# Homework 5

EEL 6935 Data Analytics

Due: 4/27/19, 11:59 PM

## 1 Experiment: Clustering

[50 pts] Download "perfume-data.txt". This dataset contains 560 odor-meter measurements in the first column. The task is to find natural groupings which potentially correspond to different perfume types. There are multiple measurements for each type. Cluster using K-means and GMM for different number of clusters $k = 2, \ldots, 20$. The ground truth is also given in the second column for performance evaluation. Use Rand Index and Adjusted Rand Index to compare the K-means and GMM performances for different $k$ values. Finally initialize GMM with K-means, and compare its results with the others. *(Note: You can use built-in functions.)*

## 2 Experiment: 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 optimum number of features for classification. In each trial, use 10-fold cross validation and SVM with Gaussian kernel. Average misclassification rate over 100 trials.

b) [10 pts] What is the maximum number of uncorrelated features?