Classification and Regression Trees (CART)

Exercises

1. Which of the following are advantages of using decision trees over other models? (*Select **all** that apply*)

   a. They have the ability to use data points that have missing values for variables in the model.

   b. They tell you the lift of your population.

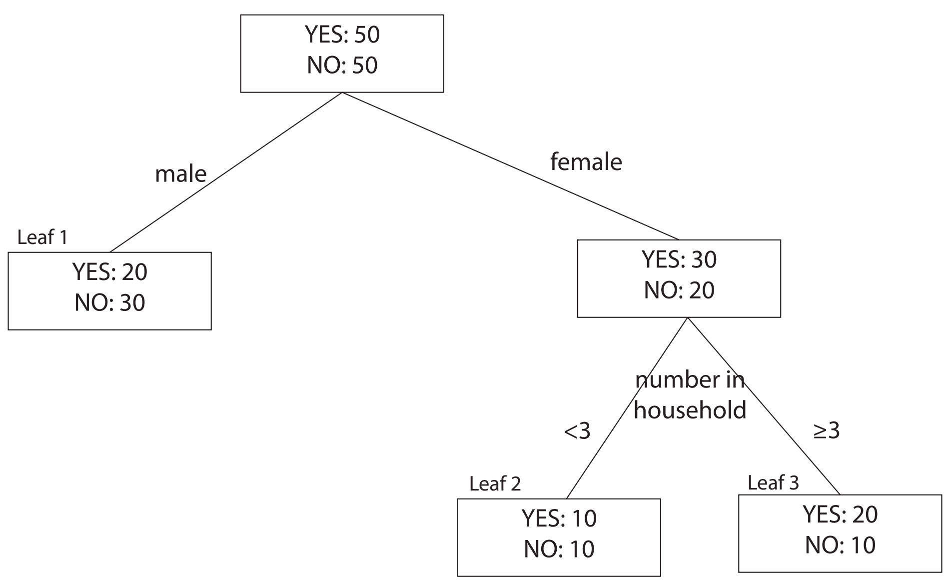
   c. They are easy to interpret so that anyone can easily see how each variable affects the target.

   d. They are almost always the best model for binary outcomes, due to their ability to fit complex patterns in a target variable.

   e. Once a decision tree model is created, it can be easy to implement, often times without the use of software.

2. I have a client base of 5000 people of whom 400 agreed to purchase an item from me. Based on the features of these clients, I got a tree whose best leaf had 500 people with 200 of them purchasing. From this, calculate the lift associated with marketing to this best 10%.

3. Suppose we have the following decision tree modeling a customer's response to an advertising campaign.



   a) What is the predicted probability of response for a male with 4 household members?

   b) Assuming a cutoff probability of 0.51, what is the misclassification rate of this tree?

   c) The concordance statistic (ROC statistic, area under the curve, etc.) involves tied pairs. How many tied pairs do we have here?

4. I have 2 predictor variables. One is *gender* with 2 possible values (M, F) and one is *age* in years taking on 36 different values in my data. My target is binary and I am building a decision tree using the Chi-square criterion. When I split on *gender*, my Chi-square p-value is 0.0100 and when I try splits on *age*, the (uncorrected) Chi-square p-value for the best *age* split is smaller, namely 0.0020.

   a) What is the logworth of *gender*?

   b) Which variable (*age* or *gender*) would I choose to split on if I used logworth with no Bonferroni (Kass) adjustment?

   c) If we use a Bonferroni correction on these p-values, The new p-values for *age* would be _____ and the new p-value for *gender* would be _____.

   d) Which variable (*age* or *gender*) would I use if I do a Bonferroni correction?

