Lasso, Ridge, and Elastic Net

David Rosenberg

New York University

February 4, 2016

A Very Simple Model

- Suppose we have one feature $x_1 \in \mathbf{R}$.
- Response variable $y \in \mathbf{R}$.

A Very Simple Model

- Suppose we have one feature $x_1 \in \mathbf{R}$.
- Response variable $y \in \mathbf{R}$.
- Got some data and ran least squares linear regression.
- The ERM is
$$\hat{f}(x_1) = 4x_1.$$

A Very Simple Model

- Suppose we have one feature $x_1 \in \mathbf{R}$.
- Response variable $y \in \mathbf{R}$.
- Got some data and ran least squares linear regression.
- The ERM is
$$\hat{f}(x_1) = 4x_1.$$
- What happens if we get a new feature x_2,
    - but we always have $x_2 = x_1$?

Duplicate Features

- New feature x_2 gives no new information.

Duplicate Features

- New feature x_2 gives no new information.
- ERM is still

$$\hat{f}(x_1, x_2) = 4x_1.$$

- Now there are some more ERMs:

$$\begin{aligned}\hat{f}(x_1, x_2) &= 2x_1 + 2x_2 \\ \hat{f}(x_1, x_2) &= x_1 + 3x_2 \\ \hat{f}(x_1, x_2) &= 4x_2\end{aligned}$$

Duplicate Features

- New feature x_2 gives no new information.
- ERM is still
$$\hat{f}(x_1, x_2) = 4x_1.$$
- Now there are some more ERMs:
$$\begin{aligned}
\hat{f}(x_1, x_2) &= 2x_1 + 2x_2 \\
\hat{f}(x_1, x_2) &= x_1 + 3x_2 \\
\hat{f}(x_1, x_2) &= 4x_2
\end{aligned}$$
- What if we introduce ℓ_1 or ℓ_2 regularization?

Duplicate Features: ℓ_1 and ℓ_2 norms

- $\hat{f}(x_1, x_2) = w_1 x_1 + w_2 x_2$ is an ERM iff $w_1 + w_2 = 4$.

Duplicate Features: ℓ_1 and ℓ_2 norms

- $\hat{f}(x_1, x_2) = w_1 x_1 + w_2 x_2$ is an ERM iff $w_1 + w_2 = 4$.
- Consider the ℓ_1 and ℓ_2 norms of various solutions:

w_1	w_2	$\|w\|_1$	$\|w\|_2^2$
4	0	**4**	16
2	2	**4**	**8**
1	3	**4**	10
-1	5	6	26

Duplicate Features: ℓ_1 and ℓ_2 norms

- $\hat{f}(x_1, x_2) = w_1 x_1 + w_2 x_2$ is an ERM iff $w_1 + w_2 = 4$.
- Consider the ℓ_1 and ℓ_2 norms of various solutions:

w_1	w_2	$\|w\|_1$	$\|w\|_2^2$
4	0	**4**	16
2	2	**4**	**8**
1	3	**4**	10
-1	5	6	26

- $\|w\|_1$ doesn't discriminate, as long as all have same sign
- $\|w\|_2^2$ minimized when weight is spread equally

Duplicate Features: ℓ_1 and ℓ_2 norms

- $\hat{f}(x_1, x_2) = w_1 x_1 + w_2 x_2$ is an ERM iff $w_1 + w_2 = 4$.
- Consider the ℓ_1 and ℓ_2 norms of various solutions:

w_1	w_2	$\|w\|_1$	$\|w\|_2^2$
4	0	**4**	16
2	2	**4**	**8**
1	3	**4**	10
-1	5	6	26

- $\|w\|_1$ doesn't discriminate, as long as all have same sign
- $\|w\|_2^2$ minimized when weight is spread equally
- Picture proof: Level sets of loss are lines of the form $w_1 + w_2 = c$...

Duplicate Features: Take Away

- For identical features
  - ℓ_1 regularization spreads weight arbitrarily (all weights same sign)
  - ℓ_2 regularization spreads weight evenly
- Extrapoloation to correlated variables:
  - ℓ_1 regularization may choose just one variable from a group and ignore the rest
  - ℓ_2 tends to spread weight roughly equally among correlated variables

Example with highly correlated features

- Model in words:
  - y is a linear combination of z_1 and z_2
  - But we don't observe z_1 and z_2 directly.

Example based on Section 4.2 in Hastie et al's *Statistical Learning with Sparsity*.

Example with highly correlated features

- Model in words:
  - y is a linear combination of z_1 and z_2
  - But we don't observe z_1 and z_2 directly.

  - We get 3 noisy observations of z_1.
  - We get 3 noisy observations of z_2.

Example based on Section 4.2 in Hastie et al's *Statistical Learning with Sparsity*.

