Lecture Erasmus+ course, October 16, 2023

ordinary least squares (OLS)

$$C(\beta) = \frac{1}{n} \sum_{i=0}^{n-1} \left(y_i - \sum_j x_{ij} \beta_j \right)^2$$

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min}\, C(\beta) \quad (= MSE)$$

Design matrix $X = \begin{bmatrix} x_{00} & \cdots & x_{0p-1} \\ \vdots & & \\ x_{n-10} & \cdots & x_{n-1p-1} \end{bmatrix}$

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

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

in case $X^T X$ is singular

$$X^T X + \lambda I_{p \times p} \qquad \lambda \geq 0$$

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

with MSE as $C(\beta)$

$$\frac{\partial^2 C}{\partial \beta \partial \beta^T} = \frac{2}{n} X^T X$$

Hessian matrix

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

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

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

$$\Sigma = \begin{bmatrix} \sigma_0 & & & \\ & \sigma_1 & & \\ & & \ddots & \\ & & & \sigma_{p-1} \\ & & & \\ & & 0 & \end{bmatrix}$$

$\sigma_0 > \sigma_1 > \sigma_2 > \cdots \sigma_{p-1} > 0$

$$X^T X = V \Sigma^T \underbrace{U^T U} \Sigma V^T$$

$$= V \underbrace{\Sigma^T \Sigma}_{p \times p} V^T = V \Sigma^2 V^T$$

$$\Sigma^2 = \begin{bmatrix} \sigma_0^2 & & \\ & \ddots & \\ & & \sigma_{p-1}^2 \end{bmatrix}$$

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

$$V = \begin{bmatrix} v_0 & v_1 & \cdots & v_{p-1} \end{bmatrix}$$

$$v_i^T v_j = \delta_{ij}$$

$$(X^T X) v_i = \sigma_i^2 v_i$$

$X^T X + \lambda \mathbb{1} \Rightarrow \lambda \geq 0$

Ridge Regression

$$C(\beta) = \frac{1}{n} \sum_{i=0}^{n-1} \left(y_i - \sum_j x_{ij} \beta_j\right)^2 + \lambda \sum_{j=0}^{p-1} \beta_j^2$$

it is common to leave out β_0 ($j=0$), the intercept.

$$\frac{\partial C}{\partial \beta} = 0 \Rightarrow \hat{\beta}_{Ridge} = \left(X^T X + \lambda I\right)^{-1} X^T y$$

$$MSE = \frac{1}{n} \sum_{i=0}^{n-1} (y_i - \tilde{y}_i)^2$$

$$\tilde{y} = X\hat{\beta}$$

$$C(\beta)_{Ridge} = \frac{1}{n} \|(y - X\beta)\|_2^2 + \lambda \|\beta\|_2^2$$

(common to leave out β_0)

$$C(\beta)_{Lasso} = \frac{1}{n} \|(y - X\beta)\|_2^2 + \lambda \|\beta\|_1$$

$$\|\beta\|_1 = \sum_{j=0}^{p-1} |\beta_j|$$

Lasso regression we have

$$\frac{d|\beta_j|}{d\beta_j} = \begin{cases} +1 & \beta_j > 0 \\ -1 & \beta_j < 0 \end{cases}$$

$\Rightarrow$ optimization with constraint

(no analytical solution for $\vec{\beta}$)

$$\tilde{y}_{OLS} = \left[\sum_{j=0}^{p-1} u_j u_j^T\right] y$$

$$\tilde{y}_{Ridge} = \left[\sum_{j=0}^{p-1} \frac{\sigma_j^2}{\sigma_j^2 + \lambda} u_j u_j^T\right] y$$

with a given λ, we can shrink the role of a specific parameter β_j

$$X = \begin{bmatrix} 2 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \qquad \vec{\beta}_{as} = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$$

$$y = \begin{bmatrix} 2 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 2 \\ 2 \end{bmatrix} = \begin{bmatrix} 4 \\ 2 \\ 0 \end{bmatrix}$$

$$\left(y = \begin{bmatrix} 4 \\ 2 \\ 3 \end{bmatrix} \right)$$

$$MSE = \frac{1}{n=3} \sum_{i=0}^{2} (y_i' - y_i^2) = 3$$

Basic assumptions

$$y = f(x) + \Sigma$$

$f(x)$: non-stochastic and continuous function

Σ: random noise

$$\Sigma \sim N(0, \sigma^2)$$

$p(y) = ?$

$E[y] = ?$

$$E[y] = \int_{y \in D} p(y)\, y\, dy$$

$$\left\{ = \sum_{y_i \in D} p(y_i)\, y_i \right.$$

in most cases we do not know $p(y)$.

