STAT 435
SPRING QUARTER 2018

Homework # 8
Due Friday, June 1, 2018 at 12:00 PM (Noon)
Online Submission Via Canvas

Instructions: You may discuss the homework problems in small groups, but you must write up the final solutions and code yourself. Please turn in your code for the problems that involve coding. However, for the problems that involve coding, you must also provide written answers: you will receive no credit if you submit code without written answers. You might want to use Rmarkdown to prepare your assignment.

1. In this problem, you will fit some models to a data set of your choice.
   (a) Find a very large data set of your choice (large n, possibly large p). Select one quantitative variable to be your response, $Y \in \mathbb{R}$. Describe the data.
   (b) Grow a very big regression tree to the data. Plot the tree, and report its residual sum of squares (RSS) on the (training) data.
   (c) Now use cost-complexity pruning to prune the tree to have 6 leaves. Plot the pruned tree, and report its RSS on the (training) data. How does this compare to the RSS obtained in (b)? Explain your answer.
   (d) Perform cross-validation to estimate the test error, as the tree is pruned using cost-complexity pruning. Plot the estimated test error, as a function of tree size. The tree size should be on the x-axis and the estimated test error should be on the y-axis.
   (e) Plot the "best" tree (with size chosen by cross-validation in (d)), fit to all of the data. Report its RSS on the (training) data.
   (f) Perform bagging, and estimate its test error.
   (g) Fit a random forest, and estimate its test error.
   (h) Which method (regression tree, bagging, random forest) results in the smallest estimated test error? Comment on your results.
2. In this problem, we will consider fitting a regression tree to some data with $p = 2$.
   (a) Find a data set with n large, $p = 2$ features, and $Y \in \mathbb{R}$. It's OK to just use the data from Question 1 with just two of the features.

(b) Grow a regression tree with 8 terminal nodes. Plot the tree.

(c) Now make a plot of feature space, showing the partition corresponding to the tree in (b). The axes should be X_1 and X_2. Your plot should contain vertical and horizontal line segments indicating the regions corresponding to the leaves in the tree from (b). Superimpose a scatterplot of the n observations onto this plot. This should look something like Figure 8.2 in the textbook. Label each region with the prediction for that region.

Note: If you want, you can plot the horizontal and vertical line segments in (c) by hand (instead of figuring out how to plot them in `R`*).*

3. This problem has to do with bagging.

(a) Consider a single regression tree with just two terminal nodes (leaves). Suppose that the single internal node splits on $X_1 < c$. If $X_1 < c$ then a prediction of 13.9 is made; if $X_1 \geq c$ then a prediction of 3.4 is made. Write out an expression for $f(\cdot)$ in the regression model $Y = f(X_1, \ldots, X_p) + \epsilon$ corresponding to this tree.

(b) Now suppose you bag some regression trees, each of which contain just two terminal nodes (leaves). Show that this results in an *additive* model, i.e. a model of the form

$$Y = \sum_{j=1}^{p} f_j(X_j) + \epsilon.$$

(c) Now suppose you perform bagging with larger regression trees, each of which has at least three terminal nodes (leaves). Does this result in an additive model? Explain your answer.

4. If you've paid attention in class, then you know that in statistics, there is no free lunch: depending on the form of the function $f(\cdot)$ in the regression model

$$Y = f(X_1, \ldots, X_p) + \epsilon,$$

a given statistical machine learning algorithm might work very well, or not well at all. You will now demonstrate this in a simulation with $p = 2$ and $n = 1000$.

(a) Generate X_1, X_2, and ϵ as

```
x1 <- sample(seq(0,10,len=1000))
x2 <- sample(seq(0,10,len=1000))
eps <- rnorm(1000)
```

If you generate Y according to the model $Y = f(X_1, X_2) + \epsilon$, then what will be the value of the irreducible error?

(b) Give an example of a function $f(\cdot)$ for which a *least squares regression model* fit to $(x_1, y_1), \ldots, (x_n, y_n)$ can be expected to outperform a *regression tree* fit to $(x_1, y_1), \ldots, (x_n, y_n)$, in terms of expected test error. Explain why you expect the least squares regression model to work better for this choice of $f(\cdot)$.

(c) Now calculate $Y = f(X_1, X_2) + \epsilon$ in `R` using the `x1`, `x2`, `eps` generated in (a), and the function $f(\cdot)$ specified in (b). Estimate the test error for a least squares regression model, and the test error for a regression tree (for a number of values of tree size), and display the results in a plot. The plot should show tree size on the horizontal axis and estimated test error on the vertical axis; the estimated test error for the linear model should be plotted as a horizontal line (since it isn't a function of tree size). Your result should agree with your intuition from (b).

(d) Now repeat (b), but this time find a function for which the regression tree can be expected to outperform the least squares model.

(e) Now repeat (c), this time using the function from (d).