DTI 5126: Fundamentals for Applied Data Science

Fall 2021

Assignment 2

Submission Deadline: 4th Nov 2021 on Brightspace.

This assignment should be **completed individually using R**. Upon completion, present your result in one submission, including the answers generated or plots **(Note: not more than 15 pages**). Where applicable, submit the source codes used to generate your results as a separate attachment.

Part A: Classification (50 points)

Customer churn rate is an important performance metric in the Telecoms industry due to the highly competitive markets. The churn rate enables companies to understand why their customers are leaving. You are hereby provided with a *churn dataset* containing randomly collected data from a telecom company's database. Develop ML models that can help the retention team analyze high risk churn customers before they leave by completing the following:

i. Ensure data is in the correct format for downstream processes and address missing data (5 points)
ii. Generate a scatterplot matrix to show the relationships between the variables and a correlation matrix to determine correlated attributes (5 points)
iii. Split the dataset into 80 training /20 test set and fit a decision tree to the training data. Plot the tree, and interpret the results. (10 points)
iv. Describe the first few splits in the decision tree. Extract some rules. (5 points)
v. Try different ways to improve your decision tree algorithm (e.g., use different splitting strategies, prune tree after splitting). Does pruning the tree improves the accuracy? (5 points)
vi. Train an XGboost model using 10-fold cross-validation repeated 3 times and a hyperparameter grid search to train the optimal model. Evaluate the performance. (5 points)
vii. Build a multilayer perceptron with 5 nodes at the hidden layer. *Use a standard or normalization to scale the variables.* Try changing the activation function, varying the neurons, learning rate, epochs or removing the bias. What effects does any of these have on the result? With a confusion matrix, evaluate the performance of the NN model based on sensitivity, specificity & accuracy (10 points)
viii. Carry out a ROC analysis to compare the performance of the DT, XGboost & NN techniques. Plot the ROC graph of the models. (5 points)

Part B: Clustering (50 points)

I. You are hereby provided with the *Shopping* dataset, which contains some basic information on customers. Complete the following using R:
   a. Perform *k*-means clustering, specifying $k = 2$ clusters and plot. Determine the attribute that is most correlated with the clusters (*Hint: save the labels of the cluster into a data frame and use heat map to find the most correlated attribute*). (10 points)
   b. Apply the elbow method to determine the best *k* and plot. (10 points)
   c. Evaluate the quality of the clusters using the Silhouette Coefficient method. (10 points)
   d. Apply hierarchical clustering (single & complete linkage) to the dataset using Euclidean-based distance, and plot the dendrogram. Do your results depend on the type of linkage used? (10 points)

II. Complete this problem **without using R** to make sure that you understand the details of the hierarchical clustering algorithm.

Consider the following "data" to be clustered: **10 20 40 80 85 121 160 168 195.**

For each part of the problem, assume that Euclidean distance will be used to measure the distance between the data points. Use hierarchical agglomerative clustering with single linkage to cluster the data. Draw a dendrogram to illustrate your clustering and include a vertical axis with numerical labels indicating the height of each parental node in the dendrogram (10 points)