Lagrange Multipliers and Optimization

Dorian Goldman

February 21, 2017

1 Introduction to Constrained Optimization

Let $\beta = (\beta_1, \beta_2)$ be the desired coefficients in a linear regression so that we seek to minimize

$$\mathcal{L}(\beta) := \frac{1}{N}\sum_{i=1}^{N}(y_i - \beta_1 x_{i1} - \beta_2 x_{i2})^2. \tag{1.1}$$

Recall that we wish to penalize the size of the coefficients, so we impose a constraint on the size of β. More precisely, we seek to solve

$$\min_{\beta} \mathcal{L}_\lambda(\beta) \tag{1.2}$$

$$|\beta_1|^p + |\beta_2|^p \leq C, \tag{1.3}$$

for $p = 1, 2$. Let's define $g(\beta) = |\beta_1|^p + |\beta_2|^p$, and for now, focus on $p = 2$. Referring to the figure on the right below, the level sets of g are spheres. ie.

$$\beta_1^2 + \beta_2^2 \leq C.$$

How do we maximize this? Imagine to fix ideas that $C = 1$ so that $\beta_1^2 + \beta_2^2 \leq 1$, so that we seek to solve

$$\min_{\beta} \mathcal{L}_\lambda(\beta) \tag{1.4}$$

$$\beta_1^2 + \beta_2^2 \leq 1, \tag{1.5}$$

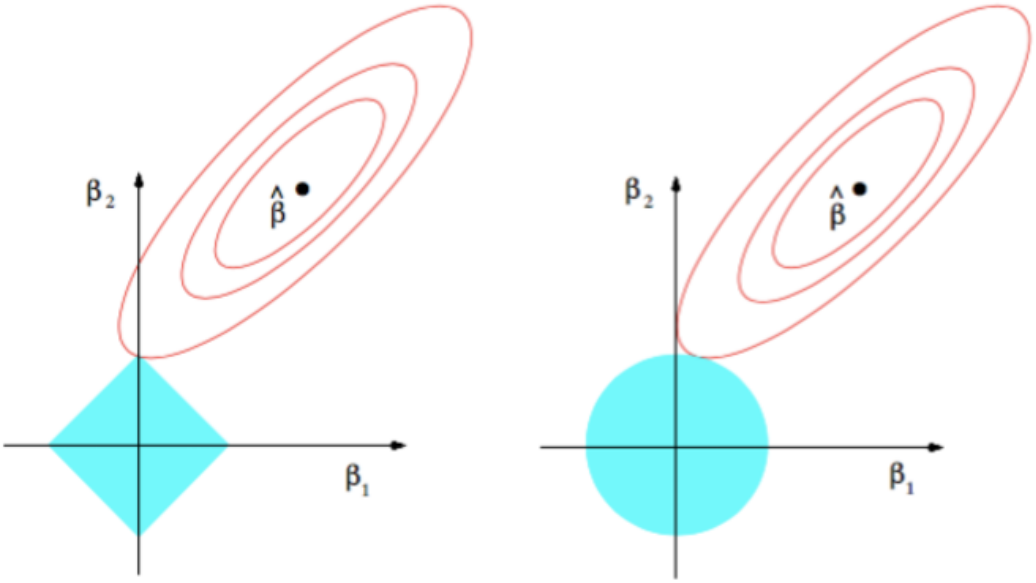

FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions* $|\beta_1| + |\beta_2| \leq t$ *and* $\beta_1^2 + \beta_2^2 \leq t^2$*, respectively, while the red ellipses are the contours of the least squares error function.*

Figure 1.1: L^1 and L^2 regularizaiton.

2 Derivation of Lagrange Multipliers

The following facts will make the above clear:

- $\beta \mapsto \mathcal{L}(\beta)$ is constant along level sets by definition.
- $\beta \mapsto \mathcal{L}(\beta)$ changes only in the direction **orthogonal** to the level sets, and this is given by $\nabla_\beta \mathcal{L}(\beta)$.
- **Case 1:** If $\nabla \mathcal{L}(\beta_0) = 0$ for some $\beta^{\mathbf{0}}$ in $\beta_1^2 + \beta_2^2 < 1$ then we solve as we do in normal calculus.
- **Case 2:** If $\nabla \mathcal{L}(\beta_0) \neq 0$ in $\beta_1^2 + \beta_2^2 < 1$. Then the minimum occurs on the boundary of $\beta_1^2 + \beta_2^2$. This is the Lagrange multiplier case.
- Recall the vector orthogonal to the level set is the gradient vector. So if $g(\beta_1, \beta_2) = \beta_1^2 + \beta_2^2$, then the orthogonal vector to the surface $\beta_1^2 + \beta_2^2$ is in the direction of $\nabla g = 2\langle \beta_1, \beta_2 \rangle$.
- **Main Point:** The minimum of $\mathcal{L}$ has to occur at a point where $\nabla \mathcal{L}$ is in the same direction as ∇g. If it weren't, then we could move along the surface $\beta_1^2 + \beta_2^2$ a bit to decrease the value (try drawing a picture or looking at the figures), so it wouldn't be a minimum!.

From the above points, we conclude that the minimum occurs at some point $\langle \beta_1^*, \beta_2^* \rangle$ such that

$$\nabla \mathcal{L}(\beta_1^*, \beta_2^*) = \lambda \nabla g(\beta_1^*, \beta_2^*).$$

3 Interpreting Lasso and Ridge regression

Observing the figure on the right, when we have $\beta_1^2 + \beta_2^2 \leq C$, we see the minimum has an equal chance of hitting the level set of $\beta_1^2 + \beta_2^2 = C$ at any point. As a result, the errors are generally equally distributed amongst the coefficients β_1 and β_2.

On the other hand, when $|\beta_1| + |\beta_2| = C$, we see that the level set of $\mathcal{L}$ is most likely to be tangent to the level sets (diamonds) at a corner (ie. where $\beta_1 = 0$ or $\beta_2 = 0$), **since there are only 4 possible directions where both of the coefficients are non-zero, making it highly unlikely that the level sets of $\mathcal{L}$ has a tangent parallel to any one of these directions.**

Conclusion: As a result, Lasso tends to result in *sparser* coefficients (ie. many zero coefficients), while Ridge generally distributes the error more evenly among the coefficients.