Programming Assignment 2

Name – Amarjit Jadhav

Report

Description-

In this programming assignment I have written a Naïve Bayes classifier for classification of spam base database, which consist of spam and non-spam mails. The classification or the labeling of the classes is as follows :

- spam mail class is 1
- non-spam mail is class 0 is classic as 1

We are dividing the data set into two sets 50% training data and 50% testing data. The probability of the spam emails is 40% and probability of the non-spam emails is 60%.

In this we are calculating the standard deviation and mean for the training set for each and every attribute and then use obtained values to calculate probabilities on the testing data. Depending the probabilities, using logarithmic sum of probabilities, we are classifying the class. At the end, these values are used for calculating accuracy, precision, recall and for the creation of the confusion matrix. Depending upon the standard deviation values if it is 0 for any of the attribute, then that is replaced by lowest value .0001. We will use Gaussian Naïve Bayes algorithm to compute the probabilities.

Output -

```
Confusion matrix:
 [[917 501]
 [ 36 847]]
Accuracy Value:          0.7666232073011734
Precision Value:         0.6283382789317508
Recall Value:    0.9592298980747452
```

Spam mail probability is: 0.3760869565217391

Non-spam mail probability is: 0.6239130434782609

Programming Assignment 2

Result-

From the above screen it is evident that the accuracy of the Naïve Bayes is not that good and is below expectation.

With the help of confusion matrix from the above screenshot we can get the number of examples which are classified correctly and incorrectly. The number of examples classified correctly and incorrectly can be seen in the confusion matrix. Precision and recall values can also be spotted.

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

As per the assumption we made, the attributes which are considered are independent. Although, we can have an anomalous behavior where presence of a word indicates the presence of other word with probability of 90 % which indicates dependence. for example we can consider the word discussion in the mail, have high probability of having word meeting in the same mail or health insurance mail will have medical in that mail. So this way there could be dependence at same level.

Does Naïve Bayes do well on this problem despite the independence assumption? Speculate on other reasons Naïve Bayes might do well or poorly on this problem.

As it is clearly seen in results/outputs that Naïve Bayes does not perform well for the given model. Although we considered independence assumption, the obtained accuracy is still not that good.

There is a lot scope for of improvement. Naïve Bayes can perform well and output better result if we will only consider features/attributes which have some statistical meaning and are more important i.e those have more presence in the data. If we use only these attributes in for calculation of mean, standard deviation and probabilities which are required for classification. By using this technique there is scope of improvement in performance of this classifier.