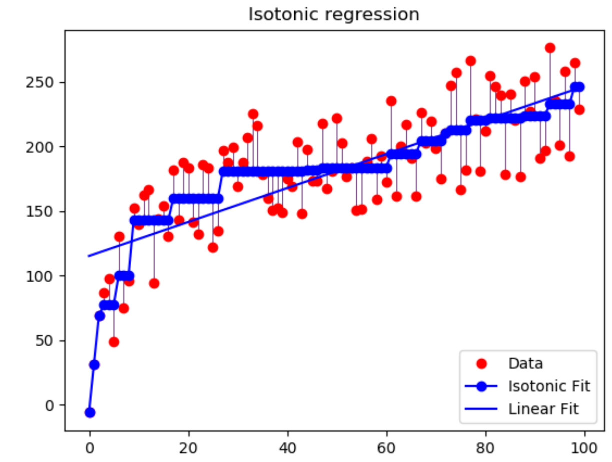
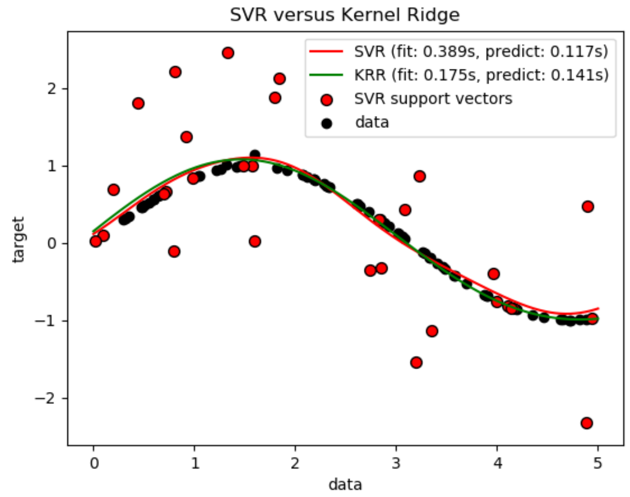
迴歸與分類

黃志勝（Tommy Huang）
義隆電子 人工智慧研發部
國立陽明交通大學 AI學院 合聘助理教授
國立台北科技大學 電資學院合聘助理教授

迴歸: Regression

- In the last presentation, we brief introduce ML topic.
- Regression: predicting a continuous-valued attribute associated with an object.
- What to do?
- How to do?

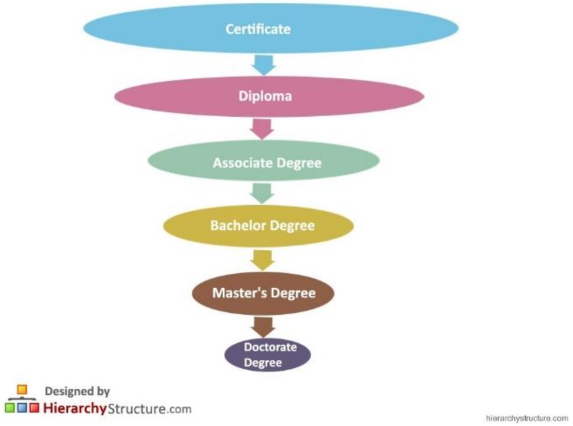




Regression

●What to do?

independent variables



predict

dependent variables

Which are dependent variables?

Depend on your problem : specific definition (salary prediction or bodyfat prediction)

Which are independent variables?

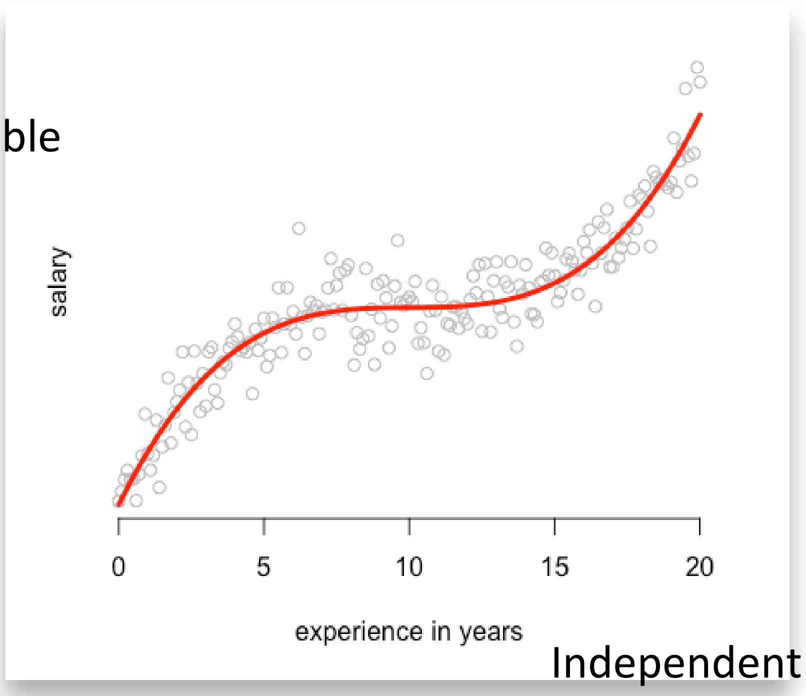
Depend on your collecting data.

Regression

●How to do?

Finding the curve that best fits your data is called regression.

Dependent variable
(y)



Independent variable
(x)

$$y = f(x)$$

f is a linear function：
linear regression

f is non-linear function：
nonlinear regression

Regression

y: salary, x: experience in years

$$y = f(x) = \beta_0 + \beta_1 x$$ → Simple linear regression

β_0: intercept
β_1: Slope

Regression

If there are more than one independent variables.

y: salary

x_1: experience in years

x_2: career

$$y = f(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \longrightarrow \text{Multiple linear regression}$$

Regression

- How to do nonlinear?
- Let your independent variables as a other independent variable by

1. polynomial.

$$y = f(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2$$

2. Interact.

$$y = f(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

3. Nonlinear function (ϕ): sigmoid function,...

$$y = f(x) = \phi(\beta_0 + \beta_1 x_1 + \beta_2 x_2)$$

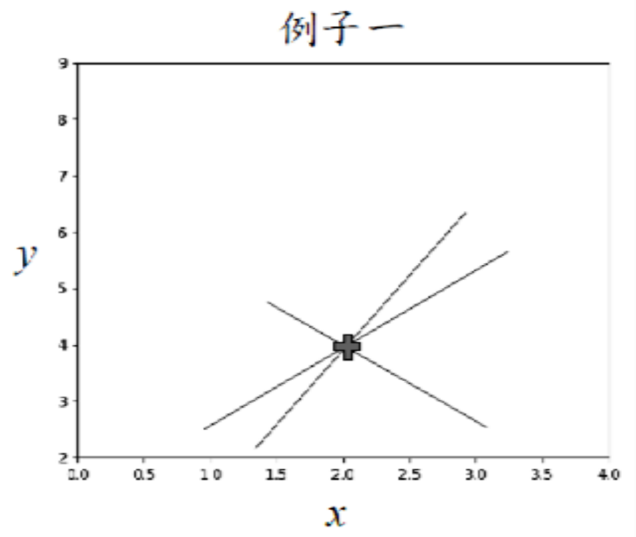
Regression(Example)

$$y = f(x) = \beta_0 + \beta_1 x$$

- 訓練資料只有一筆資料(*x*, *y*) = {(2, 4)}，我們將此資料代入方程式
- 內：

$$4 = \beta_0 + 2\beta_1$$

- β_0和β_1的解有無限多組。

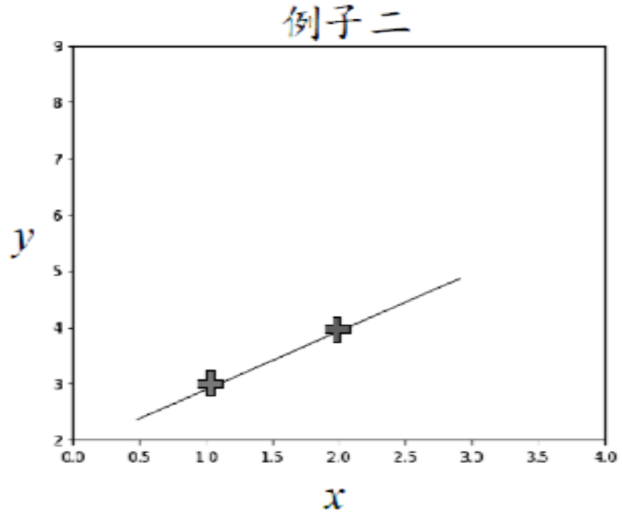


Regression(Example)

$$y = f(x) = \beta_0 + \beta_1 x$$

訓練資料只有一筆資料(*x*, *y*) = {(2, 4),(1,3)}，我們將此資料代入方程式內：

$$\begin{cases} 4 = \beta_0 + 2\beta_1 \\ 3 = \beta_0 + 1\beta_1 \end{cases} \Rightarrow \begin{cases} \beta_0 = 2 \\ \beta_1 = 1 \end{cases}$$

