CS 545: Programming #2

Kirtan Patel PSU ID: 973582669

Classification using Naïve Bayes

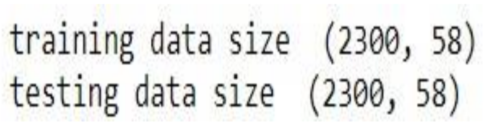
Description:

Classification on “spam dataset” is performed using Naïve Bayes classification.

The dataset contain 4601 instances (with 40% spam and 60% non-spam messages in them), there are total 57 attributes and the last column of “spambase.data” denotes whether the e-mail is spam (1) or not spam (0).

Creating Training and testing set

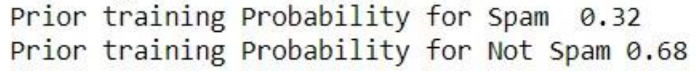
For training and testing of the model we divide the 4601 instances into half to have 2300 in both training and testing dataset, each contains 40% spam and 60% non-spam messages to replicate the statistics of original dataset.



```
training data size  (2300, 58)
testing data size  (2300, 58)
```

Creating probabilistic model

Prior probability of each class (spam and not spam) for training dataset is calculated

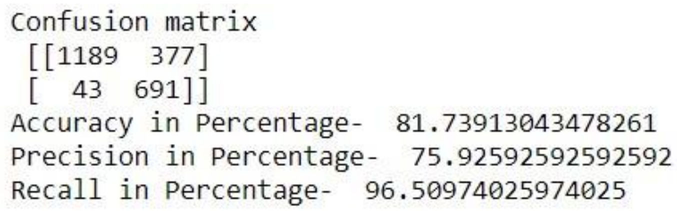


```
Prior training Probability for Spam  0.32
Prior training Probability for Not Spam 0.68
```

The Mean standard deviation for each feature (57) is calculated, and if any of the feature has a standard deviation of 0, that feature will be assigned with minimal standard deviation (0.0001) to avoid divide-by-zero error in Gaussian Naïve Bayes.

Naïve Bayes on the test dataset:

After performing classification on testing dataset using Naïve Bayes the confusion matrix, accuracy, precision, and recall obtained are as follows.



```
Confusion matrix
 [[1189  377]
 [  43  691]]
Accuracy in Percentage-  81.73913043478261
Precision in Percentage-  75.92592592592592
Recall in Percentage-  96.50974025974025
```

Attributes independence assumption by Naïve Bayes

Bayes theorem assumes every feature to be independent without any correlation with other attributes, but this assumption has limitations, as in real life it's not possible to get complete independence.

As for this "spam database" in which most attributes indicated how frequently the words have been displayed in the email and considering each independent frequency of words for classification, this would neglect the context of the message, which could be derived using the combination of the words used in the mail. So, I think there could be some level of dependency in this dataset.

Naïve Bayes Performance

The overall accuracy of the model is 81.7% which is not bad considering the fact that if performed classification based on feature independence. But this accuracy could be improved if correct data features are given, like instead of giving word frequency, if it's given string frequency the model could perform better.