Lecture September 4

- Bias-variance tradeoff
- Resampling techniques
  - Bootstrap (Efron 1979)
  - Cross-validation
- Ridge Regression

Bootstrap

resampling with replacement

$\underline{Z} = (z_1, z_2, z_3, \ldots, z_n) \quad Z \subset A$

$$P_n = \frac{1}{n} \sum_{i=1}^{n} I(z_i \in A)$$

algorithm

(i) Draw a new sample

$Z_1^* = (z_1^*, z_2^*, \ldots, z_n^*)$

compute $\hat{\theta}_i = g(z_1^*, z_2^*, \ldots, z_n^*)$

(ii) Repeat (i) B times, we have then

$\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_B$

(iii) compute standard deviation

$$STD = \sqrt{\frac{1}{} \sum^{B} (\hat{\theta}_i - \bar{\theta})^2}$$

$$\bar{\theta} = \frac{1}{B} \sum_{j=1}^{B} \hat{\theta}_j$$

Discrete PDF:

$$\mathbb{E}[x] = \sum_i x_i \, p(x_i) = \mu$$

$$\left(\int dx \, x \, p(x) \right)$$

sample mean $\bar{\mu} = \frac{1}{n} \sum_i x_i$

$$p(x_i) = \frac{1}{n}$$

CV = cross validation

LOOCV = leave one out CV

$X = \{x_1 \, x_2 \, \ldots \, x_n\}$

$y = \{y_1, y_2 \, \ldots \, y_n\}$

$MSE = 0$

for $i = 1, n$

$X_{cv} = \{x_1 \, x_2 \ldots x_{i-1}, x_{i+1}, \ldots x_n\}$

$Y_{cv} = \{y_1, y_2 \ldots y_{i-1}, y_{i+1}, \ldots y_n\}$

$x_{out} = x_i$

$y_{out} = fit(\underbrace{x_{cv}, y_{cv}}_{fit}, \overset{predict}{x_{out}})$

$MSE = MSE + (y_i - y_{out})^2$

$(y_i - \hat{y}_{oi})$

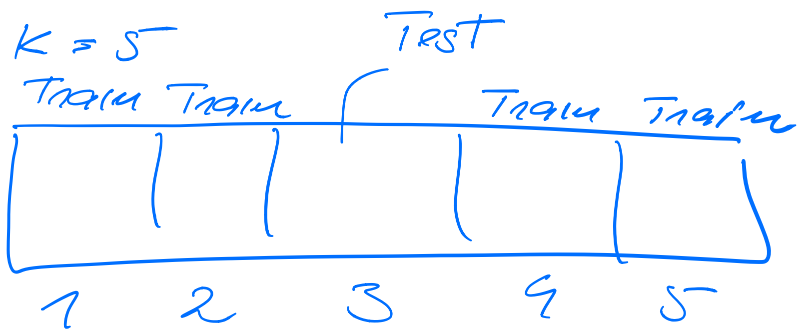
END FOR

$MSE = MSE/n$

K-Fold CV

Randomly partition data
K - equally sized sub samples

$K = 5 - 10$



The CV is then repeated K - times, with each of the K subsamples used just once as test data.

$K = 2$

Model 1 : d_0, d_1 (equal size)

Train on d_0 and test

on d_1 (Err_1)

Model 2 : Train on d_1
and test on d_0
(Err_0)

Total error = err = $\frac{1}{2}(err_0 + err_1)$

Ridge Regression

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

if singular, no $\hat{\beta}^{OLS}$

$$X^T X = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ \vdots & & & \\ a_{n1} & \cdots & \cdots & a_{nn} \end{bmatrix}$$

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

$(X^T X + \lambda I)$ $\qquad \lambda \sim 10^{-10}$

$$I = \begin{bmatrix} 1 & & 0 \\ & 1 & \\ 0 & & \ddots \\ & & & 1 \end{bmatrix}$$ $\qquad I \in \mathbb{R}^{p \times p}$

Corresponds to minimizing

$$C(\beta) = \frac{1}{n} \sum_{i=0}^{n-1} (y_i - X_{i*}\beta)^2 + \lambda \underbrace{\sum_{j=0}^{p-1} (\beta_j)^2}_{\beta^T\beta}$$

$$= \frac{1}{n} \| y - X\beta \|_2^2 + \lambda \|\beta\|_2^2$$

norm.