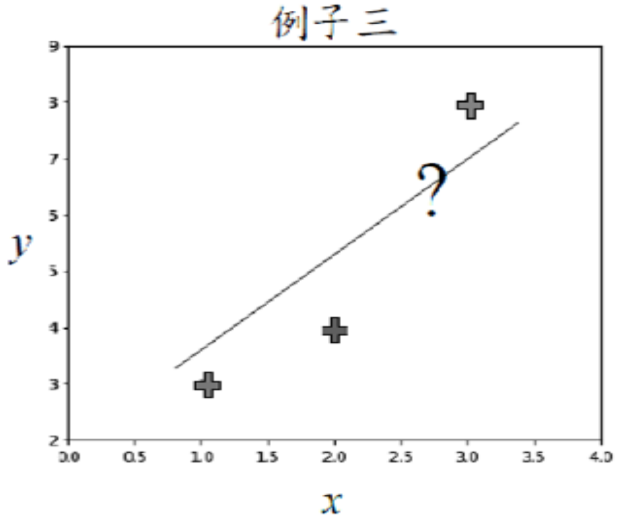


Regression(Example)

$$y = f(x) = \beta_0 + \beta_1 x$$

訓練資料只有一筆資料(*x*, *y*) = {(2, 4),(1,3),(3,8)}，我們將此資料代入方程式內：

$$\begin{cases} 4 = \beta_0 + 2\beta_1 \cdots (1) \\ 3 = \beta_0 + 1\beta_1 \cdots (2) \\ 8 = \beta_0 + 3\beta_1 \cdots (3) \end{cases}$$



Regression

- For now, we clearly understand what is regression.

Recall: How to do?

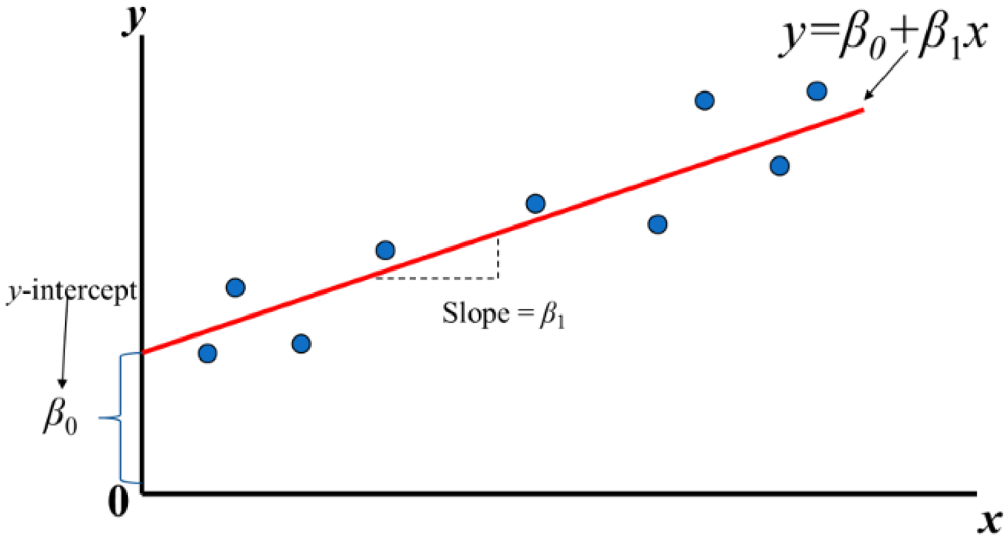
Finding the curve that best fits your data is called regression.

Two key points: **1. data, 2. curve.**

Data is the blue point

Curve is the red line

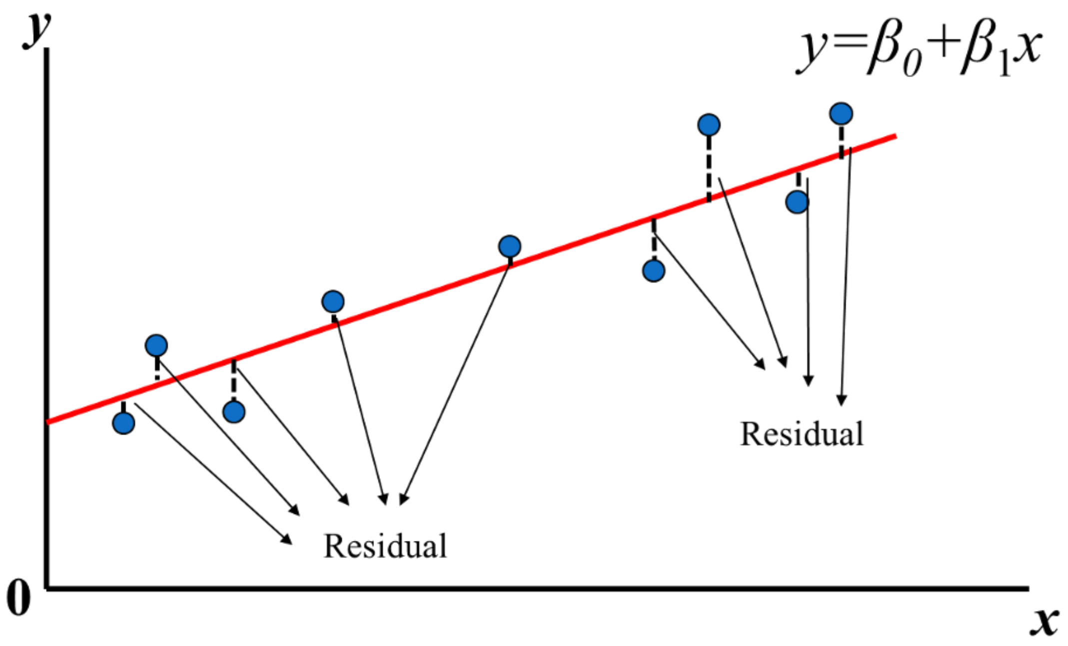
Using the data to find the β_0 and β_1



Regression

- Using the data to find the β_0 and β_1 .

How to achieve this goal?



Ideal:

All the data can fix on this line.

Real:

Fix on the line as best as possible.

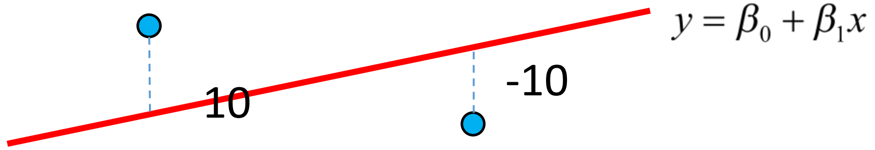
Residuals are as small as possible.

Regression

- Residuals are as small as possible.

$$residual = \hat{y} - y$$

- Residuals can be positive and negative.



$$sum\ error = \sum_i \left(\hat{y}_i - y_i\right) = 10 - 10 = 0$$

$$sum\ square\ error = \sum_i \left(\hat{y}_i - y_i\right)^2 = 100 + 100 = 200$$

Regression

- We usually hope the can let the sum square error as small as possible.

$$sum\ square\ error(SSE) = \sum_i (\hat{y}_i - y_i)^2$$

$$mean\ square\ error(MSE) = \frac{1}{n}\sum_{i=1}^{n} (\hat{y}_i - y_i)^2$$

- SO in regression, the objective/loss function is MSE.

$$\min_{\beta_0,\beta_1}\left\{ loss(\beta_0, \beta_1) = \frac{1}{n}\sum_{i=1}^{n} (\hat{y}_i - y_i)^2 = \frac{1}{n}\sum_{i=1}^{n} ((\beta_0 + \beta_1 x) - y_i)^2 \right\}$$

Regression

- In calculation, using derivative to find the minima.

$$\min_{\beta_0,\beta_1}\left\{loss(\beta_0,\beta_1)=\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i-y_i)^2=\frac{1}{n}\sum_{i=1}^{n}((\beta_0+\beta_1 x)-y_i)^2\right\}$$

$$\frac{\partial loss(\beta_0,\beta_1)}{\partial \beta_0}=0$$

$$\frac{\partial loss(\beta_0,\beta_1)}{\partial \beta_1}=0$$

Regression

Find β_0 (intercept)

$$\frac{\partial loss(\beta_0,\beta_1)}{\partial \beta_0} = \frac{\partial \frac{1}{n}\sum_{i=1}^{n}(\beta_0+\beta_1 x_i - y_i)^2}{\partial \beta_0} = 0$$

$$\Rightarrow \frac{2}{n}\sum_{i=1}^{n}(\beta_0+\beta_1 x_i - y_i) = 0$$

$$\Rightarrow \sum_{i=1}^{n}(\beta_0) + \sum_{i=1}^{n}(\beta_1 x_i - y_i) = 0$$

$$\Rightarrow n\beta_0 = \sum_{i=1}^{n}(y_i - \beta_1 x_i)$$

$$\Rightarrow \beta_0 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_1 x_i) = \frac{1}{n}\sum_{i=1}^{n}(y_i) - \beta_1 \frac{1}{n}\sum_{i=1}^{n}(x_i) = \overline{y} - \beta_1 \overline{x}$$

