Lecture October 25

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

$\tilde{y}_i = \sum_{j=0}^{p-1} x_{ij}\beta_j$

$= x_{i*}\beta$

$\beta = [\beta_0, \beta_1 \,--\, \beta_{p-1}]^T$

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

$y, \tilde{y} \in \mathbb{R}^n \quad \Rightarrow \hat{\beta} = (X^T X)^{-1} X^T y$

Ridge

$C(\beta) = \frac{1}{n}\sum_{i=0}^{n-1}(y_i - x_{i*}\beta)^2$

$+ \lambda \sum_{j=0}^{p-1} \beta_j^2$

$\lambda > 0 \wedge \sum_{j=0}^{p-1}\beta_j^2 < t$

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

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

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

Lasso

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

Scaling/normalization of data.

Scikit-learn:

Ridge and Lasso do by default not include β_0 in the fitting

$$\lambda \sum_{j=1}^{p-1} \beta_j^2 \quad \text{Ridge}$$

$$\lambda \sum_{j=1}^{p-1} |\beta_j| \quad \text{Lasso.}$$

$$\tilde{y}_i = \sum_{j=0}^{p-1} \beta_j x_i^j \quad \rightarrow$$

$$X = \begin{bmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^{p-1} \\ 1 & x_1 & x_1^2 & \cdots & x_1^{p-1} \\ \vdots & & & & \vdots \\ 1 & & & & x_{n-1}^{p-1} \end{bmatrix}$$

1st example includes the intercept.

2nd example

$$\tilde{X} = \begin{bmatrix} x_0 & x_0^1 & \cdots & x_0^{p-1} \\ x_1 & & & \\ x_2 & & & \\ \vdots & & & \\ x_{n-1} & \cdots & & x_{n-1}^{p-1} \end{bmatrix}$$

Singular value decomp (SVD)

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

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

$$U^T U = U U^T = I_n$$

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

$$\mathcal{U} = \begin{bmatrix} | & | & & \\ u_0 & u_1 & \cdots & u_{n-1} \\ | & | & & \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} \sigma_0 & & & 0 \\ & \sigma_1 & & \\ 0 & & \ddots & \\ & & & \sigma_p \\ & & & \vdots \\ & & & 0 \end{bmatrix}$$

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

$$\sigma_0 > \sigma_1 > \cdots > \sigma_p > 0$$

then for $\sigma_{p-1} \cdots \sigma_{n-1} = 0$

$$V^T V = V V^T = I_p$$

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

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

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

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

Simple example to illu-
strate Lasso, Ridge and OLS

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

$$X = I = \begin{bmatrix} 1 & & 0 \\ & 1 & \\ 0 & & \ddots & 1 \end{bmatrix}$$

OLS (skip $1/n$)

$$C(\beta) = \sum_{i=0}^{n-1} (y_i - \beta_i)^2$$

$$= \sum_{i=0}^{p-1} (y_i - \beta_i)^2$$

$$\hat{\beta}^{OLS} = y_i$$

Ridge

$$C(\beta) = \sum_{i=0}^{p-1} (y_i - \beta_i)^2 + \lambda \sum_{i=0}^{p-1} \beta_i^2$$

$$C(\beta) = \sum_{i=0} (y_i - \beta_i) + \lambda \sum_{i=0} \beta_i$$

$$\frac{\partial C}{\partial \beta} = 0 \quad \Rightarrow$$

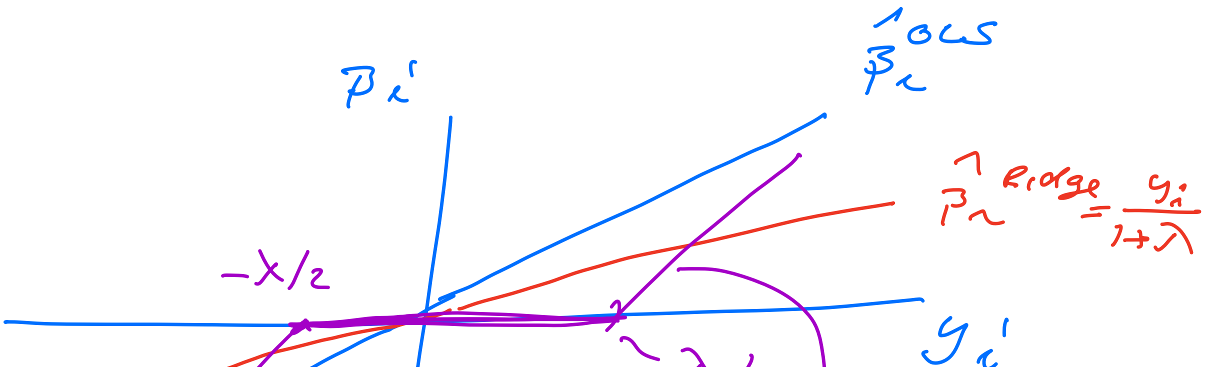
$$\hat{\beta}_i^{Ridge} = \frac{y_i}{1+\lambda} \qquad \lambda > 0$$

