Notes on Regularization and Coefficient Shrinkage

Signal Data Science

We can see that L^1 regularization successfully drives coefficient estimates to 0 as λ increases while L^2 regularization does not. Why does this happen? We can get more insight into what's going on by looking at the underlying mathematics.

Note: The reasoning below will consider only the single variable case, where we have a single regression coefficient β. However, the reasoning applies equally as well to higher-dimensional cases – the notation just get a little bit more cluttered.

Suppose we have a vector of true values $\mathbf{y}$ and a predictor variable $\mathbf{x}$, and consider an L^p regularized linear model for $\mathbf{y}$ in terms of $\mathbf{x}$ with regularization hyperparameter λ, coefficient estimate β, and intercept term I. That is, our model is given by

$$\mathbf{y} = \beta\mathbf{x} + I.$$

Call the sum of squared errors $\text{SSE} = S(\beta) = \sum_i (y_i - \beta x_i + I)^2$. Then our total cost function for the model is given by

$$C_p(\beta) = \text{SSE} + \lambda|\beta|^p = S(\beta) + \lambda|\beta|^p.$$

L^2 regularization

First, let's consider the case when we perform L^2 regularization. In that situation, $p = 2$ so $|\beta|^p = \beta^2$, and our cost function is

$$C_2(\beta) = S(\beta) + \lambda\beta^2.$$

What value of β minimizes $C_2(\beta)$? Since $C_2(\beta)$ is the sum of two quadratic functions of β, it is smooth (being a quadratic function of β itself) and therefore the minimum is achieved when $C_2'(\beta) = 0$, *i.e.*, when

$$S'(\beta) + 2\lambda\beta = 0.$$

Remember that we're interested in the situation where regularization causes the coefficient estimate β to be driven to 0. It's then natural to ask: what needs to be true for $C_2(\beta)$ to be minimized at $\beta = 0$? The condition $C_2'(0) = 0$ must hold. Substituting $\beta = 0$ into our expression above, we obtain the condition

$$S'(0) + 2\lambda \cdot 0 = S'(0) = 0.$$

We can conclude that L^2 regularization will drive the coefficient estimate β to 0 *if and only if* the condition $S'(0) = 0$ holds. Since the sum of squared errors $S(\beta)$ is a smooth quadratic function of β, the condition $S'(0) = 0$ is equivalent to saying that the sum of squared errors is minimized at $\beta = 0$, *i.e.*, that $\mathbf{y}$ is absolutely uncorrelated with $\mathbf{x}$.

Therefore: L^2 regularization drives the coefficient estimates to 0 if and only if the target variable is completely uncorrelated with its predictors. This is essentially *never* the case, so L^2 regularization will *never* drive coefficient estimates to 0.

We can also think about L^2 regularization in the following fashion: The cost function $C_2(\beta)$ is the sum of two convex quadratics, and the minimum of a sum of two convex quadratics has a minimum somewhere in between the minima of the two convex quadratics. *I.e.*, if the two convex quadratics are minimized at β_1 and β_2, their sum will be minimized for some β *between* but *not equal to* β_1 and β_2. If $\mathbf{y}$ and $\mathbf{x}$ have nonzero correlation, then the sum of squared errors is minimized at some value $\beta \neq 0$, whereas the quadratic regularization parameter is minimized at $\beta = 0$. As such, it is *impossible* for their sum to be minimized at $\beta = 0$ precisely; only in the *infinite limit* of $\lambda \to \infty$, where the regularization term *completely dominates* the sum of squared errors, does the minimum of their sum approach $\beta = 0$.

L^1 regularization

Now, let's consider L^1 regularization, where $p = 1$ so

$$C_1(\beta) = S(\beta) + \lambda|\beta|.$$

This is the sum of a quadratic function of β and a scaled absolute value function of β. Each of the two functions has a single local minimum, so the global minimum of $C_1(\beta)$ must be located at *either* (1) at the smooth local minimum of $C_1(\beta)$, where $C_1'(\beta) = 0$, *or* at (2) the minimum of the regularization parameter, where $\beta = 0$.

Taking the deriative of $C_1'(\beta)$, we obtain

$$C_1'(\beta) = S'(\beta) + \lambda \frac{|\beta|}{\beta}.$$

Since $S(\beta)$ is a quadratic function of β, $S'(\beta)$ is a linear function of β. Without any regularization (at $\lambda = 0$), $S'(\beta)$ is guaranteed to be 0 for *some* value of β (any straight line on the x–y axis will pass through $y = 0$ eventually).

