CS 545: Machine Learning, Spring 2020

Name: Sree Vandana Nadipalli

Classification using Naïve Bayes

Description:

Classification on "Spam dataset" is performed using Naïve Bayes classification.
The Spam dataset contains 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 test set:

- For training and testing of the model the entire data is divided, to have 2320 instances in both train data and test data, each containing 40% spam and 60% not spam to replicate the statistics of original dataset.

```
training data size  (2320, 58)
testing data size  (2320, 58)
```

Creating probabilistic model:

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

```
Prior training Probability for Spam  0.4
Prior training Probability for Not Spam 0.6
```

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

Naïve Bayes on the test data:

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

```
Confusion matrix
 [[1067  326]
 [  62  865]]
Accuracy in Percentage-  83.27586206896552
Precision in Percentage-  76.597272074659
Recall in Percentage-  94.50841452612931
```

Attributes independence assumption by Naïve Bayes

The naïve Bayes assumes, each attribute exists independently 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 dataset" in which, most attributes indicated how frequently the words has 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 83% which is ok, considering the fact that it 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 (contains similar strings and strings with same meaning implying the context) the model could perform better.