Erasmus School of Economics

Machine Learning

FEM31002

Regularization

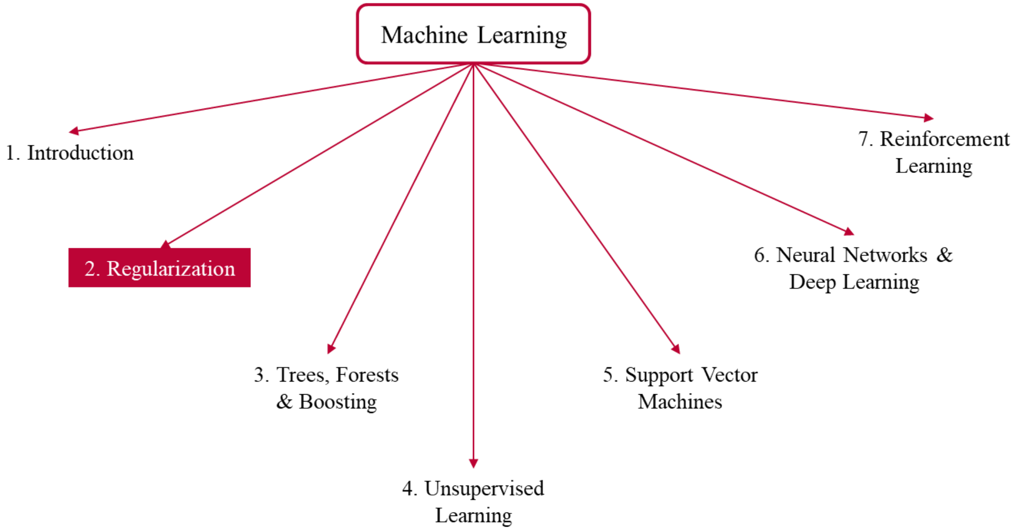
Part 2

Ilker Birbil

birbil@ese.eur.nl

ERASMUS UNIVERSITEIT ROTTERDAM

Outline



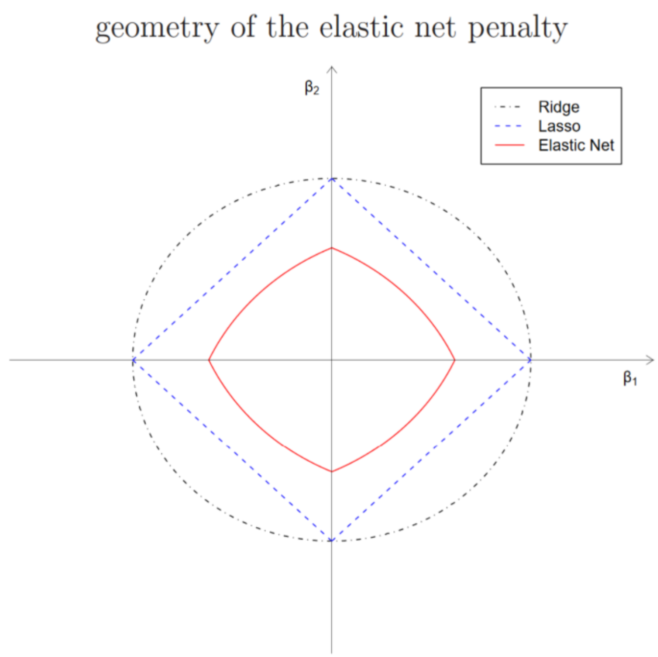
- Shrinkage: Ridge Regression and Lasso
- Elastic Net
- Least Angle Regression
- Integer Programming models

Elastic Net

$$\hat{\boldsymbol{\beta}}_{\text{EN}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p}(\alpha\beta_j^2 + (1-\alpha)|\beta_j|)$$

$$0 \leq \alpha \leq 1$$

*



* *Regularization and variable selection via the elastic net*, H. Zou and T. Hastie, 2005, pg 5.
(available online, last access date 20 August 2019)

