PRML Assignment-3 Report

SPAM or HAM

Aim: Spam Classifier from scratch using SVM

Training data:

The dataset is taken from Kaggle (UCI Machine learning repository). It contains one set of SMS messages in English of 5,574 messages, tagged accordingly being ham (legitimate) or spam.

Procedure:

Step-1 : Importing the data.

Step-2 : Converting the class labels from string to numeric form

Step-3 : Data Cleaning

1) Removing the punctuations from the emails
2) Converting all the emails to lower case
3) Splitting the words of emails and creating a dictionary with unique words. There are a total of 9483 unique words in the data.

Step-4 : Feature extraction – Here the feature is the number of times a word appears in a mail.

Step-5 : Vectorizing the dataset

Step-6 : Splitting the data into training and testing for cross validation

Step-7 : Analyzing the percentage of spam and ham messages in test and training data to see both are nearly equal.

Step-8 : Import SVM from sklearn and use linear and rbf kernels to train the model

Step-9 : Test the model using testing data.

Step-10: Calculating the score gives 99 % accuracy for linear kernel and 97.8 % accuracy for rbf kernel.

Testing the model's performance:

The given email1.text(rtf) and email2.txt (rtf) are converted to text using striprtf in python.

Then the text is stored in the form of csv. The emails can be checked by giving the path of the file in function2.