PATTERN RECOGNITION AND MACHINE LEARNING (CS6015)

ASSIGNMENT 3

Submitted by:

Deepak Mori

CS22M039

MTech (CSE)

TASK:

To build a spam classifier from scratch. No training data will be provided. You are free to use whatever training data that is publicly available/does not have any copyright restrictions (You can build your own training data as well if you think that is useful). You are free to extract features as you think will be appropriate for this problem. The final code you submit should have a function/procedure which when invoked will be able to automatically read a set an email from a folder titled test in the current directory. Each file in this folder will be a test email and will be named 'email#.txt' ('email1.txt','email2.txt', etc). For each of these emails, the classifier should predict +1 (spam) or 0 (non-Spam). You are free to use whichever algorithm learnt in the course to build a classifier (or even use more than one). The algorithms (except SVM) need to be coded from scratch. Your report should clearly detail information relating to the data-set chosen, the features extracted and the exact algorithm/procedure used for training including hyperparameter tuning/kernel selection if any. The performance of the algorithm will be based on the accuracy on the test set

Solution:

I have used Support Vector Machine (SVM) model to classify the email as spam or ham.

Support Vector Machine:

Support vector machines (SVMs) are supervised machine learning algorithms that can be used for both classification and regression tasks**.** However, it is mainly used in classification problems. The SVM algorithm plots each data item as a point in n-dimensional space (where n is the number of features it possesses) where the value of each feature is the value of a specific coordinate. The dimension of the hyperplane depends on the number of features. If the number of input features is 2, the hyperplane is just a line. If the number of input features is 3**,** the hyperplane will be a 2D plane.

Classification is then done by finding the hyperplane that distinguishes the two classes very well.

Advantages of SVM:

- Effective in high dimensional cases.
- Storage is efficient because it uses a subset of training points in a decision function called support vectors.
- The decision function can be a different kernel function and can be a custom kernel.

PROCEDURE:

- Firstly, I imported the required libraries.
- Load the dataset.
- Checking the information of dataset and cleaning the data.
- Now pre-processing the data
    1) The entire email is converted into lower case.
    2) Removing the special characters.
    3) Removing the stop words and punctuations.
    4) Stemming.
- Now model the data
  Splitting the data into X and Y i.e., training and testing and converting text into integer using count vectorizer ().
- Now apply SVM algorithm on the training data and test the email that are in the test folder for testing emails as spam or ham.