Least Angle Regression

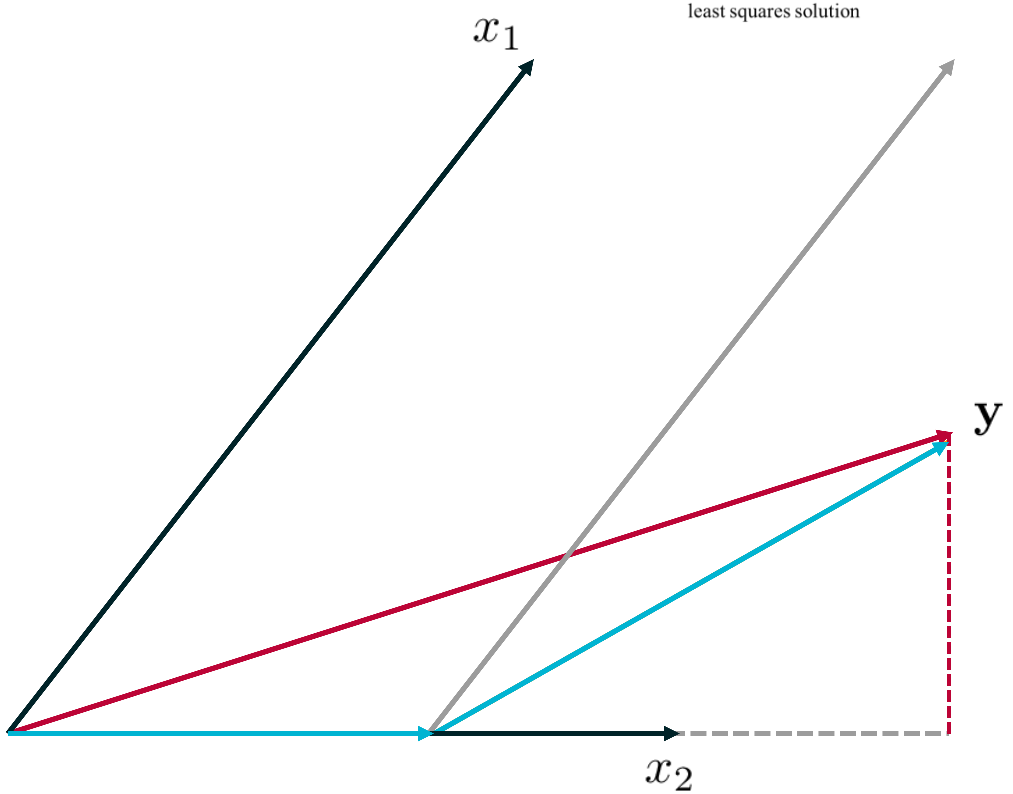
1. Scale the **predictors**: $\sum_{i=1}^{n} y_i = 0, \quad \sum_{i=1}^{n} x_{ij} = 0, \quad \sum_{i=1}^{n} x_{ij}^2 = 1 \text{ for } j = 1, \ldots, p$
2. Set $\boldsymbol{\beta} = \mathbf{0}, \mathbf{r} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta} = \mathbf{y}$
3. Select the predictor that is most correlated with residual: Say x_j
4. Move β_j towards its least squares coefficient of the current residual until some predictor x_k has the same correlation with the residual as x_j
5. Move β_j and β_k in the direction defined by their joint least squares coefficient of the residual until another predictor has the same correlation with the residual as x_j and x_k
6. Continue like this until all predictors join. In $\min\{n-1, p\}$ steps, obtain the full least squares solution

Lasso modification:

