BTMA 531 Assignment 3

Due March 11, 2020 by noon on D2L

Instructions

- You should create a single R script called {firstname}_{lastname}_Asgn3.r, which has the required code for all parts of the assignment.
- Make sure to use commenting (#) so that your R file is readable by someone else. Yo do not need to comment on what you are trying to do in each line, but it should be clear where the answer for each question is.
- Make sure that your R script is executable from top to bottom on another computer. A good approach is to test it before submission on a lab computer. Make sure all requirements of questions are done using R (e.g. do not use calculator, excel, . . . to calculate things).
- The assignment submission should be done through D2L dropbox. Upload the R file to the assigned dropbox folder in D2L.
- The purpose of the assignments is to help you learn through practice. I recommend working on assignments in groups. You may want to use R help or search online for answers. However, note that this is an individual assignment. The work you submit should be 100% yours. Do not copy, share, or ask for files, chunks of code, or answers. Refer to the course outline for some examples of what to do and what not to do, and to Code of Student Conduct for more information on cheating and plagiarism. If you are note sure about a behavior, please ask.

Questions

1 [30] The attached "CarEvals.csv" dataset includes data on conditions and evaluations of second hand cars. There are 6 input variables, and an outcome variable called "Class" (This is a modified dataset based on https://archive.ics.uci.edu/ml/datasets/Car+Evaluation).

a) [10] Create a classification tree for classifying the "class" variable based on the other variables. Plot the tree.

b) [5] Create another tree using only a 1000 observations from the dataset, selected randomly. Predict the classes for the rest of the observations (719 observations).

c) [5] Calculate the accuracy of the predictions and draw the confusion matrix.

d) [5] Use cross-validation to find the best size of the tree. Plot the cross-validation error and the tuning parameter versus tree size. What is the best tree size?

e) [5] Prune the tree to the best size found in part d. Calculate the accuracy of the newly created model with the rest of the observations (719 observations).

2 [40] Use the "mtcars" dataset for this question. Only use columns 3 to 5 of the data.

a) [10] Use K-means clustering to cluster the data. Use 20 starting points to find the best clusters. For the number of clusters, once use 3, and once use 4.

b) [5] Plot the observations on all four dimensions, showing each cluster with a different color.

c) [5] Calculate the ratio of within-cluster errors to total errors for each cluster in the K=4 case. Plot the errors.

d) [10] Use hierarchical clustering to cluster the data, using complete linkage. Plot the dendrogram. Cluster the data into 4 clusters.

e) [10] Cluster the data using the dendrogram from previous part, and using a dissimilarity level of h=100. How many clusters does this clustering have?

3 [30] Use the included "Grocieries" dataset (within the arules package) for this question.

a) [15] Use the Apriori algorithm to find the association rules with a minimum support of 0.02 and minimum confidence of 0.4.

b) [5] Remove the redundant rules from your set of rules.

c) [5] Plot the rule performance measures based on support (on x axis) and confidence (on y axis) and lift (as shading).

d) [5] Create the frequency plot of items with minimum support of 0.1.