Regression

Find β_1 (Slope)

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

$$\frac{\partial loss(\beta_0, \beta_1)}{\partial \beta_1} = \frac{\partial \frac{1}{n} \sum_{i=1}^{n} (\beta_0 + \beta_1 x_i - y_i)^2}{\partial \beta_1} = 0$$

$$\Rightarrow \frac{2}{n} \sum_{i=1}^{n} (\beta_0 + \beta_1 x_i - y_i) x_i = 0$$

$$\Rightarrow \sum_{i=1}^{n} (\bar{y} - y_i) x_i + \beta_1 \sum_{i=1}^{n} (x_i - \bar{x}) x_i = 0$$

$$\Rightarrow \beta_1 \sum_{i=1}^{n} (x_i - \bar{x}) x_i = \sum_{i=1}^{n} (y_i - \bar{y}) x_i$$

$$\Rightarrow \beta_1 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Details

$$\beta_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})x_i}{\sum_{i=1}^{n}(x_i - \bar{x})x_i}$$

分母：

$$\sum_{i=1}^{n}(x_i - \bar{x})x_i = \sum_{i=1}^{n}(x_i x_i - \bar{x}x_i) = \sum_{i=1}^{n}x_i^2 - \sum_{i=1}^{n}\bar{x}x_i = \sum_{i=1}^{n}x_i^2 - \bar{x}\sum_{i=1}^{n}x_i = \sum_{i=1}^{n}x_i^2 - n\bar{x}^2 \ldots(1)$$

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}x_i^2 - 2\bar{x}\sum_{i=1}^{n}x_i + \sum_{i=1}^{n}\bar{x}^2 = \sum_{i=1}^{n}x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 = \sum_{i=1}^{n}x_i^2 - n\bar{x}^2 \ldots(2)$$

$$\sum_{i=1}^{n}(x_i - \bar{x})x_i = \sum_{i=1}^{n}(x_i - \bar{x})^2$$

分子：

$$\sum_{i=1}^{n}(y_i - \bar{y})x_i = \sum_{i=1}^{n}(x_i y_i - \bar{y}x_i) = \sum_{i=1}^{n}x_i y_i - \bar{y}\sum_{i=1}^{n}x_i = \sum_{i=1}^{n}x_i y_i - n\bar{x}\bar{y} \ldots(3)$$

$$\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) = \sum_{i=1}^{n}x_i y_i - \bar{x}\sum_{i=1}^{n}y_i - \bar{y}\sum_{i=1}^{n}x_i + \sum_{i=1}^{n}\bar{x}\bar{y} = \sum_{i=1}^{n}x_i y_i - n\bar{x}\bar{y} - n\bar{x}\bar{y} + n\bar{x}\bar{y} = \sum_{i=1}^{n}x_i y_i - n\bar{x}\bar{y} \ldots(4)$$

$$\sum_{i=1}^{n}(y_i - \bar{y})x_i = \sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})$$

Ordinary Least Square Estimation (OLSE)

We hope the loss as small as possible, so this approach is called ordinary least square estimation.

Recall:

$$\min_{\beta_0,\beta_1}\left\{ loss(\beta_0,\beta_1) = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2 \right\}$$

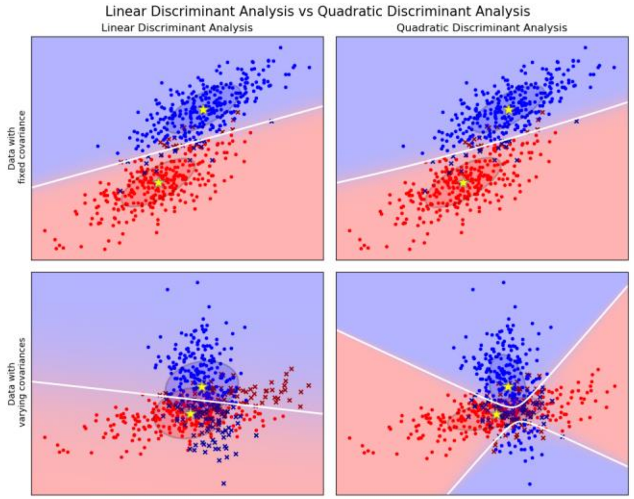
$$\frac{\partial loss(\beta_0,\beta_1)}{\partial \beta_0} = 0 \Rightarrow \beta_0 = \bar{y} - \beta_1\bar{x}$$

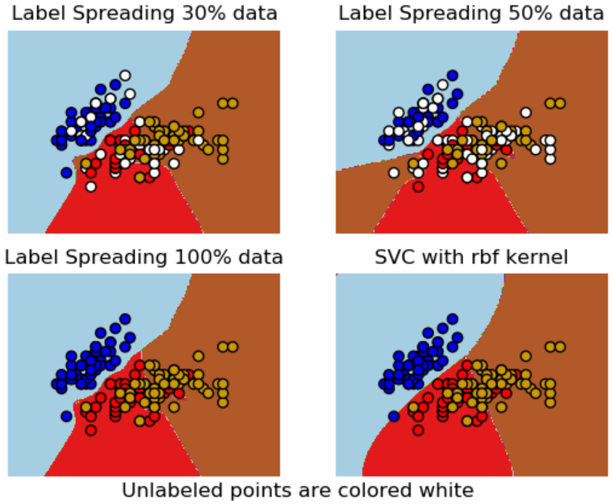
$$\frac{\partial loss(\beta_0,\beta_1)}{\partial \beta_1} = 0 \Rightarrow \beta_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

分類: Classification

Identifying to which category an object belongs to.

- **Logistic Regression**
- **Linear and Quadratic Discriminant Analysis**
- Support Vector Machine
- Nearest neighbors
- Random forest
- Neural Network





Classification

A Very simple classification problem

"How to classify {male or female} by a measured feature (body fat)?"

<u>Collected data (body fat(%))</u>

Female:{22, 25, 30, 33, 35}

Male:{ 10, 15, 20, 25, 30}

Classification

Female: {22, 25, 30, 33, 35}

Male: { 10, 15, 20, 25, 30}

The simplest way:

Using mean value as decision rule.

$$\frac{\text{Mean value (Female)} + \text{Mean value (Male)}}{2} = \frac{29+20}{2} = 24.5$$

Body fat>24.5 → Female

Body fat<24.5 → Male

Classification

Female: {22, 25, 30, 33, 35}

Male: { 10, 15, 20, 25, 30}

The simplest way:

Using mean value as decision rule.

$$\frac{\text{Mean value (Female)} + \text{Mean value (Male)}}{2} = \frac{29+20}{2} = 24.5$$

Body fat>24.5 → Female

Body fat<24.5 → Male

Male	**10**	**15**	**20**	**25**	**30**	
Female			22	25	30	35
分布	-15	15-20	20-25	25-30	30-35	35-

Male

Female

24.5

Classification

Female with 100 data, Male with 100 data (Body fat).

Visualization by histogram.

