Jashaswimalya Acharjee
CS22E005
November 17, 2022

CS 5691: Pattern Recognition and Machine Learning
Assignment: 3
Course Instructor : Arun Rajkumar

Question: SPAM or HAM?

In this assignment, you will 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 a emails 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.

Answer:

Dataset Details: For the Dataset, I have used a freely available dataset for spam or ham classification task, having no copyright infringements — called Enron-Spam Classification Dataset. The dataset has `3672` Legitimate or non-spam emails, while `1500` emails are spam in the dataset. Total number of emails (legitimate + spam): 5975

Using the knowledge of Basic NLP, I have performed some basic level Text Preprocessing, Tokenizing and filtering of StopWords using suitable libraries and functions to build a dictionary of features and transform documents (containing emails) to feature vectors. I have removed the stop words in order to improve the analytics. Also, I have split the training dataset in 80:20 fashion to validate.

Executing the Code: Make sure the host PC has all the required libraries in order to run the given piece of code which is attached inside the `.zip` file uploaded on the Moodle Instance.

Evaluation Metrics: For classification we have incorporated **Support Vector Machines (SVMs)** and have tested our downloaded custom dataset on several kernels and comparitively here are the results of the same:

Accuracy obtained for Kernel —

- **Linear:** 0.9816363338252172
- **Polynomial:** 0.6205935399245778
- **Rbf;** 0.9627807837350385
- **Sigmoid:** 0.9404820462370881