Fjasdf

Subtitle

Johan Larsson

Department of Statistics, Lund University

April 2, 2024

Preliminaries

General Setup

- Data consists of a fixed matrix of features $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ and a response vector $\boldsymbol{y} \in \mathbb{R}^n$.
- $\boldsymbol{y}$ comes from a linear model, that is,

$$y_i = \beta_0^* + \boldsymbol{x}_i^\intercal \boldsymbol{\beta}^* + \varepsilon_i \quad \text{for} \quad i \in 1, \ldots, n,$$

  where $\boldsymbol{\beta}^*$ is the vector of *true* coefficients.
- ε_i is the measurement noise, generated from some random variable[1].

[1] No assumption on normality (yet).

The Elastic Net

Linear regression plus a combination of the ℓ_1 and ℓ_2 penalties:

$$\hat{\boldsymbol{\beta}} = \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} \left(\frac{1}{2} \|\boldsymbol{y} - \beta_0 - \tilde{\boldsymbol{X}}\boldsymbol{\beta}\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_1 + \frac{\lambda_2}{2} \|\boldsymbol{\beta}\|_2^2 \right).$$

The Elastic Net

Linear regression plus a combination of the ℓ_1 and ℓ_2 penalties:

$$\hat{\boldsymbol{\beta}} = \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} \left(\frac{1}{2} \|\boldsymbol{y} - \beta_0 - \tilde{\boldsymbol{X}}\boldsymbol{\beta}\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_1 + \frac{\lambda_2}{2} \|\boldsymbol{\beta}\|_2^2 \right).$$

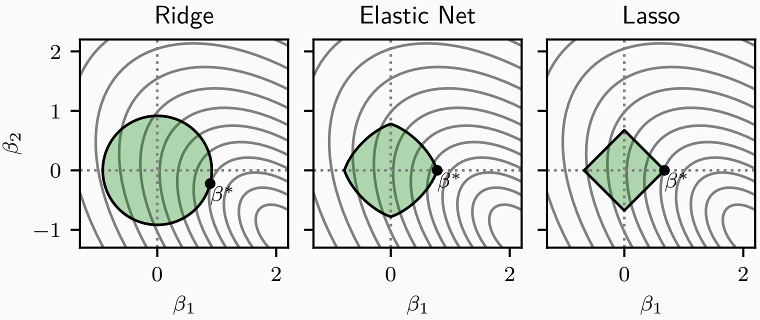

Figure 1: The elastic net penalty is a combination of the lasso and ridge penalties

Sensitivity to Scale

Since both the lasso and ridge penalties are norms, they are sensitive to the scale of the input features.

But what is the *optimal* scaling?

If the features are "normal", then most people would agree that stanardizing them (i.e., subtracting the mean and dividing by the standard deviation) is a good idea.

Normalization

Let $\boldsymbol{S}$ be the *scaling matrix*, which is a $p \times p$ diagonal matrix with entries $s_1, s_2, \ldots, s_p$. Let $\boldsymbol{C}$ be the *centering matrix*, which is an $n \times p$ matrix with each row equal to $[c_1, c_2, c_n]^\intercal$. Then the *normalized design matrix* $\tilde{\boldsymbol{X}}$ is defined as $\tilde{\boldsymbol{X}} = (\boldsymbol{X} - \boldsymbol{C})\boldsymbol{S}^{-1}$.

Table 1: Common ways to normalize a matrix of features

Normalization	Centering (c_{1j})	Scaling (s_j)
Standardization	$\frac{1}{n}\sum_{i=1}^{n} x_{ij}$	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2}$
Min–Max	$\min_i(x_{ij})$	$\max_i(x_{ij}) - \min_i(x_{ij})$
Unit Vector (L2)	0	$\sqrt{\sum_{i=1}^{n} x_{ij}^2}$
Max–Abs	0	$\max_i(\lvert x_{ij}\rvert)$
Adaptive Lasso	0	β_j^{OLS}

Binary Features

Let's say we have a binary feature $\boldsymbol{x}_j$, such that $x_{ij} \in \{0, 1\}$.

What is the "best" way to scale this feature?

Solution for Binary Features

We assume the that normalized features are orthogonal, that is

$$\tilde{\boldsymbol{X}}^\intercal \tilde{\boldsymbol{X}} = \operatorname{diag}(\dots)$$

Class Imbalance

Mixed Data

Mixed Data

Experiments