Statistics 5525: Homework 2

Due on Thurs., Oct. 13

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.

Yes, CART is invariant to using $\tilde{x}$ instead of x. When the covariates are standardized, the splitting threshold $t_1, t_2, t_3, ...$ will be rescaled (standardized) automatically by the algorithm of CART in order to minimize the in-group sum of squares, according to what dimension that t_i belongs to. As a result, the classification label for each standardized data point is the same as the classification label for the data point before the standardization. In other words, the classification results are equivalent before and after the standardization of the covariates when using CART.

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.

No, LASSO regression is not invariant to using $\tilde{x}$ instead of x. If we set D to be the transformation matrix with which we have

$$\tilde{X} = XD.$$

Plug the above expression into the formula for LASSO regression, we have

$$\min_{\boldsymbol{\beta}} ||Y - XD\boldsymbol{\beta}||_2 + \lambda||\boldsymbol{\beta}||_1 = \min_{\boldsymbol{\alpha}=D\boldsymbol{\beta}} ||Y - X\boldsymbol{\alpha}||_2 + \lambda||D^{-1}\boldsymbol{\alpha}||_1$$

This means, the LASSO after standardization is equivalent to the LASSO of the original covariate with a modified constraint $||D^{-1}\boldsymbol{\alpha}||_1 < s$. The shape of this new constraint in the parameter space will be different from the original one (for the covariates without standardization), which will result in a different set of parameter estimations. Thus, the LASSO won't give the equivalent answer before and after the standardization of the covariates.

Problem 2

Prove that the LASSO formulation

$$\min_{\beta} ||Y - X\beta||_2 \text{ subject to } \sum_k |\beta_k| < s, \tag{1}$$

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

$$\min_{\beta} ||Y - X\beta^c||_2 + \lambda \sum_{i=1}^{p} |\beta_i^c|. \tag{2}$$

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

Proof: Using the Lagrange multiplier, the constraint formulation of the LASSO can be expressed as follows. First, we add the term with Lagrange multiplier to the original sum of square of errors:

$$L = ||Y - X\boldsymbol{\beta}||_2 + \eta(||\boldsymbol{\beta}||_1 - s)$$

where L is the quantity we want to minimize. Then we obtain the LASSO estimator by

$$\min_{\boldsymbol{\beta}} L = \min_{\boldsymbol{\beta}}[||Y - X\boldsymbol{\beta}||_2 + \eta(||\boldsymbol{\beta}||_1 - s)]$$

Since s is a constant, the above minimization process can be simplified as

$$\min_{\boldsymbol{\beta}} L = \min_{\boldsymbol{\beta}}[||Y - X\boldsymbol{\beta}||_2 + \eta||\boldsymbol{\beta}||_1] = \min_{\boldsymbol{\beta}}[||Y - X\boldsymbol{\beta}||_2 + \eta \sum_{i=1}^{p} |\beta_i|] \tag{3}$$

for finding the LASSO estimators. Compare Eq.(3) with Eq.(2), it is not hard the constraint formulation of LASSO (Eq.(3) is derived from the constraint formulation shown in (1)) is equivalent to the Lagrange formulation of LASSO (shown in (2) directly). Thus, the β_k^c's and the original β_k's should be the same.

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.

Here I set the number of terminal nodes to be 110, the overall classification error rate is 0.03478. See below for the cross validation plot, and the decision tree. See Appendix for code.

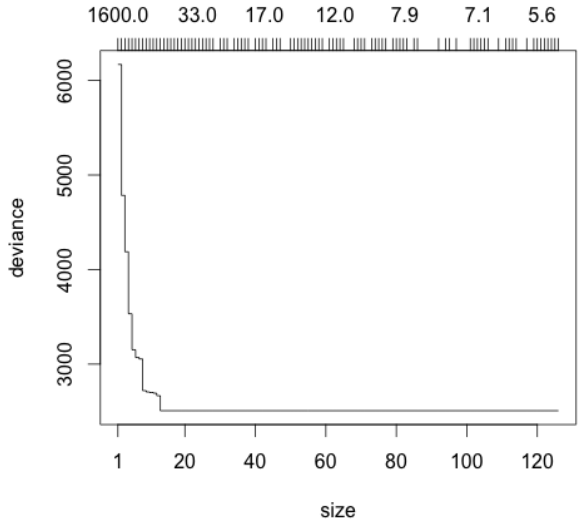

Figure 1: Cross validation plot for finding a classification tree with size larger than 100.

