Lagrange Multipliers and Optimization

Dorian Goldman

February 22, 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 in order to not over fit our model, so we impose a constraint on the size of β. More precisely, we seek to solve the **constrained optimization problem:**

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

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

for $p = 1, 2$.

Common Question: Why do we choose $p = 1$ or $p = 2$? Why not some other p?

Answer:

- The norm L^2 (ie. $p = 2$) is very well behaved and is related to the equation for a sphere ($x^2 + y^2 = r^2$). Recall that we have an exact solution to the linear regression problem, ie. (1.1), when we choose our norm to be L^2 (as we have above, known

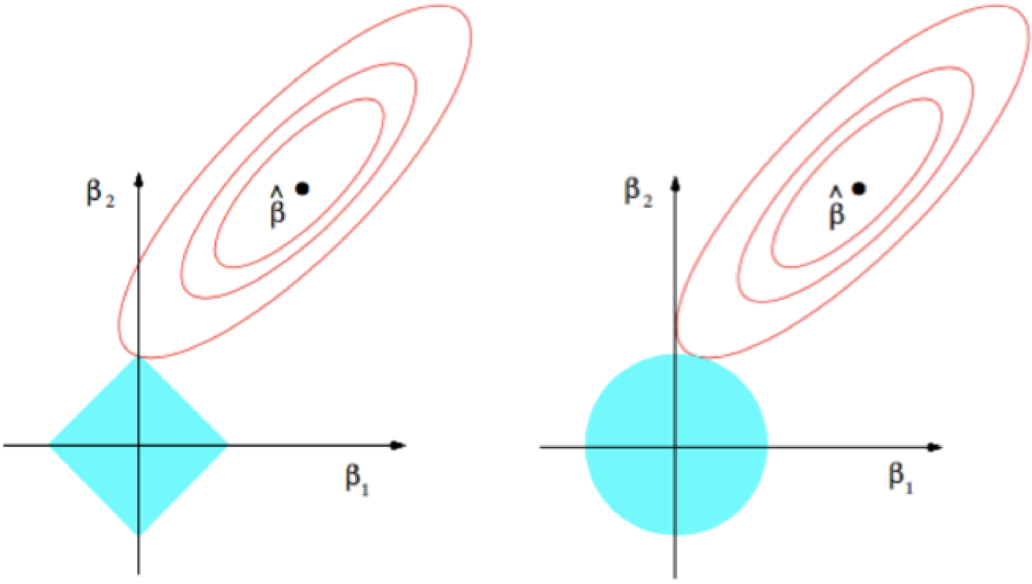

FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions $|\beta_1| + |\beta_2| \leq t$ and $\beta_1^2 + \beta_2^2 \leq t^2$, respectively, while the red ellipses are the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

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

as ordinary least squares - there is no exact solution if we replace the exponent 2 in (1.1) with a 1). For L^1 we can have many solutions to the same problem. ie. $|1/2| + |1/2| = 1$ and $|1| + |0| = 1$.

- The constant surfaces of L^2 (level surfaces, ie. where $x^2+y^2 = r^2$) are the surfaces of constant distance from the origin, meaning that distances are rotationally invariant (why would we want to count certain directions more than others?)
- Thinking of a convex function $f(x) = x^2$, this has a constant second derivative $f''(x) = 2$ so it's **uniformly convex** - meaning we have stability, uniqueness, etc when minimizing with respect to this norm generally.
- $p = 1$ is rather unique in that it provides a way of *regularizing which results in sparse coefficients* (ie. many zero coefficients, allowing you to choose the most important features). This is explained in these notes below.
- There isn't really any other advantage to using some other L^p space for $p > 2$, yet many disadvantages (such as degeneracy).

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, we seek to minimize (1.1) with some constraint

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

The size of the constraint depends on how strong we want our regularization to be, and we choose the constant C which gives the best performance on test data (as in lecture). More on this below.

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

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

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

2 Derivation of Lagrange Multipliers

The following facts will make the above clear:

- $\beta \mapsto \mathcal{L}(\beta)$ is constant along *level sets* (ie. where $\mathcal{L}(\beta) =$ constant) by definition.
- $\beta \mapsto \mathcal{L}(\beta)$ changes only in the direction **orthogonal** to the level sets, and this is given by $\nabla_\beta \mathcal{L}(\beta)$. This is clear because in any direction along the surface, $\mathcal{L}$ is constant.
- **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.

4 How much of a constraint do we use?

Recall from lecture that we want to train a model by solving the problem:

$$\min_{\beta} \mathcal{L}_\lambda(\beta|(x_i, y_i) = \text{ training data }) \tag{4.1}$$

$$|\beta_1|^p + |\beta_2|^p \leq C. \tag{4.2}$$

Let's denote β_C as the solution to the above constrained optimization problem (note it depends on C), so that our model is

$$f_C(x) = \beta_C \cdot x.$$

Then the optimal C, denoted C^* is determined by

$$C^* = \text{argmin}_C \frac{\sum_{i=1}^N (y_i - f_C(x_i))^2}{\sum_{i=1}^N (y_i - \bar{y})^2}. \tag{4.3}$$