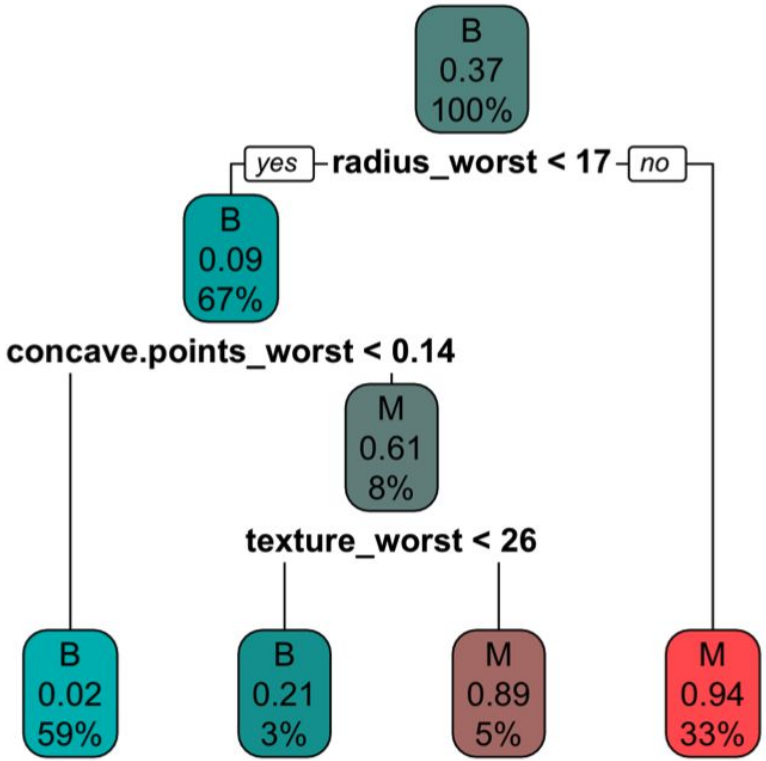
Chapter 7: Building a Decision Tree

Modern Clinical Data Science
Chapter Guides
Bethany Percha, Instructor

How to Use this Guide

- Read the corresponding notes chapter first
- Try to answer the discussion questions on your own
- Listen to the chapter guide (should be 15 min, max) while following along in the notes

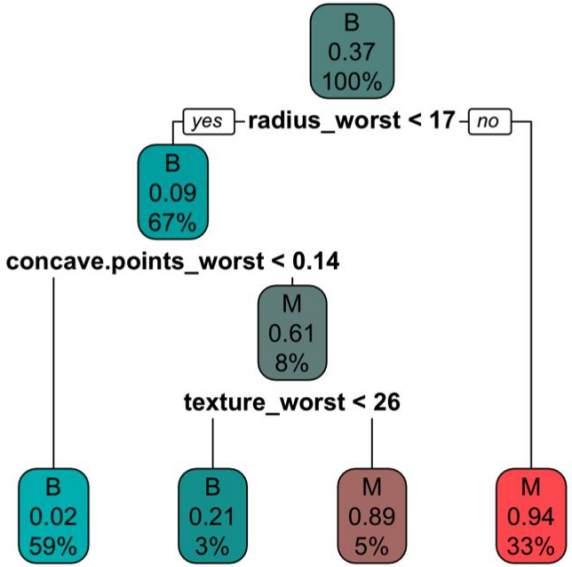

B
0.37
100%
yes
radius_worst < 17
no
B
0.09
67%
concave.points_worst < 0.14
M
0.61
8%
texture_worst < 26
B
0.02
59%
B
0.21
3%
M
0.89
5%
M
0.94
33%