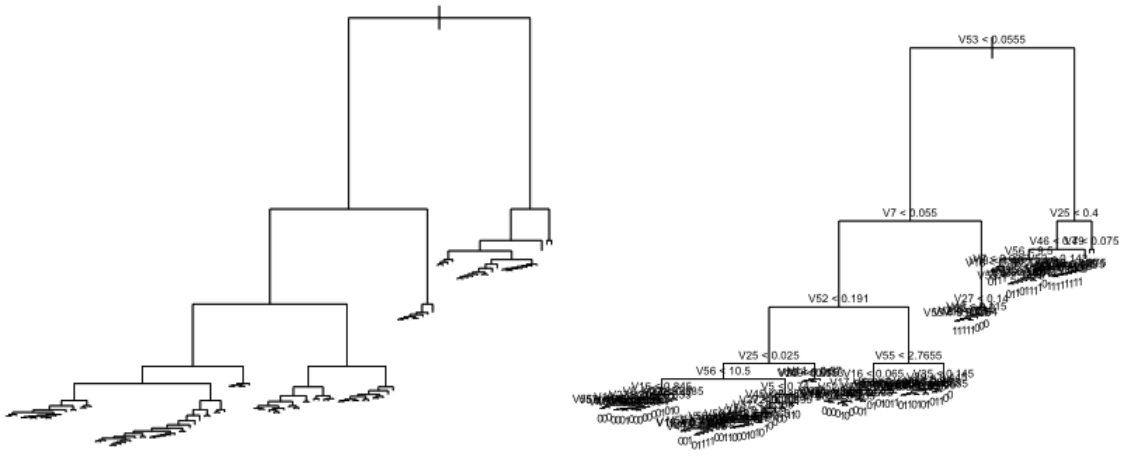

Figure 2: A classification tree with size of 110.

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.

Here I pruned the tree and let the final number of terminal nodes to be 20. The overall classification error rate is 0.07826 (, which is larger than the one in Part a due to simplification of the tree). See below for the cross validation plot, and the decision tree. See Appendix for code.

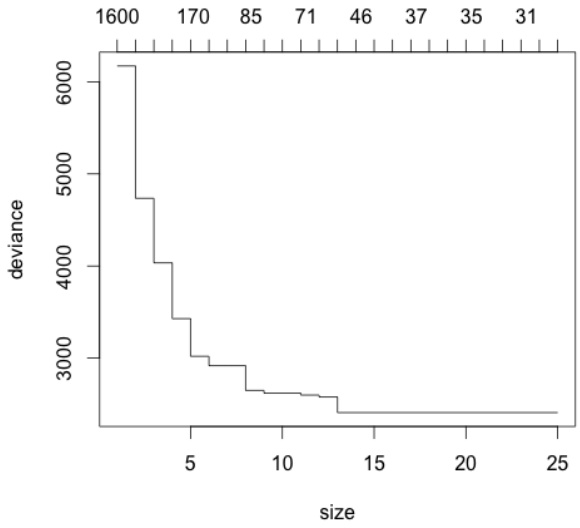

Figure 3: Cross validation plot for finding a classification tree with size smaller than 100.

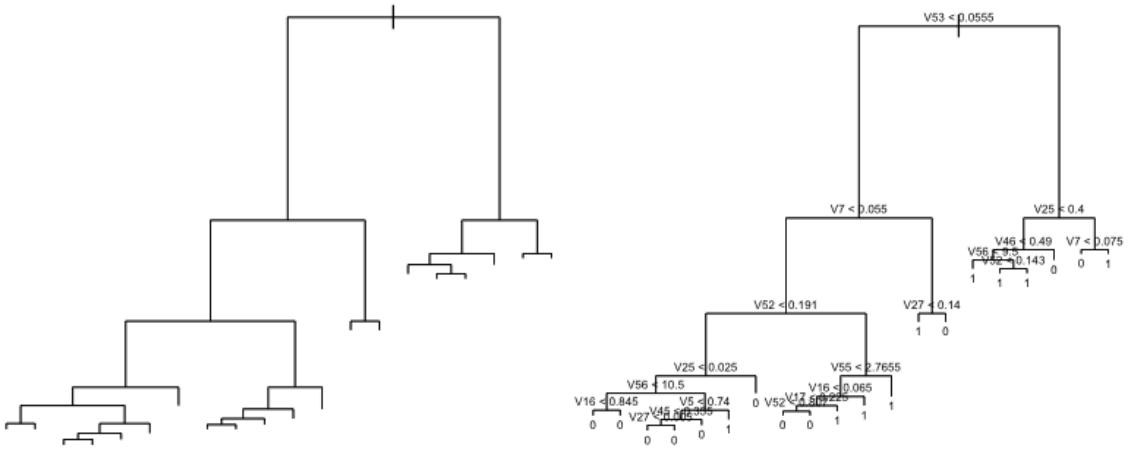

Figure 4: A classification tree with size of 20.

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.

In the plot of cross validation, I found the deviance (cross validation error) stops dropping when the size of the tree reaches 13. Then I pruned the tree with size=13. The pruned tree has overall error rate as 0.08261, slightly larger than that of Part b. See below the cross validation plot and the decision tree. See Appendix for code.

