DS 303 Homework 9
Due: Nov. 15, 2021 on Canvas by 11:59 pm (CT)

Instructions: Homework is to be submitted on Canvas by the deadline stated above. Please clearly print your name and student ID number on your HW.

Show your work (including calculations) to receive full credit. Please work hard to make your submission as readable as you possibly can - **this means no raw `R` output or code** (unless it is asked for specifically or needed for clarity).

Code should be submitted with your homework as a separate file (for example, a .R file, text file, Word file, or .Rmd are all acceptable). You should mark sections of the code that correspond to different homework problems using comments (e.g. `##### Problem 1 #####`).

Problem 1: ROC Curve

Use the `Spam` data set, from HW 7, for this problem. Repeat your code from Problem 4 parts (a) and (b).

a. What type of mistake do we think is more critical here: reporting a meaningful email as spam (false positive) or a spam email as meaningful (false negative)?

b. Fit a logistic regression model here and apply it to the test set. Based on your answer to part (a), plot the ROC curve of true positive rate vs. false positive rate or true negative rate vs. false negative rate.

c. Output the confusion matrix. What is the false positive and false negative rate when we set the threshold to be 0.5?

d. Adjust the threshold such that your chosen error (false positive or false negative) is no more than 0.03. You should choose the threshold carefully so that the true positive and true negative rate are also maximized.Report that threshold here.

e. Implement LDA and repeat parts (b) -(d).

Problem 2: Basics of Decision Trees

Use the `OJ` data set, which is part of the `ISLR2` package, for this problem.

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

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

c. 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. Create a plot of the tree, and interpret the results.

e. Predict the response on the test set and report the confusion matrix. What is the test error?

f. Apply `cv.tree()` to determine the optimal tree size. Produce a plot with tree size on the x-axis and cross-validated classification error rate on the y-axis.

g. What tree size corresponds to the lowest cross-validated classification error rate?

h. 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.

i. Compare the training error rates between the pruned and un-pruned trees. Which is higher? Is this what you expect? Explain.

j. Compare the test errors rates between the pruned and un-pruned trees. Which is higher? Is this what you expect? Explain.

Problem 3: Conceptual Review

(a) You would like to show your friend a visual example of how decision trees work. Draw a toy-example of a partition of a two-dimensional predictor space that could result from recursive binary splitting. Your predictors can only take values from $[0, 10]$ and your example should contain at least six regions. Also draw the decision tree corresponding to this partition. Be sure to label all aspects of your drawings, including the regions R_1, R_2, $\ldots$, the cut-points and so forth.

(b) Continuing with this toy-example, suppose you then implemented bagging. Construct 3 bagged trees. Choose 2 values of X_1 and X_2 (for example $X_1 = 4$ and $X_2 = 8$) and explain how you would predict the response for a test observation given those specific values of X_1, X_2.

(c) Suppose we obtained ten bootstrapped samples from a data set where Y can take two values: red or green. We then apply a classification tree to each bootstrapped sample, and for a specific value of X, produce 10 estimates of $P(Y = \text{red}|X)$:

$$0.1, 0.15, 0.2, 0.2, 0.55, 0.6, 0.6, 0.65, 0.7, 0.75.$$

There are two common ways to combine these results together into a single class prediction. One is the majority vote approach discussed in lecture. The second approach is to classify based on the average probability. In this example, what is the final classification under each of these two approaches?