ML Assignment-1

Rishab Nahar-2018A7PS0173H

Sarvesh Khetan-2018A4PS0947H

Samkit Jain- 2017B2A71723H

Spam detection by Naive Bayes Classifier

1.1 Objective:

The aim of this assignment is to implement a simple Naive Bayes classifier to classify mails as spam or not. The model is tested and trained by using the 7-cross validation method. Tokenization is done as remove the stop words and punctuations from the given dataset as they don't contribute to the sentiment of the text. Laplace smoothing is performed to avoid the problem of division with zero.

Hence when a new message comes in, our Naive Bayes algorithm will make the classification based on the results it gets to these two equations below, where "w_1" is the first word, and w_1,w_2, ..., w_n is the entire message:

$$P(w_i|Spam) = \frac{N_{w_i|Spam} + \alpha}{N_{Spam} + \alpha \cdot N_{Vocabulary}}$$

Where N_{wi} is the frequency of a word given it is a spam mail, N_{Spam} is the number of spam mails and $N_{vocabulary}$ is no of words present in the dataset. Also alpha is the smoothening parameter whose value is set to 1 in out model.

1.2 Model Explanation

Naive Bayes spam filtering is a baseline technique for dealing with spam that can tailor itself to the email needs of individual users and give low false positive spam detection rates that are generally acceptable to users.

The dataset given is firstly preprocessed by removing all the stop words and punctuation marks. In our model we divide the given dataset into 7 parts using the concept of ‘k-fold classification’. After all the processing the model is trained individually by each of the 7 datasets. A Numpy array is used which stores the frequency of each word given it is spam or not. This frequency is used to calculate the Bayesian probability while testing our model. Both probabilities (given a word whether a mail is spam or not) are calculated. Both the calculated probabilities are then compared and a mail is classified as spam or not based on the factor which probability is greater.

1.3 Results

The results from the 7-fold classification are shown below: -

Fold No	True Positive	False Positive	Accuracy
1	115	33	77.70%
2	117	25	82.39%
3	124	18	87.32%
4	115	27	80.99%
5	109	33	76.76%
6	111	31	78.17%
7	115	27	80.99%

Results of overall 7-Fold Cross Validation are as follow:

Rightly Classified Mails :-806.0
Wrongly classified Mails: -194.0
Accuracy: - 80.62%

1.4 Limitations

- Bayesian spam filtering may be susceptible to Bayesian Poisoning, a technique used by spammers in an attempt to degrade the effectiveness of spam filters that rely on Bayesian filtering.
- Another technique used to try to defeat Bayesian spam filters is to replace text with pictures, either directly included or linked. The whole text of the message, or some part of it, is replaced with a picture where the same text is "drawn". The spam filter is usually unable to analyze this picture, which would contain the sensitive words.
- Words that normally appear in large quantities in spam may also be transformed by spammers. The recipient of the message can still read the changed words, but each of these words is met more rarely by the Bayesian filter, which hinders its learning process. As a general rule, this spamming technique does not work very well, because the derived words end up recognized by the filter just like the normal ones.