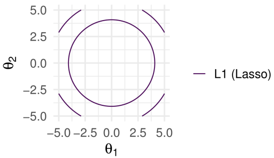
Introduction to Machine Learning

Elastic Net and Regularization for GLMs



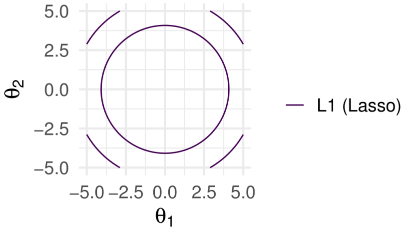
Learning goals

- Know the elastic net as compromise between Ridge and Lasso regression
- Know regularized logistic regression

ELASTIC NET

Elastic Net combines the L_1 and L_2 penalties:

$$\mathcal{R}_{\text{elnet}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} (y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{(i)})^2 + \lambda_1 \|\boldsymbol{\theta}\|_1 + \lambda_2 \|\boldsymbol{\theta}\|_2^2.$$



- Correlated predictors tend to be either selected or zeroed out together.
- Selection of more than n features possible for $p > n$.

ELASTIC NET

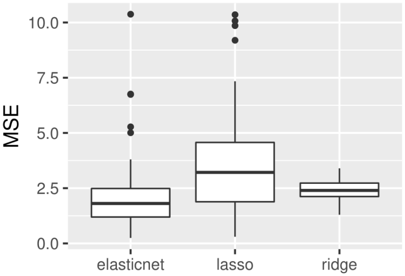
Simulating two examples with each 50 data sets and 100 observations each:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \sigma\epsilon, \quad \epsilon \sim N(0,1), \quad \sigma = 1$$

Ridge performs better for:

$$\boldsymbol{\beta} = (\underbrace{2,\ldots,2}_{5}, \underbrace{0,\ldots,0}_{5})$$

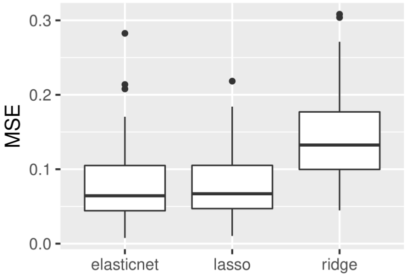
$$\text{corr}(\mathbf{X}_i, \mathbf{X}_j) = 0.8^{|i-j|} \text{ for all } i \text{ and } j$$



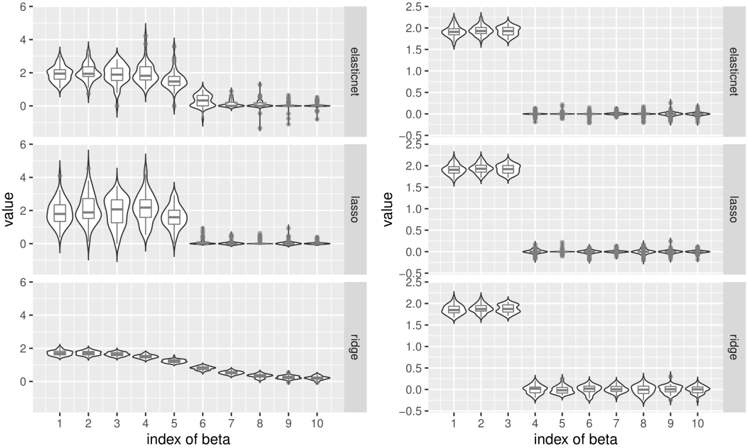
Lasso performs better for:

$$\boldsymbol{\beta} = (2, 2, 2, \underbrace{0,\ldots,0}_{7})$$

$$\text{corr}(\mathbf{X}_i, \mathbf{X}_j) = 0 \text{ for all } i \neq j, \text{ otherwise } 1$$



ELASTIC NET



Since Elastic Net offers a compromise between Ridge and Lasso, it is suitable for both data situations.

REGULARIZED LOGISTIC REGRESSION

Regularizers can be added very flexibly to basically any model which is based on ERM.

Hence, we can, e.g., construct L_1- or L_2-penalized logistic regression to enable coefficient shrinkage and variable selection in this model.

$$
\begin{aligned}
\mathcal{R}_{\text{reg}}(\boldsymbol{\theta}) &= \mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) + \lambda \cdot J(\boldsymbol{\theta}) \\
&= \sum_{i=1}^{n} \log \left[1 + \exp\left(-2y^{(i)} f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)\right)\right] + \lambda \cdot J(\boldsymbol{\theta})
\end{aligned}
$$

REGULARIZED LOGISTIC REGRESSION

We fit a logistic regression model using polynomial features for x_1 and x_2 with maximum degree of 7. We add an L_2 penalty. We see for

- $\lambda = 0$: The unregularized model seems to overfit.
- $\lambda = 0.0001$: Regularization helps to learn the underlying mechanism.
- $\lambda = 1$: The real data-generating process is captured very well.