BTMA 531 Assignment

Due March 13, 2024 by noon on D2L

Instructions

- You should create a single R script (or R Markdown if you wish) 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 script file can be understood 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. Provide the written answers in comments (#) within the R script file.
- Make sure that your R script is executable from top to bottom on another 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. 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.
- The focus is on the response to the question, and not on the approach you use to get there. However, unless otherwise stated, you should use only code to get to the response (that is, you cannot use a calculator for example to do some intermediate calculations, should not hard code numbers, and everything needs to be done in R).
- You are allowed to use online resources to find answers or chunks of code. However, make sure that you reference the source in your script using commenting (#). This is generally a good practice in coding, and for the purpose of this course, removes any confusion when several students use the same online source and therefore have similar codes in their scripts.

Questions

1 [25] Use the attached "ad.csv" dataset for this question. This dataset includes advertising data in TV, radio, and newspapers, with every observation's corresponding sales. We plan to cluster these observations into 3 or 4 different clusters, as we believe this would allow us to better understand the different advertising campaigns.

a) [2.5] Do we need to separate the data to training and test sets? Why?

b) [5] Use set.seed(1) for this part. Use K-means clustering to cluster the data. Use 10 starting points to find the best clusters. Scale the data if necessary. For the number of clusters, once use 3, and once use 4.

c) [7.5] Plot the observations on all three dimensions (or scaled dimensions) for the case with 3 clusters, showing each cluster with a different color. What is your observation of these clusters with respect to the three variables?

d) [5] Calculate the ratio of within-cluster errors to total errors for each cluster in the K=4 case. Plot the errors. Which cluster is the most homogeneous cluster?

e) [2.5] Use hierarchical clustering to cluster the data, using complete linkage. Plot the dendrogram. Cluster the data into 4 clusters.

f) [2.5] Cluster the data using the dendrogram from previous part, and using a dissimilarity level of h=3. How many clusters does this clustering have?

2 [15] Use the included "Boston" dataset (within the MASS package) for this question.

a) [10] 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. Use seed=1 before sampling. Scale the data first using min-max scaling. Plot the neural network.

b) [5] 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?

3 [40] Use the attached "online_shoppers_intention2.csv" dataset for this question. This data is a slightly modified version of the data from https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset.

a) [15] Create a deep learning neural network to predict the shoppers intention (the "Revenue" column indicates whether a revenue was generated) based on the rest of variables. Use a random set of 2,000 observations for training the neural network. Use seed=1 before sampling. Scale the numerical variables of the data first using min-max scaling. Use at least three layers, each with at least 10 neurons. Plot the neural network. Note that the model might take several minutes to train.

b) [5] Predict the shopper intention for the rest of the data (test set). What is the accuracy of the model in terms of MSE? Draw the confusion matrix.

c) [15] Now create a boosted tree (XGBoost) model to predict the shoppers intention based on the rest of variables. Use the same random set of 2,000 observations for training as in part (a). There is no need to scale the data for this model. Use the GBM package and for model parameters set n.trees=1000, interaction.depth = 2, and shrinkage = 0.01.

d) [5] Predict the shopper intention for the rest of the data (test set) using the XGBoost model you created in the last part. What is the accuracy of the model in terms of MSE? Draw the confusion matrix. How do the results compare to the NN model created in part (a)?

4 [20] Use the attached "accent-mfcc-data-1.csv" dataset for this question. This dataset includes 11 different Mel-frequency cepstrum (MFC) attributes on soundtracks of different people reading words. The MFC data is often used in sound processing. In this application, the goal is to predict the native language of the speaker from six European languages (ES, FR, GE, IT, UK, US) based on the MFCs. This dataset is based on https://archive.ics.uci.edu/ml/datasets/Speaker+Accent+Recognition.

a) [15] Create a SVM classification model for classifying the "language" variable based on the other variables. Use a random sample of 280 observations to create the SVM model. Use a seed=1 before sampling. Use the radial kernel with cost=4 and scaled data.

b) [5] Predict the classes for the remaining data (test set). Calculate the accuracy of the model and draw the confusion matrix.