Report: Assignment 3 (CS5691: Pattern Recognition and Machine Learning)

CS22M008

The objective is to build a spam classifier from scratch i.e. to classify a set of emails as spam or ham.

Dataset
I have used a publicly available dataset: link
The dataset has emails in two folders namely spam containing the spam emails, and ham containing the ham emails..

I have used **Naive Bayes Classifier** to classify the emails.
A Naive Bayes classifier is a probabilistic algorithm and is based on the Bayes theorem.
It comes under generative models.

Data Preprocessing
First I cleaned the dataset by removing all '\n' and replacing them with " ". Then I tokenized the emails.
Then for every token I checked whether all characters in the token are alphabets.
I converted all the tokens obtained into lowercase.

Training the Model
I created a super dictionary which contains all the words in the dataset and two other dictionaries each containing the words in the ham and the spam dataset respectively.
I computed the frequencies of each word in the corresponding dictionary.

Feature Engineering
Feature extraction is done using the top 2200 words from the super dictionary (dictionary containing all the words in the dataset).

Prediction
To classify each email, I calculated the probability: (using Bayes Theorem)
P(Spam/Mail)=P(Mail/Spam)*P(Spam)/P(Mail)
and P(Ham/Mail)=P(Mail/Ham)*P(Ham)/P(Mail)

P(Spam) and P(Ham) are the prior probabilities of a spam and ham mail respectively.
P(Mail) is the Evidence, which we do not need to calculate as it would get cancelled when we find the ratio of P(Spam/Mail) and P(Ham/Mail).

P(Spam/Mail) is the conditional probability of a given email being spam and P(Ham/Mail) is the conditional probability of a given email being non-spam.

If the ratio of P(Spam/Mail) and P(Ham/Mail) is less than 1 we classify the test email as Ham else we classify it as spam.

Observation:
The **accuracy** of Naive Bayes Classifier is 80% to 95%.