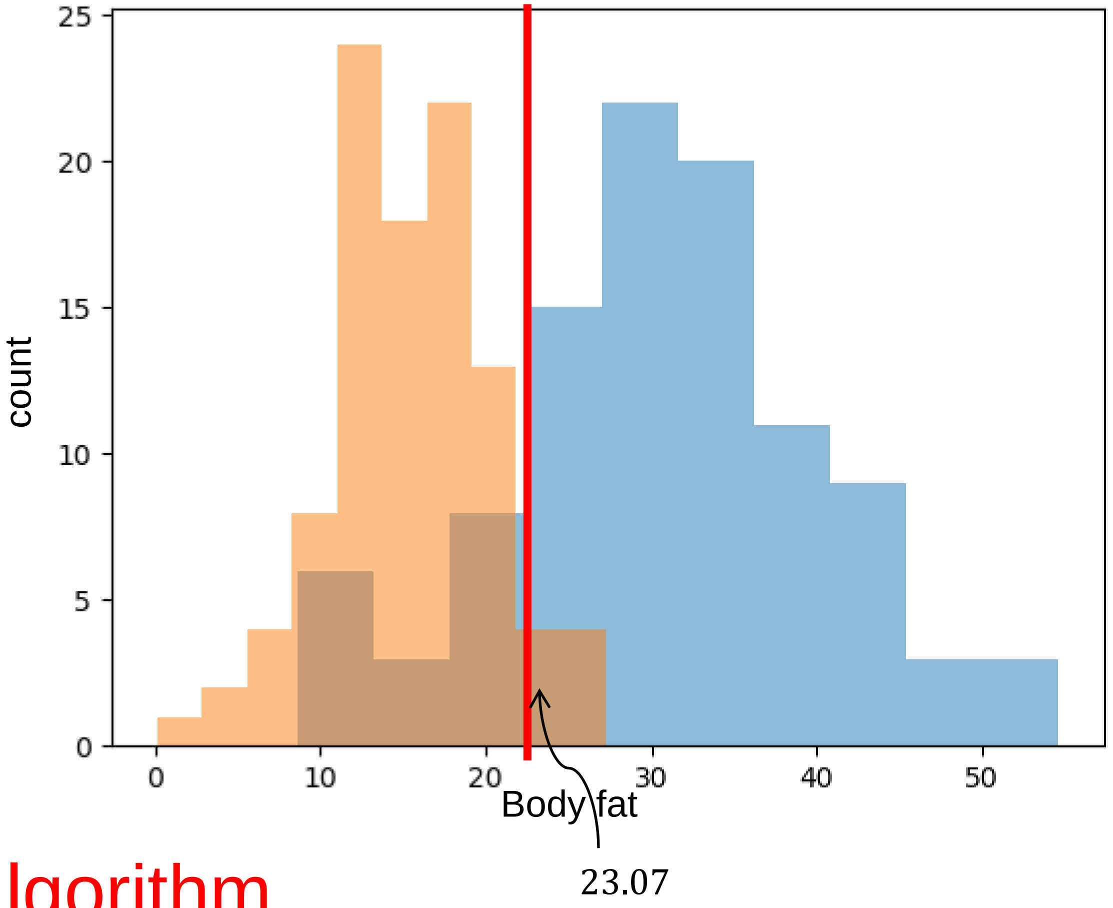
Blue: Male

Red: Female

$$\frac{\text{Mean value (Female)} + \text{Mean value (Male)}}{2}$$

$$= \frac{30.79 + 15.35}{2} = 23.07$$

You Just learn a classification algorithm



Classification (平均數法)

$\{x_i\}, \forall i, x: baby\ fat$

$$\mu_c = \frac{1}{n_c}\sum_{i=1}^{n_c} x_i\,, c = \{male, female\,\}$$

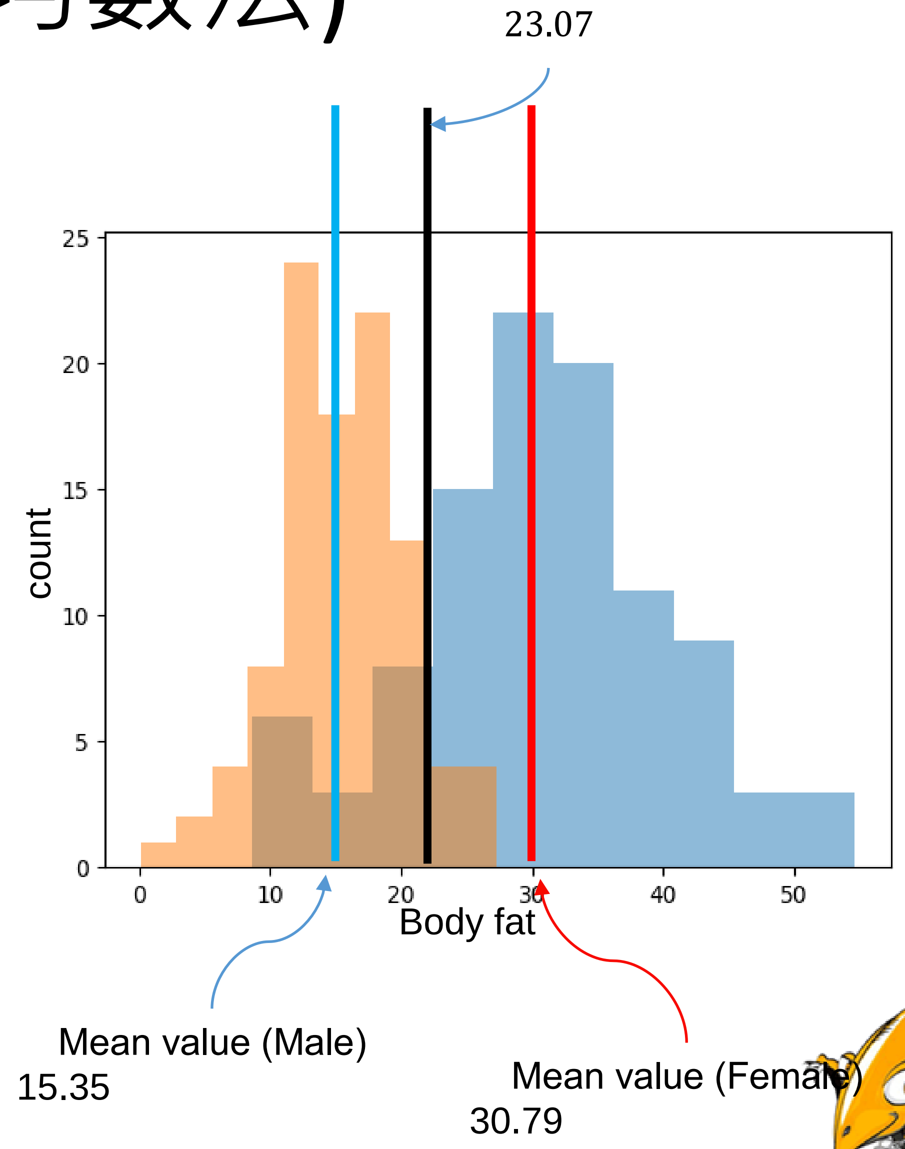
$$f_{male}(x) = x - \mu_{male}$$

$$f_{female}(x) = x - \mu_{female}$$

Decision rule: feature value(x) is closed to which class, and classify this x to which class.

Decision rule:

$$Decision(x) = \begin{cases} female & f_{male}(x) - f_{female}(x) \geq 0 \\ male & f_{male}(x) - f_{female}(x) < 0 \end{cases}$$



Likelihood function(Single variable)

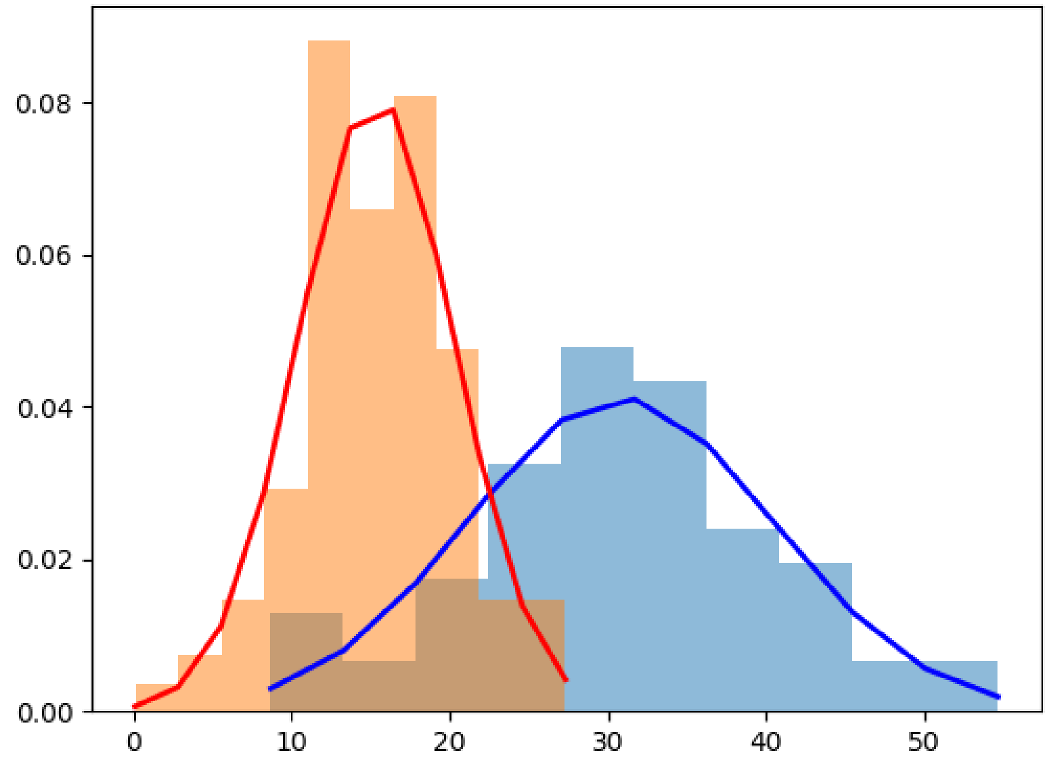
We can assume the histogram (density) is a **<u>Gaussian</u>** (normal)-like distribution.