4. If a nonzero coefficient becomes zero, remove its corresponding predictor

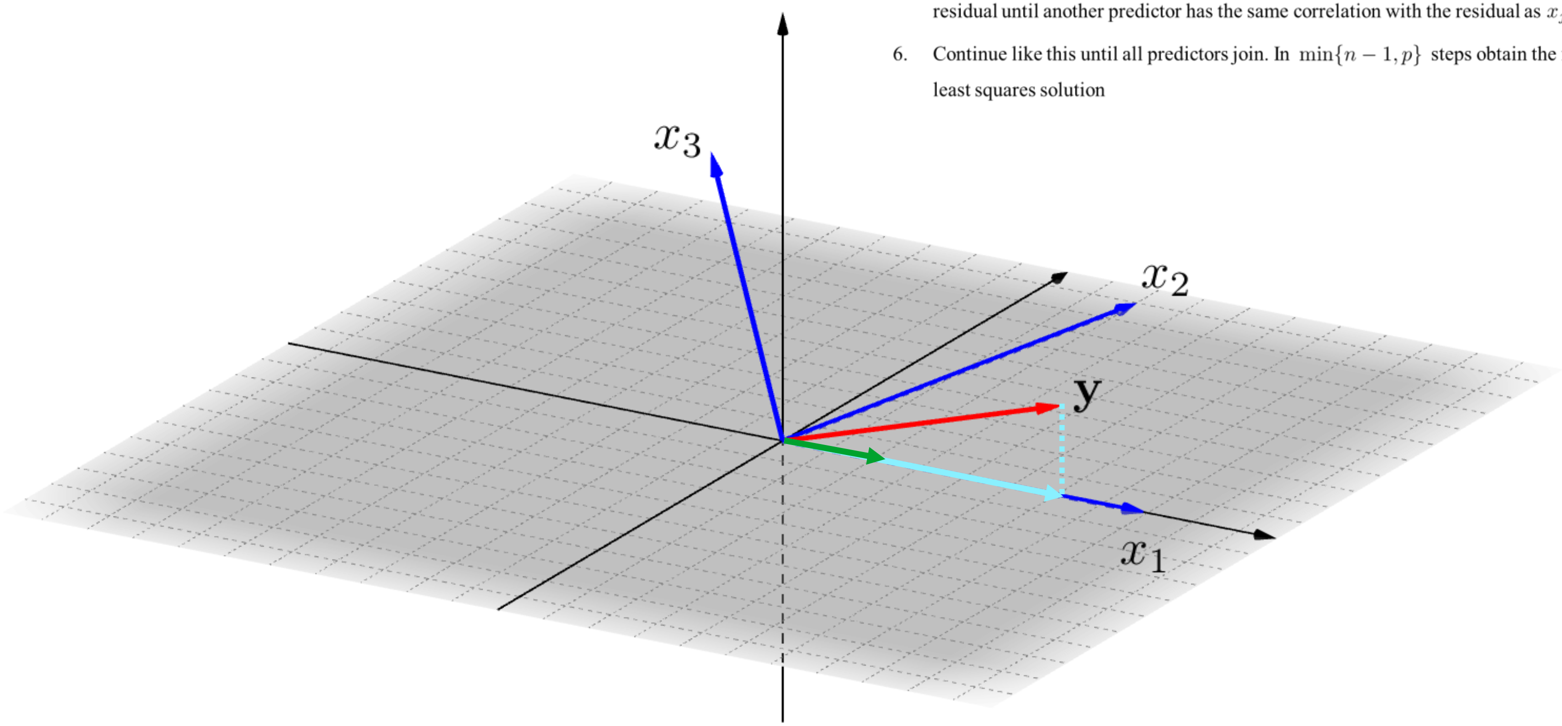
Least Angle Regression

1. Scale the **predictors**: $\sum_{i=1}^{n} y_i = 0, \quad \sum_{i=1}^{n} x_{ij} = 0, \quad \sum_{i=1}^{n} x_{ij}^2 = 1$ for $j = 1, \ldots, p$
2. Set $\boldsymbol{\beta} = \mathbf{0}, \mathbf{r} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta} = \mathbf{y}$
3. Select the predictor that is most correlated with residual: Say x_j
4. Move β_j towards its least squares coefficient of the current residual until some predictor x_k has the same correlation with the residual as x_j
5. Move β_j and β_k in the direction defined by their joint least squares coefficient of the residual until another predictor has the same correlation with the residual as x_j and x_k
6. Continue like this until all predictors join. In $\min\{n-1, p\}$ steps obtain the full least squares solution



