DTI 5126: Fundamentals for Applied Data Science

Summer 2021

Assignment 3

Submission Deadline: 28th June 2021 on Brightspace.

This assignment should be completed individually. Upon completion, present your result in one submission, including the answers generated or plots. Where applicable, submit the source codes used to generate your results as a separate attachment.

Part A: Clustering (50 points)

1) K-Means Clustering

You are hereby provided with the *Framingham* data set. Using only the *Sex* and *Age* fields (ensure you standardize *Age),* complete the following:

a) Perform *k*-means clustering on the selected attributes, specifying $k = 4$ clusters and plot.
b) Apply the elbow method to determine the best *k* and plot.
c) Evaluate the quality of the clusters using the Silhouette Coefficient method.

2) Hierarchical Clustering

Complete this problem without the use of a computer to make sure that you understand the details of the clustering algorithms. Consider the following "data" to be clustered as described below.

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.

a) 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.
b) Repeat part (a) using hierarchical agglomerative clustering with complete linkage.

Part B: Model Evaluation & Performance Improvement (50 points)

Customer churn is a huge problem for telecoms providers, considering an annual churn rate of 15-20% in some markets. To keep a low churn rate, telecoms providers need to predict which customers are likely to churn. You are provided with the *customer_churn* dataset from a telecoms company, complete the following:

a) Partition the data set using the holdout method, so that 67% of the records are included in the training data set and 33% are included in the test data set. Use a bar graph to confirm your proportions.

b) Identify the total number of records in the training data set and how many records in the training data set have a churn value of true (or 1). Calculate how many true churn records you need to resample in order to have 20% of the rebalanced data set have true churn values.
c) Perform the rebalancing described in (b) and confirm that 20% of the records in the rebalanced data set have true churn values.
d) Create a decision tree model that can predict Churn using the data set given. Use predictors you think are appropriate and obtain the predicted value.
e) Use an ensemble method (e.g., Random Forest, Adaboost) to obtain the predicted value of Churn. Tune the hyper-parameters (e.g., node size, max depth, max terminal nodes, etc.) of the ensemble model and compare against the initial model.
f) Using a confusion matrix, compare the evaluation measures from the ensemble method with the decision tree model based on the following criteria: Accuracy, Sensitivity and Specificity. Identify the model that performed best and worst according to each criterion.
g) Carry out a ROC analysis to compare the performance of the ensemble method with the decision tree technique. Plot the ROC graph of the models.