Lecture January 21

what have we done?

$$y = f(x) + \varepsilon \quad | \quad y_i = x_{i*}\beta + \varepsilon_i$$

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

$$\mathbb{E}[y_i] = x_{i*}\beta$$

$$y_i \sim N(x_{i*}\beta, \sigma^2)$$

OLS

$$\beta^{opt}_{ols} = (x^T x)^{-1} x^T y$$

Ridge

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

Lasso

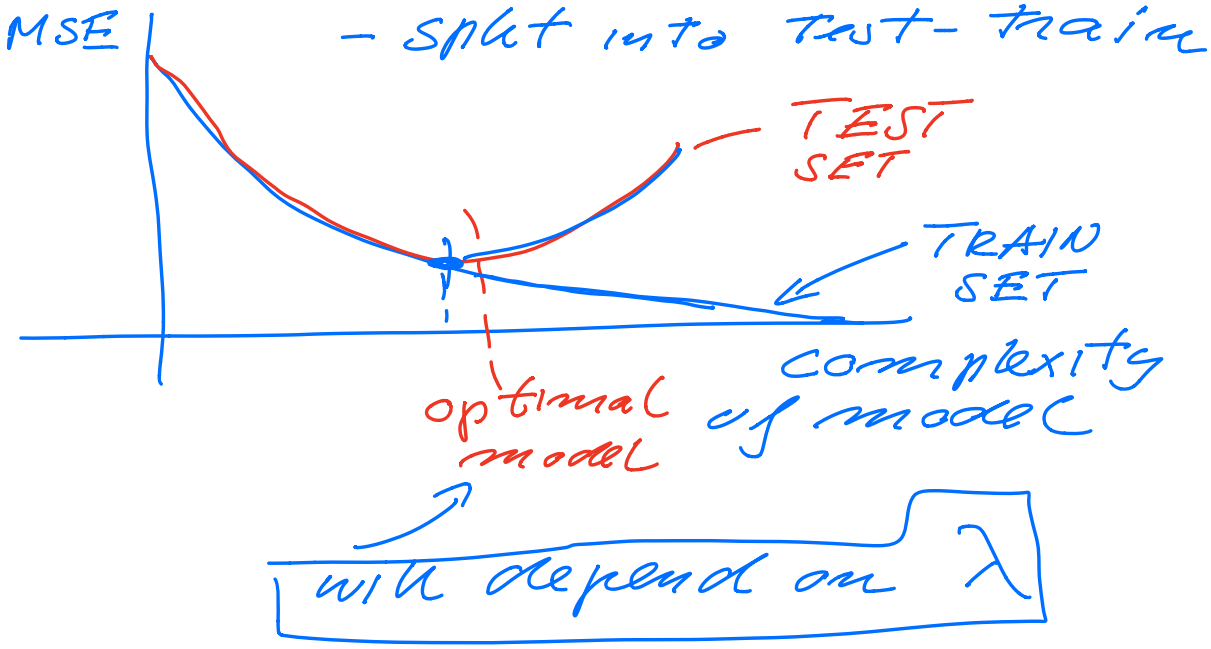
$$C(\beta) = \frac{1}{n} \|(y - x\beta)\|_2^2 + \lambda \|\beta\|_1$$

λ: **hyperparameter**.

Tuning λ ($\lambda \geq 0$) can yield a MSE for the

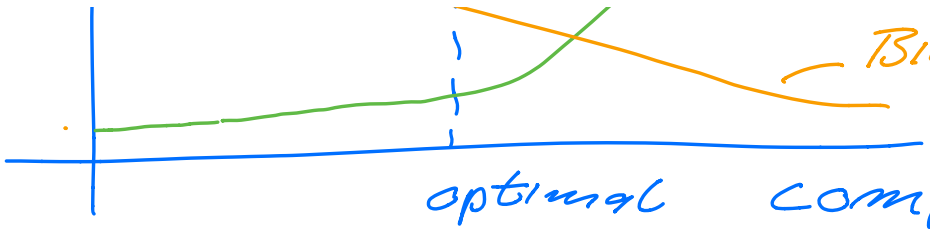
test data which is smaller than $MSE_{\substack{OLS \\ \lambda=0}}$ when $\lambda > 0$

- split into test-train



Bias - variance Tradeoff

$$MSE = \underbrace{\mathbb{E}\left[(y - \mathbb{E}[\tilde{y}])^2\right]}_{\text{Bias}} + \underbrace{\text{var}[\tilde{y}]}_{\text{variance}} + \sigma^2$$



- Resampling (Reliable expected values)
  - Bootstrap
  - Cross-validation.

$$\bar{\mu}(\text{sample}) = \frac{1}{n}\sum_{i=0}^{n-1} x_i' \neq$$

Mittwe

cross-validation:

k-fold CV

k = 3

- 1) [TRAIN | TRAIN | TEST] → Predicted error on Test ε_1
  0 — 1/3 — i — 1/3 — j — 1/3 — n-1
- 2) [TRAIN | TEST | TRAIN] → ε_2
  i — j
- 3 [TEST | TRAIN | TRAIN] → ε_3

$$\bar{\varepsilon} = \frac{\varepsilon_1 + \varepsilon_2 + \varepsilon_3}{k} \qquad k = 3$$

