IEOR 4650: Business Analytics
Spring 2022 Professor Adam Elmachtoub

Assignment 4

Due date: March 28, 8am

Attention: Each team should submit on Gradescope, only one submission per team. If you have issues with your team, email the professor directly. Please prepare one file for each homework assignment: a .pdf file for your answers including relevant figures,as well as any relevant R scripts. You may use R Markdown for your convenience. Your submissions must be based on your own original work.

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 deviance as your splitting criterion. Let the minimum leaf size be 4 and minimum improvement be at least 0.001. Plot the resulting tree.

   (b) Run 7-fold cross-validation to prune the tree from the previous part. Use misclassification error as your criterion for pruning. Plot the final pruned tree.

   (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 females and non-females equally. Pick 2 fairness measures from class and see how close we are to achieving those fairness measures.

   (e) Which fairness measure do you think is most important in this setting? Least important?

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 K-means clustering on all the futures, for $K = 1, \ldots, 50$. Graph 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.

   (b) 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.

   (c) How would you chose the right number of clusters for hierarchical clustering? Give a concrete algorithm.