Spambase Classification
By: Samarth Kedilaya

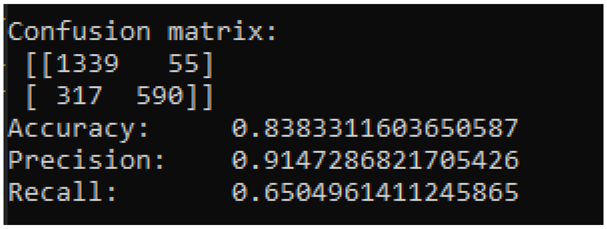
Description:

This project is based on Gaussian Naïve Bayes algorithm. Here, we try to classify the spambase data with the help of Gaussian Naïve Bayes and Logical regression model. The word “spam” concept here is diverse which refers to various advertisements for products/websites, fast money schemes, chain letters or porn etc.

The spambase data mentioned here consists of a collection of spam and non-spam mails which are identified with labels 1 and 0 respectively. The spambase dataset consists of 4601 instances in total which are then split into 50% train data and remaining 50% as test data. Both sets have 2300 instances each with 40% spam and 60% non-spam mails. We determine that the prior probability of spam as 40% and non-spam as 60%.

We try to calculate the mean and standard deviation here for both spam and non-spam data based on the 57 attributes in the train dataset. Also, we change the standard deviation is changed to 0.0001 whenever it is encountered as 0 so as to avoid division by zero error thereby assigning it a minimal standard deviation. Then, we use the Gaussian Naïve Bayes algorithm to obtain the required probabilities.

Results:



```
Confusion matrix:
 [[1339   55]
 [ 317  590]]
Accuracy:     0.8383311603650587
Precision:    0.9147286821705426
Recall:       0.6504961411245865
```

We found that both the train and test data consist of approximately 40% spam and 60% non-spam data and the accuracy obtained is 83.83%. With the help of confusion matrix, we also found that there were 372(317+55) mails which were classified incorrectly. Precision and Recall were also calculated from the confusion matrix. Although Gaussian Naïve Bayes accuracy takes less time to train the data, accuracy is not that great.

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

The presumption made by Naïve Bayes was that every one of the attributes were independent. However, it proved to be incorrect. For example, the recurrence of single word may not be totally independent of the other. Two words can be firmly related, for example, synonyms or one pursued by another portrays their atypical behavior. Similarly, there can be dependency at a same level. It might be wrong to calculate the accuracy based just on the recurrence of words. It results in reduction of the accuracy.

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

Regardless of the independence presumption, Naïve Bayes does a moderate job. Because of the recurrence of words, one gets more significance over another. Here, the importance of the sentences is not considered. Anyway, there is a considerable scope for improvement. Naïve Bayes can be improved if it considers features that have more statistical significance thereby resulting in better probability, mean and standard deviation, hence improving the classification.