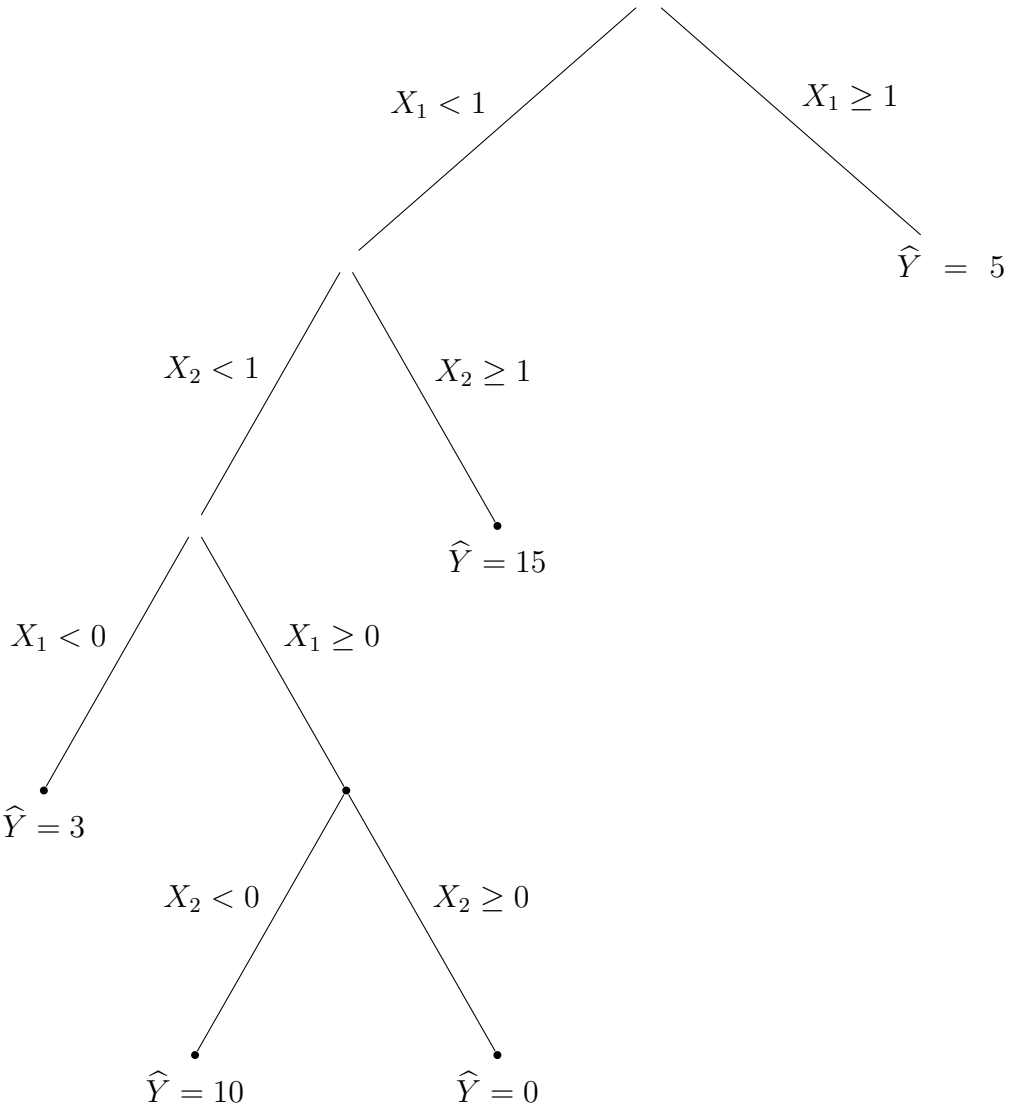
COM S 474/574 Spring 2020
Introduction to Machine Learning

