Normalization and Binary Features

Intro Presentation

Johan Larsson

Department of Mathemathical Sciences, University of Copenhagen

September 9, 2024

The Elastic Net

$$\boldsymbol{\beta}^* = \underset{\beta \in \mathbb{R}^p}{\operatorname{minimize}} \left(\frac{1}{2} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \underbrace{\lambda_1 \|\boldsymbol{\beta}\|_1}_{\text{lasso}} + \underbrace{\frac{\lambda_2}{2} \|\boldsymbol{\beta}\|_2^2}_{\text{ridge}} \right)$$

The Elastic Net

$$\boldsymbol{\beta}^* = \operatorname*{minimize}_{\beta \in \mathbb{R}^p} \left(\frac{1}{2} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \underbrace{\lambda_1 \|\boldsymbol{\beta}\|_1}_{\text{lasso}} + \underbrace{\frac{\lambda_2}{2} \|\boldsymbol{\beta}\|_2^2}_{\text{ridge}} \right)$$

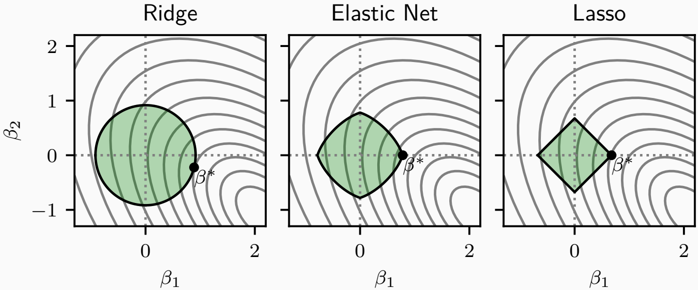

Figure 1: The elastic net penalty is a combination of the lasso and ridge penalties. Here shown as a constrained problem.

Sensitivity to Scale

Lasso and ridge penalties are **norms**—feature scales matter!

Sensitivity to Scale

Lasso and ridge penalties are **norms**—feature scales matter!

Example

Assume

$$\boldsymbol{X} \sim \mathrm{Normal}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}4 & 0\\0 & 1\end{bmatrix}\right), \qquad \boldsymbol{\beta}^* = \begin{bmatrix}\frac{1}{2}\\1\end{bmatrix}.$$

Sensitivity to Scale

Lasso and ridge penalties are **norms**—feature scales matter!

Example

Assume

$$\boldsymbol{X} \sim \mathrm{Normal}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}4 & 0\\0 & 1\end{bmatrix}\right), \qquad \boldsymbol{\beta}^* = \begin{bmatrix}\frac{1}{2}\\1\end{bmatrix}.$$

Model	$\hat{\boldsymbol{\beta}}$	$\hat{\boldsymbol{\beta}}_{\mathsf{std}}$
OLS	$\begin{bmatrix}0.50 & 1.00\end{bmatrix}^\intercal$	$\begin{bmatrix}1.00 & 1.00\end{bmatrix}^\intercal$

Sensitivity to Scale

Lasso and ridge penalties are **norms**—feature scales matter!

Example

Assume

$$\boldsymbol{X} \sim \mathrm{Normal}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}4 & 0\\0 & 1\end{bmatrix}\right), \qquad \boldsymbol{\beta}^* = \begin{bmatrix}\frac{1}{2}\\1\end{bmatrix}.$$

Model	$\hat{\boldsymbol{\beta}}$	$\hat{\boldsymbol{\beta}}_{\mathsf{std}}$
OLS	$\begin{bmatrix}0.50 & 1.00\end{bmatrix}^\intercal$	$\begin{bmatrix}1.00 & 1.00\end{bmatrix}^\intercal$
Lasso	$\begin{bmatrix}0.38 & 0.50\end{bmatrix}^\intercal$	$\begin{bmatrix}0.74 & 0.50\end{bmatrix}^\intercal$

Sensitivity to Scale

Lasso and ridge penalties are **norms**—feature scales matter!

Example

Assume

$$\boldsymbol{X} \sim \mathrm{Normal}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}\right), \qquad \boldsymbol{\beta}^* = \begin{bmatrix} \frac{1}{2} \\ 1 \end{bmatrix}.$$

Model	$\hat{\boldsymbol{\beta}}$	$\hat{\boldsymbol{\beta}}_{\mathsf{std}}$
OLS	$\begin{bmatrix} 0.50 & 1.00 \end{bmatrix}^\intercal$	$\begin{bmatrix} 1.00 & 1.00 \end{bmatrix}^\intercal$
Lasso	$\begin{bmatrix} 0.38 & 0.50 \end{bmatrix}^\intercal$	$\begin{bmatrix} 0.74 & 0.50 \end{bmatrix}^\intercal$
Ridge	$\begin{bmatrix} 0.37 & 0.41 \end{bmatrix}^\intercal$	$\begin{bmatrix} 0.74 & 0.41 \end{bmatrix}^\intercal$

Sensitivity to Scale

Lasso and ridge penalties are **norms**—feature scales matter!

