Lecture Notes September 17

OLS, Ridge & Lasso

OLS : $\|(y - X\beta)\|_2^2$.

$$= \frac{1}{n}\sum_{i=0}^{n-1}(y_i - x_{i*}\beta)(y_{i'} - x_{i'*}\beta)$$

Ridge : $\|(y - X\beta)\|_2^2 + \lambda\|\beta\|_2^2$

$$\|\beta\|_2^2 = \sum_{i=0}^{p-1}\beta_i^2 \leq t$$

Lasso Regression :

$$\|(y - X\beta)\|_2^2 + \lambda\|\beta\|_1$$

$$\|\beta\|_1 = \sum_{i=0}^{p-1}|\beta_i| \leq t$$

$\lambda \geq 0$

$$\hat{\beta}^{OLS} = (X^T X)^{-1} X^T y$$

$$\hat{\beta}^{Ridge} = (X^T X + \lambda I)^{-1} X^T y$$

$\hat{\beta}^{Lasso}$: not analytical solution

$$\frac{d|x|}{dx} = \text{sign}(x) = \begin{cases} 1 & x > 0 \\ 0 & x = 0 \end{cases}$$

$-1 \quad x < 0$

$$\frac{\partial C^{Lasso}(\beta)}{\partial \beta} = -2X^T(y - X\beta) + \lambda \, sign(\beta)$$

$$\frac{\partial C^{Ridge}}{\partial \beta} = -2X^T(y - X\beta) + \lambda 2 \cdot \beta$$

$$\frac{\partial C^{OLS}}{\partial \beta} = -2X^T(y - X\beta)$$

$$\|v\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$$

$$\|v\|_2^2 = \sum x_i^2$$

$$\|v\|_1 = \sum_i |v_i|$$

Hessian

$$\frac{\partial^2 C^{OLS}}{\partial \beta \partial \beta^T} = 2X^TX = 2 \cdot (H)$$

$$var(\hat{\beta}^{OLS}) = \sigma^2 (X^TX)^{-1}$$

$$var(\hat{\beta}_i^{OLS}) = \sigma^2 (X^TX)^{-1}_{ii}$$

$$\vec{\varepsilon} \sim N(0, \sigma^2)$$

covariance matrix $= C[x]$

$$= \frac{1}{n} X^T X$$

$$= \begin{bmatrix} \mathrm{var}[x_0] & \mathrm{cov}[x_0 x_1] & \cdots & \mathrm{cov}[x_0 x_{p-1}] \\ \vdots & & & \\ \mathrm{cov}[x_{p-1} x_0] & - & - & \mathrm{var}[x_{p-1}] \end{bmatrix},$$

$$\mathrm{var}[x_0] = \frac{1}{n} \sum_{i=0}^{n-1} (x_{i0} - \bar{x}_0)^2$$

$\hat{H} \propto C[x]$ a positive definite matrix

$\Rightarrow \hat{H} > 0 \Rightarrow \hat{\beta}$ is a true minimum of $C(\beta)$

$$\frac{\partial^2 C^{\text{Ridge}}}{\partial \beta \partial \beta^T} = 2H + 2\lambda\beta$$

$$\frac{\partial^2 C^{\text{Lasso}}}{\partial \beta \partial \beta^T} = 2H$$

$$\frac{\partial C^{lasso}}{\partial \beta} = -2X^T y + 2H\beta + \lambda \, sign(\beta)$$

$\mathbb{C}[X]$ = covariance matrix

$c(\beta) = c(y|X\beta) =$

cost function

$= \mathcal{L}(y|X\beta)$

loss function

$= R(y|X\beta)$

Risk function.

$\begin{cases} var[x] & \mathbb{E}[x] \\ cov[x,y] & \text{statistical properties.} \end{cases}$

$$(X^T X) V = V \Sigma^2$$

$$X = U \Sigma V^T$$

$X \in \mathbb{R}^{n \times p}$

$U \in \mathbb{R}^{n \times n}$

$\Sigma \in \mathbb{R}^{n \times p}$

$V \in \mathbb{R}^{p \times p}$

$UU^T = U^T U = \mathbb{1}$

$V^T V = VV^T = \mathbb{1}$

$$(XX^T)U = U\Sigma\Sigma^T = U\Sigma^2$$

$$\tilde{y}^{OLS} = X\hat{\beta}^{OLS} = \sum_{i=0}^{p-1} u_i u_i^T y$$

$$\tilde{y}^{Ridge} = X\hat{\beta}^{Ridge} = \sum_{i=0}^{p-1} u_i \frac{\sigma_i^2}{\sigma_i^2 + \lambda} u_i^T y$$

$\sigma_i \geq \sigma_{i+1}$

Ridge shrinks the singular values, reducing more (for a given λ) those σ_i which are smaller. These correspond to eigenvalues with small variance.

