DS 303 Homework 10
Due: Nov. 13, 2023 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: Conceptual Review

(a) 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?

(b) See Figure 1. Sketch the tree corresponding to the partition of the predictor space illustrated on the left-hand side of Figure 1. The numbers inside the boxes indicate the mean of Y within each region.

(c) See Figure 1. Create a diagram showing how the predictor space is partitioned (similar to the left-hand size of Figure 1) based on the tree on the right-hand side of Figure 1. You should divide up the predictor space into the correct regions, and indicate the prediction of Y for each region.

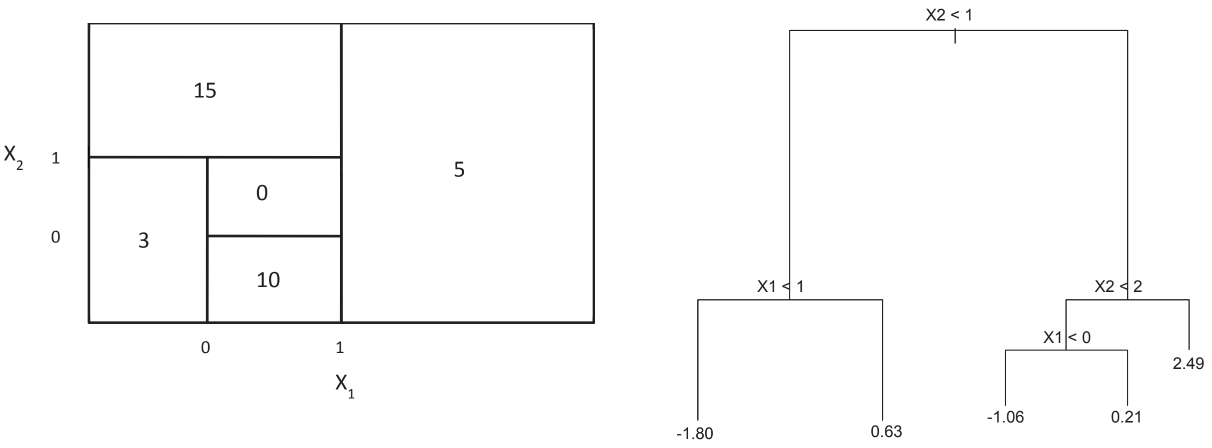

Figure 1: Figure corresponding to Problem 1 (b) and (c)

(d) Can bagging ever result in a higher variance than an individual tree? Justify your answer by a simple argument using statistical ideas or theory.

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: Bagging and Random Forests

We'll use the `Carseats` for this problem; it is part of the `ISLR2` library. Convert `Sales` to a qualitative response, the same way we did in class.

a. Split the data set into a training and test set.

b. Fit a classification tree to the training set. Use gini index as your splitting criteria. Plot the tree here and interpret the results. Report your training and test error.

c. Implement cross-validation to obtain the optimal level of tree complexity. What size tree is optimal? What is the test error for the pruned tree?

d. Implementing bagging on the training set. Set $B = 500$, where B is the number of trees. What test error do you obtain? Use the `importance()` function to determine which variables are the most important and report them here.

e. Implement random forests on the training. Experiment with different values of m and report the test error for different values of m in a table.

f. Looking at your table from part (e), would it be appropriate to choose the m that gives us the smallest test error? Explain. (Hint: the answer is no.)

g. Technically m is a tuning parameter. Implement a data-driven approach to decide on the appropriate m. Report that value here.

h. Obtain the OOB error estimation from implementing random forests. Set seed to be 1, $B = 500$, and $m = 6$. Write your own code to do so. Since this is a classification problem, the OOB error estimation will be calculated as the misclassification error (not the MSE). As general advice, you will want to run your code line-by-line and check the output. It can be frustrating to troubleshoot your code if you run chunks of code all at once. Report the following:

    i. What is the total number of bootstrapped trees the 4th observation appears?

    ii. What is the OOB classification for the 10th observation (based on majority vote). What are the OOB proportions of "No" and "Yes" for observation 10?

    iii. Report your OOB error estimation. Copy/paste any relevant `R` code here.

End of assignment