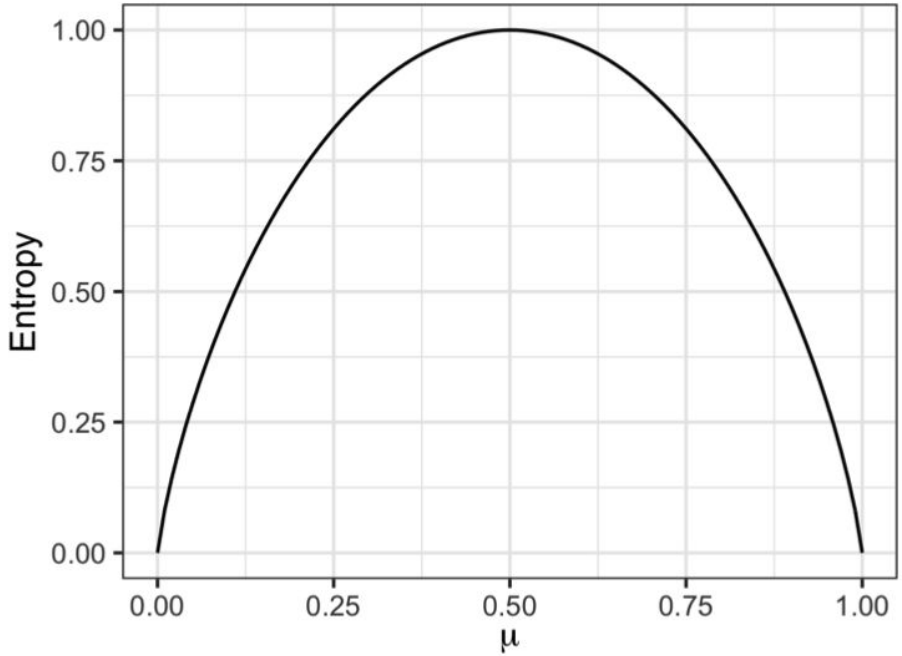
Question 7.1

What is the most important feature, as identified by the decision tree learning algorithm, for determining whether a breast mass is benign or malignant? What two other features are considered important by the tree? Which features are ignored completely?

Question 7.2

Looking at the decision tree for the Wisconsin Breast Cancer dataset, what do you think the advantages of a decision tree are for this problem over other classification methods, such as logistic regression and KNN?





Question 7.3

At which value(s) of μ are we maximally uncertain about the outcome? At which value(s) of μ are we completely certain about the outcome? This should make intuitive sense if you consider the meaning of μ.

Question 7.4

Say you have a dataset where the outcome is $Y = [0,1,0,1,0,1]$ and there are two predictors: $X = [0,0,1,1,2,2]$, and $Z = [1,2,1,2,1,2]$. Intuitively, which predictor would make the better splitting variable and why? Calculate $Gain(Y,X)$ and $Gain(Y,Z)$. Which value is higher?

Outcome (Y)	X	Z
0	0	1
1	0	2
0	1	1
1	1	2
0	2	1
1	2	2

$$\begin{aligned}\text{Gain}(Y,X) &= H(Y) - \sum_{x \in \text{Values(X)}} \frac{|Y(X=x)|}{|Y|} H(Y(X=x)) \\ &= H(Y) - H(Y|X)\end{aligned}$$

Datapoint ID	friends (X_1)	money (X_2)	free time (X_3)	happy (Y)
1	1	1	0	0
2	1	1	1	0
3	0	1	1	0
4	0	0	0	0
5	1	0	0	0
6	0	0	0	0
7	1	2	1	1
8	1	0	1	1
9	0	0	1	1
10	1	0	0	1

$$x_1 = \begin{cases} 0 & \text{no friends} \\ 1 & \text{friends} \end{cases} \qquad x_2 = \begin{cases} 0 & \text{poor} \\ 1 & \text{enough money} \\ 2 & \text{rich} \end{cases} \qquad x_3 = \begin{cases} 0 & \text{no free time} \\ 1 & \text{some free time} \end{cases}$$

Question 7.5

Build a decision tree for this dataset using the ID3 algorithm. To get started, you need to know the entropy of the overall outcome distribution. It is:

$$H(Y) = -\frac{4}{10}\log_2\frac{4}{10} - \frac{6}{10}\log_2\frac{6}{10} = \mathbf{0.971}$$

Datapoint ID	friends (X_1)	money (X_2)	free time (X_3)	happy (Y)
1	1	1	0	0
2	1	1	1	0
3	0	1	1	0
4	0	0	0	0
5	1	0	0	0
6	0	0	0	0
7	1	2	1	1
8	1	0	1	1
9	0	0	1	1
10	1	0	0	1

(a) Perform the initial split at the tree root to determine which variable to split on first. Update the tree with this information.

$$\frac{|Y(X_1=0)|}{|Y|}H[Y(X_1=0)] = \frac{4}{10}\left(-\frac{1}{4}\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4}\right) = 0.325$$

$$\frac{|Y(X_1=1)|}{|Y|}H[Y(X_1=1)] = \frac{6}{10}\left(-\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6}\right) = 0.600$$

$$\text{Gain}(Y, X_1) = 0.971 - 0.325 - 0.600 = \boxed{0.046}$$

$$\frac{|Y(X_2=0)|}{|Y|}H[Y(X_2=0)] = \frac{6}{10}\left(-\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6}\right) = 0.600$$

$$\frac{|Y(X_2=1)|}{|Y|}H[Y(X_2=1)] = \frac{3}{10}\left(-\frac{0}{3}\log_2\frac{0}{3} - \frac{3}{3}\log_2\frac{3}{3}\right) = 0$$

$$\frac{|Y(X_2=2)|}{|Y|}H[Y(X_2=2)] = \frac{1}{10}\left(-\frac{1}{1}\log_2\frac{1}{1} - \frac{0}{1}\log_2\frac{0}{1}\right) = 0$$

