Homework 2

EEL 4774 & 6777 Data Analytics

1 Dimensionality Reduction with PCA

[50 pts] Download “url-data.txt”. This dataset contains 1000 websites as instances with 64 numeric features. The first column contains the labels as 0 for benign and 1 for malicious. The other 64 columns hold the features.

a) [40 pts] Using PCA determine the minimum number of principal components that maintain 95% and 99% of the variance.

{Hint: You can compute the maintained variance ratio using $\frac{\sum_{i=1}^{M}\lambda_i}{\sum_{i=1}^{D}\lambda_i}$, where λ_i is the ith greatest eigenvalue, M is the number of principal components used, and D is the number of original features. You can also use built-in functions from common Python libraries such as numpy, scipy, sklearn, etc.}

b) [10 pts] Show the scree plot for eigenvalues.

2 Clustering with K-means

[50 pts] Download “iris.csv”. The task is to find natural groupings which potentially correspond to different flower types. Apply PCA with 1, 2, 3, and 4 principal components. For each number of principal components, cluster the projected data into 3 clusters using K-means. Compare the 4 cases (i.e., PCA with 1, 2, 3, and 4 components) in terms of:

a) [25 pts] the average silhouette score of clustering results

c) [25 pts] the adjusted rand index of clustering results

{Hint: In each part, you should explain what number of principal components gives the best clustering result according to the considered criterion. For the adjusted rand index, you need the actual class labels, which you can find in the last column of the dataset.}