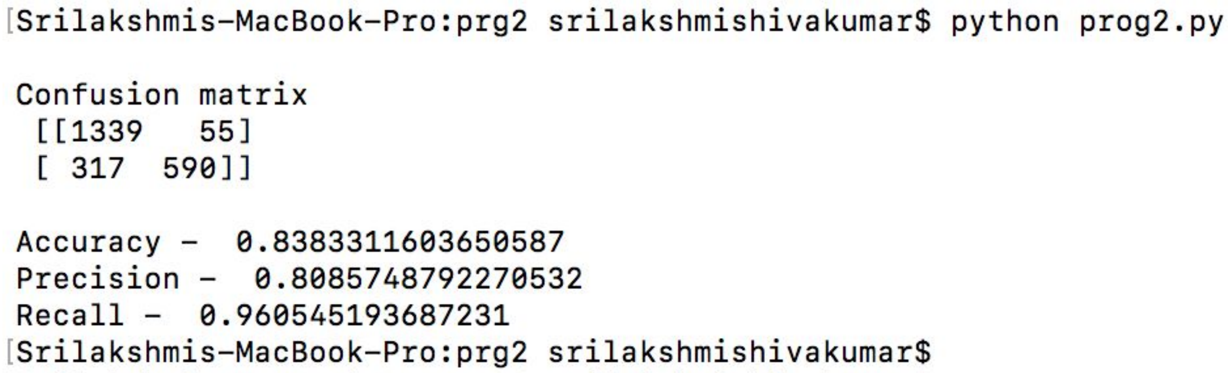
CS 545: Machine Learning, Spring 2018
Programming Assignment #2

Description

In the programming assignment, we classify Spambase data using Gaussian Naïve Bayes and Logistic Regression. The data consists of spam and non-spam mails which are have labels 1 and 0 respectively. The entire dataset consists of 4601 instances which are split into 50% train and test data, each having 2300 instances, 40% spam and 60% non-spam. The prior probability of spam is 40% and non-spam is 60%. Based on the 57 features in the train dataset, we calculate the mean and standard deviation for spam and non-spam data. The standard deviation is changed to 0.0001 if it's 0, to avoid division by zero. Next, we will use Gaussian Naïve Bayes algorithm to obtain the probabilities.

Result



```
[Srilakshmis-MacBook-Pro:prg2 srilakshmishivakumar$ python prog2.py

 Confusion matrix
  [[1339   55]
  [ 317  590]]

 Accuracy -  0.8383311603650587
 Precision -  0.8085748792270532
 Recall -  0.960545193687231
[Srilakshmis-MacBook-Pro:prg2 srilakshmishivakumar$
```

The train and test data consists of approximately 40% spam and 60% non-spam data. The accuracy is 83%. From the confusion matrix, 372 mails are classified incorrectly. Precision and recall are also calculated from the confusion matrix. Gaussian Naïve Bayes accuracy takes less time to train the data however, accuracy is not very good.

Do you think the attributes here are independent, as assumed by Naïve Bayes?

The initial assumption made by Naïve Bayes is that all the attributes are independent. However, it is not the case. For example, the frequency of one word may not be completely independent of the other. Two words can be closely related such as synonyms or one followed by another which describes their atypical behavior. In this way, there can be dependency at the same level. It is incorrect to calculate accuracy based on the frequency of words. It decreases the accuracy.

Does Naïve Bayes do well on this problem in spite of the independence assumption?

In spite of the independence assumption, Naïve Bayes does a moderate job. Due to the frequency of words, one gets more importance over another. Here, the meaning of the sentences are not taken into account. However there is always scope for improvement. Naïve Bayes can be improved if it considers features which have more statistical meaning and thus results in better probability, mean and standard deviation. And hence improve classification.