IEOR 290A – Lecture 8
Lasso Regression

1 Lasso Regression

The M-estimator which had the Bayesian interpretation of a linear model with Laplacian prior

$$\hat{\beta} = \arg\min_{\beta} \|Y - X\beta\|_2^2 + \lambda\|\beta\|_1,$$

has multiple names: Lasso regression and $L1$-penalized regression.

1.1 Computation of Lasso Regression

Computation of this estimator is a complex topic because the objective is not differentiable, but for pedagogy we talk about how the corresponding optimization can be rewritten as a constrained quadratic program (QP). If we use an epigraph formulation, then we can rewrite the optimization as

$$\hat{\beta} = \arg\min_{\beta} \|Y - X\beta\|_2^2 + t$$

$$\text{s.t. } t \geq \lambda\|\beta\|_1.$$

But because $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$ (by definition), we can rewrite the above optimization as a constrained QP

$$\hat{\beta} = \arg\min_{\beta} \|Y - X\beta\|_2^2 + t$$

$$\text{s.t. } t \geq \lambda \sum_{j=1}^{p} \mu_j$$

$$- \mu_j \leq \beta_j \leq \mu_j, \forall j = 1, \ldots, p.$$

It is worth stressing that this is not an efficient way to compute Lasso regression.

1.2 Soft Thresholding

The Lasso regression estimate has an important interpretation in the bias-variance context. For simplicity, consider the special case where $X'X = \mathbb{I}_p$. In this case, the objective of the Lasso regression decouples

$$\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1 = Y'Y + \beta'X'X\beta - 2Y'X\beta + \lambda\|\beta\|_1$$

$$= Y'Y + \sum_{j=1}^{p} \left(\beta_j^2 - 2Y'X_j\beta_j + \lambda|\beta_j| \right),$$

where X_j is the j-th column of the matrix X. And because it decouples we can solve the optimization problem separately for each term in the summation.

Note that even though each term in the objective is not differentiable, we can break the problem into three cases. In the first case, $\beta_j > 0$ and so setting the derivative equal to zero gives

$$2\hat{\beta}_j - 2Y'X_j + \lambda = 0 \Rightarrow \hat{\beta}_j = Y'X_j - \lambda/2.$$

In the second case, $\beta_j < 0$ and so setting the derivative equal to zero gives

$$2\hat{\beta}_j - 2Y'X_j - \lambda = 0 \Rightarrow \hat{\beta}_j = Y'X_j + \lambda/2.$$

In the third case, $\beta_j = 0$.

For reference, we also compute the OLS solution in this special case. If we define $\hat{\beta}_j^0$ to be the OLS solution, then a similar calculation to the one shown above gives that $\hat{\beta}_j^0 = Y'X_j$. And so comparing to OLS solution to the Lasso regression solution, we have that

$$\hat{\beta}_j = \begin{cases} \hat{\beta}_j^0 + \lambda/2, & \text{if } \hat{\beta}_j = \hat{\beta}_j^0 + \lambda/2 < 0 \\ \hat{\beta}_j^0 - \lambda/2, & \text{if } \hat{\beta}_j = \hat{\beta}_j^0 - \lambda/2 > 0 \\ 0, & \text{otherwise} \end{cases}$$

This can be interpreted as a soft thresholding phenomenon, and it is another approach to balancing the bias-variance tradeoff.