Lecture November 15

Resampling techniques

- Bootstrap (ideal for iid)
- Cross-validation (CV)
- Jackknife
- Blocking

Bootstrap

$$D = \{ x_0, x_1, x_2, \dots x_{n-1} \}$$

(i) shuffle randomly with replacement

$$D^{(i)} = \{ x_0^{(i)}, x_1^{(i)} \dots x_{n-1}^{(i)} \}$$

compute expected value $\mu^{(i)} = \frac{1}{n} \sum_{i=0}^{n-1} x_i^{(i)}$

(ii) repeat (i) B-times

(iii) evaluate final μ

$$\mu = \frac{1}{B} \sum^{B-1} \mu^{(i)}$$

Train-test split: resampling only on training

Cross-validation

- Folds (blocks of data)
  K-Fold
- Example $K = 5$

$D =$

TRAIN	TRAIN	TRAIN	TRAIN	TEST

(i) Train model and compute μ on test data. $\mu^{(1)}$

(ii)

TRAIN	TRAIN	TRAIN	TEST	TRAIN

$\mu^{(2)}$

(iii)

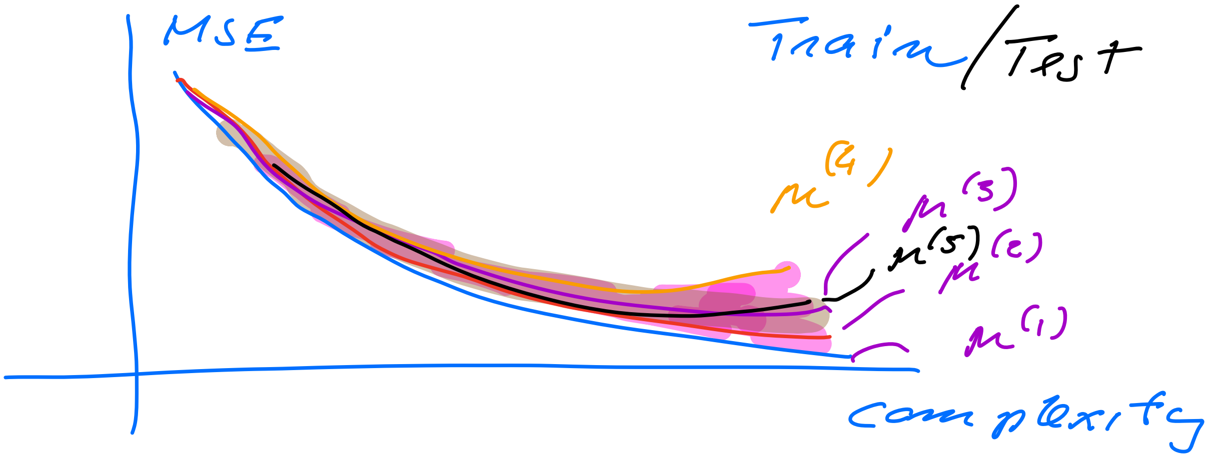
Train	Train	Test	Train	Train

$\mu^{(3)}$

⋮

(v)