Example

Assume

$$\boldsymbol{X} \sim \mathrm{Normal}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}4 & 0\\0 & 1\end{bmatrix}\right), \qquad \boldsymbol{\beta}^* = \begin{bmatrix}\frac{1}{2}\\1\end{bmatrix}.$$

Model	$\hat{\boldsymbol{\beta}}$	$\hat{\boldsymbol{\beta}}_{\text{std}}$
OLS	$\begin{bmatrix}0.50 & 1.00\end{bmatrix}^\intercal$	$\begin{bmatrix}1.00 & 1.00\end{bmatrix}^\intercal$
Lasso	$\begin{bmatrix}0.38 & 0.50\end{bmatrix}^\intercal$	$\begin{bmatrix}0.74 & 0.50\end{bmatrix}^\intercal$
Ridge	$\begin{bmatrix}0.37 & 0.41\end{bmatrix}^\intercal$	$\begin{bmatrix}0.74 & 0.41\end{bmatrix}^\intercal$

Large scale means less penalization because the size of β_j can be smaller for an equivalent effect (on $\boldsymbol{y}$).

Normalization

- Scale sensitivity can be mitigated by normalizing the features.

Normalization

- Scale sensitivity can be mitigated by normalizing the features.
- Let $\tilde{\boldsymbol{X}}$ be the normalized feature matrix, with elements

$$\tilde{x}_{ij} = \frac{x_{ij} - c_j}{s_j}.$$

Normalization

- Scale sensitivity can be mitigated by normalizing the features.
- Let $\tilde{\boldsymbol{X}}$ be the normalized feature matrix, with elements

$$\tilde{x}_{ij} = \frac{x_{ij} - c_j}{s_j}.$$

- After fitting, we transform the coefficients back to their original scale via

$$\hat{\beta}_j = \frac{\hat{\beta}_j^{(n)}}{s_j} \quad \text{for} \quad j = 1, 2, \ldots, p.$$

Type of Normalization Matters

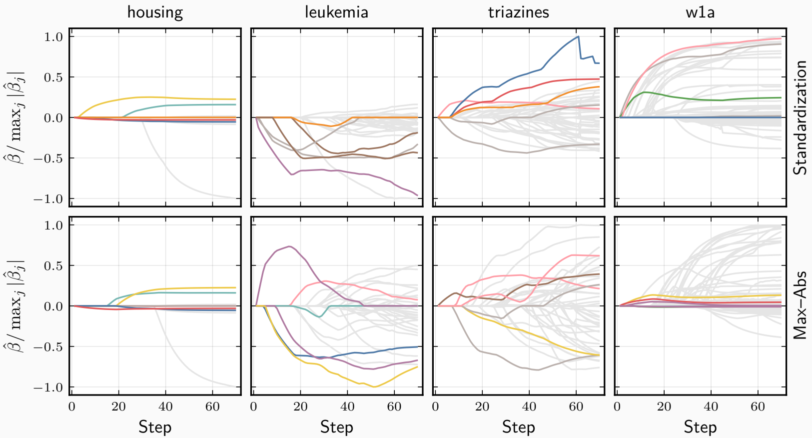

Figure 2: Lasso paths under two different types of normalization (standardization and max–abs normalization). The union of the first five features selected in any of the schemes are colored.

Binary Features

For binary features (values 0 and 1 only), we have for the noiseless case

$$\hat{\beta}_j = \frac{\mathrm{S}_{\lambda_1}\left(\frac{\beta_j^* n(q-q^2)}{s_j}\right)}{s_j\left(\frac{n(q-q^2)}{s_j^2} + \lambda_2\right)}.$$

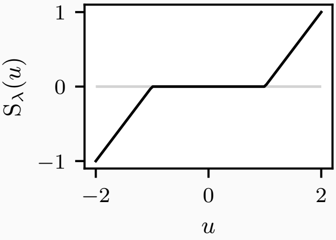

Figure 3: Soft-thresholding

Binary Features

For binary features (values 0 and 1 only), we have for the noiseless case

$$\hat{\beta}_j = \frac{\mathrm{S}_{\lambda_1}\left(\frac{\beta_j^* n(q-q^2)}{s_j}\right)}{s_j\left(\frac{n(q-q^2)}{s_j^2} + \lambda_2\right)}.$$

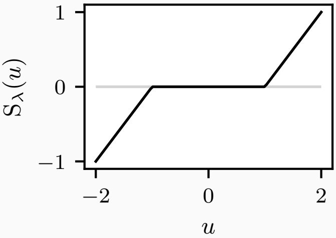

Figure 3: Soft-thresholding

Conclusions

- The elastic net estimator depends on class balance (q).

