STA 545 Statistical Data Mining I, Fall 2020

Homework 9, due: Wednesday 11/18/2020 (1PM)

Please submit the pdf file generated by R markdown in UBlearns. Please use tables, figures, or a few sentences to answer data analysis questions.

1. (10 points) Consider the following example discussed in class,

	Class1	Class2	p_1	p_2
Parent	7	3	7/10	3/10
Split 1: Left	3	0	3/3	0
Split 1: Right	4	3	4/7	3/7
Split 2: Left	2	1	2/3	1/3
Split 2: Right	5	2	5/7	2/7

   (a) Which split is better if we use entropy as the impurity measure?

   (b) Which split is better if we use the misclassification error as the impurity measure?

2. (55 points) This problem involves the OJ data set which is part of the ISLR package.

   (a) (5 points) Create a training set containing a random sample of 800 observations, and a test set containing the remaining observations.

   (b) (5 points) Fit a tree to the training data, with Purchase as the response and the other variables as predictors. Use the summary() function to produce summary statistics about the tree, and describe the results obtained. What is the training error rate? How many terminal nodes does the tree have?

(c) (5 points) Type in the name of the tree object in order to get a detailed text output. Pick one of the terminal nodes, and interpret the information displayed.

(d) (5 points) Create a plot of the tree, and interpret the results.

(e) (5 points) Predict the response on the test data, and produce a confusion matrix comparing the test labels to the predicted test labels. What is the test error rate?

(f) (5 points) Apply the cv.tree() function to the training set in order to determine the optimal tree size.

(g) (5 points) Produce a plot with tree size on the x-axis and cross-validated classification error rate on the y-axis.

(h) (5 points) Which tree size corresponds to the lowest cross-validated classification error rate?

(i) (5 points) Produce a pruned tree corresponding to the optimal tree size obtained using cross-validation. If cross-validation does not lead to selection of a pruned tree, then create a pruned tree with five terminal nodes.

(j) (5 points) Compare the training error rates between the pruned and unpruned trees. Which is higher?

(k) (5 points) Compare the test error rates between the pruned and unpruned trees. Which is higher?

3. (35 points) Apply the tree-based classification method to a dataset of your choice (e.g., a dataset from your research projects or the UCI Machine Learning Repository). Be sure to fit the models on a training set and to evaluate their performance on a test set. How accurate are the results compared to simple methods like LDA or logistic regression? Which of these approaches yields the best performance?