Example with highly correlated features

- Model in words:
  - y is a linear combination of z_1 and z_2
  - But we don't observe z_1 and z_2 directly.
  - We get 3 noisy observations of z_1.
  - We get 3 noisy observations of z_2.
- We want to predict y from our noisy observations.

Example based on Section 4.2 in Hastie et al's *Statistical Learning with Sparsity*.

Example with highly correlated features

- Suppose (x, y) generated as follows:

$$
\begin{aligned}
z_1, z_2 &\sim \mathcal{N}(0,1) \text{ (independent)} \\
\varepsilon_0, \varepsilon_1, \ldots, \varepsilon_6 &\sim \mathcal{N}(0,1) \text{ (independent)}
\end{aligned}
$$

Example based on Section 4.2 in Hastie et al's *Statistical Learning with Sparsity*.

Example with highly correlated features

- Suppose (x, y) generated as follows:

$$
\begin{aligned}
z_1, z_2 &\sim \mathcal{N}(0,1) \text{ (independent)} \\
\varepsilon_0, \varepsilon_1, \ldots, \varepsilon_6 &\sim \mathcal{N}(0,1) \text{ (independent)} \\
y &= 3z_1 - 1.5z_2 + \varepsilon_0
\end{aligned}
$$

Example based on Section 4.2 in Hastie et al's *Statistical Learning with Sparsity*.

Example with highly correlated features

- Suppose (x, y) generated as follows:

$$
\begin{aligned}
z_1, z_2 &\sim \mathcal{N}(0,1) \text{ (independent)} \\
\varepsilon_0, \varepsilon_1, \ldots, \varepsilon_6 &\sim \mathcal{N}(0,1) \text{ (independent)} \\
y &= 3z_1 - 1.5z_2 + \varepsilon_0 \\
x_j &= \begin{cases} z_1 + \varepsilon_j/5 & \text{for } j = 1,2,3 \\ z_2 + \varepsilon_j/5 & \text{for } j = 4,5,6 \end{cases}
\end{aligned}
$$

Example based on Section 4.2 in Hastie et al's *Statistical Learning with Sparsity*.

Example with highly correlated features

- Suppose (x, y) generated as follows:

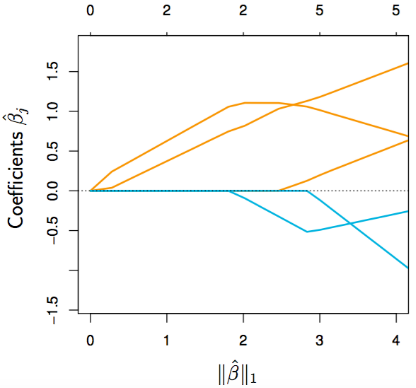
$$
\begin{aligned}
z_1, z_2 &\sim \mathcal{N}(0,1) \text{ (independent)} \\
\varepsilon_0, \varepsilon_1, \ldots, \varepsilon_6 &\sim \mathcal{N}(0,1) \text{ (independent)} \\
y &= 3z_1 - 1.5z_2 + \varepsilon_0 \\
x_j &= \begin{cases} z_1 + \varepsilon_j/5 & \text{for } j = 1,2,3 \\ z_2 + \varepsilon_j/5 & \text{for } j = 4,5,6 \end{cases}
\end{aligned}
$$

- Generated a sample of (x, y) pairs of size 100.
- Correlations within the groups of x's were around 0.97.

Example based on Section 4.2 in Hastie et al's *Statistical Learning with Sparsity*.

Example with highly correlated features

- Lasso regularization paths:



- This is not a good outcome – why?

From Figure 4.1 of Hastie et al's *Statistical Learning with Sparsity*.

Hedge Bets When Variables Highly Correlated

- When variables are highly correlated,
  - we want to give them roughly the same weight.
- Why?
  - robustness: what if one of the input variables has large error

From Figure 4.1 of Hastie et al's *Statistical Learning with Sparsity*.

Hedge Bets When Variables Highly Correlated

- When variables are highly correlated,
    - we want to give them roughly the same weight.
- Why?
    - robustness: what if one of the input variables has large error
- How can we get the weight spread more evenly?

From Figure 4.1 of Hastie et al's *Statistical Learning with Sparsity*.

Elastic Net

- The **elastic net** combines lasso and ridge penalties:

$$\hat{w} = \underset{w \in \mathbf{R}^d}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \left\{ w^T x_i - y_i \right\}^2 + \lambda_1 \|w\|_1 + \lambda_2 \|w\|_2^2$$

Elastic Net

- The **elastic net** combines lasso and ridge penalties:

$$\hat{w} = \underset{w \in \mathbf{R}^d}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \left\{ w^T x_i - y_i \right\}^2 + \lambda_1 \|w\|_1 + \lambda_2 \|w\|_2^2$$

- We expect correlated random variables to have similar coefficients.