$$\text{Gain}(Y, X_2) = 0.971 - 0.600 - 0 - 0 = \boxed{0.371}$$ ← split here first

$$\frac{|Y(X_3=0)|}{|Y|}H[Y(X_3=0)] = \frac{5}{10}\left(-\frac{1}{5}\log_2\frac{1}{5} - \frac{4}{5}\log_2\frac{4}{5}\right) = 0.361$$

$$\frac{|Y(X_3=1)|}{|Y|}H[Y(X_3=1)] = \frac{5}{10}\left(-\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}\right) = 0.485$$

$$\text{Gain}(Y, X_3) = 0.971 - 0.361 - 0.485 = \boxed{0.125}$$

(b) We see that two of the leaves of our tree are "pure", meaning that all of the training examples that arrive there are of one outcome class. For those two leaves, we're done. However, for the third ($X_2 = 0$, or poor), we need to perform another split. Perform the split at the $X_2 = 0$ node to find which variable to split on next and update the tree with this information.

$$H[\mathcal{T}(X_2 = 0)] = -\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6} = 1.0$$

$$\frac{|\mathcal{T}(X_2 = 0, X_1 = 0)|}{|\mathcal{T}(X_2 = 0)|} H[\mathcal{T}(X_2 = 0, X_1 = 0)] = \frac{3}{6}\left(-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}\right) = 0.459$$

$$\frac{|\mathcal{T}(X_2 = 0, X_1 = 1)|}{|\mathcal{T}(X_2 = 0)|} H[\mathcal{T}(X_2 = 0, X_1 = 1)] = \frac{3}{6}\left(-\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}\right) = 0.459$$

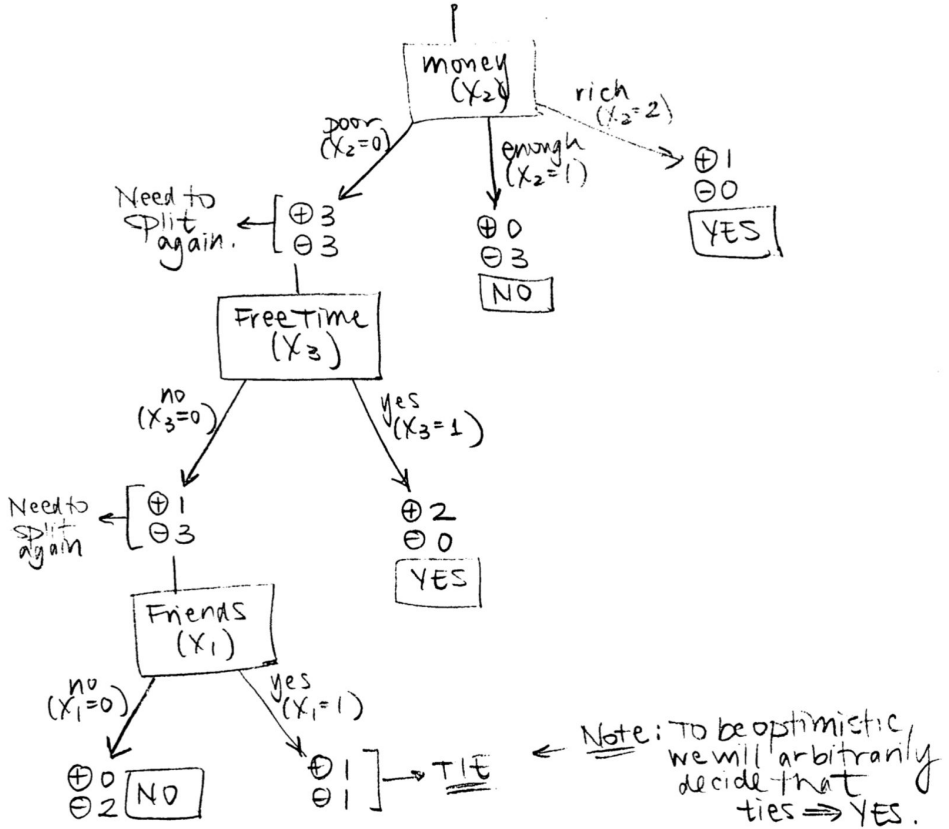
$$\text{Gain}(\mathcal{T}(X_2 = 0), X_1) = 1.0 - 0.459 - 0.459 = \boxed{0.082}$$

$$\frac{|\mathcal{T}(X_2 = 0, X_3 = 0)|}{|\mathcal{T}(X_2 = 0)|} H[\mathcal{T}(X_2 = 0, X_3 = 0)] = \frac{4}{6}\left(-\frac{1}{4}\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4}\right) = 0.541$$

$$\frac{|\mathcal{T}(X_2 = 0, X_3 = 1)|}{|\mathcal{T}(X_2 = 0)|} H[\mathcal{T}(X_2 = 0, X_3 = 1)] = \frac{2}{6}\left(-\frac{2}{2}\log_2\frac{2}{2} - \frac{0}{2}\log_2\frac{0}{2}\right) = 0$$

