IE615: Data Analytics in Operations Research **Jul-Nov 2017**

Assignment 3: Oct 22

Instructions: *You are free to code in any language of your choice. Your code should ask to input training set and output respective learners upon execution. Discussion among the class participants is highly encouraged. But please write your own code to make sure that you understand the algorithms. Submit the assignment by* 1^{st} *Nov* 2017.

Question 1 (Alpha Boost) *Apply AlphaBoost and AdaBoost on the binary dataset that is provided. Plot a graph of train error and test error and train score and test score for various rounds of the algo. Also plot the test error graph for various values of* α*. Note that the values of* α *and* T *need to choosen together for good results. The output of the algorithm should be a confusion matrix and the performance in terms of precision, recall, f1-score and support. You can see the sample output in the attached jupyter notebook.*

Question 2 (PCA) *Apply PCA on digits dataset that is provided using both SVD and eigenvectors based schemes. Plot a SCREE graph and proportion of variance explained graph. Report the first* 10 *principal components for both the methods. Reconstruct the images using the features limited to these* 10 *principal components, for both the methods.*

Question 3 (Visualizing clusters) *Apply PCA on digits dataset to reduce the dimension to* $\mathcal{R}^2$*. Apply agglomerative clustering (HAC algo) on this dimensionality reduced dataset to form cluster and plot the cluster. Agglomerative clustering should use 'complete', 'single' and 'average' link functions.*

Question 4 (kMeans) *Also apply K-means on both the original feature space as well as the the dimensionally reduced dataset for different values of* $K = 5, 6, \cdots, 15$*.*

Plot the clustering performance using silhouette (averaged) for different number of clusters.

Note: please review the attached jupyter notebook to understand the implementation and output format.

You are expected code each algorithm. You can use only packages for optimization modules like, LP or Convex programs, but not packages for the algorithms itself.

Your report should contain all the results similar to output of jupyter notebook