36-350: Data Mining

Homework 7
Date: October 12, 2001 **Due: start of class October 26, 2001**

---

1. We want to make a regression tree for the following table:

f1	f2	response
0	1	49
1	1	62
0	0	46
1	0	57

   Assuming the response is normal with constant variance, which predictor, f1 or f2, should we pick for the root of the tree? If we stop at that split, what are the leaf values? Draw the tree.

2. The file `hw7.dat` contains a data frame relating a single categorical predictor to a numerical response. You can read it via `x <- read.table("hw7.dat")`.

   (a) Using `break.factor`, find the best abstraction of the factor into two bins, assuming the response variance is the same for all categories. Remember that you can access variables in a data frame via `x$predictor` or `x$response`.

   (b) Repeat part (a) without assuming the response variance is the same for all categories. Explain any differences in the result.

3. In this problem and the next, you will analyze the factors that determine house prices. The file `Boston.dat` contains a data frame. Read in the frame via `Boston <- read.table("Boston.dat")`. The homework page has a description of the variables.

   (a) Make a regression tree to predict `medv` from two predictors, `rm` and `lstat`. We know from earlier in the class that these are the dominant factors. Do not use any special options to control the tree size. Make a `cplot` on these two factors which shows the data and partitions found by the tree. The plot does not have to be printed in color.

   (b) Using the above tree, what would be the predicted `medv` for a housing group with `rm = 7` and `lstat = 10`?

4. Now you will perform an analysis of the residuals of the Boston housing tree.

   (a) Compute the residuals for the tree in problem 2. Construct a tree to model them. Hint: Copy the table into a new variable and change `medv` to be the residuals, for example:

```
x <- Boston
x$medv <- res
```

(b) For the tree in part (a), find the terminal node (leaf of the tree) with largest mean. This is the node which corrects most of the error in the first tree. What are the two predictors which define this node?

(c) Construct an alternative tree for the residuals in part (a) which only uses `dis` and `lstat`. Make a `cplot`.

(d) Using the `cplot`, describe the relationship between `dis` and `lstat` in this dataset.

(e) Explain the second-order effect found by the tree in part (c). Focus on the cells with the largest means. In light of (d), is the data in these cells typical or atypical?