CS5691: Pattern Recognition and Machine Learning

Assignment 3

Amishi Panwar

Roll Number – CS22M009

REPORT

Dataset Used:

https://www.kaggle.com/code/vermichel/intro-to-nlp-spam-classifiier

For the spam classifier, I have used the data that is publicly available on kaggle, link given above.

Data Preprocessing:

First of all, we need to preprocess data as this dataset contains irrelevant information. For the spam classifier the words and symbols which are not useful in classification of spam and non spam mails can be removed from the dataset, so I cleaned the data by removing all symbols except letters, replaced all gaps with spaces, removed 'b' from the beginning of each text, converted words into lowercase. I have also used PorterStemmer for stemming `words.`

Algorithm:

I have used Naive Bayes for classification. First I calculated the number of times a word appears in the spam mails and counted the word only once if it appears in one mail multiple times. We get the probability of each word appearing in spam mail or non spam mail by dividing the word count by the number of spam or non spam mails we have.
To test any mail for spam or non spam, we calculate probability of mail being spam and nonspam, whichever is greater accordingly we classify it as spam or non spam.
It is calculated by multiplying the probability of a word if it is present in our test mail and complement of the probability if it is not present in our test mail.

I have also tried other algorithms but naives bayes is giving highest accuracy.

I used Svm algorithm in which training data was divided by a hyper-plane which has maximum margin. And two separations was classified as spam and non spam.

Decision tree resulted in overfitting and thus the accuracy was less.