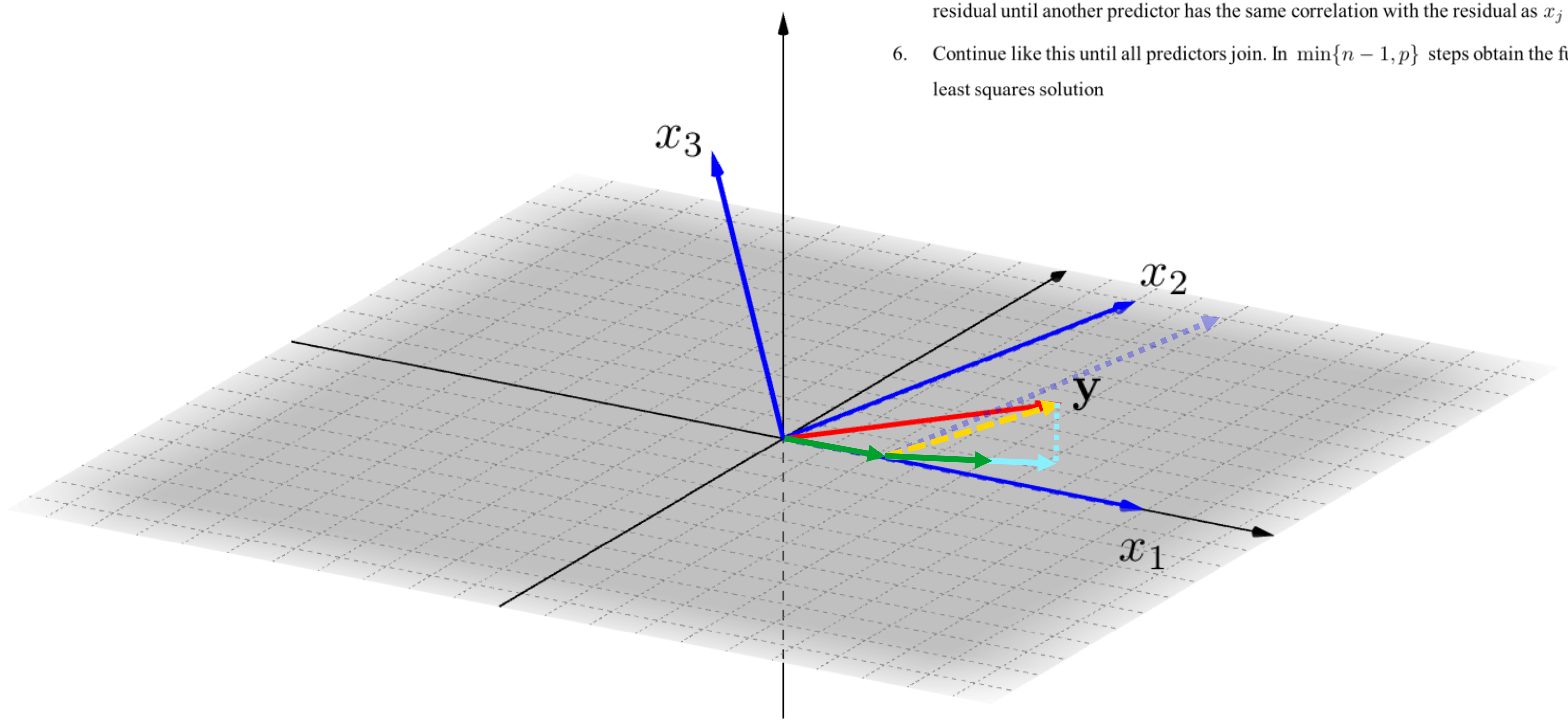
Least Angle Regression

1. Scale the **predictors**: $\sum_{i=1}^{n} y_i = 0, \quad \sum_{i=1}^{n} x_{ij} = 0, \quad \sum_{i=1}^{n} x_{ij}^2 = 1$ for $j = 1, \ldots, p$
2. Set $\boldsymbol{\beta} = \mathbf{0}, \mathbf{r} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta} = \mathbf{y}$
3. Select the predictor that is most correlated with residual: Say x_j
4. Move β_j towards its least squares coefficient of the current residual until some predictor x_k has the same correlation with the residual as x_j
5. Move β_j and β_k in the direction defined by their joint least squares coefficient of the residual until another predictor has the same correlation with the residual as x_j and x_k
6. Continue like this until all predictors join. In $\min\{n-1, p\}$ steps obtain the full least squares solution



