Assignment 3 Report

The dataset for training was used from the github repository vincechan/**spam-detection**. **The dataset is provided as a series of emails and the corresponding labels spam or ham. This dataset as nearly equal amount of spams and hams.**

total no. of samples: 2100
total no. of spam samples: 1043
total no. of ham samples: 1057

The training dataset is also chosen randomly from the whole dataset to keep equal numbers of spams and hams during the training. The split for training to testing can be set up in any value, preferably 80% training and 20% testing is good.

The dataset is available along with code submitted.

I have used the Naïve Bayes algorithm for classification purpose. For Naïve bayes algorithm we need data to be available in a vector of numbers format where each number represents whether a particular word is present in the email or not.

Data pre-processing is done using the sklearn's feature_extraction.text.TfidfVectorizer library. This library converts the text data to vector of frequency of each word. The frequency of words are then weighted and the words that are generally used are dropped as part of feature extraction.

We have extracted about 650 words for our model to work with according to the data available. These words will act as the vocabulary for our model.

Naive Bayes algorithm:

Estimate the models parameters:

: find the probability of an email being spam - p . This is the fraction of spam emails in training data.

: find the probability of a word appearing in the email given the email is spam or not. This is the fraction of spam/ham labelled emails that contain the jth world

:After finding the parameter of the model, we can now check how the model works for the testing data.

:For doing this we will find 2 quantities for each test email:

Probability of spam given the test email and probability of ham given the test email.

:We see which quantity is greater and accordingly predict the mail is spam or not

Results

The error in testing comes to about 99%. The naïve bayes model works quite well for email spam detection and has been a popular choice from the early 90's. We assume in naïve bayes that he words appearing the emails are conditionally independent, even though this may not be true in practical scenarios.

I have created a function that reads the emails and predicts whether the email is spam(1) or ham(0). It return the list of all predictions.

Multiple other algorithms can also be used to do the classification task especially SVM and Ada Boost.