University of Texas at Austin

Homework Assignment 11

More on trees.

Please, provide your **complete solutions** to the following problems. Final answers only, even if correct will earn zero points for those problems.

Problem 11.1. (10 points) Solve Problem **8.4.2** (pp.361) from the textbook.

Solution: We can start with the formula in the *boosting* algorithm:

$$f(x) = \sum_{b=1}^{B} \lambda \hat{f}^b(x)$$

Since for $d = 1$ every tree considered is a ”stump”, it has only one split which - by definition - must be based on only one variable. If one considers equation (8.9) from the book, one realizes that that equation becomes exactly

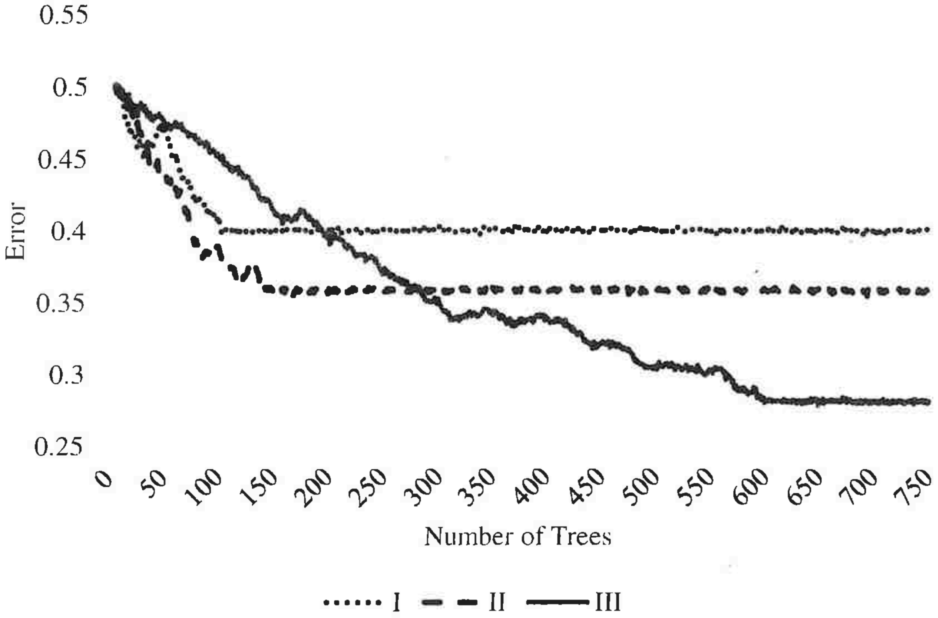
$$f(x) = c\mathbb{I}_{\{x_j > t_j\}} + c'\mathbb{I}_{\{x_j \leq t_j\}}$$

for **one specific** j for each tree in the boost.

Problem 11.2. (5 points) Solve Problem **8.4.5** (pp.362) from the textbook.

Solution: By majority vote, we get **red**. On the other hand, the average of the probabilities is 0.45 which implies **green**.

Problem 11.3. (5 points) *Source: MAS-II, Fall 2018.*
An actuary creates three tree-based models using bagging, boosting, and random forests. The error on the test data set, as a function of the number of trees in each model, is plotted on the graph below:



Determine the type of model most likely to have created each of the lines on the graph.

(a) I: Boosting, II: Bagging, III: Random forest

(b) I: Bagging, II: Boosting, III: Random forest

(c) I: Bagging, II: Random forest, III: Boosting

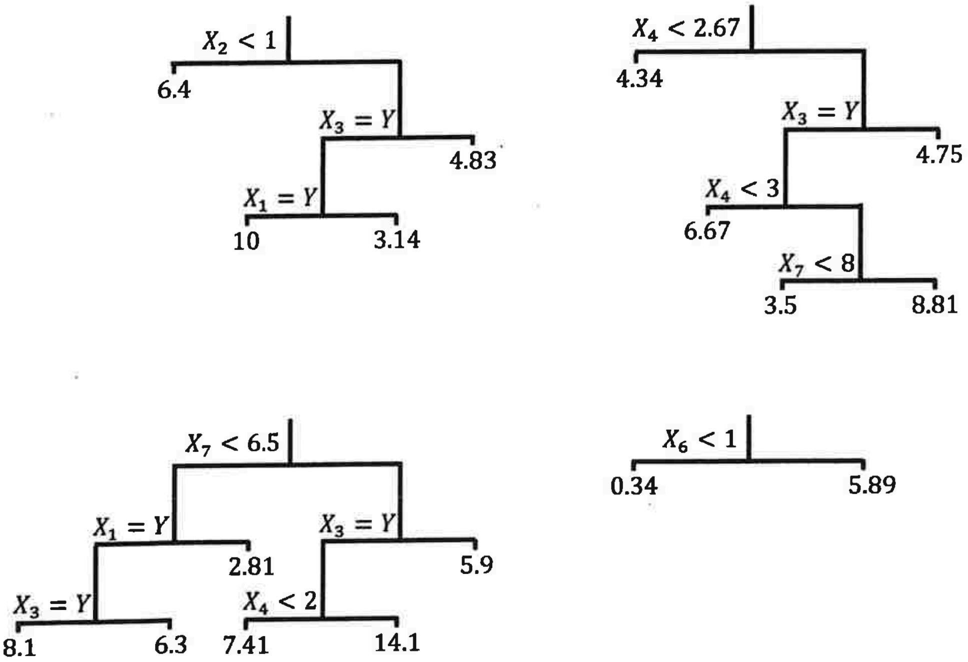
(d) I: Random forest, II: Bagging, III: Boosting