Homework 4 - Solutions 1-3

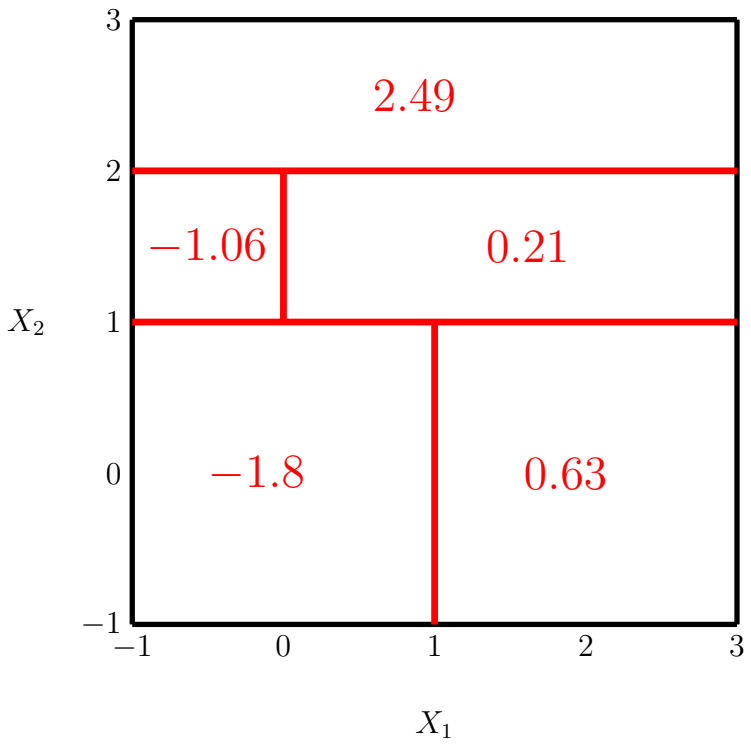
1 Solutions

Problem 1. [10 points] Ch. 8 #4 "This question relates ..." Note, in the book's Figure 8.12, the split has <threshold corresponding to the left branch and ≥threshold corresponding to the right branch.

A.



B.



Problem 2. [5 points] Ch. 8 #5 "Suppose we produce . . . "

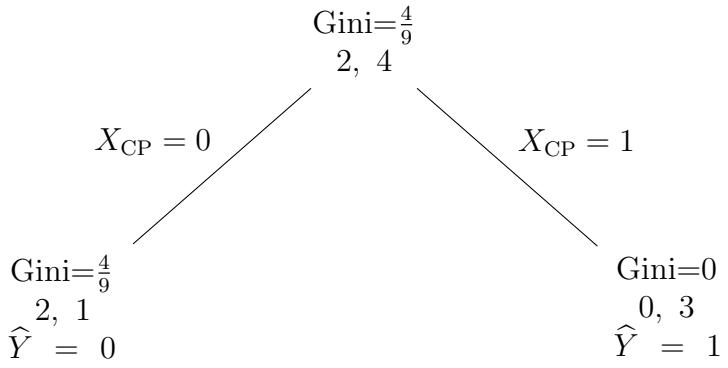
If we use a majority vote, we count how many estimates are above 0.5. There are 6 above with 4 below, so using majority vote, we would predict Red.

If we average the probability, then the average is 0.45, so the prediction would be Green.

Problem 3. [15 points] Suppose we construct a decision tree (shown below) for the data set in the following table, using the Gini index to select which features to split on. We want to predict $Y_{\text{Heart attack}}$.

Fill in the missing values, including the branches (what feature is split on, such as X_{male}), the Gini index at each node, the number of samples at each node with $Y = 0$ and $Y = 1$, and the predictions at the leaf nodes.

Show your calculations. Also report the training accuracy.



