Parameters for Correlated Features in Elastic Net

David S. Rosenberg

Recall the Elastic Net Lasso objective function

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

and let $\hat{w} = (\hat{w}_1, \ldots, \hat{w}_d)$ 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. As we would in practice, let's assume the data are standardized so that for 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 $\rho_{ij} = \rho(x_i, x_j) = \frac{1}{n}x_i^T x_j$. 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. [1]*Under the conditions described above, if $\hat{w}_i\hat{w}_j > 0$, then*

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

[original also assumes that y is centered, but I don't see why we need that... although it will make the bound better.]

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,

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

[1] Theorem 1 in "Regularization and variable selection via the elastic net": https://web.stanford.edu/~hastie/Papers/B67.2%20(2005)%20301-320%20Zou%20&%20Hastie.pdf

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

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

Putting things together,

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

□