Elastic Net

- The **elastic net** combines lasso and ridge penalties:

$$\hat{w} = \operatorname*{arg\,min}_{w \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^{n} \left\{ w^T x_i - y_i \right\}^2 + \lambda_1 \|w\|_1 + \lambda_2 \|w\|_2^2$$

- We expect correlated random variables to have similar coefficients.

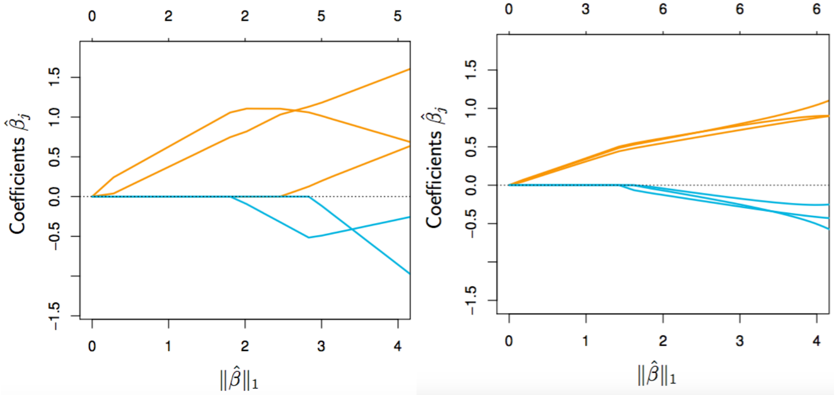
Theorem

[a] *Let $\rho_{ij} = \widehat{corr}(x_i, x_j)$. Suppose $\hat{w}_i$ and $\hat{w}_j$ are selected by elastic net. If $\hat{w}_i \hat{w}_j > 0$, then*

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

[a] https://web.stanford.edu/~hastie/TALKS/enet_talk.pdf

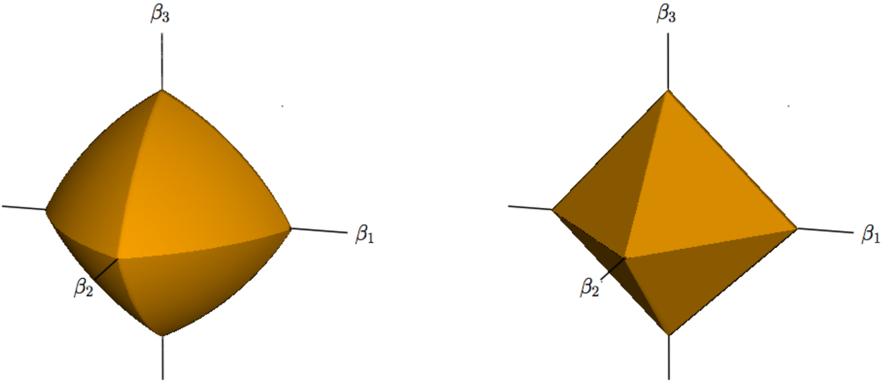
Elastic Net Results on Model



- Lasso on left; Elastic net on right.
- Ratio of ℓ_2 to ℓ_1 regularization roughly $2:1$.

From Figure 4.1 of Hastie et al's *Statistical Learning with Sparsity*.

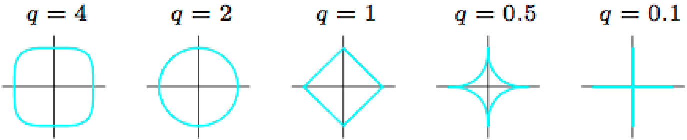
Elastic Net vs Lasso Norm Ball



From Figure 4.2 of Hastie et al's *Statistical Learning with Sparsity*.

The $(\ell_q)^q$ Norm Constraint

- Generalize to ℓ_q norm: $(\|w\|_q)^q = |w_1|^q + |w_2|^q$.
- $\mathcal{F} = \{f(x) = w_1x_1 + w_2x_2\}$.
- Contours of $\|w\|_q^q = |w_1|^q + |w_2|^q$:



$\ell_{1.2}$ vs Elastic Net

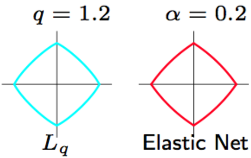


FIGURE 3.13. *Contours of constant value of* $\sum_j |\beta_j|^q$ *for* $q = 1.2$ *(left plot), and the elastic-net penalty* $\sum_j (\alpha\beta_j^2 + (1-\alpha)|\beta_j|)$ *for* $\alpha = 0.2$ *(right plot). Although visually very similar, the elastic-net has sharp (non-differentiable) corners, while the* $q = 1.2$ *penalty does not.*

From Hastie et al's *Elements of Statistical Learning*.