36-350: Data Mining

Homework 11
Date: November 6, 2002 **Due: start of class November 11, 2002**

---

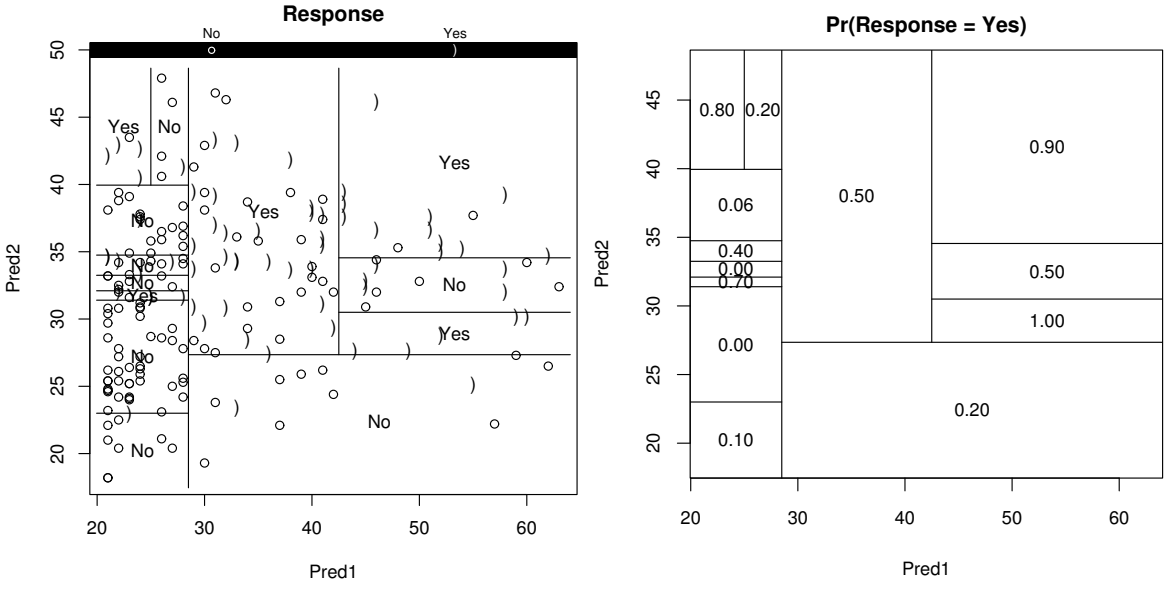
1. You want to construct a classification tree on a dataset with two predictors whose values are "Yes" and "No". The number of samples in class 1 and class 2 is given below for each combination of predictor values:

Predictor1	Predictor2	Class 1	Class 2
Yes	Yes	1	49
Yes	No	44	6
No	Yes	20	30
No	No	35	15

   Which predictor should be used at the top of the tree? Draw the tree that results from using this one predictor, including the class probabilities for each branch.

2. A data miner trains a classification tree on a dataset with 64 predictors. The tree ends up using 3 different predictor variables and has high accuracy on a holdout set. Does this mean that data in different classes do not differ significantly with respect to the other predictors? Explain.

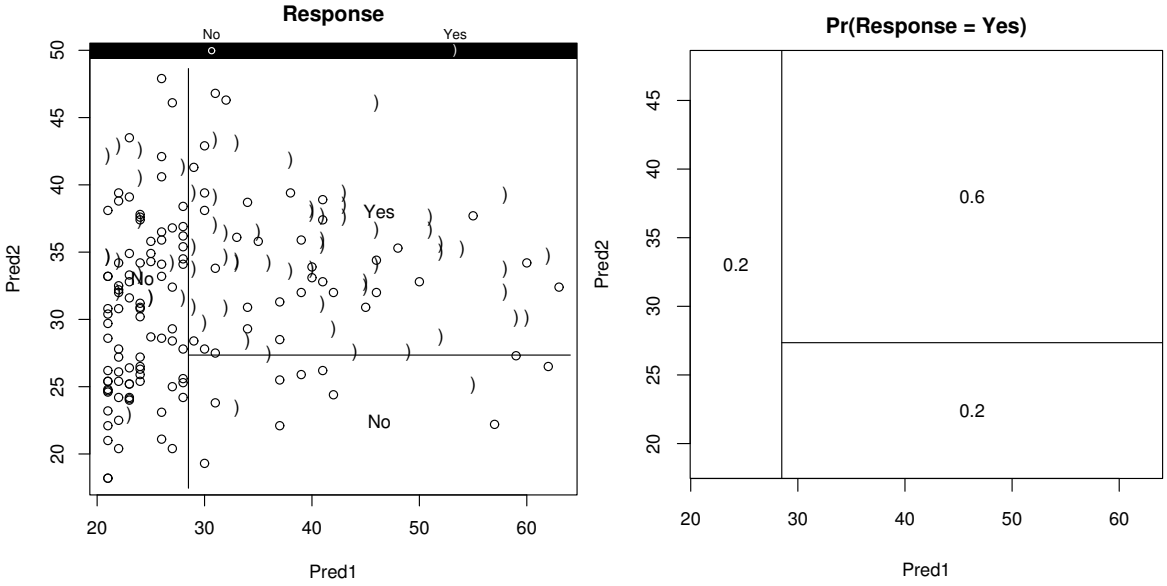
3. A classification tree is built from a dataset with two numerical predictors. The first plot below shows the data, the partitions made by the tree, and the most frequent response in each partition. The second plot shows the probability of a "Yes" response in each partition.



   (a) Some of the splits separate regions which make the same decision ("Yes" or "No"). Do these splits have any effect on the misclassification rate of the tree? Why are these splits made?

(b) Suppose the cost of a correct classification is zero, the cost of a false negative classification ("No" when truth is "Yes") is 3, and the cost of a false positive classification is 1. A new individual has `Pred1 = 30, Pred2 = 25`. Based on the tree, what is the expected cost of classifying this individual as "Yes"? What is the expected cost of classifying this individual as "No"? Which classification of the individual has minimum expected cost?

4. Below is a pruned version of the tree in the previous problem. It was pruned to 3 leaves based on 10-fold cross-validation.



(a) Which of the two trees, the smaller or the larger, would you expect to have better performance on the sample?

(b) Which of the two trees would you expect to have better performance on future data?

5. In the computer lab, you constructed a classification tree and a nearest-neighbor classifier on a dataset.

(a) Describe four benefits of using a classification tree on this dataset versus a nearest-neighbor classifier.

(b) This dataset is biased, i.e. it does not represent an independent random sample of people who apply for loans. Explain why this is so and what affect it will have on the classifier. Does this make the classifier useless?

6. Cross-validation is useful for reducing the complexity of a model. In previous labs, you reduced the complexity of linear models via the `step` function, which is based on AIC scoring, not cross-validation. Describe how cross-validation could be used to determine which predictor to drop from a linear model.

7. For the dataset depicted below, explain why a classification tree would be a poor choice. (Hint: try drawing the partitions.)