(e) None of the above.

Solution: (c)
Bagging is going to converge before random forest and the lines I and II are the only ones exhibiting a ”flattening”. That leaves III for boosting. The shape does make sense as there seems to be a gradual trend downwards that does **not** level off. On the other hand, the book does warn against the potential for overfitting when there are too many trees used in boosting. So, one would imagine that the error in the test set would start creeping up. We don’t see this in the plot provided, but then again $750 \neq \infty$.

Problem 11.4. (15 points) *Source: MAS-II, Spring 2019.*
A boosted tree model is defined by:

- $\lambda = 0.2$
- The following four trees:



You are given the following record:

X_1	X_2	X_3	X_4	X_5	X_6	X_7
N	6	Y	4	0.5	0.25	6

Calculate the prediction of the boosted tree model for this record.

Solution: In the first tree, we have that - in the given record - $X_2 \geq 1$, then $X_1 = N$ (which means $\neq Y$). We land at the terminal node with the value 3.14.

In the second tree, we have that - in the given record - $X_4 \geq 2.67$, then $X_3 = Y$, then $X_4 \geq 3$, and, finally, $X_7 < 8$. We land at the leaf with the value 3.5.

In the third tree, we have that - in the given record - $X_7 < 6.5$, followed by $X_1 = N$ (i.e., $\neq Y$). We land at the leaf with the value 2.81.

Finally, in the fourth tree, we have that - in the given record - $X_6 < 1$. Thus, the terminal node is the one with 0.34.

Overall, the prediction of the boosted model is

$$0.2(3.14 + 3.5 + 2.81 + 0.34) = 1.958.$$

Problem 11.5. (10 points) A classification tree was constructed in order to predict the value of a categorical random variable with levels I, II, and III. A split of a specific node in the tree yielded these two regions:

Region	Count of I	Count of II	Count of III
R_1	40	10	10
R_2	5	25	10

Calculate the Gini index, the cross-entropy, and the classification error for this split.

Solution: For the Gini index, we have

$$0.6\left(\left(\frac{2}{3}\right)\left(\frac{1}{3}\right)+\left(\frac{1}{6}\right)\left(\frac{5}{6}\right)+\left(\frac{1}{6}\right)\left(\frac{5}{6}\right)\right)+0.4\left(\left(\frac{1}{8}\right)\left(\frac{7}{8}\right)+\left(\frac{5}{8}\right)\left(\frac{3}{8}\right)+\left(\frac{1}{4}\right)\left(\frac{3}{4}\right)\right)=0.5125$$

For the cross-entropy, we have

$$-\left(0.6\left(\left(\frac{2}{3}\right)\ln\left(\frac{2}{3}\right)+\left(\frac{1}{6}\right)\ln\left(\frac{1}{6}\right)+\left(\frac{1}{6}\right)\ln\left(\frac{1}{6}\right)\right)+0.4\left(\left(\frac{1}{8}\right)\ln\left(\frac{1}{8}\right)+\left(\frac{5}{8}\right)\ln\left(\frac{5}{8}\right)+\left(\frac{1}{4}\right)\ln\left(\frac{1}{4}\right)\right)\right)=0.8806404$$

For the classification error, we have

$$\frac{10+10+5+10}{100}=0.35.$$

Problem 11.6. (10 points) Consider a node in a classification tree. You are considering making a split at that node. Right now, there are 30 lapses and 10 non-lapses at that node. What would be the total reduction in the Gini index if this node were split into two regions so that one of them has 20 lapses and 4 non-lapses? Was this a meaningful split in and of itself?

Solution: The current value of the Gini index is

$$2(0.75)(0.25)=0.375$$

After the split, there are 24 data points total in one region and the remaining 16 in the other region. So, the weight for the first region is 0.6 and the weight for the other region is 0.4. The total Gini index is

$$0.6(2)\left(\frac{5}{6}\right)\left(\frac{1}{6}\right)+0.4(2)\left(\frac{5}{8}\right)\left(\frac{3}{8}\right)=0.3541667$$

While there is a reduction in the value of the Gini index equal to 0.02083333, if these are to be terminal nodes it's not clear why as they all go to *lapses*.