ERASMUS+, Oct 10, 2022

standard scaler

$$X = \begin{bmatrix} x_{00} & x_{01} & x_{02} & \cdots & x_{0p-1} \\ x_{10} & & & & \vdots \\ \vdots & & & & \vdots \\ x_{n-1\,0} & & & & x_{n-1\,p-1} \end{bmatrix}$$

$$= \left[\mathbf{x}_0, \mathbf{x}_1 \ \cdots \ \mathbf{x}_{p-1} \right]$$

$$\mu_0 = \frac{1}{n} \sum_{i=0}^{n-1} x_{i0}$$

$$X \rightarrow \bar{X} = \left[\underbrace{(\mathbf{x}_0 - \mu_0)}_{\bar{x}_0} \ \bar{x}_1 \ \bar{x}_2 \ \cdots \ \bar{x}_{p-1} \right]$$

OLS

$$y = f(x) + \varepsilon \qquad \varepsilon \sim N(0,1)$$

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

$H = X^T X$ is not invertible

$H \to X^T X + \mathbb{I}_p \lambda$

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

$\mathbb{1} \to \mathbb{1}_p$

Ridge regression:

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

$$\left(\|x\|_2 = \sqrt{\sum_{i=0}^{n-1} x_i^2} \right)$$

$$\frac{1}{n} \|(y - \tilde{y})\|_2^2 + \lambda \|\beta\|_2^2$$

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

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

λ = hyperparameter / regularization parameter

$$\lambda \|\beta\|_2^2 = \text{regularizer}$$

Lasso regression:

$$\lambda \|\beta\|_2^2 \Rightarrow \lambda \|\beta\|_1 =$$

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

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

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

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

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

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

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

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

$$u^T u = u u^T = \mathbb{1}$$
$$v^T v = v v^T = \mathbb{1}$$

$$U = [u_0 \; u_1 \; \dots \; u_{n-1}]$$

Left singular vectors

$$V = [v_0 \; v_1 \; v_2 \; \dots \; v_p]$$

Right singular vectors

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

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

$$\Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} \tilde{\Sigma} \\ 0 \end{bmatrix}$$

$$\Sigma^T \Sigma = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}$$ only non-zero values

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

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

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

$$= V \Sigma^T \Sigma V^T \in \mathbb{R}^{p \times p}$$

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

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

if A and B are square matrices and A and B are invertible

$$(AB)^{-1} = B^{-1} A^{-1}$$

$$\Rightarrow \tilde{y}_{ols} = U \Sigma V^T V (\Sigma^T \Sigma)^{-1} V^T V \Sigma^T U^T y$$

$$= U U^T y = \left(\sum_{i=0}^{p-1} u_i u_i^T \right) y$$

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

$$(X^T X) V = V \Sigma^T \Sigma \underbrace{V^T V}_{I}$$

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

$$V = [v_0 \; v_1 \; \ldots \; v_{p-1}]$$

v_i are the eigenvectors of $X^T X$ with eigenvalues

$$= \{\sigma_0, \sigma_1, \ldots \sigma_{p-1}\}$$

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

since $\sigma_0 > \sigma_1 \ldots > \sigma_{p-1} > 0$

$\Rightarrow$ H is positive definite and has a global min.

Same analysis with Ridge:

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

$$\hat{y}_{Ridge} = X \hat{\beta}_{Ridge} =$$

$$= \left(\sum_{j=0}^{p-1} u_j u_j^T \frac{\sigma_j^2}{\sigma_j^2 + \lambda} \right) y$$

$$\hat{y}_{OLS} = \left(\sum_{j=0}^{p-1} u_j u_j^T \right) y$$

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

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

$$\hat{\beta}_{OLS} = \frac{1}{\mathbb{I}} X^T y = X^T y$$