CS5691: Pattern recognition and machine learning
Programming Assignment 3
Course Instructor : Arun Rajkumar.
Release Date : December-4, 2020
Submission Date: On or before 5 PM on December-24,2020

SCORING: There is 1 question in this assignment which contributes to 20 points towards your final grade.

DATASETS Check the README file for description.

WHAT SHOULD YOU SUBMIT? You should submit a zip file titled 'Solutions_rollnumber1 _rollnumber2.zip' where rollnumber1 and rollnumber2 are roll numbers of the members of the group. Your assignment will NOT be graded if it does not contain all of the following:

- A text file titled 'Participants.txt' with names and roll numbers of members.
- A PDF file which includes explanations regarding each of the solution as required in the question. Title this file as 'Report.pdf'
- Source code for all the programs that you write for the assignment clearly named.

CODE LIBRARY: You are expected to code all algorithms from scratch. You cannot use standard inbuilt libraries for **computations**. The only allowed library are those that compute the Eigenvectors and Eigenvalues of matrices. If your code calls any other library function for computation, it will fetch 0 points. You are free to use inbuilt libraries for plots. You can code using either Python or Matlab or C.

GUIDELINES: Keep the below points in mind before submission.

- Plagiarism of any kind is unacceptable. These include copying text or code from any online sources. These will lead to disciplinary actions according to institute guidelines.
- Any graph that you plot is unacceptable for grading unless it labels the x-axis and y-axis clearly.
- Don't be vague in your explanations. The clearer your answer is, the more chance it will be scored higher.

LATE SUBMISSION POLICY You are expected to submit your assignment on or before the deadline to avoid any penalty. Late submission incurs a penalty equal to the number of days your submission is late by.

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). Two sample emails your classifier will be tested on can be found in the folder *test*. 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.