IEOR 4650: Business Analytics
Spring 2019 Professor Adam Elmachtoub

Assignment 4

Due date: April 5, 11:59am

Attention: Please prepare two files for each homework assignment: a .pdf file for your answers including relevant figures, and a .R file for your relevant R scripts. File names should be `Last_First_hw.pdf` and `Last_First_hw.R`, e.g., `Obama_Barack_4.pdf` and `Obama_Barack_4.R`. Your submissions must be based on your own original work. Late submissions will be penalized at 10% per hour.

1. We shall analyze the `Tahoe_Healthcare_Data.csv` data set.

   (a) Build a tree to predict the probability of a patient being readmitted in less than 30 days on the entire dataset. Use the default parameters. Plot the resulting tree.

   (b) Run 10-fold cross-validation to prune the tree from the previous part. Use deviance as your criterion for pruning.

   (c) Now suppose we assign CareTracker to anyone with a probability of at least .15 of being readmitted, according to the tree you found in the previous part. What percentage of patients will receive CareTracker?

   (d) We would like to see if the CareTracker assignment in the previous part is fair to men and women equally. Pick 2 fairness measures from class and see how close we are to achieving those fairness measures.

2. In this problem you will need to analyze the `CollegeData.csv` data set. This data set from 2013 represents all colleges that grant graduate degrees in the United States. I have already significantly trimmed down the original data set which can be found online on data.gov. In this assignment, you will try to figure out what factors can be used to predict the quality of a school, using SAT_AVG as our measure of quality. You can look up the meanings of the columns in `CollegeDataDictionary.csv`. Download `CollegeData.csv` to your computer and read it into R. Be aware that the rows and columns are labeled. Remove any rows that have missing entries. The function `na.omit(...)` is useful.

   (a) Run a linear regression model to predict SAT_AVG and report the coefficient estimates and standard errors.

   (b) Run K-means clustering on all the futures, for $K = 1, \ldots, 50$. Report the total within-cluster variation (WSS- Within-cluster Sum of Squares) for each K. What value of K would you chose? Remember to scale your data first.

   (c) Focus on the first 100 universities that appear when sorted in alphabetical order. Run hierarchical clustering using average linkage, and create 4 clusters. What is the centroid of each cluster? Remember to scale the data first.

   (d) Run PCA on the dataset, and plot the cumulative PVE. Remember to scale the data.

   (e) Using the first two principal components, generate a biplot and try interpret the meaning of the two principal components. You may need to explore the plot options to make it look nice.

   (f) Run PCA regression and Partial Least Squares using 5-fold cross validation. Which technique works better? What value for M did you choose?