DTI 5126: Fundamentals for Applied Data Science

Summer 2021

Assignment 2

Submission Deadline: 14th June 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. Where applicable, submit the source codes used to generate your results as a separate attachment.

Part A: Decision Trees (50 points)

A benchmark dataset to predict the type of thyroid disease a patient has is hereby provided in "*hypothyroid.csv*". Complete the following:

a. There are some missing values in the dataset. Several strategies can be used to handle them, e.g., remove cases with unknowns. Apply one of these methods to address the missing values.
b. Perform attribute selection on the dataset and state briefly why attribute selection is sometimes important.
c. Split the dataset into a train and test set using k-fold cross-validation (k= 10). Create a decision tree model using the selected attributes from your dataset that can predict the type of thyroid disease a patient has.
d. Visualize and describe the first few splits in the decision tree. Extract some rules.
e. Try different ways to improve your decision tree algorithm, e.g., use different splitting strategies, prune tree after splitting

Part B: Support Vector Machines (50 points)

A sample of the dataset for women of the Pima Indian heritage living near Phoenix, Arizona that tested for diabetes, according to World Health Organization, is hereby provided in *"diabetes.csv"*. Develop a Feed-Forward Neural Network to predict the outcome of the test.

a. Some data points are not available, handle the missing data by applying central measure of tendency to derive the missing value.
b. Partition the dataset into a train dataset (75%) and test dataset (25%). Use the train dataset to build the Neural Network and the test dataset to evaluate how well the model generalizes to future results.
c. Neural networks work best when the input data are scaled to a narrow range around zero. Rescale the data with a normalizing (e.g., *min_max normalization*) or standardization (e.g., *z_score standardization*) function.

d. Train & plot a simple Neural Network with only 2 hidden nodes (not layer). Then, train & plot a multilayer perceptron with 2 layers & 5 nodes. What impact does the change in the number of layers & nodes have on the accuracy of your model?
e. Try changing the activation function, varying the learning rate, epochs or removing the bias. What effects does any of these have on the result?