Handout 07: Lasso Regression

Last time, we looked at adding a penalty term to our loss function to prefer smaller regression vectors over larger ones. Adding an ℓ_2-penalty leads to the ridge regression, which has some nice properties. For example, we can write down an analytic expression for the form of the regression vector and could prove (though we did not) that it does an ideal job of minimizing the variance of estimated regression vector.

Today we will look at two other penalties that could be added to the sum of squared residuals. The first is called the ℓ_0-norm, though it is not in fact a vector norm. It counts the number of non-zero terms in a vector:

$$||b||_0 = \#\{j \quad \text{s.t.} \quad b_j \neq 0\}. \tag{7.1}$$

Adding this to the least squares estimator leads to best subset regression:

$$\beta_\lambda^{BSR} = \arg\min_b \left\{||y - Xb||_2^2 + \lambda||b||_0\right\} \tag{7.2}$$

As another alternative, we can use the ℓ_1-norm, given by the sum of absolute values of the coordinates:

$$||b||_1 = \sum_j |b_j|. \tag{7.3}$$

This is a proper vector norm. Adding it to the square errors leads to the lasso regression vector:

$$\beta_\lambda^{LASSO} = \arg\min_b \left\{||y - Xb||_2^2 + 2\lambda||b||_1\right\} \tag{7.4}$$

Best subset regression is useful when you have only a small number of variables. For large datasets it is computationally intractable because the optimization problem is not convex. The only way to find a solution is to check every single combination of variables; the number of possibilities explodes beyond just a few variables. The lasso regression does not have an analytic solution but can be approximated using iterative methods; it is a convex optimization task. What makes it so attractive is that it will do a form of subset selection that, in practice, is nearly as good as the best subset selection.

Deriving the iterative solutions for the lasso regression problem is fairly extensive and not applicable to many other applications. We will not get into the details in this course. Today you are going to work with the simple case where the columns of X are uncorrelated:

$$X^t X = 1_p. \tag{7.5}$$

In this particular example it is possible to find analytic solutions to both best subset selection and the lasso regression. I think it yields a lot of motivation for understanding the behavior of the lasso in the more general case.

LAB QUESTIONS

1. This was a question from last lab, but in case you did not get there, make sure that you can derive a formula for ridge regression under the uncorrelated assumption in Equation 7.5.

2. Write the best subset selection loss function as a sum over j from $j = 1$ to $j = p$. That is, you can write the loss function as a sum of independent terms, where each elements depends only on the j'th column of X and the j'th component of β.

3. Repeat the previous question for lasso regression, showing that it also decouples over the individual variables.

4. Understand why your answers to the last two questions allow us to minimize the loss function individually for each component β_j.

5. Assume that $\beta_j \neq 0$. What would be its optimal value under best subset regression? When is the loss at this point better than setting $\beta_j = 0$? Put these conditions together to get a general formula for β_j under best subset selection.

6. Use a similar approach to find a solution for β_j for lasso regression. This requires a little bit more work but is very doable with simple one-dimensional calculus!

7. Consider a dataset for $n = 100$ where Equation 7.5 holds and we have:

$$X^t y = \begin{bmatrix} 10 \\ 3 \\ -5 \\ 1 \end{bmatrix} \tag{7.6}$$

   Draw a sketch with λ on the x-axis and β_k for the best subset selection on the y-axis. That is, you'll have four different curves showing the values of β_1, β_2, β_3, and β_4.

8. Repeat the previous question for lasso regression.

9. Finally, repeat the sketch for ridge regression.