$$\mathbb{E}[y] = \underbrace{\frac{1}{n}\sum_{i=0}^{n-1} y_i}_{\text{sample mean}} \neq \text{True } \mathbb{E}[y]$$

$$= \int_{y \in D} p(y)\, y\, dy$$

$$\mathbb{E}[y] = \mathbb{E}[f(x)] + \mathbb{E}[\varepsilon] = 0$$

$$\mathbb{E}[f(x)] = \int p(y)\, f(x)\, dx =$$

$$f(x) \int \underbrace{p(y)\, dx}_{=1}$$

$$= f(x)$$

$$\mathbb{E}[y] = f(x) \simeq X\beta$$

$$\boxed{\mathbb{E}[y_i] = \sum_{j=0}^{p-1} x_{ij}\beta_j = x_i^* \beta}$$

$$\mathrm{var}[y_i] = \mathbb{E}\left[(y_i - \mathbb{E}[y_i])^2\right]$$

$$= \mathbb{E}[y_i^2] - (\mathbb{E}[y_i])^2$$

$$= \mathbb{E}\left[(x_i^* \beta + \varepsilon_i)^2\right] - (x_i^* \beta)^2$$

$$= \mathbb{E}\left[(x_i^* \beta)^2 + 2 x_i^* \beta \varepsilon_i + \varepsilon_i^2\right]$$

$$- (x_i^* \beta)^2$$

$2 x_i^* \beta \mathbb{E}(\varepsilon_i) = 0$

$$= E[\varepsilon_i^2] = \mathrm{var}[\varepsilon_i] = \sigma^2$$

we can assume that

$$y_i \sim N(x_i' * \beta, \sigma^2)$$

($x_i' * \beta$: mean value)

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[- \frac{(y_i - x_i' * \beta)^2}{2\sigma^2} \right]$$

(From Bayes' theorem)

$\Rightarrow$ Derivation of OLS

$$y_i \in D = \{(x_0, y_0), (x_1, y_1), \dots, (x_{n-1}, y_{n-1})\}$$

y_i are independent and identically distributed

(i.i.d)

$$p(\underbrace{x, y}_{D} | \beta) = \prod_{i=0}^{n-1} p(y_i, x | \beta)$$

$$= \prod_{i=0}^{n-1} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[- \frac{(y_i - x_i * \beta)^2}{2\sigma^2} \right]$$

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\max}\ p(D | \beta)$$

MLE — Max-likelihood estimator

equivalent problem

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\max} \log p(D|\beta)$$

or

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} - \log p(D|\beta)$$

$$c(\beta) = -\log p(D|\beta)$$

$$= -\sum_{i=0}^{n-1} \log p(y_i, x_i|\beta)$$

$$= \frac{n}{2} \log(2\pi\sigma^2) +$$

$$+ \sum_{i=0}^{n-1} \frac{(y_i - x_i^T \beta)^2}{2\sigma^2}$$

$$= \frac{n}{2} \log(2\pi\sigma^2) + \frac{1}{2\sigma^2} \|(y - X\beta)\|_2^2$$

$$\frac{\partial C}{\partial \beta} = 0 = -X^T(X\beta - y) = 0$$

$$\Rightarrow \boxed{\beta = (X^T X)^{-1} X^T y}$$

$$y = \underset{\sim X\beta}{f(x)} + \varepsilon$$

For Ridge & Lasso, we need Bayes' theorem

$$p(\beta|D) \propto p(D|\beta)\,p(\beta)$$

posterior distribution

likelihood

prior