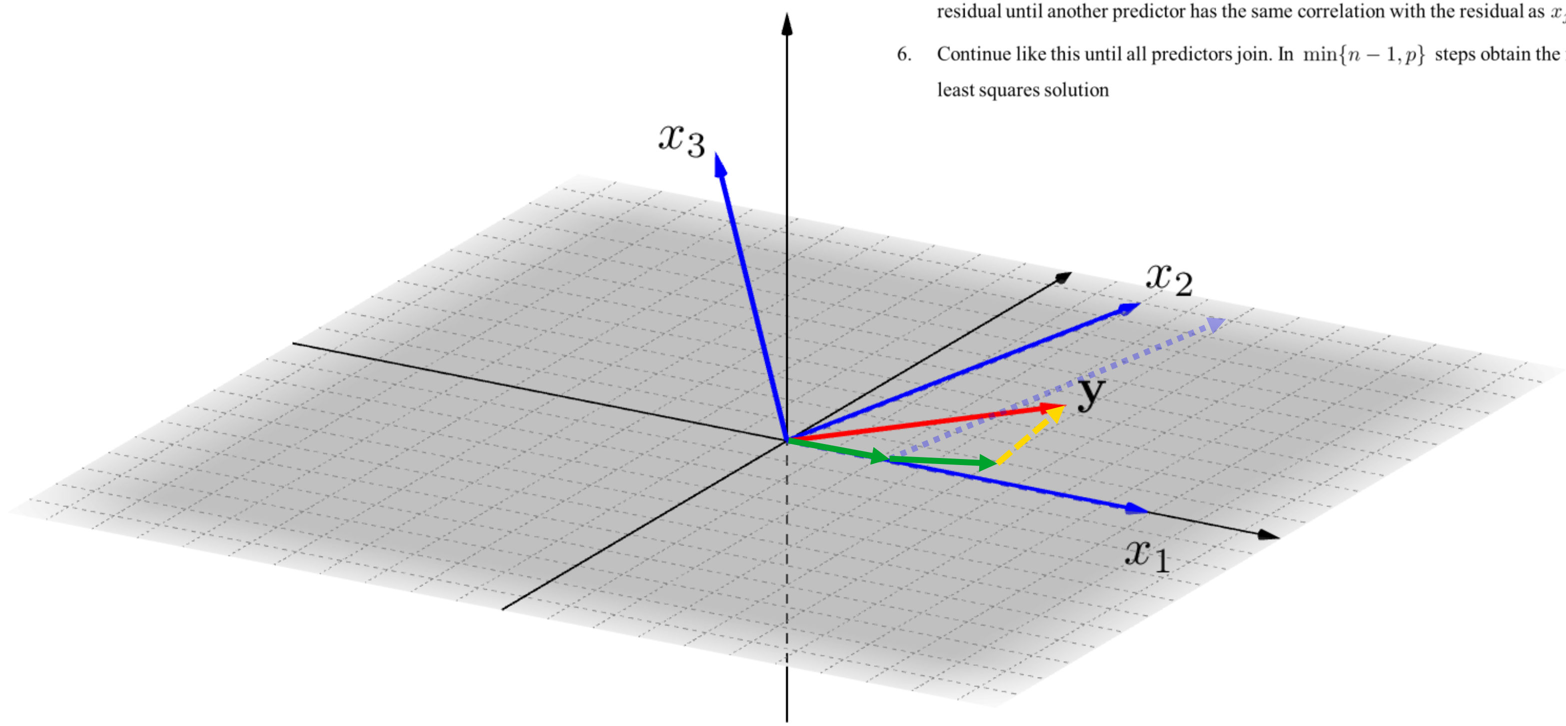
Least Angle Regression

1. Scale the **predictors**: $\sum_{i=1}^{n} y_i = 0, \quad \sum_{i=1}^{n} x_{ij} = 0, \quad \sum_{i=1}^{n} x_{ij}^2 = 1$ for $j = 1, \ldots, p$
2. Set $\boldsymbol{\beta} = \mathbf{0}, \mathbf{r} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta} = \mathbf{y}$
3. Select the predictor that is most correlated with residual: Say x_j
4. Move β_j towards its least squares coefficient of the current residual until some predictor x_k has the same correlation with the residual as x_j
5. Move β_j and β_k in the direction defined by their joint least squares coefficient of the residual until another predictor has the same correlation with the residual as x_j and x_k
6. Continue like this until all predictors join. In $\min\{n-1, p\}$ steps obtain the full least squares solution



