University of Texas at Austin

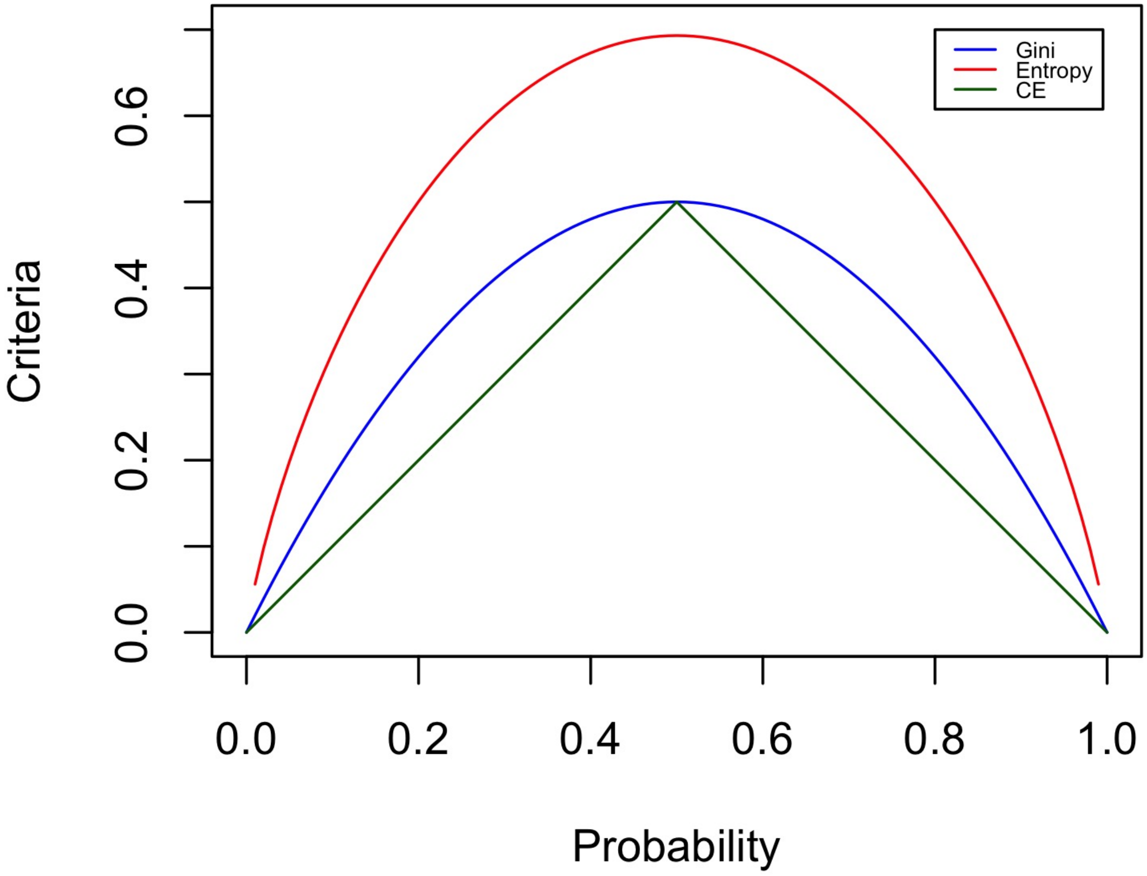
Homework Assignment 8

Classification trees.

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

Problem 8.1. (10 points) Solve Problem **8.4.3** (pp.361-362) in the textbook.

Solution: Here is the plot:



Problem 8.2. (10 points) A classification tree is constructed to predict whether a student will pass *Predictive Analytics*. There are two categorical predictors:

- X_1 indicating whether the student passed *Linear Algebra* prior to enrolling in *Predictive Analytics*, and

- X_2 indicating whether the student passed *Mathematical Statistics* prior to enrolling in *Predictive Analytics.*

Here is the table from a data set of students:

X_1	X_2	outcome
0	0	5 passed, 20 didn't
1	0	10 passed, 10 didn't
0	1	15 passed, 20 didn't
1	1	15 passed, 5 didn't

What's the first split made using the Gini index? Be careful to calculate the **weighted** average.

Solution: Since both X_1 and X_2 are binary, we can make the first split in a unique way with respect to X_1 or with respect to X_2. If we split along X_1, we have that

$X_1 = 0$ 20 passed and 40 didn't;
$X_1 = 1$ 25 passed and 15 didn't.

