DS 303 Homework 10
Due: Nov. 29, 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: Concept Review

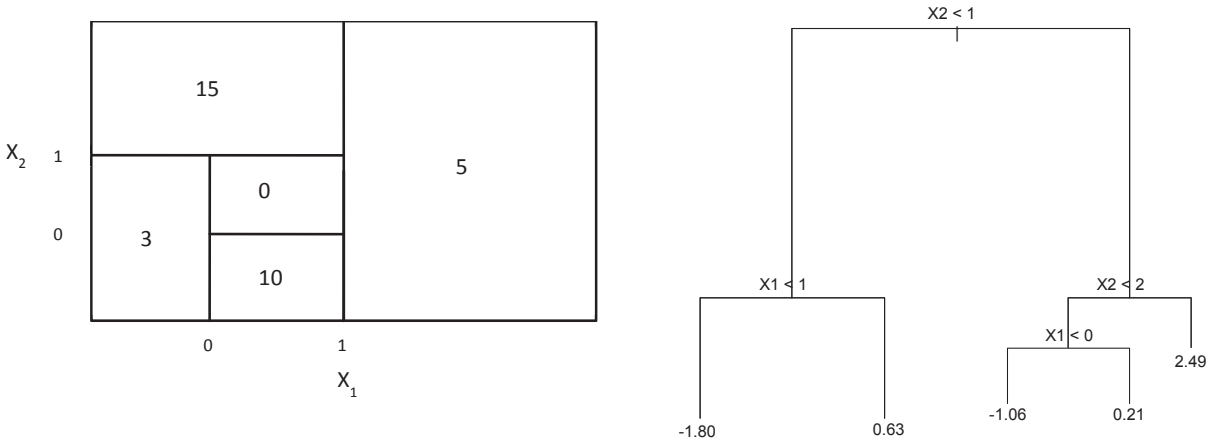

Figure 1: Figure corresponding to Problem 1

- 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.
- See Figure 1. Create a diagram (similar to the left-hand size of Figure 1) using the tree illustrated in the right-hand side of Figure 1. You should divide up the predictor space into the correct regions, and indicate the mean for each region.

Problem 2: 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. What splitting criteria did you use? Plot the tree here and interpret the results. What test MSE did you obtain?

c. Use cross-validation in order to obtain the optimal level of tree complexity. What size tree is optimal? What is the test MSE for the pruned tree?

d. Implementing bagging on the training set. Set $B = 500$, where B is the number of trees. What test MSE 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 MSE 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 MSE? Explain. (Hint: the answer is no.)

g. 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.

h. Compare your OOB error estimation to the value reported by extracting the attribute `$err.rate[400]`. Do the values match? Note, there may be a very slight rounding error.

Problem 3: Boosting

We'll use the `Hitters` for this problem; it is part of the `ISLR2` library.

a. Remove the observations for whom the salary information is unknown, and then log-transform the salaries.

b. Create a training set consisting of the first 200 observations, and a test set consisting of the remaining observations.

c. Perform a grid search on the training set to decide the optimal number of trees, the optimal λ, and optimal depth. To keep things computationally simple, only considered 3 different values for each of the tuning parameters. You may decide what those values are.

d. Implement boosting on the training set with the tuning parameters you have selected from the grid search. Which variables appear to be the most important predictors in the boosted model?

e. What is the test MSE of the boosted model from part (d)?

f. Now apply bagging to the training set. What is the test set MSE for this approach?

Problem 4: MNIST handwritten digit database

Load the handwritten digits (MNIST) dataset into R using the scripts provided from HW 8. Repeat the pre-processing that you did to obtain a training and test set. Implement random forests. Experiment with different values of m (usually $m = \sqrt{p}$). Report the following:

a. The confusion matrix.

b. Specific misclassification rates for each digit.

c. The overall misclassification rate.

d. Comments on this compares to the results you obtained for KNN.

Extra Credit

This extra credit is all or nothing. You cannot receive partial credit.

a. Can bagging ever result in a higher variance than an individual tree? Justify your answer by a simple argument using statistics.

b. Suppose you have some observations: $y_i, i = 1, \ldots, n$ that are iid (independent and identically distributed) with mean μ and variance σ^2. Let $\bar{y}_1^\star$ be the sample mean from a bootstrapped sample and $\bar{y}_2^\star$ be the sample mean from a different bootstrap sample. Show the steps to obtain

$$\text{var}(\bar{y}_1^\star) = \frac{2n-1}{n^2}\sigma^2 \text{ and } \text{Cov}(\bar{y}_1^\star, \bar{y}_2^\star) = \frac{\sigma^2}{n}.$$