ESI 4606: Analytics I - Foundations of Data Science
Final Exam - Part I
Due: December 5th (11:30PM), 2022

R Programming Problem (30 points)

Import data set "Final_Exam_Data_2022.RData". Let y represent response variable and the rest of variables (i.e., x_1 to x_{13}) represent input variables. Training data set and test data set consist of 200 and 300 observations, respectively. **Please answer all questions underlined below clearly and include R codes in the appendix sections to get full credits.** (Note: training or test error refers to the mean squared error (MSE))

(1) **(4 points)** Fit multiple linear regression with best subset selection using training data. Report variable selection results. Based on variable selection results, which input variable(s) is/are selected using BIC in the final best model? With selected input variable(s), train and test the final best model and report training error and test error.

(2) **(4 points)** Fit multiple linear regression with backward subset selection using training data. Report variable selection results. Based on variable selection results of both backward and best subset selection, are their corresponding best models with 8 input variables the same? Also, explain why.

(3) **(4 points)** Fit ridge regression using training data and report optimal tuning parameter value (using “lambda.1se", “grid=10^(seq(-2,5,0.1))" and setting random seed as 2022) based on 10-fold cross-valuation. Based on the optimal tuning parameter, train and test the ridge regression model and report training error and test error. Also, which variable(s) is/are selected in ridge regression results? (Hint: don’t forget to standardize input variables before fitting ridge regression).

(4) **(4 points)** Fit LASSO regression using training data and report optimal tuning parameter value (using “lambda.1se", “grid=10^(seq(-2,5,0.1))" and setting random seed as 2022) based on 10-fold cross-valuation. Based on the optimal tuning parameter, train and test the LASSO regression model and report training error

and test error. Also, which variable(s) is/are selected in LASSO regression results? (Hint: don’t forget to standardize input variables before fitting LASSO regression).

(5) **(3 points)** Train and test the Classification and Regression Tree (CART) model and report training error and test error. (Setting random seed as 2022).

(6) **(3 points)** Based on prediction results from (1), (3)-(5), which model(s) give(s) the best prediction?

(7) **(4 points)** Using 13 input variables in training data set, perform principle component analysis and draw scree plot. To keep at least 80% variability of data, how many principle components will be selected at least?

(8) **(4 points)** Using x_1 and x_2 of training data set, perform K-means clustering (by setting “center=3", “nstart=50" and random seed as 2022). What are cluster centroids? Also, visualize data observations, i.e., (x_1, x_2)’s, and add red color points of cluster centroids onto the plot.