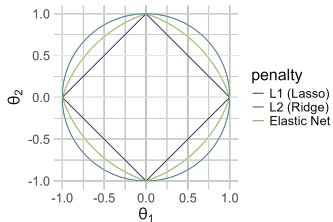
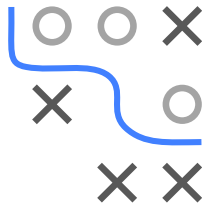


Introduction to Machine Learning

Regularization

Elastic Net and regularized GLMs

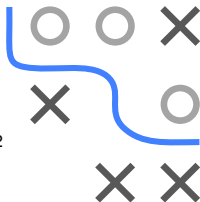
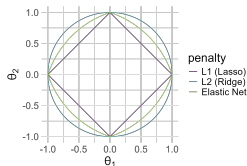


Learning goals

- Compromise between L1 and L2
- Regularized logistic regression

► Zou and Hastie 2005

$$\begin{aligned}\mathcal{R}_{\text{elnet}}(\boldsymbol{\theta}) &= \sum_{i=1}^n (y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{(i)})^2 + \lambda_1 \|\boldsymbol{\theta}\|_1 + \lambda_2 \|\boldsymbol{\theta}\|_2^2 \\ &= \sum_{i=1}^n (y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{(i)})^2 + \lambda \left((1 - \alpha) \|\boldsymbol{\theta}\|_1 + \alpha \|\boldsymbol{\theta}\|_2^2 \right), \quad \alpha = \frac{\lambda_2}{\lambda_1 + \lambda_2}, \lambda = \lambda_1 + \lambda_2\end{aligned}$$



- 2nd formula is simply more convenient to interpret hyperpars;
 λ controls how much we penalize, α sets the “L2-portion”
- Correlated features tend to be either selected or zeroed out together
- Selection of more than n features possible for $p > n$

SIMULATED EXAMPLE

5-fold CV with $n_{train} = 100$ and 20 repetitions with $n_{test} = 10000$ for setups:

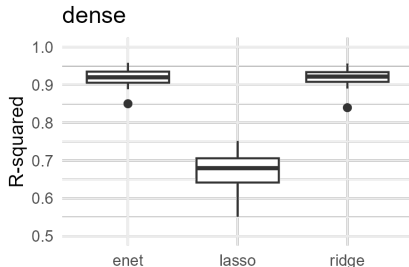
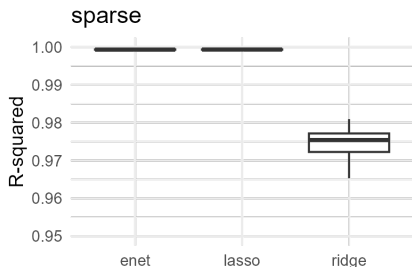
$$y = \mathbf{x}^T \boldsymbol{\theta} + \epsilon; \quad \epsilon \sim N(0, 0.1^2); \quad \mathbf{x} \sim N(0, \Sigma); \quad \Sigma_{k,l} = 0.8^{|k-l|}:$$

Lasso better for sparse features:

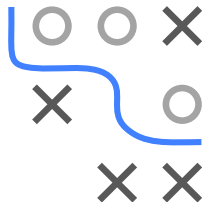
$$\boldsymbol{\theta} = (\underbrace{1, \dots, 1}_5, \underbrace{0, \dots, 0}_{495})$$

Ridge better for dense features:

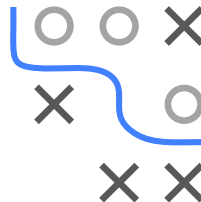
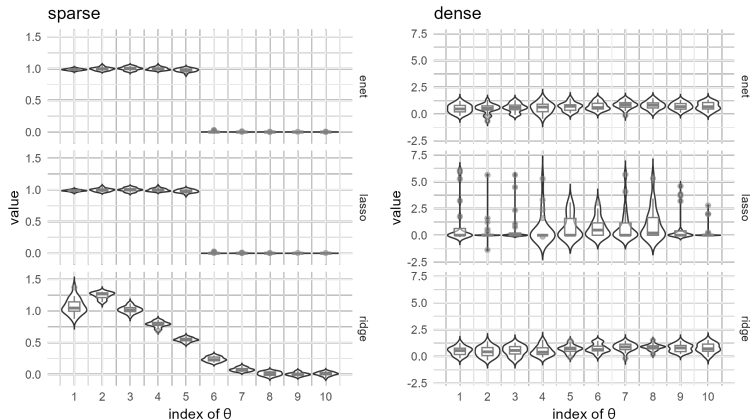
$$\boldsymbol{\theta} = (\underbrace{1, \dots, 1, 1, \dots, 1}_{500})$$



⇒ elastic net handles both cases well



SIMULATED EXAMPLE / 2



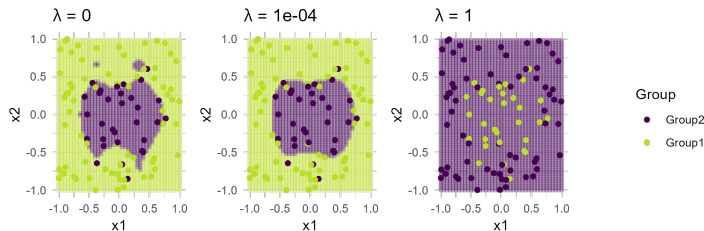
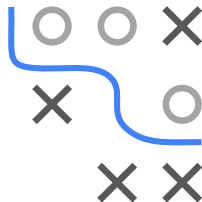
LHS: ridge estimates of noise features hover around 0 while lasso/e-net produce 0s.

RHS: ridge cannot perform variable selection compared to lasso/e-net.

Lasso more frequently ignores relevant features than e-net (longer tails in violin plot).

REGULARIZED LOGISTIC REGRESSION

- Penalties can be added very flexibly to any model based on ERM
- E.g.: L_1 - or L_2 -penalized logistic regression for high-dim. spaces and feature selection
- Now: LR with polynomial features for x_1, x_2 up to degree 7 and L_2 penalty on 2D “circle data” below



- $\lambda = 0$: LR without penalty seems to overfit
- $\lambda = 0.0001$: We get better
- $\lambda = 1$: Fit looks pretty good