DTI 5126: Fundamentals of Data Science

Summer 2023

Assignment 2

Submission Deadline: 12th July 2023 on Brightspace.

This assignment should be completed by a team with 2 members using R. Upon completion, present your result in one submission, including the answers generated or plots. **Note: You can submit your R notebook. If submitting a PDF, it should not exceed 10 pages**.

Part 1: Classification (60 points)

Bay clinic is a medical centre that operates with a unique mission of blending research and education with clinical and hospital care. The medical center has a huge head force of 25,000 employees, and as a result of the combined effort of those employees, the medical center has been able to handle approximately 3 million visits so far. In recent times, the hospital was incurring losses despite having the finest doctors available and not lacking scheduled appointments. To investigate the reason for the anomaly, a sample data dump of appointments ***medicalcentre.csv*** is hereby presented. The collected data provides information on the patient's age, gender, appointment date, various diseases, etc. To cut costs, predict if a patient will show up on the appointment day or not (i.e., predict the appointment status) by completing the following:

A. **Feature Engineering (20 points):**

1. Prepare the data for downstream processes, e.g., dealing with missing values
2. Initialize a function to plot all features within the dataset to visualize for outliers
3. Count the frequency of negative Age feature observations, and remove them
4. The values within AwaitingTime are negative, transform them into positive values
5. ML algorithm requires the variables to be coded into its equivalent integer codes. Encode the string categorical values into an integer code
6. Separate the date features into date components
7. ML algorithms work best when the input data are scaled to a narrow range around zero. Rescale the age feature with a normalizing (e.g., *min_max normalization*) or standardization (e.g., *z_score standardization*) function.
8. Conduct variability comparison between features using a correlation matrix & drop correlated features

B. **Model Development I (10 points):**

Develop a SVM and Decision Tree classifier to predict the outcome of the test. The performance of the classifier should be evaluated by partitioning the dataset into a train dataset (70%) and test dataset (30%). Use the train dataset to build the models and the test dataset to evaluate how well the model generalizes to future results.

C. **Model Development II (10 points):**

Train a Deep Neural Network using Keras (Ensure to determine the best configuration that provides the best accuracy). Try changing the activation function or dropout rate. What effects does any of these have on the result?

D. **Model Evaluation & Comparison (20 points):**

1) Write a Function to detect the Model's Accuracy by applying the trained model on a testing dataset to find the predicted labels of Status. Was there overfitting?
2) Tune the model using GridSearchCV
3) Evaluate the performance of the SVM, Decision tree and Deep Neural Network classifier on the dataset based on the following criteria: Accuracy, Sensitivity and Specificity. Identify the model that performed best and worst according to each criterion.
4) Carry out a ROC analysis to compare the performance of the SVM model with the Decision Tree model. Plot the ROC graph of the models.

Part 2: Unsupervised Learning (40%)

A. **K-Means Clustering (20 points):**

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

(1) Perform k-means clustering on the selected attributes, specifying k = 4 clusters and plot.
(2) Apply the elbow method to determine the best k and plot.
(3) Evaluate the quality of the clusters using the Silhouette Coefficient method.

B. **Hierarchical Clustering (20 points):**

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.

(1) 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.
(2) Repeat part (a) using hierarchical agglomerative clustering with complete linkage.