 0.452747682781


Test Set Results:
Total emails classified : 1000.0
Number of emails correctly classified : 843.0
Accuracy of classifier : 84.3 %
Precision : 86.9109947644 %
Recall is : 85.8620689655 %
```