So, the Gini index is

$$\frac{60}{100} \cdot 2 \cdot \left(\frac{20}{60}\left(1 - \frac{20}{60}\right)\right) + \frac{40}{100} \cdot 2 \cdot \left(\frac{25}{40}\left(1 - \frac{25}{40}\right)\right) = 0.4541667$$

If we split along X_2, we have that

$X_2 = 0$ 15 passed and 30 didn't;
$X_2 = 1$ 30 passed and 25 didn't.

So, the Gini index is

$$\frac{45}{100} \cdot 2 \cdot \left(\frac{15}{45}\left(1 - \frac{15}{45}\right)\right) + \frac{55}{100} \cdot 2 \cdot \left(\frac{30}{55}\left(1 - \frac{30}{55}\right)\right) = 0.4727273$$

Since the first Gini index is smaller, we should split along X_1.

Problem 8.3. (30 points) A classification tree is fitted to be used to classify drivers into one of three categories: *Good, Medium* and *Bad*. There are three terminal nodes in the classification tree designating the regions R_1, R_2 and R_3. Here is the breakdown of categories in each region:

Region	Good	Medium	Bad
R_1	70	20	10
R_2	30	50	10
R_3	20	25	35

Calculate the overall classification error rate, Gini index, and cross-entropy using **weighted averages**.

Solution: There are $n_1 = 100$ observations in R_1, $n_2 = 90$ observations in R_2, and $n_3 = 80$ observations in R_3. The total number of observations is $n = 270$. So, in each calculation, the weight for R_1 is $w_1 = \frac{100}{270} = \frac{10}{27}$, the weight for R_2 is $w_2 = \frac{90}{270} = \frac{1}{3}$, and the weight for R_3 is $w_3 = \frac{80}{270} = \frac{8}{27}$.

(i) In R_1, the most common category is *Good*, so the classification error rate equals

$$\frac{20 + 10}{100} = 0.3.$$

In R_2, the most common category is *Medium*, so the classification error rate equals

$$\frac{30 + 10}{90} = \frac{4}{9}.$$

In R_3, the most common category is *Bad*, so the classification error rate equals

$$\frac{20 + 25}{80} = 0.5625.$$

Overall, we obtain

$$\frac{10}{27} \cdot \frac{3}{10} + \frac{1}{3} \cdot \frac{4}{9} + \frac{8}{27} \cdot \frac{9}{16} = 0.4259259.$$

Equivalently, and way more easily, we could have calculated

$$\frac{20+10+30+10+20+25}{100+90+80}=\frac{115}{270}=0.4259259.$$

(ii) In R_1, the Gini index is

$$\frac{70}{100}\cdot\frac{30}{100}+\frac{20}{100}\cdot\frac{80}{100}+\frac{10}{100}\cdot\frac{90}{100}=0.46$$

In R_2, the Gini index is

$$\frac{30}{90}\cdot\frac{60}{90}+\frac{50}{90}\cdot\frac{40}{90}+\frac{10}{90}\cdot\frac{80}{90}=0.5679012$$

In R_3, the Gini index is

$$\frac{20}{80}\cdot\frac{60}{80}+\frac{25}{80}\cdot\frac{55}{80}+\frac{35}{80}\cdot\frac{45}{80}=0.6484375$$

Overall, we have

$$\frac{10}{27}(0.46)+\frac{1}{3}(0.5679012)+\frac{8}{27}(0.6484375)=0.5518004$$

(iii) In R_1, the cross-entropy is

$$-(0.7\ln(0.7)+0.2\ln(0.2)+0.1\ln(0.1))=0.8018186.$$

In R_2, the cross-entropy is

$$-\left(\frac{30}{90}\ln\left(\frac{30}{90}\right)+\frac{50}{90}\ln\left(\frac{50}{90}\right)+\frac{10}{90}\ln\left(\frac{10}{90}\right)\right)=0.9368883.$$

In R_3, the cross-entropy is

$$-\left(\frac{20}{80}\ln\left(\frac{20}{80}\right)+\frac{25}{80}\ln\left(\frac{25}{80}\right)+\frac{35}{80}\ln\left(\frac{35}{80}\right)\right)=1.07173$$

Overall, we have

$$\frac{10}{27}(0.8018186)+\frac{1}{3}(0.9368883)+\frac{8}{27}(1.07173)=0.9268156.$$