That means

$$x_{male} \sim N(\mu_{male}, \sigma_{male})$$

$$x_{female} \sim N(\mu_{female}, \sigma_{female})$$

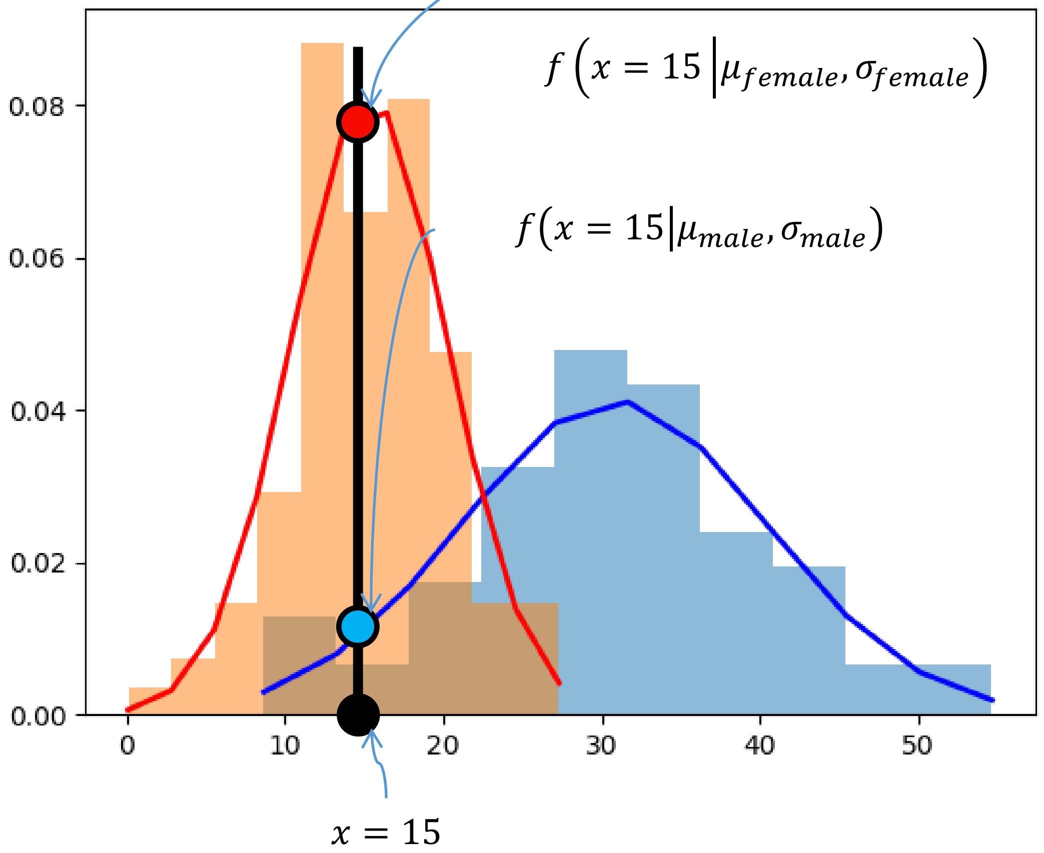
$$f(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Likelihood function(Single variable)

$x_{male} \sim N(\mu_{male}, \sigma_{male})$
$x_{female} \sim N(\mu_{female}, \sigma_{female})$

$$f(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



A unlabeled x: 15% body fat

$$f\left(x = 15 \middle| \mu_{female}, \sigma_{female}\right) > f(x = 15|\mu_{male}, \sigma_{male})$$

So this unlabeled x would be classify to Female.

$$Decision(x) = \begin{cases} female & f\left(x \middle| \mu_{female}, \sigma_{female}\right) \geq f(x|\mu_{male}, \sigma_{male}) \\ male & f\left(x \middle| \mu_{female}, \sigma_{female}\right) < f(x|\mu_{male}, \sigma_{male}) \end{cases}$$

Classification (Multi-variables)(平均數法)

If we get multi-features (i.e. body fat and height), how to do?

$$\boldsymbol{x}_i = \begin{bmatrix} x_{bodyfat} \\ x_{height} \end{bmatrix}$$

$$\boldsymbol{\mu}_c = \frac{1}{n_c}\sum_{i=1}^{n_c} \boldsymbol{x}_i \,, = \begin{bmatrix} \mu_{bodyfat} \\ \mu_{height} \end{bmatrix}_c \,, c = \{male, female\}$$

向量

$$\boxed{f(\boldsymbol{x}) = \boldsymbol{x} - \boldsymbol{\mu} = \begin{bmatrix} x_{bodyfat} - \mu_{bodyfat} \\ x_{height} - \mu_{height} \end{bmatrix}}$$

純量　　純量

We can't make decision with an array.

$$Decision(x) = \begin{cases} female & \boxed{f\left(x \middle| \mu_{female}, \sigma_{female}\right)} \geq \boxed{f\left(x \middle| \mu_{male}, \sigma_{male}\right)} \\ male & f\left(x \middle| \mu_{female}, \sigma_{female}\right) < f\left(x \middle| \mu_{male}, \sigma_{male}\right) \end{cases}$$

Classification (Multi-variables) (平均數法)

$$f(\boldsymbol{x}) = \boldsymbol{x} - \boldsymbol{\mu} = \begin{bmatrix} x_{bodyfat} - \mu_{bodyfat} \\ x_{height} - \mu_{height} \end{bmatrix}$$

Quantification $(\boldsymbol{x}, \boldsymbol{\mu})$

- Euclidean Distance (L2-norm): $\|\boldsymbol{x} - \boldsymbol{\mu}\|_{\mathrm{L2}} = \underbrace{(\boldsymbol{x} - \boldsymbol{\mu})^{\mathrm{T}}}_{1 \times 2} \underbrace{(\boldsymbol{x} - \boldsymbol{\mu})}_{2 \times 1}$
- Mahalanobis Distance

Likelihood function(Multi-variables)

If we get multi-features (i.e. body fat and height), how to do?

$$\boldsymbol{x}_i = \begin{bmatrix} x_{bodyfat} \\ x_{height} \end{bmatrix}$$

純量

$$\boxed{\begin{aligned} &f(\boldsymbol{x}|\boldsymbol{\mu},\Sigma) \\ &= (2\pi)^{-d/2}|\Sigma|^{-0.5} exp\{-0.5(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\} \end{aligned}}$$

Mahalanobis Distance = $(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$

Likelihood function(Multi-variables)

$$f(\boldsymbol{x}|\boldsymbol{\mu},\Sigma) = \boxed{(2\pi)^{-d/2}}|\Sigma|^{-0.5}exp\{-0.5(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\}$$

純量

$|\Sigma|$: 共變異數的行列式值 $\rightarrow$ 純量

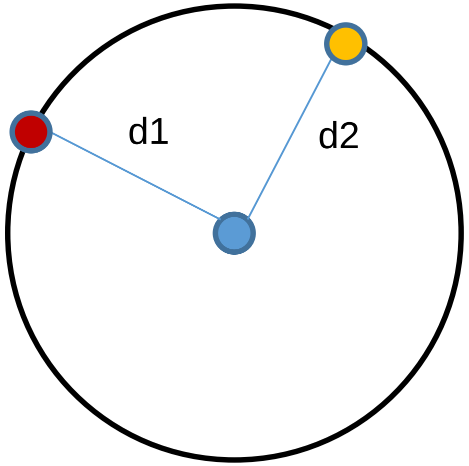
$$\boxed{\underbrace{(\boldsymbol{x}-\boldsymbol{\mu})^T}_{1\times d}\underbrace{\Sigma^{-1}}_{d\times d}\underbrace{(\boldsymbol{x}-\boldsymbol{\mu})}_{d\times 1}}$$

輸出為 $1\times 1\rightarrow$純量

Distance

Euclidean Distance

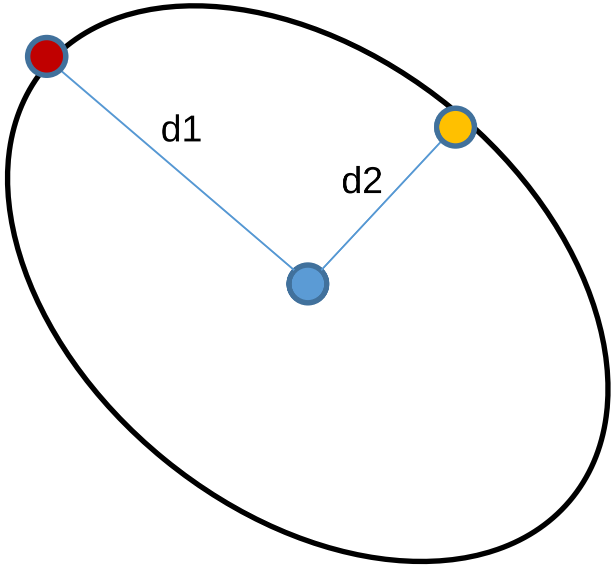
$$(\boldsymbol{x}-\boldsymbol{\mu})^T(\boldsymbol{x}-\boldsymbol{\mu})$$



d1 = d2

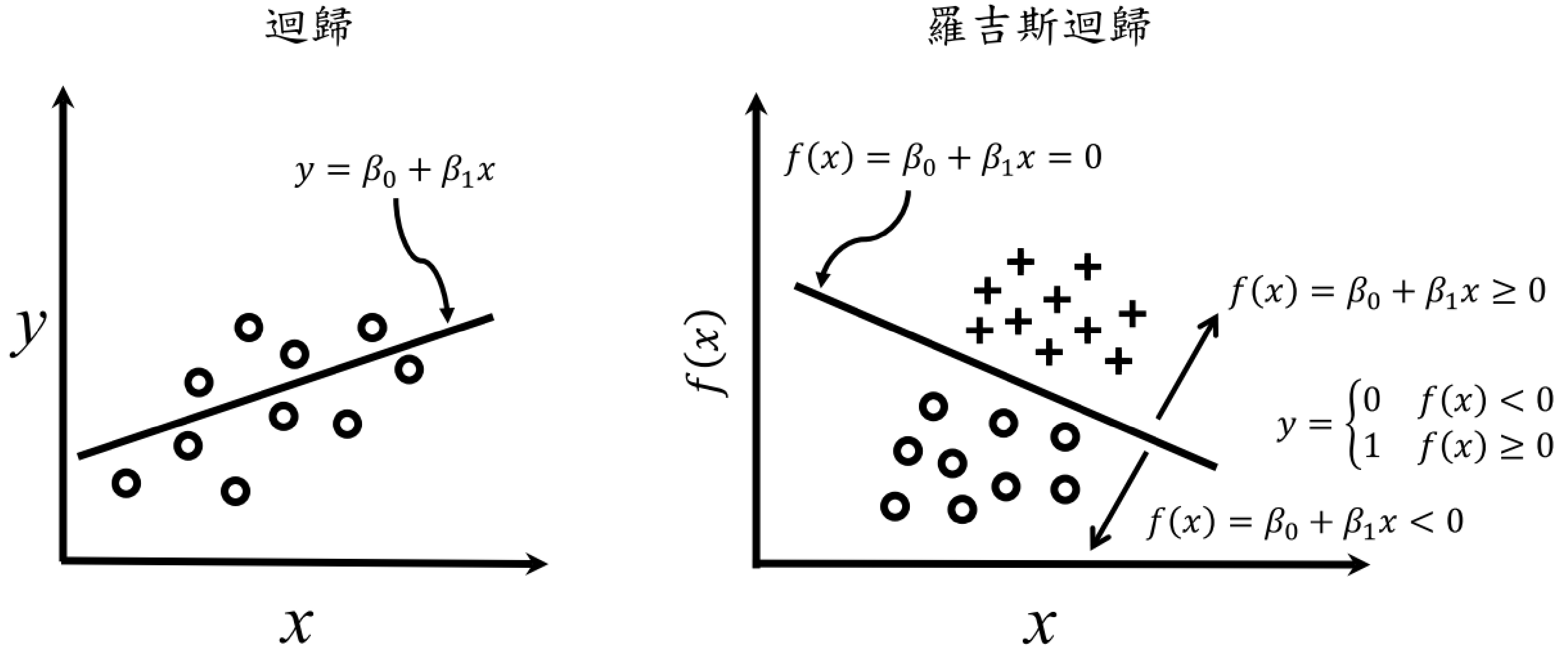
Mahalanobis Distance

$$(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$$



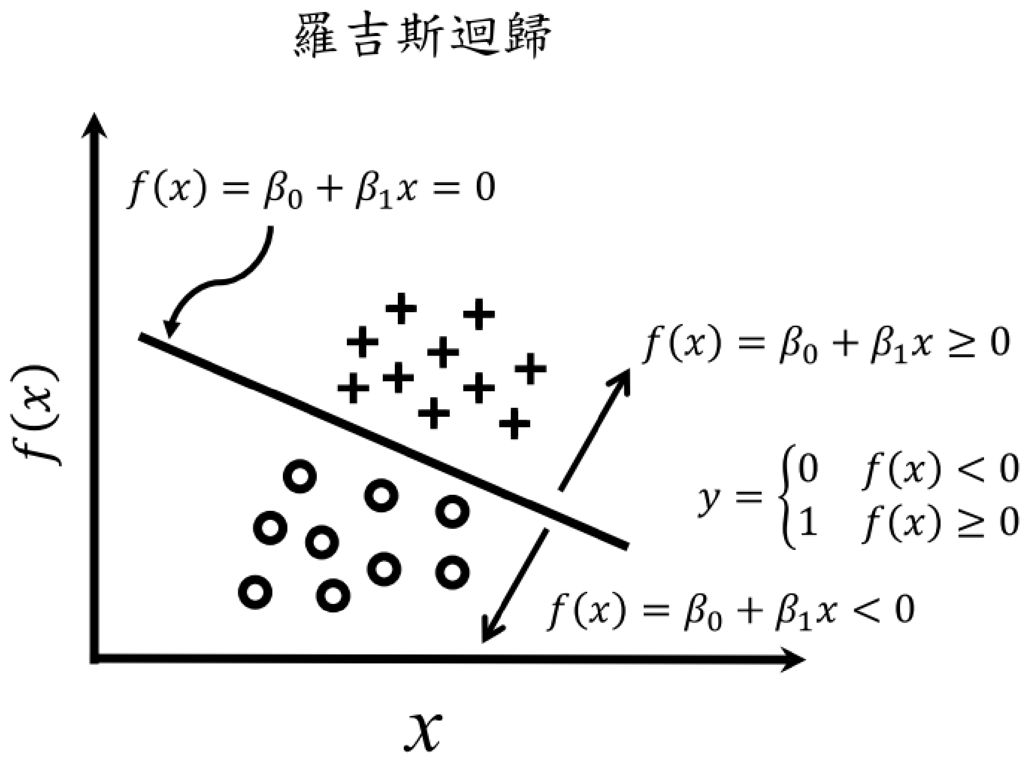
d1 = d2

羅吉斯迴歸（Logistic Regression）



- 線性迴歸跟羅吉斯迴歸公式是一樣的 (但要分清楚，前者在算出數值，後者在做分類)

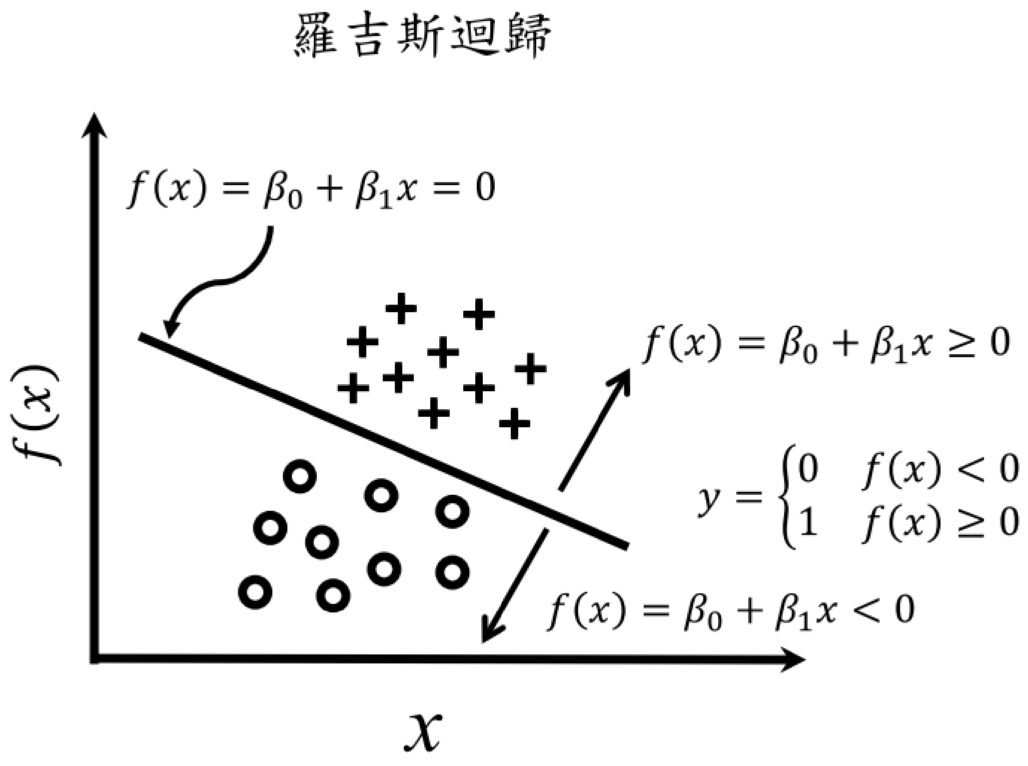
羅吉斯迴歸（Logistic Regression）



- 羅吉斯迴歸則是希望線性迴歸的輸出可以將兩類的資料能越區隔開越好。
- 最簡單的方式就是任意資料帶入迴歸方程式中判斷輸出值是否大於0，若大於0是一類(類別：1)，小於0則是另一類(類別：0)

$$y = \sigma\big(f(\mathbf{x})\big) = \begin{cases} 1 & f(\mathbf{x}) \geq 0 \\ 0 & f(\mathbf{x}) < 0 \end{cases}$$

羅吉斯迴歸（Logistic Regression）



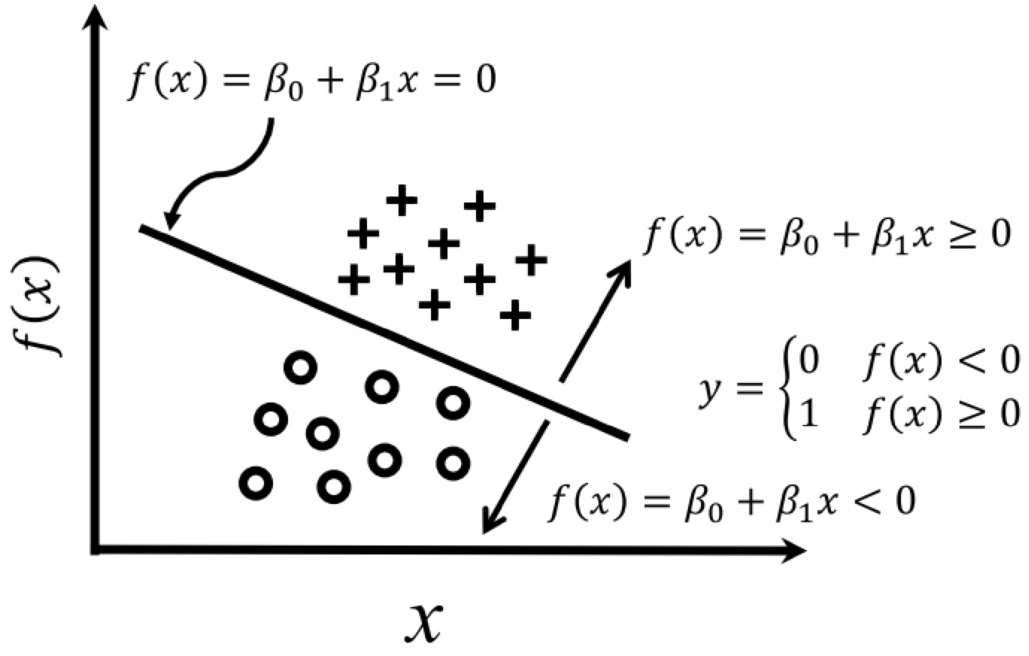
- σ(.)在機器學習上稱為單位階梯函數(unit step function)，大於一個閾值(threshold)是一類，反之為另一類，此例的閾值為0。

$$y = \sigma\left(f(\mathbf{x})\right) = \begin{cases} 1 & f(\mathbf{x}) \geq 0 \\ 0 & f(\mathbf{x}) < 0 \end{cases}$$

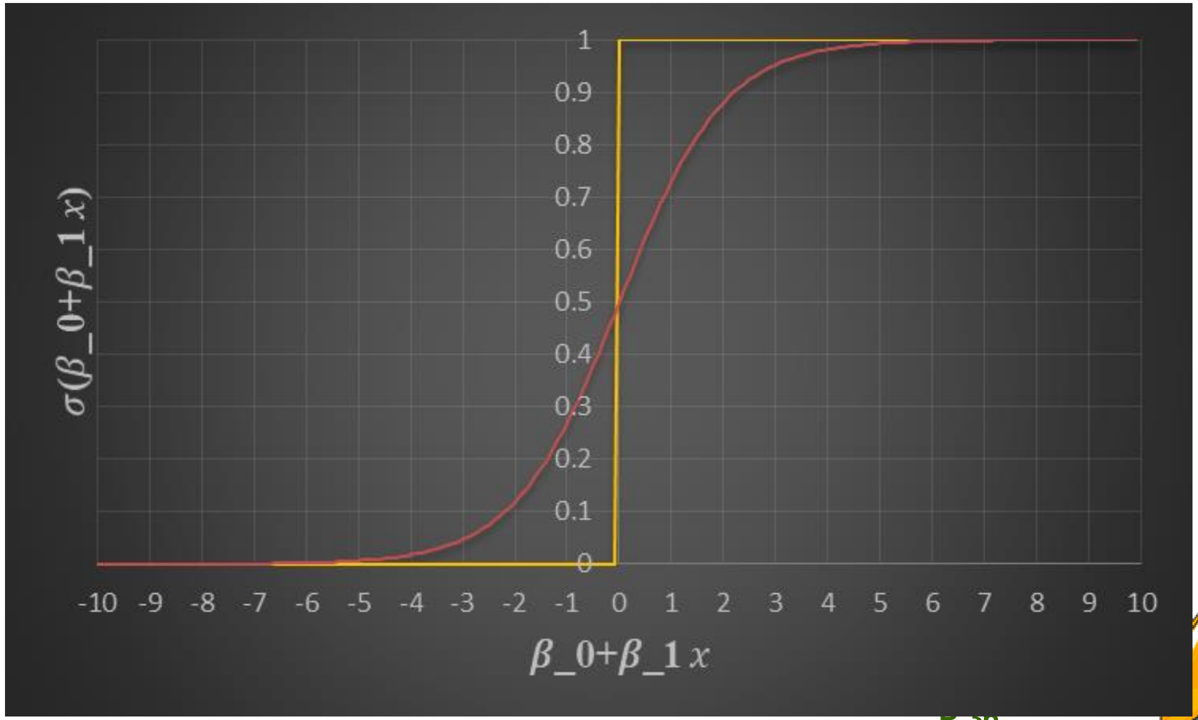
羅吉斯迴歸用Sigmoid函數限制值域

$$Sigmoid\text{函數}：s(x)=\frac{1}{1+e^{-x}}=\frac{e^{x}}{1+e^{x}},\ x\in[-\infty,\infty],\ s(x)\in[0,1]$$

羅吉斯迴歸



$$\sigma\left(f(x)\right)=\frac{1}{1+e^{-f(x)}}=\frac{e^{f(x)}}{1+e^{f(x)}}$$



羅吉斯迴歸的公式為

$$s\left(f(\mathbf{x})\right)=\frac{1}{1+e^{-f(\mathbf{x})}}=\frac{1}{1+e^{-\mathbf{x}^T\boldsymbol{\beta}}}$$

或寫成

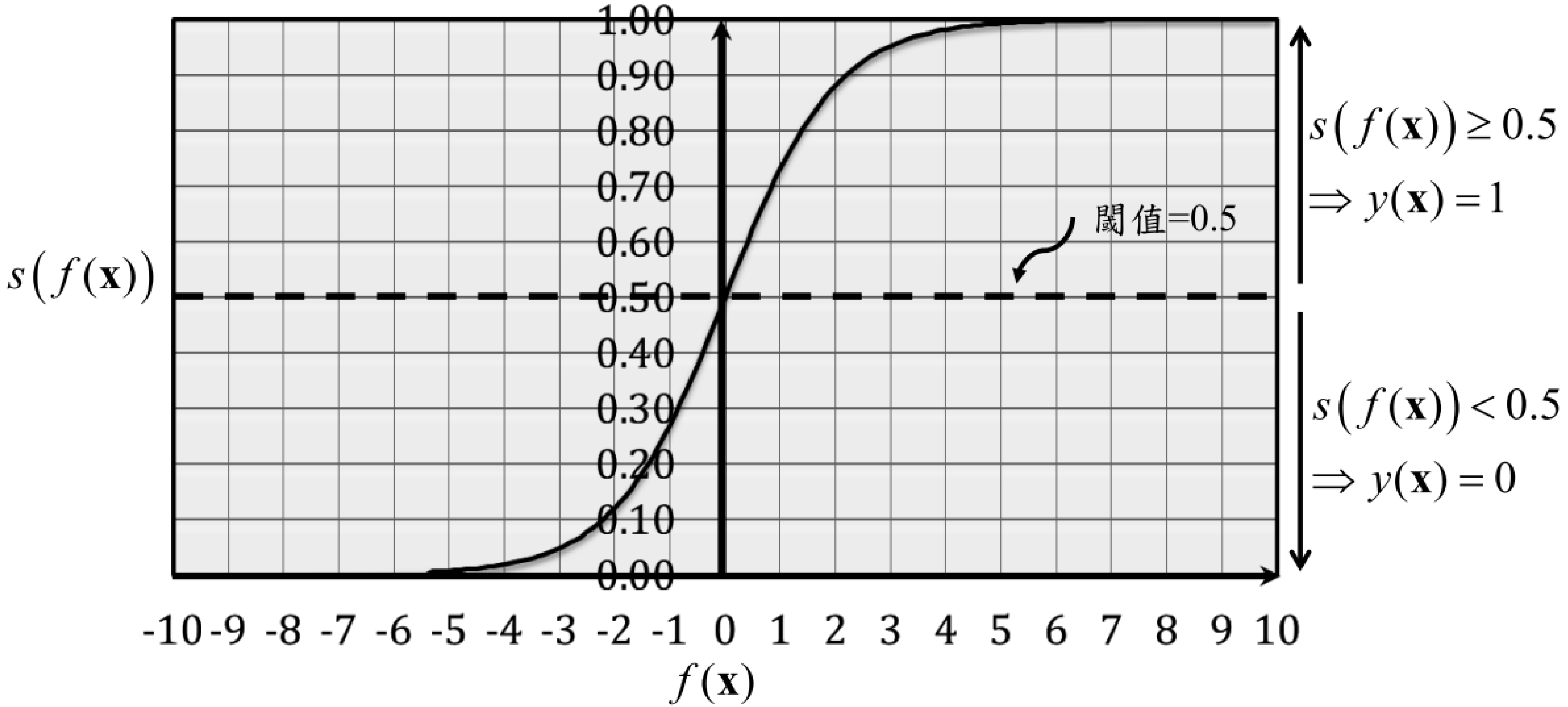
$$s\left(f(\mathbf{x})\right)=\frac{e^{f(\mathbf{x})}}{1+e^{f(\mathbf{x})}}=\frac{e^{\mathbf{x}^T\boldsymbol{\beta}}}{1+e^{\mathbf{x}^T\boldsymbol{\beta}}}$$

