FYS-STK 4155 Sept 22

Cross-validation

k-fold $k \sim 5-10$

$k = 5$

1:	TRAIN	TRAIN	TRAIN	TRAIN	TEST	$Err_1(Test)$
	1	2	3	4	5	
2:	TEST	TRAIN	TRAIN	TRAIN	TRAIN	$Err_2(Test)$
⋮						
5:	TRAIN	TRAIN	TRAIN	TEST	TRAIN	$Err_5(Test)$

$$Err_{Estimate} = \frac{1}{5}\sum_{i=1}^{5} Err_i(Test)$$

Logistic Regression

Linear regression

$$y_i = \sum_j x_{ij} \beta_j + \varepsilon_i$$

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

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

$$x_i \in (-\infty, \infty)$$

$$y_i \in (-\infty, \infty)$$

$$y_i = f(x_i) + \varepsilon_i \simeq \tilde{y}(x_i) + \varepsilon_i$$

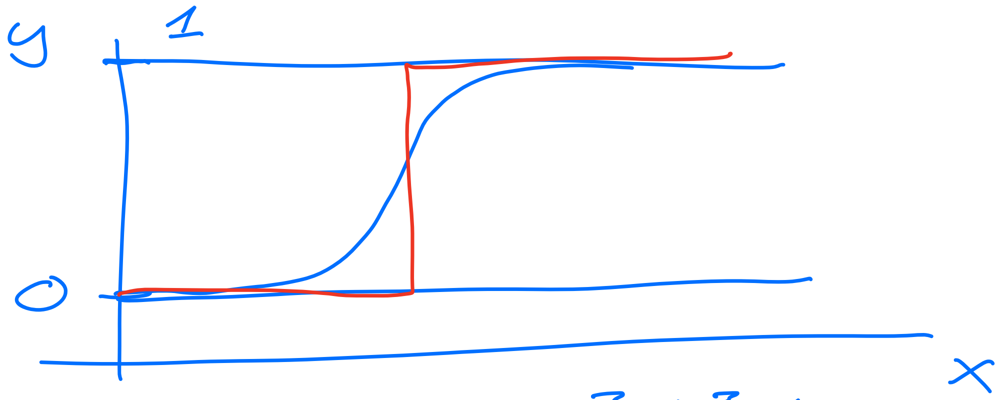
Logistic regression

$f(x_i)$ should represent discrete outputs.

Binary example

$$y_i = \{0, 1\}$$

$f(x) \rightarrow p(x)$ which is a likelihood.



$$p(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

$$\int_D p(x)\,dx = 1$$

$$p(x) = \frac{e^x}{1+e^x}$$

$$\int \frac{e^x}{1+e^x}\,dx = \ln(1+e^x)$$

assumption :

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

$$p(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

$$p(x) \to p(y_i, x_i \mid \beta)$$

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

$$P(D|\beta) = \prod_{i=0}^{n-1} \underbrace{p(y_i, x_i | \beta)}_{p(x_i)}$$

y_i are i.i.d.

$$= \prod_{i=0}^{n-1} p(x_i)$$

$$y_i = p(x_i) + \varepsilon_i$$

$y_i = \underline{1}$ then have probability $p(x_i)$

$y_i = 0$ then probability $1 - p(x_i)$

$$\sum_{i=1}^{2} p(x_i) = \underline{1}$$

$$p(x_i) = p(y_i = 1 \; x_i)$$

$$1 - p(x_i) = p(y_i = 0 \; x_i)$$

what distribution does ε follow?

$$y_i = 1 \; ; \; p(x_i)$$

$$1 = p(x_i) + \varepsilon_i \rightarrow$$

$$\varepsilon_i = 1 - p(x_i)$$

$$y_i = 0 \; ; \; 1 - p(x_i)$$

$$\varepsilon_i = -p(x_i)$$

$$y = 1 \; ; \; p(x) \; ; \; \varepsilon = 1 - p(x)$$

$$y = 0 \; ; \; 1 - p(x) \; ; \; \varepsilon = -p(x)$$

$$\mathbb{E}[\varepsilon] = (\underbrace{1 - p(x)}_{\varepsilon})\, p(x)$$

$$+ (-p(x))(1 - p(x))$$

$$= 0$$

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

$$+ (-p(x))^2 (1 - p(x))$$

$$\left((p(x) = p) \right)$$

$$= p(1 - p)$$

variance for Binomial distribution:

$$P(D \mid \beta) = \prod_{i=0}^{n-1} \underbrace{p(x_i)}_{y_i=1}^{y_i} \left(1 - p(x_i)\right)^{1-y_i}$$

Cost function

$$C(\beta) = -\ln p(D/\beta)$$

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

$$C(\beta) = -\sum_{i=0}^{n-1} \Big[y_i \ln p(x_i) + (1-y_i) \ln(1-p(x_i)) \Big]$$

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

$$= -\sum_{i=0}^{n-1} \Big[y_i (\beta_0 + \beta_1 x_i) - \ln(1+e^{\beta_0+\beta_1 x_i}) \Big]$$

$$\frac{\partial C}{\partial \beta_0} = 0 = -\sum_{i=0}^{n-1} (y_i - p(x_i)) = g_0$$

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

$$\frac{\partial C}{\partial \beta_1} = 0 = -\sum_{i=0}^{n-1} x_i (y_i - p(x_i)) = g_1$$

in general form

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

$$p, y \in \mathbb{R}^n$$

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

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

Hessian

W has only diagonal elements

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

$$w_{ii} = p(x_i)(1-p(x_i))$$

$$g = X^T(p-y) = g(\beta)$$

$$H = X^T W X = H(\beta)$$

We have g as a function of β

$$g(\beta) = X^T(p(\beta)-y) = 0$$

β?

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

Newton-Raphson (1 Dim):

$$f(s) = 0$$

Taylor expand $f(s) = f(x) + (s-x) f'(x) + \frac{(s-x)^2}{2!} f''(x) + \ldots$

$$f(s) \simeq f(x) + (s-x) f'(x)$$

$$s = x - f(x)/f'(x)$$

solve iteratively:

$$x_{n+1} = x_n - f(x)/f'(x)$$

in our case $f(x) \to g(\beta)$

$f'(x) \to H(\beta)$

with β_0 and β_1

$$H(\beta) = \begin{bmatrix} \frac{\partial g_0}{\partial \beta_0} & \partial g_0/\partial \beta_1 \\ \frac{\partial g_1}{\partial \beta_0} & \frac{\partial g_1}{\partial \beta_1} \end{bmatrix}$$

Newton's method

$$\begin{bmatrix} \beta_0^{n+1} \\ \beta_1^{n+1} \end{bmatrix} = \begin{bmatrix} \beta_0^{n} \\ \beta_1^{n} \end{bmatrix}$$

$$- H^{-1}(\beta_0^n, \beta_1^n) \times \begin{bmatrix} g_0(\beta_0^n, \beta_1^n) \\ g_1(\beta_0^n, \beta_1^n) \end{bmatrix}$$

convergence criteria

$$|\beta^{(n+1)} - \beta^{(n)}| < \varepsilon \sim 10^{-10}$$

$$\beta^{(n+1)} = \beta^{(n)} - \gamma\, g(\beta^{(n)})$$

learning rate