$$\hat{\beta}^{OLS} = (X^T X)^{-1} X^T y$$

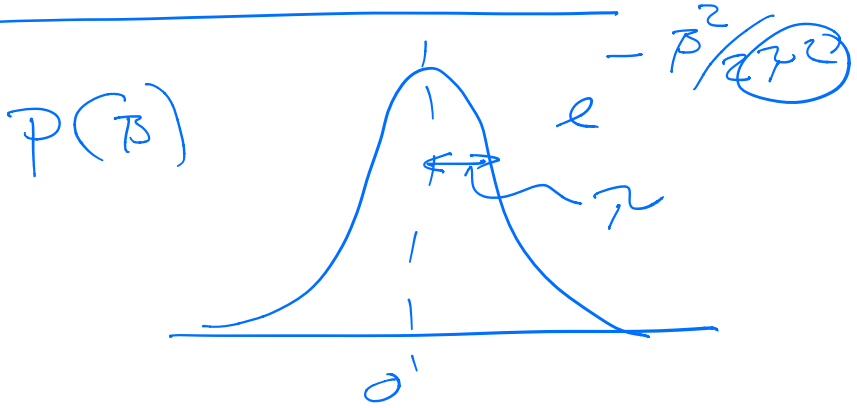
$$= (V \Sigma^T U^T U \Sigma V^T)^{-1} V \Sigma^T U^T \cdot y$$

$$= (\Sigma^T \Sigma)^{-1} V \Sigma^T U^T y$$

$$\text{var}[\hat{\beta}^{Ridge}] = \sigma^2 [X^T X + \lambda I]^{-1} \times X^T X [X^T X + \lambda I]$$

Ridge shrinks the variance of $\hat{\beta}_i^{Ridge}$ more for those β_i values with small singular values.

Bayesian approach



$$X^T X = \mathbb{1} = \begin{bmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{bmatrix}$$

$$\tilde{y}^{\text{Ridge}} = \frac{1}{1+\lambda}\, \tilde{y}^{\text{OLS}} =$$

$X^T X = \mathbb{1}$

$X^T X \neq \mathbb{1}$

$$\frac{1}{1+\lambda}\, y$$

$$\text{var}\left[\tilde{\beta}^{\text{Ridge}}\right] = \sigma^2 \frac{1}{\mathbb{P}(1+\lambda)} \mathbb{P} \frac{1}{\mathbb{P}(1+\lambda)}$$

$$= \sigma^2 \frac{1}{(1+\lambda)^2}$$

Degrees of freedom

p – predictors / features ...

$$\tilde{y}^{\text{OLS}} = \sum_{i=0}^{p-1} u_i u_i^T y$$

$$= X (X^T X)^{-1} X^T y$$

$$\text{Tr}\left[X (X^T X)^{-1} X^T\right] = p$$

= degrees of freedom,

$$\tilde{y}^{\text{Ridge}} = X \left(X^T X + \lambda \mathbb{1}_p\right)^{-1} X^T y$$

$$\text{tr}\left[X\left(X^TX+\lambda\Pi\right)^{-1}X^T\right]$$

$$= \boxed{\sum_{j=0}^{p-1}\frac{\sigma_j^2}{\sigma_j^2+\lambda}} = df(\lambda)$$

$\lambda = 0 \Rightarrow df(\lambda) = p$

$$\text{tr}[A] = \sum_{i=1}^{n} a_{ii}$$

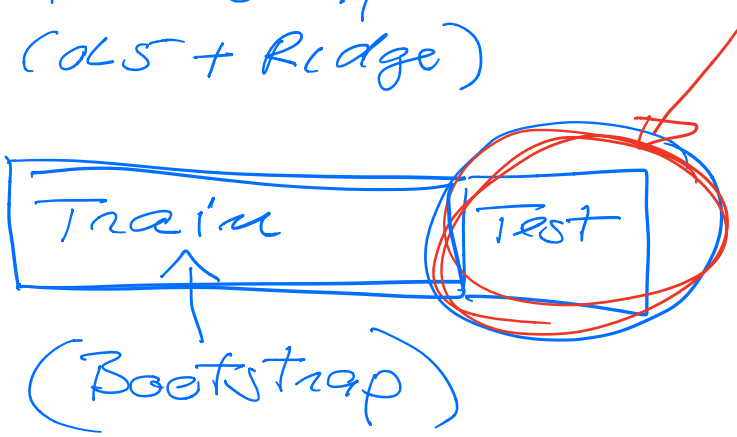
$\lambda \to \infty \Rightarrow df(\lambda) = 0$

