BTMA 531 Assignment 2

Due February 12, 2020 by noon on D2L

Instructions

- You should create and sumbit a single R script called {firstname}_{lastname}_Asgn1.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 [40] Using the "airquality" dataset from R Datasets Package:

a) [2.5] Create a new object that considers the data as a time series.

b) [2.5] Create a new object from the time series where observations that have an *NA* in any of the variables have been removed. Call this new object *airqualityFull*.

c) [5] Create a sequence plot, a lag plot, and a histogram of the *Temperature* data from *airqualityFull*. What are your observations based on these plots?

d) [5] Create a Q-Q plot and Q-Q line for comparison of *temperature* data with the normal distribution from *airqualityFull*. What do you think about the distribution of data based on this plot?

e) [5] Create the autorocrrelation plot for *temperatures* in *airqualityFull*. Based on this plot, does temperature for one day depend on its temperature the day before?

f) [5] Draw the box plots for *temperature* and *wind* variables in *airqualityFull*. Which one has outliers based on the box plot?

g) [5] Assuming the data to be normal, find the 95% confidence interval for the *temperatures* from *airqualityFull*.

h) [5] Assuming the data to be normal, test to see whether if variances and means of the first 55 observations is different from the next 56 ones. Comment on findings for each test (are they different, why).

i) [5] Create a new object that has the same data as *airqualityFull* with outliers removed (using any tool you want).

2 [20] Use the attached "activity2Sample.csv" dataset for this question. This dataset has the accelerometer data on three axes (x, y, and z) and the actual activity taking place.

a) [20] Use KNN to create clusters in the data using the X, Y and Z variables. Use the first 150 observations for training. Create two models, one with k=5 and one with k=10. What are the accuracies for these two models?

3 [40] Using the "iris" dataset:

a) [25] Create a KNN classifier to predict the *Species* of the flower based on the *Sepal.Length* and *Sepal.Width* variables. Use the first 30 observations **for each species** in the dataset as training set, and the rest as the test set (90 observations for training, 60 observations for testing). Use *K=3*.

b) [5] Calculate the accuracy of the model for the predicted species in the test dataset.

c) [10] Predict the species of flowers using the model from *a*, but changing K to take values from 1 to 6. Calculate accuracy of each model and plot the accuracies with respect to K. Which K has the best accuracy? Is it OK to use this analysis to decide on the best-performing K? Why or why not?