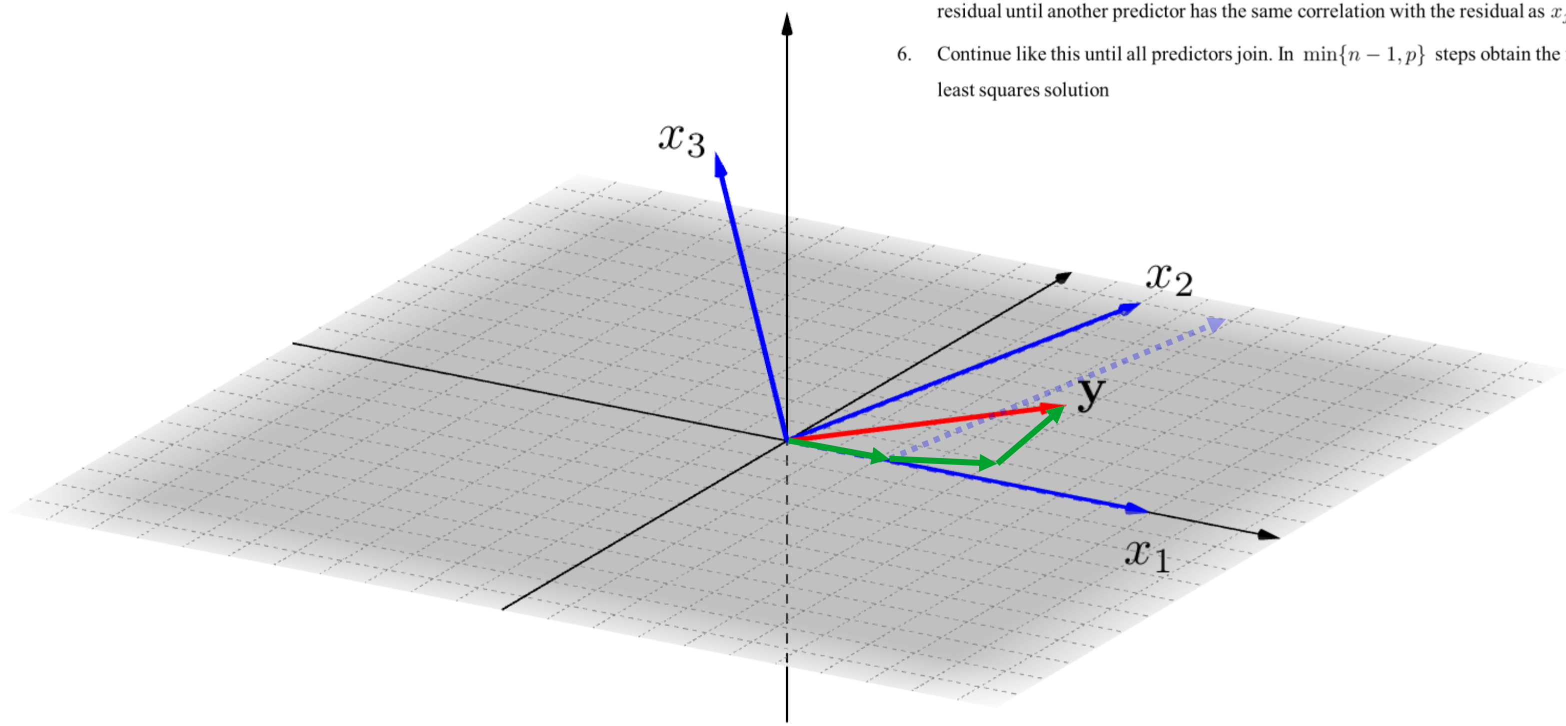
Least Angle Regression

1. Scale the **predictors**: $\sum_{i=1}^{n} y_i = 0, \quad \sum_{i=1}^{n} x_{ij} = 0, \quad \sum_{i=1}^{n} x_{ij}^2 = 1$ for $j = 1, \ldots, p$
2. Set $\boldsymbol{\beta} = \mathbf{0}, \mathbf{r} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta} = \mathbf{y}$
3. Select the predictor that is most correlated with residual: Say x_j
4. Move β_j towards its least squares coefficient of the current residual until some predictor x_k has the same correlation with the residual as x_j
5. Move β_j and β_k in the direction defined by their joint least squares coefficient of the residual until another predictor has the same correlation with the residual as x_j and x_k
6. Continue like this until all predictors join. In $\min\{n-1, p\}$ steps obtain the full least squares solution



