Statistics 5525: Homework 2

Ahmadreza Azizi and Bart Brown

October 15, 2017

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:

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

$\bar{x}_{.,j}$ and $\hat{\sigma}^2_{.,j}$ 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.
Answer:
Yes, because using either of Ginni or Cross entropy estimators, the form of Lagrangian is changed and so the optimized value is not necessarily the same. Lets consider the cross entropy loss function for classification:

$$V(\beta) = -\beta \ln(f(X, y)) - (1 - \beta) \ln(1 - f(X, y))$$

The above transformation $\tilde{x}_i = \frac{1}{\sqrt{\sigma^2}}(x_i - \bar{x})$ changes the shape of ln terms in the entropy loss function and so it changes the optimal values for βs. But we want to point out that since all points are shifted and scaled with constant values, shape of the final optimal trees does not change and stays invariant.

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.
Answer:

From the general form of LASSO regression :

$$\min_{\beta_0,\beta}\left\{\frac{1}{N}\sum_{i=1}^{N}\left(y_i-\beta_0-x_i^T\beta\right)^2\right\}\ ,\sum_{j=1}^{p}|\beta_j|\leq s$$

The transformation $\tilde{x}_i=\frac{1}{\sqrt{\sigma^2}}(x_i-\bar{x})$ has two parts. First is the translation $(x_i-\bar{x})$ and the second part is the scaling transformation $\frac{1}{\sqrt{\sigma^2}}$. The first term can be always absorbed in β_0 and so the other β_js can be unchanged, but by applying the scaling transformation β_js are also scaled and so they wont be invariant.

Problem 2

Prove that the LASSO formulation

$$\begin{aligned}\min_{\beta} \quad & ||Y-X\beta||_2\\ \text{subject to} \quad & \sum_k|\beta_k|<s,\end{aligned}$$

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.

Answer:

We want to show how two versions of the above formulations are equivalent:

In the first step we should say that if $\beta^*=arg\min_\beta||Y-X\beta^c||_2+\lambda\sum_{i=1}^{p}|\beta_i^c|$then the constraint term has play a role if only it is equal to zero since $\sum_k\beta^*-s<0$ does not have any contribution on minimizing the Lagrangian. So assume that the inequality condition can be replaced by an equality and so the general form changes to:

$$L(x)=||y-X\beta_k||^2,\qquad \sum_k|\beta_k|=s.\longmapsto C=\sum_k|\beta_k|-s=0.$$

$$\Rightarrow L(\beta,\lambda)=\frac{1}{2}||y-X\beta||^2+\lambda(C)=\frac{1}{2}||y-X\beta||^2+\lambda(\sum_k|\beta|-s)$$

where the second term λC is zero and one can add it up to the original lagrangian. Also the extra term λs is constant with respect to β and so it can be absorbed easily to the first term! and so:

$$\min_\beta L(\beta,\lambda)=\min_\beta\frac{1}{2}||y-X\beta^c||^2+\lambda(\sum|\beta^c|)$$

To show the correspondence between β and β_c, we should find the min values for both equations and find the relation between them which gives :

$$\beta^k = \beta^c + \frac{\lambda}{2X^2}$$

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.
Answer:
Using 100 leaf nodes and a 10 fold cross validation, we observed a 0.088 mean classification error with a variance of 0.0009.

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.
Answer:
We limited the number of leaf nodes to 10 in order to force a much smaller tree. The mean classification error for this case was 0.108 with a variance of 0.0011.

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.
Answer:
After studying the data and how the classification error responded to parameters such as leaf node number, tree depth, and minimum number of samples required to split an internal node, we found that leaf node number had the most profound effect on the results. Utilizing a grid search algorithm, we performed a search over the leaf node number parameter and found that 132 leaf nodes resulted in the minimum mean classification error given by 0.083. The search was performed over the range $2 : 200$ leaf nodes in steps of 1. The results of the search are shown in Fig. 1.