Now, note that $|\beta|/\beta$ is equal to 1 for $\beta > 0$, equal to -1 for $\beta < 0$, and is undefined at $\beta = 0$. As such, adding on the regularization term $\lambda|\beta|/\beta$ to $S'(\beta)$ is equivalent to *shifting* the $\beta < 0$ side of the graph of $S'(\beta)$ down λ units, shifting the $\beta > 0$ side of the graph up λ units, and making the $\beta = 0$ point undefined. Intuitively, it must be the case that after a sufficiently large shift—after λ exceeds some finite threshold—the two halves of the graph are driven completely above and below the $\beta = 0$ line, and neither one attains the value of 0 anywhere. As such, the only remaining candidate for the minimum of $C_1(\beta)$ is at the nondifferentiable corner $\beta = 0$.

Formally, let $S'(\beta) = a\beta + b$ without loss of generality, where $a > 0$ is guaranteed because $S(\beta)$ is convex. Suppose also that $\mathbf{y}$ has a nonzero correlation with $\mathbf{x}$, so $b \neq 0$ (*i.e.*, $\beta = 0$ is no tthe solution to $S'(\beta) = 0$). Then

$$C_1'(\beta) = a\beta + b + \lambda \frac{|\beta|}{\beta}.$$

We aim to show that for sufficiently large λ, $C_1'(\beta) = 0$ has no solution. Setting everything equal to 0, we obtain

$$a\beta + b + \lambda|\beta|/\beta = 0$$

Suppose that we have a solution where $\beta > 0$, meaning that $a\beta + b + \lambda = 0$. Rearranging, we obtain $\beta = -b/a - \lambda/a$. Since a is guaranteed to be positive, increasing λ will decrease the value of the entire expression; indeed, for $\lambda > -b$ we obtain $\beta < 0$, a contradiction.

Similarly, suppose that we have a solution where $\beta < 0$, meaning that $a\beta + b - \lambda = 0$. Rearranging, we obtain $\beta = \lambda/a - b/a$, and for $L > b$ we obtain $\beta > 0$, a contradiction.

As such, for $\lambda > |b|$, where the right hand side is purely a function of $\mathbf{y}$ and $\mathbf{x}$, the only possible global minimum of $C_1(\beta)$ is at the nondifferentiable cusp $\beta = 0$.

Therefore: For sufficiently large λ, L^1 regularization is *guaranteed* to drive coefficient estimates to 0, unless the target variable is completely uncorrelated with its predictors.

Here's an alternative explanation for why L^1 regularization drives coefficient estimates to 0. Consider the simpler model where our variables have been appropriately rescaled (to mean 0) and reflected such that the model is just $\mathbf{y} = \beta\mathbf{x}$ for $\beta \geq 0$. Then the L^1 regularized cost function is

$$C_1(\beta) = \sum_i (y_i - \beta x_i)^2 + \lambda\beta.$$

Expanding out the sum, we obtain

$$C_1(\beta) = \beta^2 \mathrm{Var}(\mathbf{x}) - 2\beta \mathrm{Cov}(\mathbf{x}, \mathbf{y}) + \mathrm{Var}(\mathbf{y}) + \lambda\beta.$$

Doing a bit of factoring, we arrive at

$$C_1(\beta) = \beta^2 \mathrm{Var}(\mathbf{x}) + \mathrm{Var}(\mathbf{y}) + \beta \left(\lambda - 2\mathrm{Cov}(\mathbf{x}, \mathbf{y})\right).$$

Notice that $C_1(0) = \mathrm{Var}(\mathbf{y})$, a fixed value independent of λ. Assume that $C_1(\beta)$ is minimized at some value $\beta > 0$ for all λ. However, this is a contradiction, because if $\beta > 0$ we can increase the value of $C_1(\beta)$ to arbitrarily large values by increasing λ and thereby increasing $\beta \left(\lambda - 2\mathrm{Cov}(\mathbf{x}, \mathbf{y})\right)$. It must therefore be the case that for sufficiently large λ, the only possible minimum of $C_1(\beta)$ is at $\beta = 0$.