MA 506 Probability and Statistical Inference

Lecture 16: Lasso Regression

1. Comparing different regression models ¶

Linear Regression (Ordinary Least Squares): With no penalty

$$\hat{\beta}^{OLS} = \arg\min_{\beta} \frac{1}{n}||Y - X\beta||_2^2 = (X^T X)^{-1} X^T Y$$

Ridge Regression: With L_2 penalty

$$\hat{\beta}^{ridge} = \arg\min_{\beta} \left[\frac{1}{n}||Y - X\beta||_2^2 + \lambda||\beta||_2^2\right] = (X^T X + n\lambda I)^{-1} X^T Y$$

Lasso Regression: With L_1 penalty

$$\hat{\beta}^{ridge} = \arg\min_{\beta} \left[\frac{1}{n}||Y - X\beta||_2^2 + \lambda||\beta||_1\right] =?$$

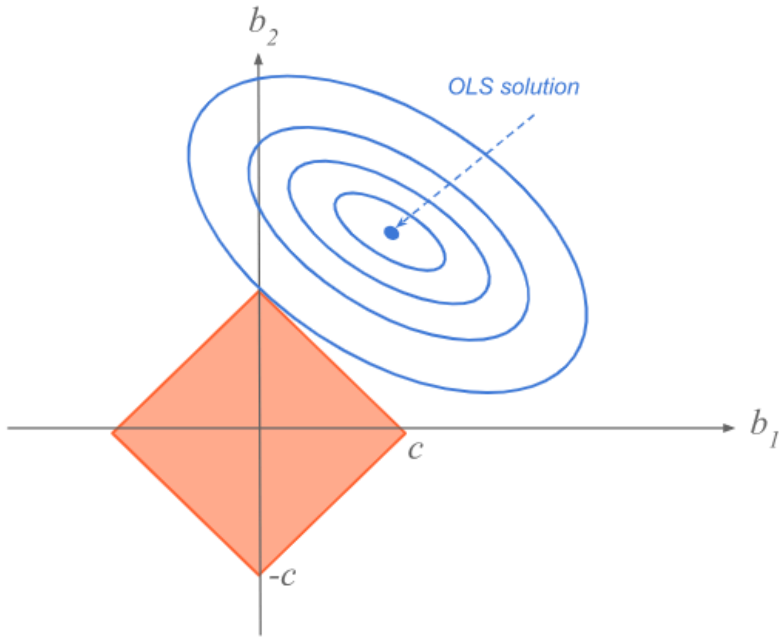
Here:

$$||\beta||_2 = \sqrt{\beta_0^2 + \beta_1^2 + \ldots + \beta_n^2}$$

$$||\beta||_1 = |\beta_0| + |\beta_1| + \ldots + |\beta_n|$$

2. More info on Lasso Regression

Lasso regression also solves the problem of large variance similar to Ridge regression. However, additionally it also helps in feature selection.



Here we are effectively solving:

$$\min_{\beta} \frac{1}{n}||Y - X\beta||_2^2, \text{ such that } ||\beta||_1 \leq c$$

The orange shape here represents the set of all β that satisfies: $||\beta||_1 \leq c$. Hence instead of choosing the OLS solution, Lasso regression chooses the first intersection of the contours (elliptical shapes) with the feasible region represented by the orange color. For L_1 norm, since there are corners across each axis so the optimal solution, is usually along some axis. This drives some other features to a 0 weight. In this lasso helps to choose some some features and delete the other ones.

For example in the above figure the optimal Lasso Regression solution at the intersection of the feasible orange region and outermost ellipse is on Y axis which makes the optimal solution as

$$\hat{\beta}^{lasso} = \begin{bmatrix} 0 \\ c \end{bmatrix}$$

And hence if we were fitting the function $y = b_1 + b_2 x$, then the optimal function is just $y = cx$ (as $b_1 = 0$)

Please note:

1. The same corner behavior of the feasible region happens in higher dimensions
2. Its not possible to find a closed form solution for Lasso regression ($\hat{\beta}^{lasso}$) as was the case for linear regression ($\hat{\beta}$) and ridge regression ($\hat{\beta}^{ridge}$). So we use an iterative algorithm to obtain that. The most commonly used algorithm is: Coordinate descent.
3. Typically weights decay with increase in λ.

In [ ]: