MEM-264 Applied Statistics

Department of Mathematics and Applied Mathematics, University of Crete

Costas Smaragdakis (kesmarag@uoc.gr)

21th lecture - 23-04-2021

Sample $\{(X_1, Y_1), \dots, (X_m, Y_m)\}_{k=1}$ iid.

X_i - independent random variable

Y_i - dependent random variable

Linear regression

In the general case we have the following random sample $\{(X_1^{(1)}, X_1^{(2)}, X_1^{(3)}, \dots, X_1^{(k)}, Y_1), \dots, (X_m^{(1)}, \dots, X_m^{(k)}, Y_m)\}$

$$Y_i = a + \sum_{j=1}^{k} b^{(j)} X_i^{(j)} + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2), \quad i = 1, \dots, n$$

$\varepsilon_1, \dots, \varepsilon_m \overset{iid}{\sim} N(0, \sigma^2)$

$$p = [a, b^{(1)}, b^{(2)}, \dots, b^{(k)}]^T \in \mathbb{R}^{k+1}$$

p ↑ models parameters

The least squares estimator

Dataset : $\{(X_1^{(1)}, \dots, X_1^{(k)}, Y_1), \dots, (X_m^{(1)}, \dots, X_m^{(k)}, Y_m)\}$ We wish to find an estimator

$\hat{p} = [\hat{a}, \hat{b}^{(1)}, \dots, \hat{b}^{(k)}]$

- Minimize the following function of p:

$$(y - Xp)^T (y - Xp)$$

$\hat{y} = \hat{a} + \sum_{j=1}^{k} \hat{b}^{(j)} X^{(j)}$

↑ the estimator

Given $(X^{(1)}, \dots, X^{(k)})$ $\xrightarrow{\text{estimator}}$ $\hat{y} \approx y$

$$\underset{\sim}{y} = [y_1, \ldots, y_n]^T, \quad \underset{\approx}{X} = \begin{bmatrix} 1 & x_1^{(1)} \ldots x_1^{(k)} \\ 1 & x_2^{(1)} \ldots x_2^{(k)} \\ \vdots & \\ 1 & x_n^{(1)} \ldots x_n^{(k)} \end{bmatrix}$$

$$(\underset{\sim}{y} - \underset{\approx}{X}\underset{\sim}{p})^T (\underset{\sim}{y} - \underset{\approx}{X}\underset{\sim}{p}) = \| \underset{\sim}{y} - \underset{\approx}{X}\underset{\sim}{p} \|_2^2$$

↓

the minimum value of the norm is taken when $\underset{\sim}{p} = (\underset{\approx}{X}^T\underset{\approx}{X})^{-1}\underset{\approx}{X}^T\underset{\sim}{y}$

or

$$(\underset{\approx}{X}^T\underset{\approx}{X})\underset{\sim}{p} = \underset{\approx}{X}^T\underset{\sim}{y}$$

k+1 equations and k+1 unknowns

By multiplying the i-th row of the matrix with $\hat{p}$

$$[1, x_i^{(1)}, \ldots, x_i^{(k)}] \begin{bmatrix} \hat{\alpha} \\ \hat{b}^{(1)} \\ \vdots \\ \hat{b}^{(k)} \end{bmatrix} = \hat{\alpha} + \sum_{j=1}^{k} \hat{b}^{(j)} x_i^{(j)}$$

exercise/ derivation ↓

$$\frac{\partial}{\partial p}(y - Xp)^T(y - Xp) = 2(X^TXp - X^Ty) = 0$$

$$\frac{\partial^2}{\partial p \partial p^T}(y - Xp)^T(y - Xp) = 2X^TX \succ 0 \text{ if } X \text{ a invertible matrix. (non singular)}$$

