Parameters for Correlated Features in Elastic Net

David S. Rosenberg

Abstract

Zou and Hastie introduce the elastic net in a 2005 paper "Regularization and variable selection via the elastic net." What they call the "naive" elastic net, is what most people just call "elastic net" these days. Theorem 1 in their paper makes precise the statement that for [naive] elastic net, the more positively correlated two features are, the closer their parameter values. The original proof is quite readable, but here we've translated it to use our notation and standard definitions.

Recall the elastic net objective function:

$$J(w) = \frac{1}{n}\|Xw - y\|_2^2 + \lambda_1\|w\|_1 + \lambda_2\|w\|_2^2,$$

where $X \in \mathbf{R}^{n\times d}$ is a design matrix and $y \in \mathbf{R}^{n\times 1}$ is the vector of responses. Let $\hat{w} = (\hat{w}_1, \ldots, \hat{w}_d)^T \in \mathbf{R}^{d\times 1}$ be an elastic net solution – that is, $\hat{w}$ minimizes $J(w)$. Let's write x_i as the i'th column of the design matrix X. (Note the change from our usual notation, in which $x_i \in \mathbf{R}^d$ is the ith training example – here $x_i \in \mathbf{R}^n$ is the ith feature, across all training data.) As we often do in practice, let's assume the data are standardized so that every column x_i has mean 0, i.e. $1^T x_i = 0$, and standard deviation 1, i.e. $\frac{1}{n}x_i^T x_i = 1$. Then we can denote the correlation between any pair of columns x_i and x_j as $\rho_{ij} = \frac{1}{n}x_i^T x_j$. In the theorem below, we find that if x_i and x_j have high correlation, then their corresponding parameters $\hat{w}_i$ and $\hat{w}_j$ are close in value, assuming they have the same sign:

Theorem 1. *Under the conditions described above, if $\hat{w}_i\hat{w}_j > 0$, then*

$$|\hat{w}_i - \hat{w}_j| \leq \frac{\|y\|_2\sqrt{2}}{\sqrt{n}\lambda_2}\sqrt{1-\rho_{ij}}.$$

In the original theorem statement from the paper, they also require that y is centered. Although this is not required for the theorem to be true, replacing y with a centered version does not change the solution[1] $\hat{w}$, though it will reduce $\|y\|_2$ in the bound.

Proof. By assumption, $\hat{w}_i$ and $\hat{w}_j$ are nonzero, and thus $J(w)$ has partial derivatives w.r.t. $\hat{w}_i$ and $\hat{w}_j$. Moreover, we must have $\frac{\partial J}{\partial w_i}(\hat{w}) = \frac{\partial J}{\partial w_j}(\hat{w}) = 0$. That is,

$$\frac{\partial J}{\partial w_i}(\hat{w}) = \frac{2}{n}(X\hat{w} - y)^T x_i + \lambda_1 \text{sign}(\hat{w}_i) + 2\lambda_2 \hat{w}_i = 0$$

and

$$\frac{\partial J}{\partial w_j}(\hat{w}) = \frac{2}{n}(X\hat{w} - y)^T x_j + \lambda_1 \text{sign}(\hat{w}_j) + 2\lambda_2 \hat{w}_j = 0.$$

Subtracting the first equation from the second, we get

$$\begin{aligned} \frac{2}{n}(X\hat{w} - y)^T (x_j - x_i) + 2\lambda_2 (\hat{w}_j - \hat{w}_i) &= 0 \\ \iff (\hat{w}_i - \hat{w}_j) &= \frac{1}{n\lambda_2}(X\hat{w} - y)^T (x_j - x_i) \end{aligned}$$

Since $\hat{w}$ is a minimizer of J, we must have $J(\hat{w}) \leq J(0)$, so

$$\frac{1}{n}\|Xw - y\|_2^2 + \lambda_1\|\hat{w}\|_1 + \lambda_2\|\hat{w}\|_2^2 \leq \frac{1}{n}\|y\|_2^2.$$

Since the regularization terms are nonnegative, we must have $\|Xw - y\|_2^2 \leq \|y\|_2^2$.

Meanwhile,

$$\|x_j - x_i\|_2^2 = x_j^T x_j + x_i^T x_i - 2x_j^T x_i.$$

Recall our standardization assumptions were that $1^T x_i = 1^T x_j = 0$ and $\frac{1}{n}x_i^T x_i = \frac{1}{n}x_j^T x_j = 1$, and the correlation between x_i and x_j is $\rho_{ij} = \frac{1}{n}x_i^T x_j$. So

$$\|x_j - x_i\|_2^2 = 2n - 2n\rho_{ij}.$$

[1] The minimizer of $J(w)$ is unchanged if we replace y by its projection onto the column space of X. Since the columns of X are centered, $X^T 1 = 0$, so 1 is orthogonal to the column space of X. Thus $y - \bar{y}1$ has the same projection onto the column space of X as y does. So centering y does not change the solution $\hat{w}$. (Thanks Brett Bernstein for suggesting this.)

Putting things together,

$$\begin{aligned}
|\hat{w}_i - \hat{w}_j| &= \frac{1}{n\lambda_2}\left|(X\hat{w} - y)^T (x_j - x_i)\right| \\
&\leq \frac{1}{n\lambda_2}\|X\hat{w} - y\|_2 \|x_j - x_i\|_2 \text{ by Cauchy-Schwarz inequality} \\
&\leq \frac{1}{n\lambda_2}\|y\|_2 \sqrt{2n(1-\rho_{ij})} \\
&= \frac{1}{\sqrt{n}}\frac{\sqrt{2}\,\|y\|_2}{\lambda_2}\sqrt{1-\rho_{ij}}
\end{aligned}$$

□