Binary Features

For binary features (values 0 and 1 only), we have for the noiseless case

$$\hat{\beta}_j = \frac{\mathrm{S}_{\lambda_1}\left(\frac{\beta_j^* n(q-q^2)}{s_j}\right)}{s_j\left(\frac{n(q-q^2)}{s_j^2} + \lambda_2\right)}.$$

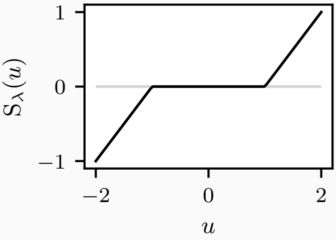

Figure 3: Soft-thresholding

Conclusions

- The elastic net estimator depends on class balance (q).
- $s_j = q - q^2$ for lasso and $s_j = \sqrt{q - q^2}$ for ridge removes effect of q.

Binary Features

For binary features (values 0 and 1 only), we have for the noiseless case

$$\hat{\beta}_j = \frac{\mathrm{S}_{\lambda_1}\left(\frac{\beta_j^* n(q-q^2)}{s_j}\right)}{s_j\left(\frac{n(q-q^2)}{s_j^2}+\lambda_2\right)}.$$

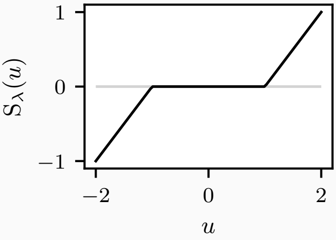

Figure 3: Soft-thresholding

Conclusions

- The elastic net estimator depends on class balance (q).
- $s_j = q - q^2$ for lasso and $s_j = \sqrt{q-q^2}$ for ridge removes effect of q.
- Suggests the parametrization

$$s_j = (q-q^2)^\delta, \qquad \delta \geq 0.$$

Mixed Data

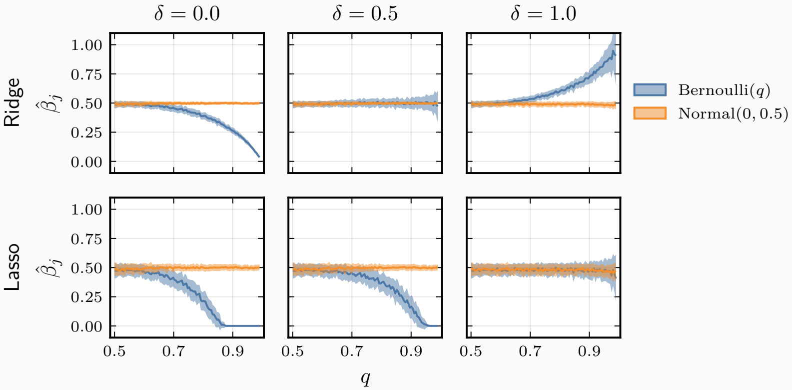

Figure 4: Comparison between lasso and ridge estimators for a data set with one binary and one quasi-normal feature.

Conclusions and Outlook

So Far

- Normalization matters

Conclusions and Outlook

So Far

- Normalization matters
- Binary features require special treatment; importance of class balance.

Conclusions and Outlook

So Far

- Normalization matters
- Binary features require special treatment; importance of class balance.
- Mixed data is tricky

Conclusions and Outlook

So Far

- Normalization matters
- Binary features require special treatment; importance of class balance.
- Mixed data is tricky

Conclusions and Outlook

So Far

- Normalization matters
- Binary features require special treatment; importance of class balance.
- Mixed data is tricky

Outlook

- Categorical features?

Conclusions and Outlook

So Far

- Normalization matters
- Binary features require special treatment; importance of class balance.
- Mixed data is tricky

Outlook

- Categorical features?
- Other types of noise and data?

Conclusions and Outlook

So Far

- Normalization matters
- Binary features require special treatment; importance of class balance.
- Mixed data is tricky

Outlook

- Categorical features?
- Other types of noise and data?
- Computational aspects of normalization?

Thank you!