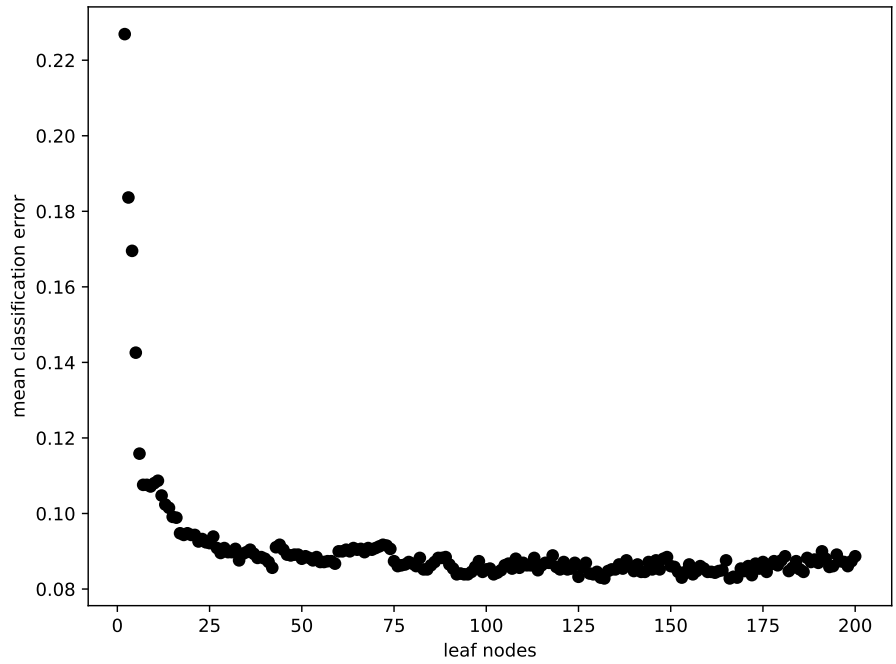

Figure 1: Problem 3 part c: The mean classification error taken over a 10-fold cross validation with respect to the leaf node number.

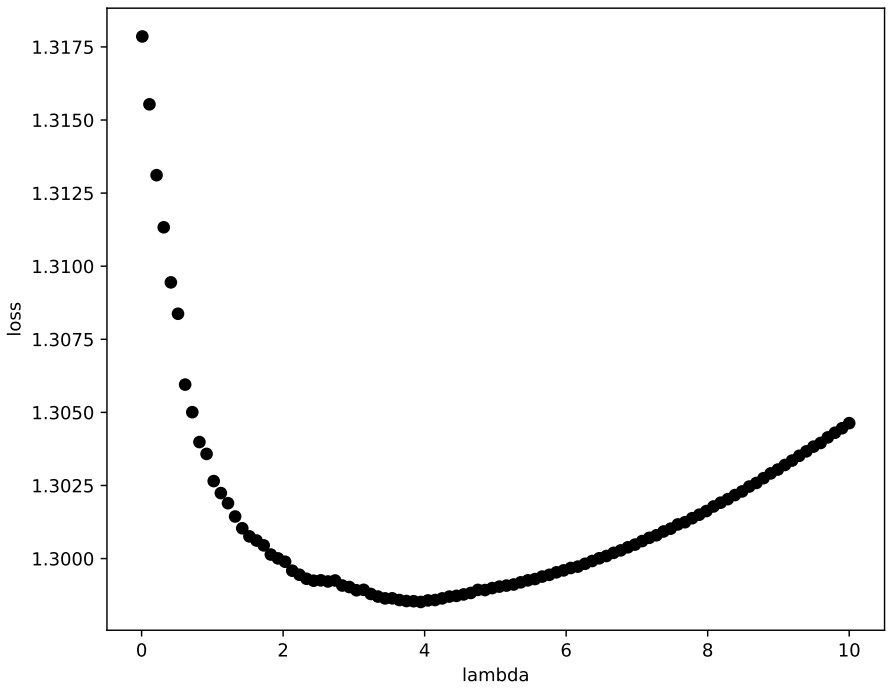

Figure 2: Problem 5: The mean loss taken over a 10-fold cross validation with respect to the l_1 penalty coefficient λ.

Part 4

Using the spam dataset, perform a logistic regression, and report the 10-fold cross validation error.
Answer:
Using logistic regression on the standardized data we observed the 10-fold cross validation to be 0.084 with a variance of 0.0009.

Part 5

Repeat the previous exercise using LASSO logistic regression, using the parameter λ that minimizes the deviance measure.
Answer:
Including an l_1 penalty with coefficient λ, we employed a grid search algorithm in order to minimize the loss function with respect to λ. We observed a minimum mean loss of 1.30 with variance 1.22×10^{-5} using a 10-fold cross validation approach. The λ which minimized the loss was found to be $\lambda = 4.05$. The results of the grid search are shown in Fig. 2. The calculated $\beta^\star$ is shown below:

$$\beta^\star = \begin{array}{cccccccc} \text{-2.2} & \text{-0.088} & \text{-0.17} & 0.070 & 0.34 & 0.35 & 0.17 & 0.90 \\ & 0.22 & 0.17 & 0.059 & \text{-0.018} & \text{-0.13} & 0.0 & 0.027 \\ & 0.23 & 0.82 & 0.39 & 0.072 & 0.13 & 0.38 & 0.26 \\ & 0.25 & 0.87 & 0.24 & 2.4 & \text{-0.74} & \text{-4.7} & 0.16 \\ & \text{-0.59} & \text{-0.11} & \text{-0.096} & 0.0 & \text{-0.32} & 0.0 & \text{-0.56} \\ & 0.28 & 0.0 & \text{-0.10} & \text{-0.27} & \text{-0.17} & \text{-0.60} & \text{-1.1} \\ & \text{-0.17} & \text{-0.68} & \text{-0.72} & \text{-1.1} & \text{-0.12} & \text{-0.50} & \text{-0.26} \\ & \text{-0.034} & \text{-0.065} & 0.33 & 1.3 & 0.53 & 0.0 & 1.6 \end{array}$$