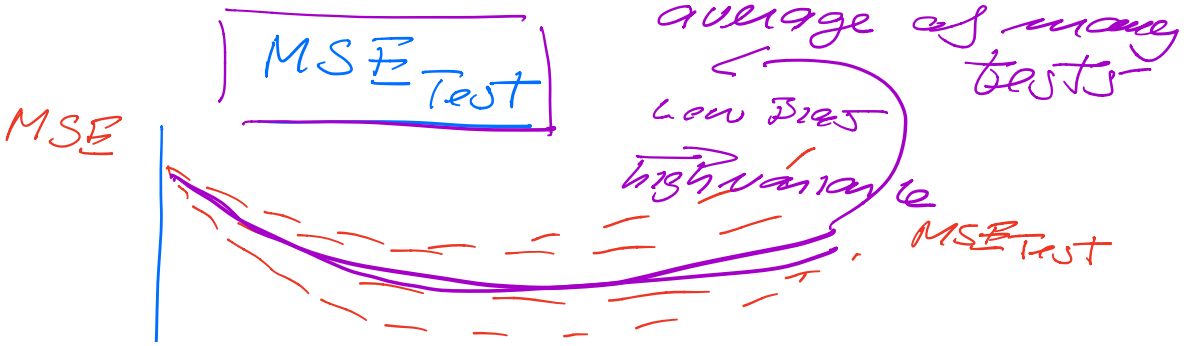
$\hat{y} = X\beta$

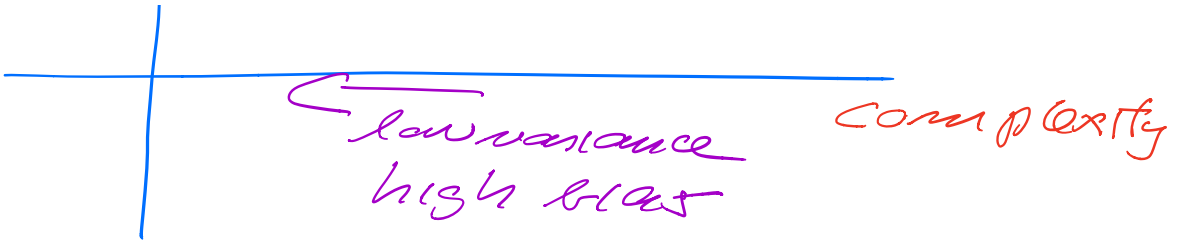
Resampling Methods

1b : Bootstrap + bias-variance (ols + Ridge)

+1d

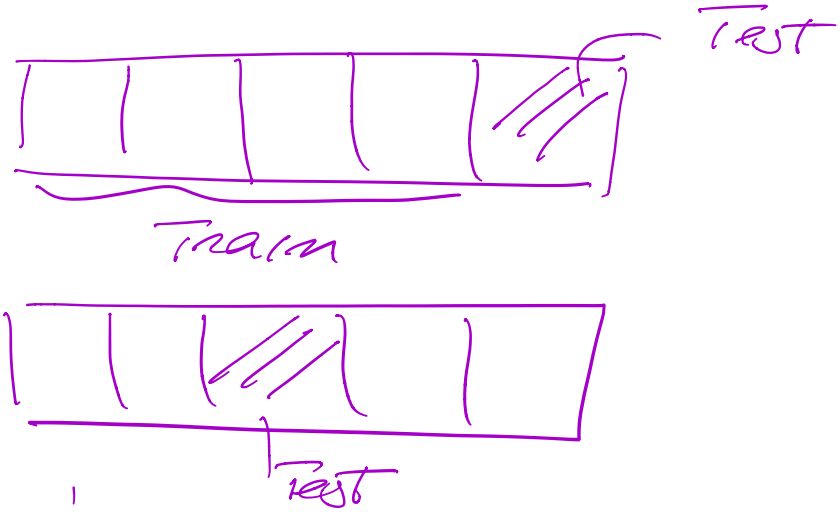






1c OLS + cross-validation

K-Folds $K = 5$

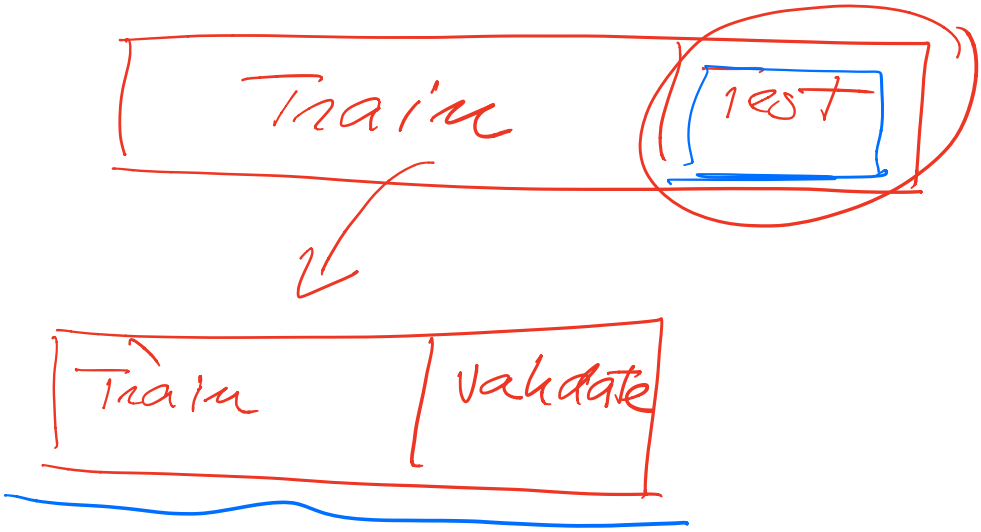


3 more times

in total 5 MSE_{Test} calculations.

— No need to do a train-test split.

Train-test split



Cross-validate here.

Ridge : polynomial complexity | λ - parameter.

validate gives optimal

MSE(λ, polynomial complexity)

↙ Final model

used on the independent test data.