HW 4 – 4803, Spring 2023
Each problem is worth 10 points

1 Part I – theoretical problems

1. Consider the weakest link pruning method we discussed in class for regression trees. Suppose that the current value of α implies that we should prune a section S_t of the current tree to node t. For the sake of notation, suppose the original tree is T_0 and the pruned tree is T_1. Show that, after this pruning, all nodes that are upstream of node t (those nodes t' for which node t is in $S_{t'}$), satisfy $g_1(t') \geq g_0(t')$, with g defined in the lecture notes. By showing this, you are proving that making a prune indicated by the current value for α does not immediately cause other nodes to require pruning before α is increased from its current value.

2-3. From the Book "An Introduction to Statistical Learning"– 8.4 Exercises 4 and 5

4-5. From the Book "An Introduction to Statistical Learning"– 9.7 Exercises 2 and 3

2 Part II – programming

1 This problem is about the Boston data set. It is an extension of Ex 7 in 8.4. Split the data into two even subsets - one for training and the other for testing.

(a) Apply regression trees to predict the median value of owner-occupied homes in \$1000's from other variables. Describe your experiments and report the test mean squared error.

(b) Apply random forests to predict the median value of owner-occupied homes in \$1000's, using $m = 6$ so that 6 random predictors are considered for each split of the tree. For the total number of trees constructed, try both 25 and 100, reporting the results separately. Describe your experiments and report the test mean squared error.

(c) Apply classification trees to predict whether a given suburb has a crime rate above or below the median from other variables. Describe your experiments and report the test classification error.

(d) Apply random forests to predict whether a given suburb has a crime rate above or below the median from other variables, using $m = 6$. Try both 25 and 100 trees. Describe your experiments and report the test classification error.

2 From the Book "The Elements of Statistical Learning"– Chapter 9 Exercise 9.5, skipping part (a).

3 From the Book "An Introduction to Statistical Learning"– 9.7 Exercise 6. Here, "cost" refers to the parameter C from equation (9.15). For part (a), one way to do this is to pick some line to represent the "true" decision boundary (essentially any line you like), then generate many data points on either side of this line, making sure that you have several points on each side that are quite close to the line.