Lecture January 19

Dataset $\{(x_0, y_0), (x_1, y_1), \ldots (x_{n-1}\, y_{n-1})\}$

Assumption

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

normally distributed

$$f(x) \simeq \tilde{y} = \text{Model} \qquad \varepsilon \sim N(0, \sigma^2)$$

$$\tilde{y} \in \mathbb{R}^n \qquad y \in \mathbb{R}^n$$

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

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

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

Error function (cost/loss..)

$$= MSE =$$

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

$$= \frac{1}{n} \sum_{i} \left(y_i - X_{i*}\beta \right)^2$$

$$= \frac{1}{n} \left(y - X\beta \right)^T \left(y - X\beta \right)$$

$$\frac{\partial C(\beta)}{\partial \beta_j} = 0$$

$$\beta^{opt} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \; C(\beta)$$

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

$$\frac{\partial C(\beta)}{\partial \beta_j} = \frac{\partial}{\partial \beta_j} \left[\frac{1}{n} \sum_{i=0}^{n-1} \left(y_i - \beta_0 x_{i0} - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots \right)^2 \right]$$

$$= -\frac{2}{n} \sum_{i=0}^{n-1} \left(\underline{x_{ij}} \left[y_i - \beta_0 x_{i0} - \beta_1 x_{i1} - \cdots \right] \right)$$

Matrix-vector form ($2/n$)

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

$$\Rightarrow X^T y = X^T X \beta^{opt} \Rightarrow$$

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

RHS is known

ORDINARY LEAST SQUARES

$$\tilde{y} = \tilde{y}_{predict} = X\beta^{opt}$$

---

splitting data in train (validation), test

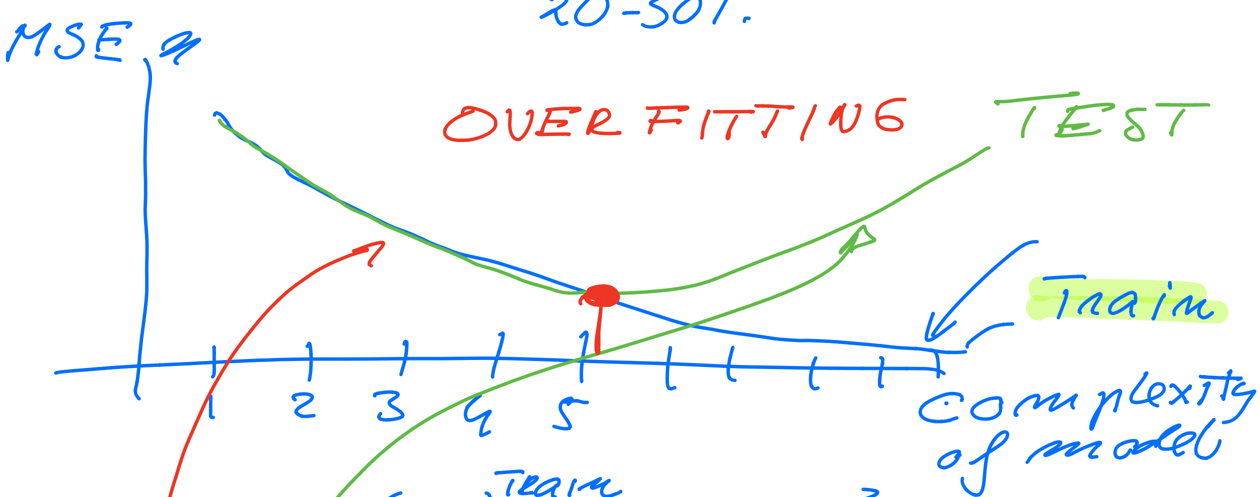
$$\beta^{opt}_{OLS} = (X^T X)^{-1} X^T y$$

$\boxed{\text{TRAIN}}$

70-80% of all data - n -

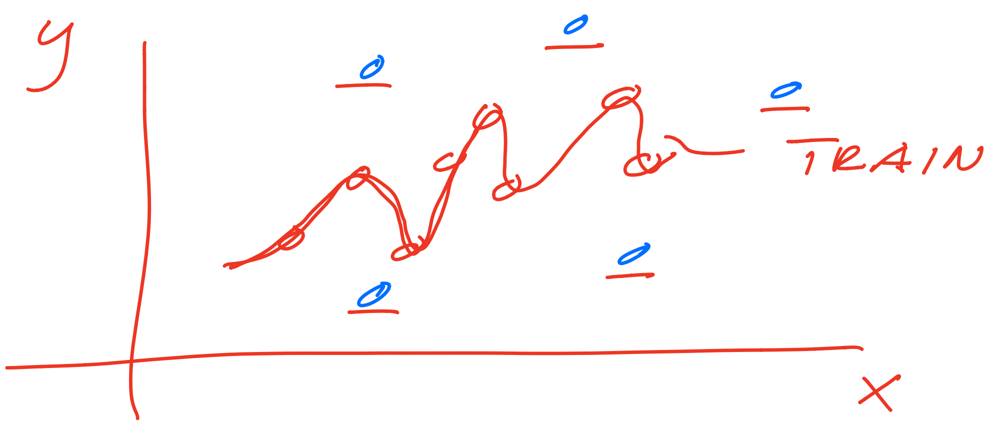
$$\tilde{y}^{predict}_{\boxed{\text{TEST}}} = X^T \beta^{opt}_{OLS}$$

20-30%.



$$MSE^{TRAIN} = \frac{1}{n_{TRAIN}} \sum_{i=0}^{(n-1)^{TRAIN}} \left(y_i - \tilde{y}_i(TRAIN) \right)^2 \quad \text{(poly-degree)}$$

$$\underline{MSE}^{Test} = \frac{1}{n_{test}} \sum_{i=0}^{n_{test}-1} \left(y_i - \tilde{y}_i(Test) \right)^2$$



Lasso & Ridge Regression;

<u>OLS</u>

$$C(\beta) = \frac{1}{n}\left[(y - X\beta)'(y - X\beta)\right]$$

norm-2 of a vector;

$$\|x\|_2 = \sqrt{\sum_i x_i^2}$$

$$\|x\|_2^2 = \sum_i x_i^2$$

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

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

$$\hat{y} = X\beta^{opt} \qquad X \in \mathbb{R}^{n \times p}$$

$$\beta \in \mathbb{R}^p$$

$(X^T X)^{-1}$ can be singular.

Ridge

$$A = \begin{bmatrix} a_{00} & a_{01} & \cdots & a_{0n} \\ \vdots & a_{11} & & \\ \vdots & & a_{22} & \\ a_{n0} & & \cdots & a_{nn} \end{bmatrix}$$

$$\tilde{A} = A + \lambda I$$

$$= \begin{bmatrix} a_{00}+\lambda & a_{01} & \cdots & a_{0n} \\ \vdots & a_{11}+\lambda & & \\ & & \ddots & \\ a_{n0} & & \cdots & a_{nn}+\lambda \end{bmatrix}$$

$$\lambda \sim 10^{-5}$$

$$(X^T X)^{-1} \longrightarrow (X^T X + \lambda I)^{-1}$$

"Ridge" :

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

Ridge - cost

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

$$\lambda \geq 0$$

$$\frac{\partial C(\beta)}{\partial \beta} = 0 \implies$$

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

Lasso

$$\|x\|_1 = \sum_i |x_i|$$

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

can reduce MSE compared to OLS by tuning λ

hyperparameter