Typical choices of k: 5-10

LOOCV = leave one out CV

1) | TRAIN | n |

2) | TRAIN | n-1 | T |

⋮

n) | 1 | TRAIN |

Bootstrap

| TRAIN | TEST |

Bootstrap — unchanged

CV - $k = 5$

1)

| TRAIN | TRAIN | TRAIN | TRAIN | TEST | → ε_1

2) TRAIN | TRAIN | TRAIN | TEST | TRA → ε_2

3) ε_3

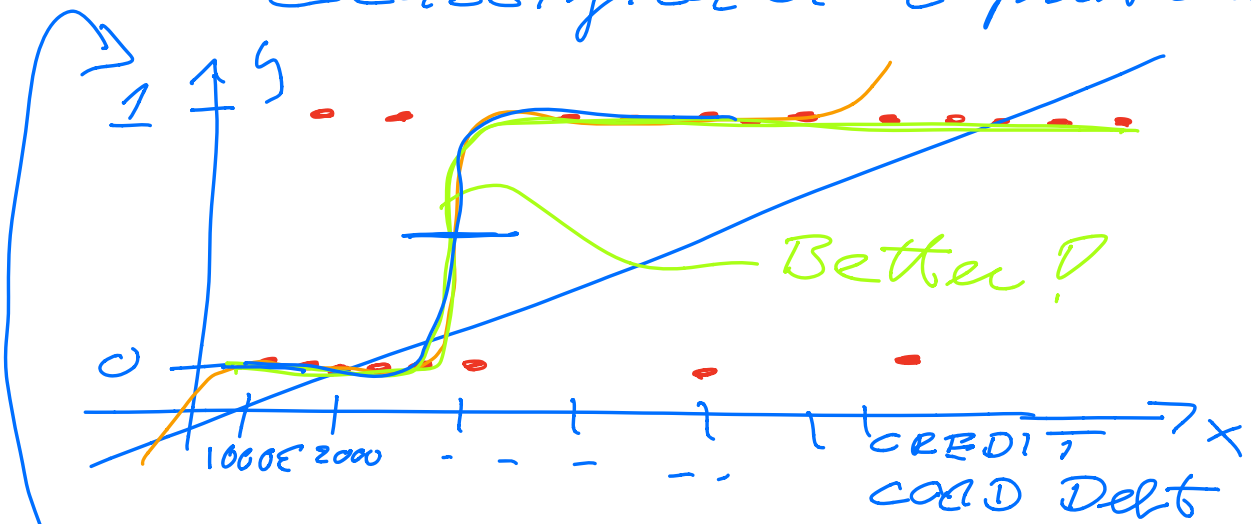
⋮

ε_4

5) ε_5

$$\bar{\varepsilon} = \frac{\varepsilon_1 + \varepsilon_2 + \varepsilon_3 + \varepsilon_4 + \varepsilon_5}{5}$$

Logistic regression = classification problem



Binary case = $\begin{cases} 1 = \text{not pay} \\ 0 = \text{pay} \end{cases}$

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

$y_i = \{0, 1\}$

$P(y_i | x_i, \beta) = ?$

$$\binom{n}{x} p^x q^{n-x}$$

Binomial Distribution?

$$P(y_i | x_i, \beta) = p(y_i=1)^{y_i} \; p(y_i=0)^{1-y_i}$$

$p = ?$ $\quad q = ?$

$$p(y_i = 0) = 1 - p(y_i = 1)$$

$$p(y_i=0) + p(y_i=1) = 1$$

Maximum likelihood:

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

$P(y_i | x_i \beta)$ | we need a model for this.

- tanh
- sigmoid.

$$p(y_i | x_i, \beta) = \frac{}{e^{\beta_0 + \beta_1 x_i}}$$

$$\frac{}{1 + e^{\beta_0 + \beta_1 x_i}}$$

$$= \frac{e^t}{1 + e^t}$$

$$= \frac{1}{1 + e^{-t}}$$

$$p(y \mid X, \beta) = \prod_{i=0}^{n-1} p(y_i = 1)^{y_i} \left(1 - p(y_i = 1)\right)^{1 - y_i}$$

$$= \sum_{i=0}^{n-1} \log\left(p(y_i = 1)^{y_i} \left(1 - p(y_i = 1)\right)^{1 - y_i} \right)$$

$$\underset{\beta \in \mathbb{R}^p}{\arg\min} \underbrace{- \sum_{i=0}^{n-1} \; \cdots \; i \; \cdots}_{\text{cost function} = C(\beta)}$$

$$= \beta^{opt}$$

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

non-linear equations in β. $\Rightarrow$ no analytical expression

for β => compute gradients and second derivatives numerically in order to minimize $C(\beta)$