λ = penalty parameter / hyperparameter / regularization.

$\lambda \geq 0$

L2 - type optimazation.

$$\frac{\partial C(\beta)}{\partial \beta} = 0 \quad -\frac{2}{n} X^T(y - X\beta) + 2\lambda\beta = 0$$

$\Rightarrow$

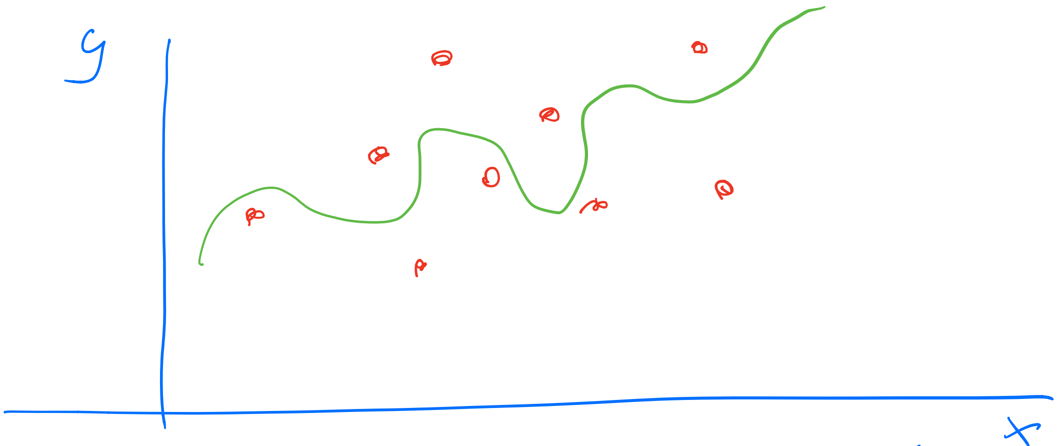
$$\hat{\beta}^{\text{ridge}} = \underbrace{(X^T X + \lambda \mathbb{I})^{-1}}_{\text{not singular.}} X^T y$$

... (ridge ...) (OLS)

$MSE(\hat{\beta}(\lambda)) \le MSE(\hat{\beta}(\lambda_{=0}))$

What does this mean?

- Intuitive understanding



- β-values may show a large scatter in values large negative and large positive.

We could try to make these parameters small using Baye's theorem:

$$P(A|B) = \frac{P(B|A)\,\underline{P(A)}}{- - - -}$$ — Prior

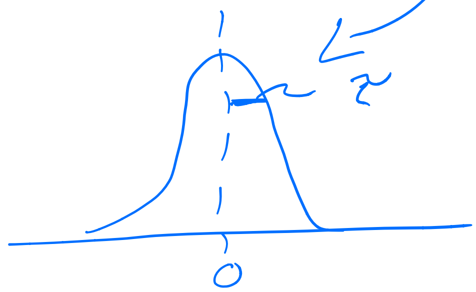
MLE

$$P(y_i|x_i\beta) \propto e^{-(y_i - x_i\beta)^2/2\sigma^2}$$

P-1

$$P(\beta) = \prod_{j=0} N(\beta_j | 0, \tau^2)$$

$$e^{-\beta_j^2/2\tau^2}$$



$$P(y|X\beta) = \prod_{i=0}^{n-1} P(y_i | x_i, \beta)$$

MLE: $P(y|X\beta)\,P(\beta)$

$$= \prod_{i=0}^{n-1} P(y_i | x_i, \beta) \prod_{j=0}^{p-1} P(\beta_j)$$

$$= \sum_{i=0}^{n-1} \underline{\log P(y_i | x_i, \beta)} + \sum_{j=0}^{p-1} \log P(\beta_j) + \text{const}$$

$$\propto \sum_{i=0}^{n-1} (y_i - x_i * \beta)^2 + \lambda \sum_{j=0}^{p-1} \beta_j^2$$

$$\lambda = \frac{1}{\tau^2} \geq 0$$