Lasso

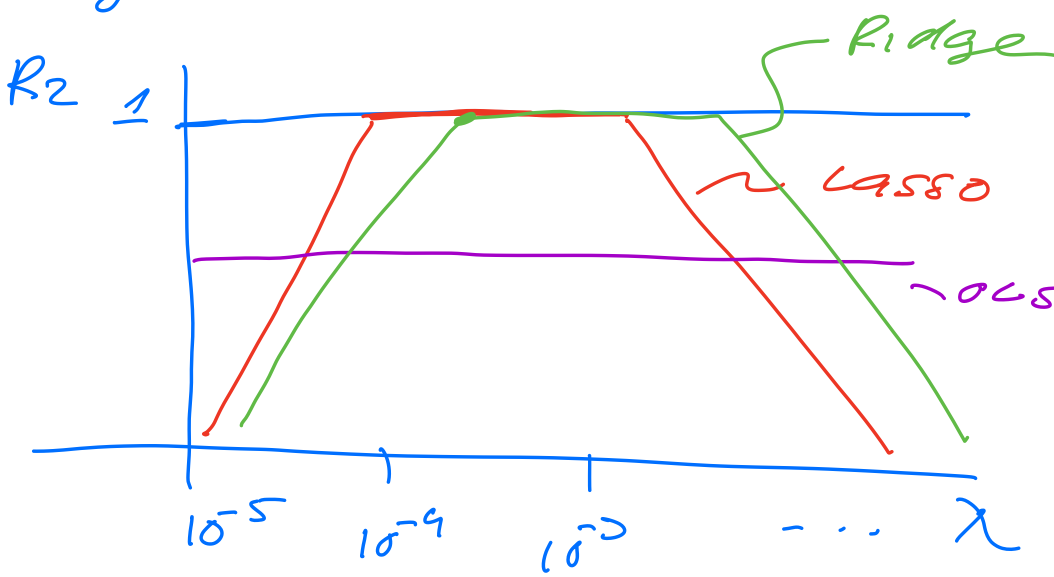
$$C(\beta) = \sum_{i=0}^{p-1} (y_i - \beta_i)^2 + \lambda \sum_{i=0}^{p-1} |\beta_i| \qquad \frac{\partial C}{\partial \beta_i} = 0$$

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

$$\hat{\beta}_i^{Lasso} = \begin{cases} y_i - \lambda/2 & \text{if } y_i > \lambda/2 \\ y_i + \lambda/2 & \text{if } y_i < -\lambda/2 \\ 0 & \text{if } |y_i| \le \lambda/2 \end{cases}$$



Mathematics of SVD and Ridge.



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

$$= V \Sigma^T \underbrace{U^T U}_{I_n} \Sigma V^T$$

$$= V \Sigma^T \Sigma V^T$$

$$= V \tilde{\Sigma}^2 V^T$$

$$\Sigma^T U^T U \Sigma = \underbrace{\Sigma^T I_n}\Sigma \in \mathbb{R}^{p \times p}$$

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

$$\tilde{\Sigma}^2 = \begin{bmatrix} \sigma_0^2 & & & \\ & \sigma_1^2 & & \\ & & \ddots & \\ & & & \sigma_{p-1}^2 \end{bmatrix}$$

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

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

$$= U \Sigma' V^T (V \tilde{\Sigma}^2 V^T)^{-1} V \Sigma^T U^T y$$

$$= \underline{U U^T} y = \sum_{i=0}^{p-1} u_i u_i^T y$$

$$U = \begin{bmatrix} | & | & | & & | \\ u_0 & u_1 & u_2 & \cdots & u_{n-1} \\ | & | & | & & | \end{bmatrix}$$

<u>Ridge</u>

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

Further properties

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

multiply from the right with V $\quad (V V^T = V^T V = I)$

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

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

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

The eigenvalues and eigenvectors of $X^T X$ are the singular values σ_i^2 and the orthogonal vectors v_i.

OLS

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

$$\frac{\partial^2 C}{\partial \beta \, \partial \beta^T} = \frac{2}{n} X^T X \quad \text{(curvature)}$$

Hessian matrix

$$\mathrm{cov}(x, y) = \frac{1}{n} \sum_{l=0}^{n-1} (x_l - \mu_x)(y_l - \mu_y)$$

$$x = [x_0, x_1, \ldots x_{n-1}]$$

$$y = [y_0, y_1, \ldots y_{n-1}]$$

$$
\begin{array}{c}
X = \begin{bmatrix} x_{00} & x_{01} & \cdots & x_{0\,p-1} \\ x_{10} & & & \\ \vdots & & & \\ x_{n-1\,0} & x_{n-1\,1} & \cdots & x_{n-1\,p-1} \end{bmatrix} \\
\quad\quad\quad x_0 \quad x_1
\end{array}
$$

$$\mathrm{cov}(x_0\, x_1) = \frac{1}{n} \sum_{l=0} (x_{l0} - \mu_{x_0})(x_{l1} - \mu_{x_1})$$

$$\Rightarrow \mathrm{cov}(X)$$

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