CS 445/545: Machine Learning
Programming Assignment #2

Dataset : Spambase dataset from UCI ML repository
https://archive.ics.uci.edu/ml/datasets/spambase

The goal of this assignment is to classify the Spambase dataset using Gaussian Naive Bayes Classification. The dataset consists of emails that are labelled as 1(spam) and 0 (not spam). The entire dataset consists of 4601 instances which is split to 50% train and 50% test data. Each has 2300 instances of 40% spam and 60% not spam emails.

We need to calculate the probability using Gaussian Naive Bayes Classification. The prior probability of spam and non-spam is 40%: 60%. Based on the 57 features in the training dataset, we calculate the mean and standard deviation for spam and non-spam data. The value of standard deviation is changed to 0.0001 if it's 0, to avoid division by zero.

When we use Gaussian Naive Bayes to calculate the accuracy, it takes less time to train the data and the accuracy is less. The train and test data consist 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.

Naive Bayes initially assumes that all the attributes are independent. However, it is not true. For example, the frequency of one word may not be completely independent of the other (in the case of synonyms). In this way, there can be dependency at the same level. Accuracy decreases if we use the frequency of words for its calculation.

Naive Bayes classifier can be improved if it considers features which have more statistical meaning and thus result in better probability, mean and standard deviation.

Results:

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

Accuracy: 0.8383311603650587
Precision: 0.9147286821705426
Recall: 0.6504961411245865