BTMA 531 Assignment 4

Due April 1, 2020 by noon on D2L

Instructions

- You should create a single R script called {firstname}_{lastname}_Asgn4.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.

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) [15] Create a SVM classification model for classifying the “class” variable based on the other variables. Use a random sample of 1500 cars to create the SVM model. Use the radial kernel with cost=5 and scaled data.

b) [15] Predict the classes for the remaining data (test set). Calculate the accuracy of the model and draw the confusion matrix.

2 [45] Use the attached “TextData” dataset for this question.

a) [15] Remove white spaces, stopwords, and numbers from the documents. Make the text all lowercase, and then stem the text.

b) [10] Create the document-term matrix. Find the frequent terms in all documents. Find highly associated terms (correlation more than 0.5) with two of the frequent terms (your choice).

c) [5] Create a word cloud of the terms.

d) [5] Cluster the terms using hierarchical clustering, with 5 clusters.

e) [5] Cluster the documents using K-means clustering, with K=5.

f) [5] Analyze the sentiment of all documents using the *syuzhet* package. Plot the sentiments as a bar plot.

3 [25] Use the included “Boston” dataset (within the MASS package) for this question.

a) [15] Create a neural network to predict the median value (medv) based on the rest of variables. Use a random set of 400 for training the neural network. Scale the data first. Plot the neural network.

b) [10] Predict the median value for the rest of the data (106 in test set). What is the accuracy of the model in terms of MSE?