$\rightarrow \quad \mu = \frac{1}{Folds} \sum_{i=0}^{Folds-1} \mu^{(i)}$



no "magic" k-folds, $k = 5-10$

LOOCV = Leave one out CV

$k = n$

Logistic Regression

Classification problem

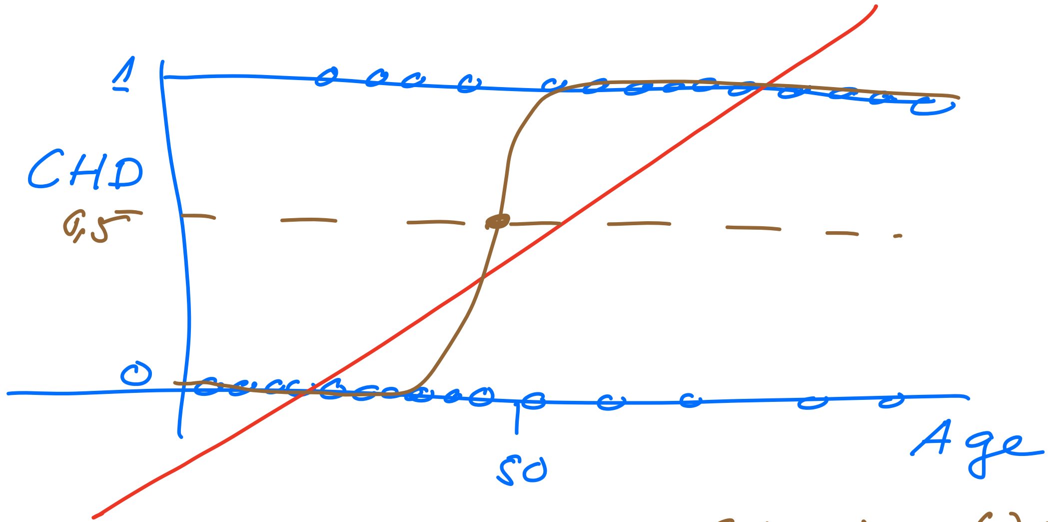
$y \in \mathbb{R}$ and $y \in (-\infty, \infty)$
for Linear regression

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

Classification: Binary
problem $\quad y \in \{0, 1\}$

Discrete outputs

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



$$f(x) \longrightarrow p(x) = \begin{cases} 1 & \text{if } p(x) \geq 0.5 \\ 0 & \text{if } p(x) < 0.5 \end{cases}$$

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

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

$$\mathbb{E}[y] = X\hat{\beta}$$

$$y \sim N(X\hat{\beta}, \sigma^2)$$

$y_i = 1$ we have a probability

$p(x_i \mid \beta) = p \quad p(y_i = 1) = p$

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

$1 = p + \varepsilon$ has probability $= \underline{p}$

$\varepsilon = 1 - p$

$0 = p + \varepsilon$ has — i — $= 1 - p$

$0 \le p \le 1$

$\varepsilon = -p$

what is the distribution of ε?

$$\mathbb{E}[\varepsilon] = \sum_i p_i \varepsilon_i$$

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

$$var[\varepsilon] = (1-p)^2 p + (-p)^2 (1-p)$$

$$= p(1-p)$$

Distribution : Binomial

$y_i = 1$ has $p(x_i | \beta) = p_i$

$y_i = 0$ has $1 - p(x_i | \beta)$
$= 1 - p_i$

y are iid. For each

y_i we have

$$p(x_i|\beta)^{y_i}\left(1-p(x_i|\beta)\right)^{1-y_i}$$

$$= p_i^{y_i}(1-p_i)^{1-y_i}$$

Model for $p(t) = \dfrac{e^t}{1+e^t}$ $(y=1)$

$y=0$ $\quad 1-p(t)$

$$t = \beta_0 + \beta_1 x$$

$$p(x_i|\beta) = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}$$

β are the parameters of model.

$$P(D|\beta) = \prod_{i=0}^{n-1} p_i^{y_i}(1-p_i)^{1-y_i}$$

$\hat{\beta}$ = Maximum likelihood estimators (MLE) estimation (MLE)

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

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

$$-\log P = -\sum_i \Big\{ y_i\big(\beta_0 + \beta_1 x_i - \log(1+e^{\beta_0+\beta_1 x_i})\big) - (1-y_i)\log(1+e^{\beta_0+\beta_1 x_i}) \Big\}$$

$$= C(\beta)$$

$$\frac{\partial C(\beta)}{\partial \beta_0} = 0 = -\sum_i (y_i - p_i)$$

$$\frac{\partial C(\beta)}{\partial \beta_1} = 0 = -\sum_i x_i(y_i - p_i)$$

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

$\mathbb{R}$

$$\frac{\partial^2 C}{\partial \beta \partial \beta^T} = X^T W X \qquad X \in \mathbb{R}^{n \times p}$$

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

$$w_{ii} = p_i(1-p_i)$$

$$w_{ij} = 0 \text{ if } i \neq j$$

Hessian matrix is positive definite

Linear Regression (OLS)

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