Data Analytics and Visualization (Fall 2019)
Rutgers Business School-Newark and New Brunswick

Instructor: Debopriya Ghosh
Final Exam
(Due December, 13, 2019 @11:59 PM)

This is a take-home exam. You are allowed to work in your respective groups. One submission per group should suffice.

Problem 1. (6*5 = 30 points)
Generate a data set with n = 500 and p = 2, such that the observations belong to two classes with a quadratic decision boundary between them.

```
> x1 = runif(500) - 50
> x2 = runif (500) - 50
> y = 1 * (x1^2 - x2^2 > 0)
```

(1) Plot the observations, colored according to their class labels. The plot should display X_1 on x-axis and X_2 y-axis.
(2) Fit a logistic regression model to the data, using X_1 and X_2 as predictors.
(3) Apply this model to the training data in order to obtain a predicted class label for each training observation. Plot the observations, colored according to the predicted class labels. The decision boundary is linear.
(4) Fit a support vector classifier to the data with X_1 and X_2 as predictors.Obtain a class prediction for each training observation. Plot the observations colored according to the predicted class labels.
(5) Fit a SVM with non linear kernel to the data. Obtain a class prediction for each training observation. Plot the observations, colored according to the predicted class labels.
(6) Comment on your results.

Problem 2. (10 + 5 + 10 + 10 = 35 points)
Consider the USArrests data, perform hierarchical clustering on the states.

(1) Using the hierarchical clustering with complete linkage and Euclidean distance, cluster the states.
(2) Cut the dendrogram at a height that results in three distinct clusters. Which states belong to which clusters?
(3) Now scale the variables to have standard deviation one. Hierarchically cluster the scaled data with complete linkage and Euclidean distance.
(4) What effect does scaling the variables have on the hierarchical clusters obtained? Should the variables be scaled before inter-observation distances are computed. Provide justification for your answer.

Problem 3 (35 points)
Apply linear regression and random forests on the Hitters dataset to predict the Salary. Remove the observations for whom the salary information is unknown and perform log-transformation of the salaries. Be sure to fit the models on a training set and evaluate using test set. How accurate are the result of random forests compared to linear regression? Use MSE to compare model accuracy.

Best of luck!