$$\boxed{y=\sigma\left(s\left(f(\mathbf{x})\right)\right)=\begin{cases}1 & s\left(f(\mathbf{x})\right)\geq 0.5\\ 0 & s\left(f(\mathbf{x})\right)<0.5\end{cases}.}$$

羅吉斯迴歸是二分類演算法

羅吉斯迴歸的公式為

$$s\left(f(\mathbf{x})\right)=\frac{1}{1+e^{-f(\mathbf{x})}}=\frac{e^{f(\mathbf{x})}}{1+e^{f(\mathbf{x})}}$$



要怎麼找(求解)羅吉斯回歸參數(**β**)?

$$y=\sigma\left(s\left(f(\mathbf{x})\right)\right)=\begin{cases}1 & s\left(f(\mathbf{x})\right)\geq 0.5\\ 0 & s\left(f(\mathbf{x})\right)<0.5\end{cases}$$

$$s\left(f(\mathbf{x})\right)=\frac{e^{f(\mathbf{x})}}{1+e^{f(\mathbf{x})}}=\frac{e^{\mathbf{x}^T\boldsymbol{\beta}}}{1+e^{\mathbf{x}^T\boldsymbol{\beta}}}$$

$$f(\mathbf{x})=\mathbf{x}^T\boldsymbol{\beta}$$

$$\mathbf{x}=\begin{bmatrix}1\\ x_1\\ \vdots\\ x_d\end{bmatrix},\quad \boldsymbol{\beta}=\begin{bmatrix}\beta_0\\ \beta_1\\ \vdots\\ \beta_d\end{bmatrix}$$

要怎麼找(求解)羅吉斯回歸參數(**β**)?

- 回顧一下伯努利機率函數，伯努利試驗結果為成功的機率為p，不成功的機率即為1-p

$$f(x) = p^x(1-p)^{1-x} = \begin{cases} p & x=1 \\ 1-p & x=0 \end{cases}$$

- 羅吉斯迴歸的輸出類別為1(成功)的機率是

$$p = p(y=1|\mathbf{x})$$

- 輸出類別為 0 的機率是

$$p(y=0|\mathbf{x}) = 1 - p(y=1|\mathbf{x}) = 1-p$$

要怎麼找(求解)羅吉斯回歸參數(β)?

- 有n組資料，其概似函數為

$$L(\boldsymbol{\beta})=\prod_{i=1}^{n}p_i^{y_i}(1-p_i)^{(1-y_i)}$$

$$p_i=p(y_i=1|\mathbf{x}_i),\ \forall i=1,\ldots.,n$$

概似函數最大化→<u>不好做</u>

我們把問題轉成-log，找最小化的問題。

$$\begin{aligned}\mathcal{L}(\boldsymbol{\beta})=-logL(\boldsymbol{\beta})&=-log\left(\prod_{i=1}^{n}p_i^{y_i}(1-p_i)^{1-y_i}\right)\\&=-\sum_{i=1}^{n}log\left(p_i^{y_i}(1-p_i)^{1-y_i}\right)\\&=-\sum_{i=1}^{n}\left(log\left(p_i^{y_i}\right)+\log\left((1-p_i)^{1-y_i}\right)\right)\\&=-\sum_{i=1}^{n}\left(y_i log(p_i)+(1-y_i)\log(1-p_i)\right)\\&=-\sum_{i=1}^{n}\left(y_i log(p_i)+\log(1-p_i)-y_i\log(1-p_i)\right)\\&=\boxed{-\sum_{i=1}^{n}\left(y_i log\left(\frac{p_i}{1-p_i}\right)+log(1-p_i)\right)}\end{aligned}$$

就是交叉熵的公式

p_i是什麼

要怎麼找(求解)羅吉斯回歸參數(**β**)?

p_i: 羅吉斯回歸的輸出

$$p_i = s\left(f(\mathbf{x}_i)\right) = \frac{e^{\mathbf{x}_i^T\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i^T\boldsymbol{\beta}}}$$

$$\ln(p_i) = \ln\left(\frac{e^{\mathbf{x}_i^T\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i^T\boldsymbol{\beta}}}\right) = \mathbf{x}_i^T\boldsymbol{\beta} - \ln\left(1+e^{\mathbf{x}_i^T\boldsymbol{\beta}}\right)$$

$$\ln(1-p_i) = \ln\left(1-\frac{e^{\mathbf{x}_i^T\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i^T\boldsymbol{\beta}}}\right) = \ln\left(\frac{1}{1+e^{\mathbf{x}_i^T\boldsymbol{\beta}}}\right) = -\ln\left(1+e^{\mathbf{x}_i^T\boldsymbol{\beta}}\right)$$

$$\mathcal{L}(\boldsymbol{\beta}) = -\sum_{i=1}^{n}\left(y_i log\left(\frac{p_i}{1-p_i}\right) + log(1-p_i)\right) = -\sum_{i=1}^{n}\left[y_i\mathbf{x}_i^T\boldsymbol{\beta} - \ln\left(1+e^{\mathbf{x}_i^T\boldsymbol{\beta}}\right)\right]$$

要怎麼找(求解)羅吉斯回歸參數(**β**)?

- 利用偏微分求得此函數的梯度(Gradient)

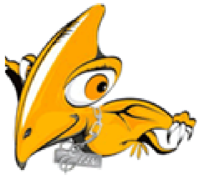
$$\frac{\partial \mathcal{L}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \frac{\partial \sum_{i=1}^{n}\left(log\left(1+e^{\boldsymbol{\beta}^T \boldsymbol{x}_i}\right) - y_i \boldsymbol{\beta}^T \boldsymbol{x}_i\right)}{\partial \boldsymbol{\beta}}$$

$$= \sum_{i=1}^{n}\left\{\frac{\partial\left(log\left(1+e^{\boldsymbol{\beta}^T \boldsymbol{x}_i}\right)\right)}{\partial \boldsymbol{\beta}} - \frac{\partial\left(y_i \boldsymbol{\beta}^T \boldsymbol{x}_i\right)}{\partial \boldsymbol{\beta}}\right\}$$

$$= \sum_{i=1}^{n}\left\{\frac{1}{\left(1+e^{\boldsymbol{\beta}^T \boldsymbol{x}_i}\right)} \times \frac{\partial\left(1+e^{\boldsymbol{\beta}^T \boldsymbol{x}_i}\right)}{\partial \boldsymbol{\beta}} - y_i \boldsymbol{x}_i\right\}$$

$$= \sum_{i=1}^{n}\left\{\frac{e^{\boldsymbol{\beta}^T \boldsymbol{x}_i}}{\left(1+e^{\boldsymbol{\beta}^T \boldsymbol{x}_i}\right)} \boldsymbol{x}_i - y_i \boldsymbol{x}_i\right\} = \sum_{i=1}^{n}\left\{p_i \boldsymbol{x}_i - y_i \boldsymbol{x}_i\right\}$$

$$= \sum_{i=1}^{n}\left\{p_i - y_i\right\} \boldsymbol{x}_i$$



要怎麼找(求解)羅吉斯回歸參數(**β**)?

- 梯度下降法

$$\boldsymbol{\beta}^{(t+1)} \leftarrow \boldsymbol{\beta}^{(t)} - \alpha \times \partial \frac{\mathcal{L}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \boldsymbol{\beta}^{(t)} + \alpha \sum_{i=1}^{n} (y_i - p_i)\mathbf{x}_i^T$$

α 為學習率

- 牛頓法（ Newton’s Method ）求羅吉斯迴歸參數

Recall

迴歸:

線性回歸: 最小平方法Ordinary Least Square Estimation (OLSE) (有closed-form solution)

分類:

線性區別分析: (有closed-form solution)

- minimum Euclidean classifier (平均數法: 歐式距離)

羅吉斯回歸: (沒有closed-form solution)→已經是神經網路的前身了。

- 梯度下降法、牛頓法找解