PATIENT ID	CHEST PAIN?	MALE?	SMOKES?	EXERCISES?	HEART ATTACK?
1.	yes	yes	no	yes	yes
2.	yes	yes	yes	no	yes
3.	no	no	yes	no	yes
4.	no	yes	no	yes	no
5.	yes	no	yes	yes	yes
6.	no	yes	yes	yes	no

First, let's compute the Gini value at the top node (that is, using the whole data set). There are 4 samples with heart attack ($Y = 1$), 2 without ($Y = 0$). So the Gini index is

$$\frac{2}{6}\left(1 - \frac{2}{6}\right) + \frac{4}{6}\left(1 - \frac{4}{6}\right) = \frac{2 * 2 * 4}{36} = \frac{4}{9}$$

Chest Pain: Next, what happens if we split using Chest Pain? When $X_{\text{CP}} = 0$, there is one sample with $Y = 1$ and two samples with $Y = 0$. So the Gini index for the left node would be

$$\frac{2}{3}\left(1 - \frac{2}{3}\right) + \frac{1}{3}\left(1 - \frac{1}{3}\right) = \frac{2 * 1 * 2}{9} = \frac{4}{9}$$

and for the right node with three $Y = 1$ samples and zero $Y = 0$ samples

$$\frac{0}{3}\left(1 - \frac{0}{3}\right) + \frac{3}{3}\left(1 - \frac{3}{3}\right) = 0$$

The weighted average would then be

$$\frac{3}{6}\left(\frac{4}{9}\right) + \frac{3}{6}(0) = \frac{2}{9}$$

Male: Next, what happens if we split using Male? When $X_{\text{M}} = 0$, there are 2 samples with $Y = 1$ and zero samples with $Y = 0$. So the Gini index for the left node would be

$$\frac{0}{2}\left(1 - \frac{0}{2}\right) + \frac{2}{2}\left(1 - \frac{2}{2}\right) = 0$$

and for the right node with two $Y = 1$ samples and two $Y = 0$ samples

$$\frac{1}{2}\left(1-\frac{1}{2}\right)+\frac{1}{2}\left(1-\frac{1}{2}\right)=\frac{1}{2}$$

The weighted average would then be

$$\frac{2}{6}(0)+\frac{4}{6}\left(\frac{1}{2}\right)=\frac{1}{3}$$

Smokes: Next, what happens if we split using Smokes? When $X_{\mathrm{S}} = 0$, there is 1 sample with $Y = 1$ and one samples with $Y = 0$. So the Gini index for the left node would be

$$\frac{1}{2}\left(1-\frac{1}{2}\right)+\frac{1}{2}\left(1-\frac{1}{2}\right)=\frac{1}{2}$$

and for the right node with three $Y = 1$ samples and one $Y = 0$ samples

$$\frac{1}{4}\left(1-\frac{1}{4}\right)+\frac{3}{4}\left(1-\frac{3}{4}\right)=\frac{2*1*3}{16}=\frac{3}{8}$$

The weighted average would then be

$$\frac{2}{6}\left(\frac{1}{2}\right)+\frac{4}{6}\left(\frac{3}{8}\right)=\frac{1}{6}+\frac{1}{4}=\frac{5}{12}$$

Exercises: Next, what happens if we split using Exercises? When $X_{\mathrm{E}} = 0$, there are 2 samples with $Y = 1$ and zero samples with $Y = 0$. So the Gini index for the left node would be

$$\frac{0}{2}\left(1-\frac{0}{2}\right)+\frac{2}{2}\left(1-\frac{2}{2}\right)=0$$

and for the right node with two $Y = 1$ samples and two $Y = 0$ samples

$$\frac{2}{4}\left(1-\frac{2}{4}\right)+\frac{2}{4}\left(1-\frac{2}{4}\right)=\frac{1}{2}$$

The weighted average would then be

$$\frac{2}{6}(0)+\frac{4}{6}\left(\frac{1}{2}\right)=\frac{1}{3}$$

Conclusion: The best feature to split on at the top node is Chest Pain. Its average Gini index is the lowest of any feature, and lower than that of the whole data set.

Using Chest Pain to split, only a single training sample is misclassified, so the training accuracy is $\frac{5}{6}$