Therefore

$$\Rightarrow \hat{p} = (X^TX)^{-1}X^Ty$$

example

$\{(1, 2, 3), (2, 5, 9)\}$

dependent: 3, 9

independent: 1, 2, 2, 5

$\underset{\sim}{y} = \begin{bmatrix} 3 \\ 9 \end{bmatrix}$ $\quad \underset{\approx}{X} = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 2 & 5 \end{bmatrix}$ $\quad p = \begin{bmatrix} \alpha \\ b^{(1)} \\ b^{(2)} \end{bmatrix}$

$$(\underset{\approx}{X}^T \underset{\approx}{X}) \underset{3\times 1}{p} = \underset{3\times 2}{\underset{\approx}{X}^T} \underset{2\times 1}{\underset{\sim}{y}}$$

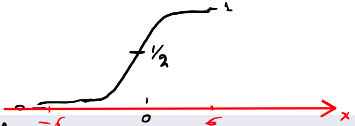
$\underset{\approx}{X}^T$: 3×2; $\underset{\approx}{X}$: 2×3; $(\underset{\approx}{X}^T \underset{\approx}{X})$: 3×3

the solution of this problem defines the estimator.

$$\hat{y} = \hat{\alpha} + \hat{b}^{(1)} x^{(1)} + \hat{b}^{(2)} x^{(2)}$$

$(1.5, 1.5, ?)$

(end of the first part of this lecture!) —



Sigmoid function (logistic function)

$$0 < \text{sigmoid}(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{e^x+1} < 1$$

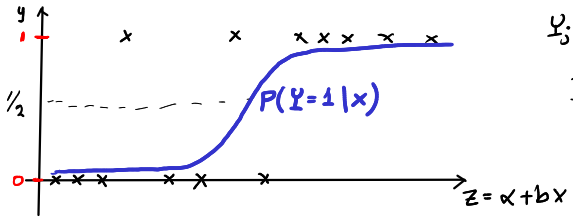
Logistic regression

$$\log \frac{p(Y_j=1|X_j^{(1)},\dots,X_j^{(k)})}{1-p(Y_j=1|X_j^{(1)},\dots,X_j^{(k)})} = a + \sum_{j=1}^{k} b^{(j)} X_j + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0,\sigma^2), \quad i = 1,\dots,n$$

$$\pi_j = P(Y_j=1|X_j^{(1)},\dots,X_j^{(k)})$$

$\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ iid (X_j, Y_j) , X_j Continuous Random Variable

$Y_j \sim \mathcal{B}e(\theta)$ Bernoulli



$$P(Y=1|X) = \text{Sigmoid}(z(x)) = \frac{e^{z(x)}}{1+e^{z(x)}}$$

$$z(x) = \log \frac{P(Y=1|x)}{\underbrace{1-P(Y=1|x)}_{P(Y=0|x)}}$$

$$\log \frac{P(Y=1|x)}{1-P(Y=1|x)} = \alpha + bx + \varepsilon$$

$$\log \frac{\pi_j}{1-\pi_j} = \alpha + bx_j + \varepsilon_j$$

$$P(\underset{\sim}{y}|\underset{\sim}{x}) = \prod_{j=1}^{n} P(y_j|x_j) = \prod_{j=1}^{n} P(y_j=1|z_j)^{y_j} P(y_j=0|z_j)^{1-y_j} =$$

$$\underset{\sim}{y} = (y_1, \dots, y_n)^T$$

$$\underset{\sim}{x} = (x_1, \dots, x_n)^T$$

$$= \prod_{j=1}^{n} \pi_j^{y_j} (1-\pi_j)^{1-y_j}$$

$$\log P(\underset{\sim}{y}|\underset{\sim}{x}) = \log \prod_{j=1}^{n} \pi_j^{y_j} (1-\pi_j)^{1-y_j} = \sum_{j=1}^{n} y_j \log \pi_j + (1-y_j)(1-\pi_j)$$

Categorical cross entropy $BCE = -\frac{1}{N} \sum_{j=1}^{n} y_j \log \pi_j + (1-y_j)(1-\pi_j)$