Statistics 5525: Homework 2

For each homework assignment, turn in at the beginning of class on the indicated due date. Late assignments will only be accepted with special permission. Write each problem up *very* neatly (LaTeX is preferred). Show all of your work.

Problem 1

Given the a dataset with covariates $X_i^t = < x_{i,1}, \ldots, x_{i,p} >$, and corresponding responses y_i $(i = 1, \ldots, N)$, consider the standardization transformation:

$$\tilde{x}_{i,j} = \frac{x_{i,j} - \bar{x}_{\cdot,j}}{\sqrt{\hat{\sigma}^2_{\cdot,j}}}.$$

$\bar{x}_{\cdot,j}$ and $\hat{\sigma}^2_{\cdot,j}$ represent the sample mean and variance across feature j, respectively.

Part a

Is CART invariant to using $\tilde{x}$ instead of x? In other words, are the answers equivalent? Explain why or why not.

Part b

Is LASSO regression invariant to using $\tilde{x}$ instead of x? In other words, are the answers equivalent? Explain why or why not.

Problem 2

Prove that the LASSO formulation

$$\begin{array}{rl} \min\limits_{\beta} & ||Y - X\beta||_2 \\ \text{subject to} & \sum\limits_k |\beta_k| < s, \end{array}$$

where $||\cdot||_2$ represents the Euclidean norm, is equivalent to the formulation:

$$\min_{\beta} ||Y - X\beta^c||_2 + \lambda \sum_{i=1}^{p} |\beta_i^c|.$$

Show the correspondence between the β_k^c's and the original β_k's. Hint: think about Lagrange multipliers.

Part 3

Load the spam dataset.

Part a

Build a Classification Tree with at least 100 terminal nodes. Using 10-fold cross validation, report the overall classification error rate.

Part b

Now determine a *simpler* tree (i.e. by pruning the tree). Again, using a 10-fold cross validation scheme, report the overall classification error rate.

Part c

Attempt to find an *optimal* tree under a 10-fold cross validation scheme. That is, try to find a tree that minimizes the cross validation error. While this is nearly an impossible task, see how close you can come. Describe your method and your overall error rate.

Part 4

Using the spam dataset, perform a logistic regression, and report the 10-fold cross validation error.

Part 5

Repeat the previous exercise using LASSO logistic regression, using the parameter λ that minimizes the deviance measure.