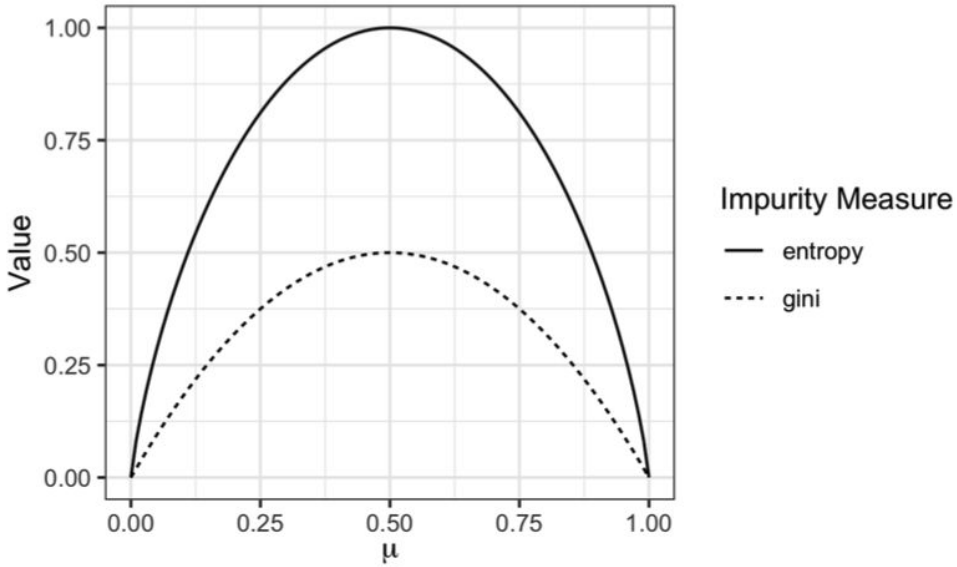
$$\text{Gain}(\mathcal{T}(X_2 = 0), X_3) = 1.0 - 0.541 - 0 = \boxed{0.459} \leftarrow \text{split here}$$

(c) We need to do one more split on the $X_2 = 0, X_3 = 0$ node. The only variable left to split on is X_1 (friends). Perform this split and add this information to the tree.



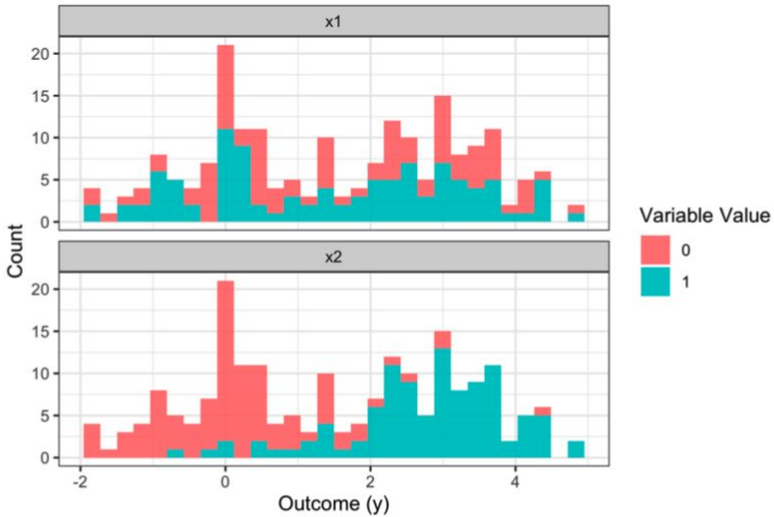
Question 7.6

Plots of the Gini impurity vs. entropy for a Bernoulli distribution are shown below. What do you notice about the value(s) of μ for which each is maximized or minimized?



Question 7.7

Imagine you have a dataset with two predictors, x_1 and x_2, each of which is binary (can only be 0 or 1). Here are the distributions of outcome values associated with x_1 and x_2:



Based on the idea of standard deviation reduction, which of these two variables, x_1 or x_2, would make the most sense for a decision tree to split on? What would such a split look like and what would the output value of the tree (the predicted value of y) be for each side of the split?

Question 7.8

Here are histograms of 10 of the predictors in the Wisconsin Breast Cancer dataset. Only the "worst" variable version of each predictor is shown for clarity. Which variable, and which threshold, appears to show the clearest division of samples into *B* and *M* groups? Compare your choice to the first split of the tree in Section 7.1.