Least Angle Regression

1. Scale the **predictors**: $\sum_{i=1}^{n} y_i = 0, \quad \sum_{i=1}^{n} x_{ij} = 0, \quad \sum_{i=1}^{n} x_{ij}^2 = 1$ for $j = 1, \ldots, p$
2. Set $\boldsymbol{\beta} = \mathbf{0}, \mathbf{r} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta} = \mathbf{y}$
3. Select the predictor that is most correlated with residual: Say x_j
4. Move β_j towards its least squares coefficient of the current residual until some predictor x_k has the same correlation with the residual as x_j
5. Move β_j and β_k in the direction defined by their joint least squares coefficient of the residual until another predictor has the same correlation with the residual as x_j and x_k
6. Continue like this until all predictors join. In $\min\{n-1, p\}$ steps obtain the full least squares solution



Integer Programming Approach

- Best subset selection problem:

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 \\ \text{subject to} & \|\boldsymbol{\beta}\|_0 \leq k. \end{array}$$

counts the number of nonzeros (a pseudo norm)

- Counting nonzeros makes the problem combinatorial
- NP-hard problem
- The same norm is also used in different domains, where sparsity of the resulting solution is important
- Here *k* is the hyperparameter

Integer Programming Approach

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 \\ \text{subject to} & \|\boldsymbol{\beta}\|_0 \leq k. \end{array}$$

A simple reformulation as a mixed integer quadratic problem:

$$\begin{array}{rlll} v_1 = & \text{minimize} & \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 & \\ & \text{subject to} & -Mz_j \leq \beta_j \leq Mz_j, & j = 1, \ldots, p, \\ & & \sum_{j=1}^{p} z_j \leq k, & \\ & & z_j \in \{0, 1\} & j = 1, \ldots, p. \end{array}$$

convex hull of the feasible region

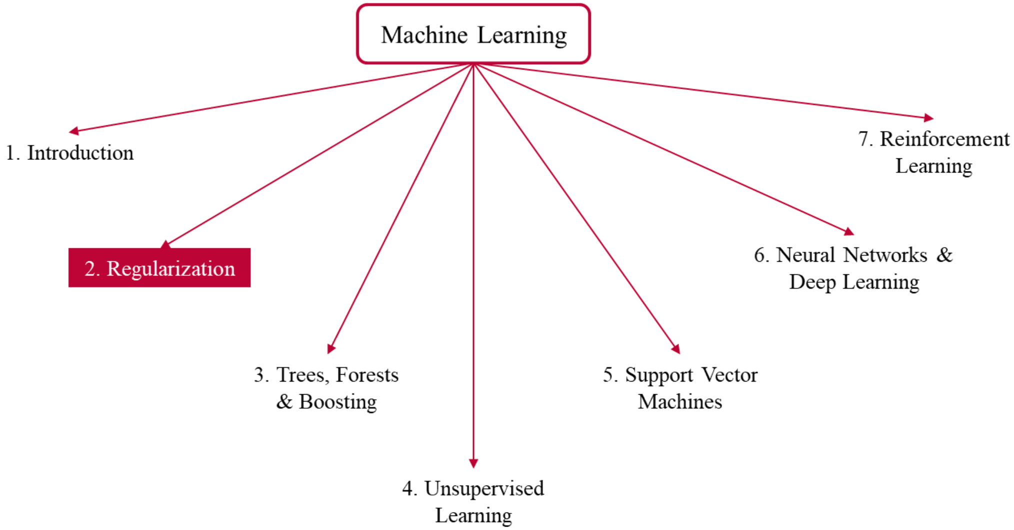
$$\{\boldsymbol{\beta} : \|\boldsymbol{\beta}\|_\infty \leq M, \|\boldsymbol{\beta}\|_1 \leq Mk\} \subseteq \{\boldsymbol{\beta} : \|\boldsymbol{\beta}\|_1 \leq Mk\}$$

$$\begin{array}{rll} v_2 = & \text{minimize} & \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 \\ & \text{subject to} & \|\boldsymbol{\beta}\|_1 \leq Mk. \end{array}$$

$$\boxed{v_2 \leq v_1}$$

Lasso

Outline



- Shrinkage: Ridge Regression and Lasso
- Elastic Net
- Least Angle Regression
- Integer Programming Models