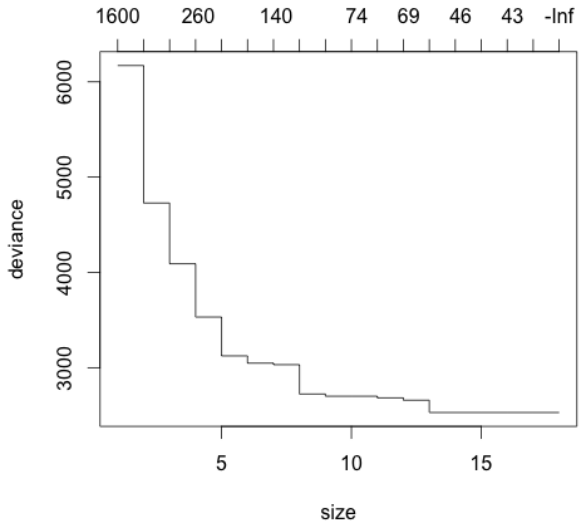

Figure 5: Cross validation plot for finding an optimized classification tree.

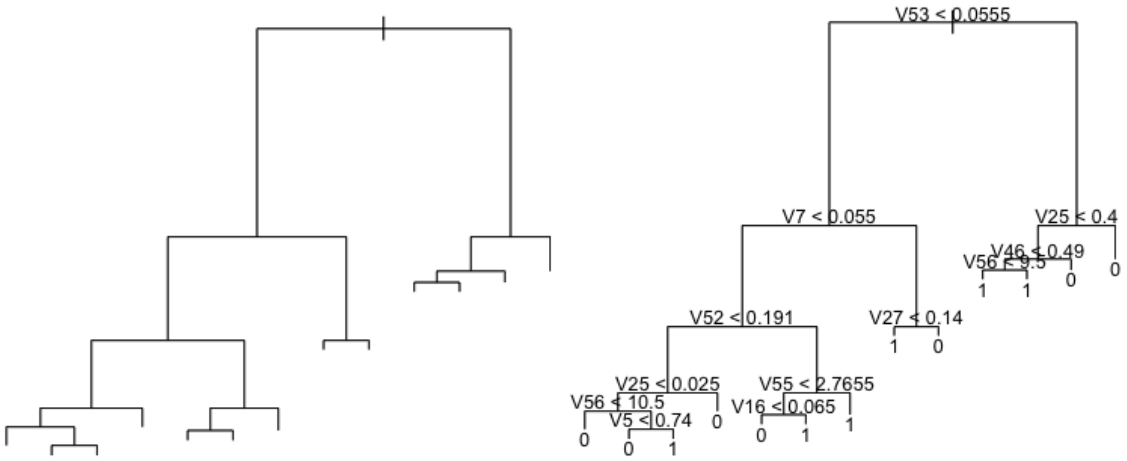

Figure 6: The optimized classification tree with size of 13.

Part 4

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

See Appendix for the implementation and detailed result for the logistic regression. The logistic regression shows that the significant predictor variables include V2, V5, V6, V7, V8, V9, V16, V17, V19, V21, V23, V24, V25, V26, V27, V28, V33, V35, V36, V39, V42, V44, V45, V46, V48, V49, V52, V53, V54, V56, V67.

The 10-fold cross validation error is found to be 0.05848635.

Part 5

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

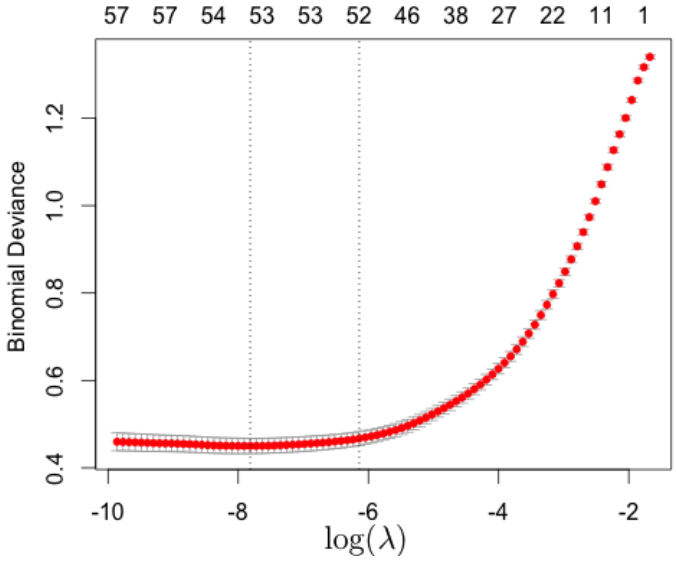

Figure 7: The cross validation plot with $\log(\lambda)$.

The parameter λ that minimizes the deviance measure is found to be 0.0004034505. See Appendix for the implementation of LASSO logistic regression with cross validation.