STAT 391 Mini Project

February 19, 2018

1

In part (a), we need to perform PCA. Based on the plot for proportion of variance explained with respect to the number of predictors, PCA with 4 components generates the best result.

2

In part (b), we need to perform K-means clustering on the data. To choose K, I first use K values from 1 to 10 and plot the sum of Euclidean distances between all data to its label centers with respect to the the K values. Using the elbow method, I determine K = 2. Then I apply the model with K = 2. For simplicity, I fit PCA with 2 predictors in order to plot the clustering results.

3

In part (c), we need to perform hierarchial clustering on the data. In order to compare with the results from part (b), I use K = 2. Then I apply the Agglomerative Clustering to data.

4

In part (d), we need to perform three linear regression methods to the data. I choose the variable " SO" to be the response. This variable denotes the Relative sulphur dioxide pollution potential, which is a key measure for air pollution. I devided the data set into two parts, training set and cross validation set, with a ratio of 0.6 and 0.4, respectively. I apply the model to the

training set, and calculate the test error using cross validation set. During the ridge and lasso models, I first use the whole data to choose a λ value, then I use this lambda value as parameter for the model.

5

In part (e), we need to perform three classification methods to the data. I still choose the variable " SO" to be the response. To convert it into a categorical variable, I split the data into two parts: those that are above the median are recorded as 1, the rest are recorded as 0. I choose LDA and QDA as other two classification methods. For SVM, I first use GridSearchCV function to find the optimal combination of C value and gamma value. In the interest of running time, I use radial kernel for SVM. Using the best combination of the parameters, I apply the model to the data, and calculate the estimated test error on cross validation set.