5. A tree for a binary response (*'Yes'* or *'No'*) has just two leaves. No decisions (i.e. profits & losses) or priors were specified. Here are the counts of *Yes*'s and *No*'s in the two leaves:

	Leaf 1	Leaf 2
Yes	400	200
No	100	300

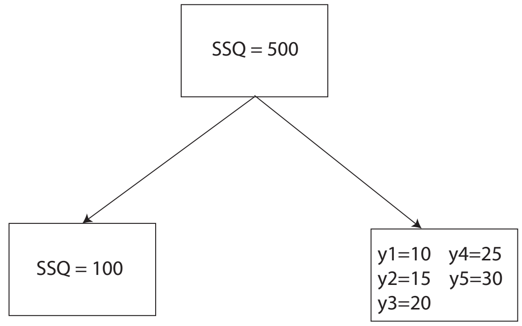
   There are only 3 points on my ROC curve in this situation, associated with varying the cutoff probability between 0 and 1:

   a) I could set the cutoff probability equal to 0, effectively calling everything a *'Yes'*. This would provide what point on the ROC curve?

   b) I could set the cutoff probability between .4 and .8, effectively calling everything in Leaf 1 a *'Yes'* and everything in Leaf 2 a *'No'*. What would be the corresponding point on the ROC curve for this interval of cutoff probabilities?

   c) I could set the cutoff probability equal to 1, effectively calling everything a *'No'*. What ROC point corresponds to this situation?

6. We build a regression tree using 10 observations. The first node (all observations) had sum of squared deviations from the mean (SSE) equal to 500. Upon splitting that node we get one leaf with $SSE = 100$ and one leaf with the observations shown below.



a) By how much did this split reduce the variation from what was in the parent node? In other words, what was the *gain* of the split in terms of SSE?

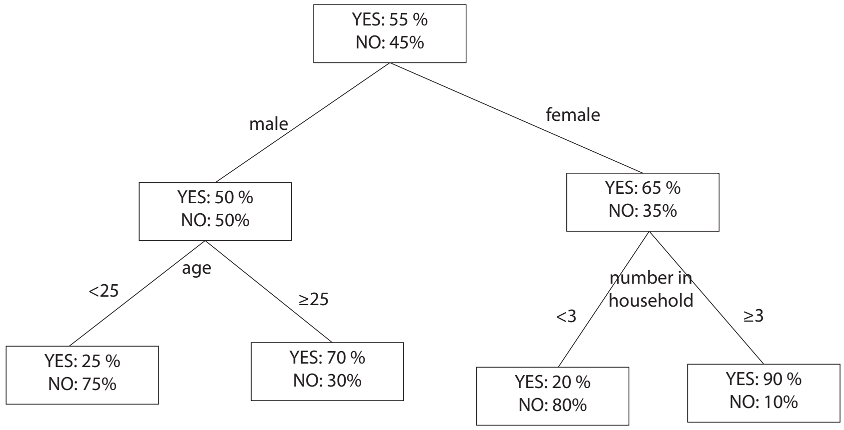
b) What would be the predicted value for observations in the leaf on the right?

7. The following table ***summarizes*** a data set with 3 binary input variables, X, Y and Z, and a binary target variable T:

X	**Y**	**Z**	**Number of Obs.**	
			$T = 1$	$T = 0$
T	T	T	10	0
T	T	F	0	0
T	F	T	10	0
F	T	T	0	20
T	F	F	25	0
F	F	T	0	10
F	T	F	0	0
F	F	F	0	5

a. Using misclassification rate as a splitting criterion, which attribute would be chosen for the first split? For each attribute, show the contingency table and the gains in misclassification rate.

8. Using the following decision/probability tree, if possible, fill in the predicted probabilities that the listed individuals will respond to the marketing campaign.



Name	Age	Gender	Number in Household	Pred. Probability
Jimbo	25	M	17	
MooMoo	22	M	1	
Madonna	20	F	3	
Batman	50	M	2	
Lulu	40	F	1	

List of Key Terms

Decision Trees

Regression Trees

'Probability Trees'

Leaf vs. Node

Pre-pruning

Post-pruning

Gain

Purity

Entropy

Gini

Average Squared Error

Bonferroni corrections

Kass adjustments

ROC curve

Sensitivity

Specificity

Lift at Depth