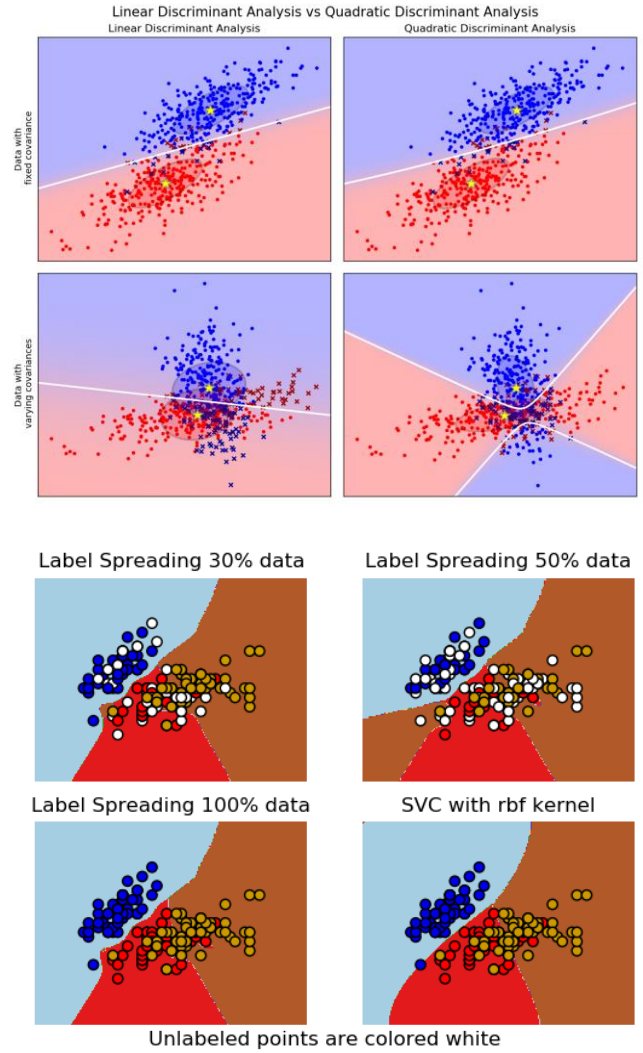
Classification

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

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)?”

Collected data (body fat(%))

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.

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

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:

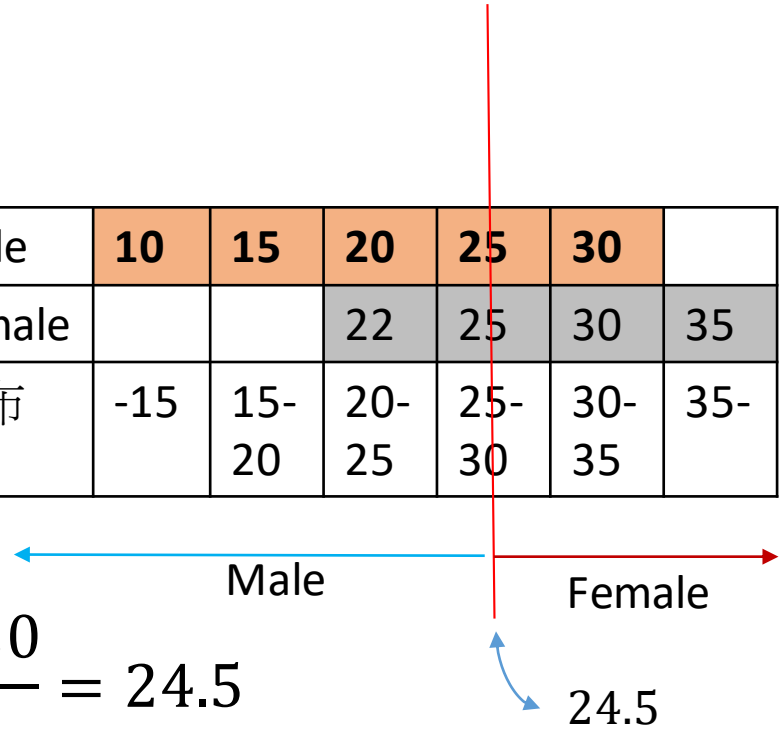
Using mean value as decision rule.

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

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-



Classification

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

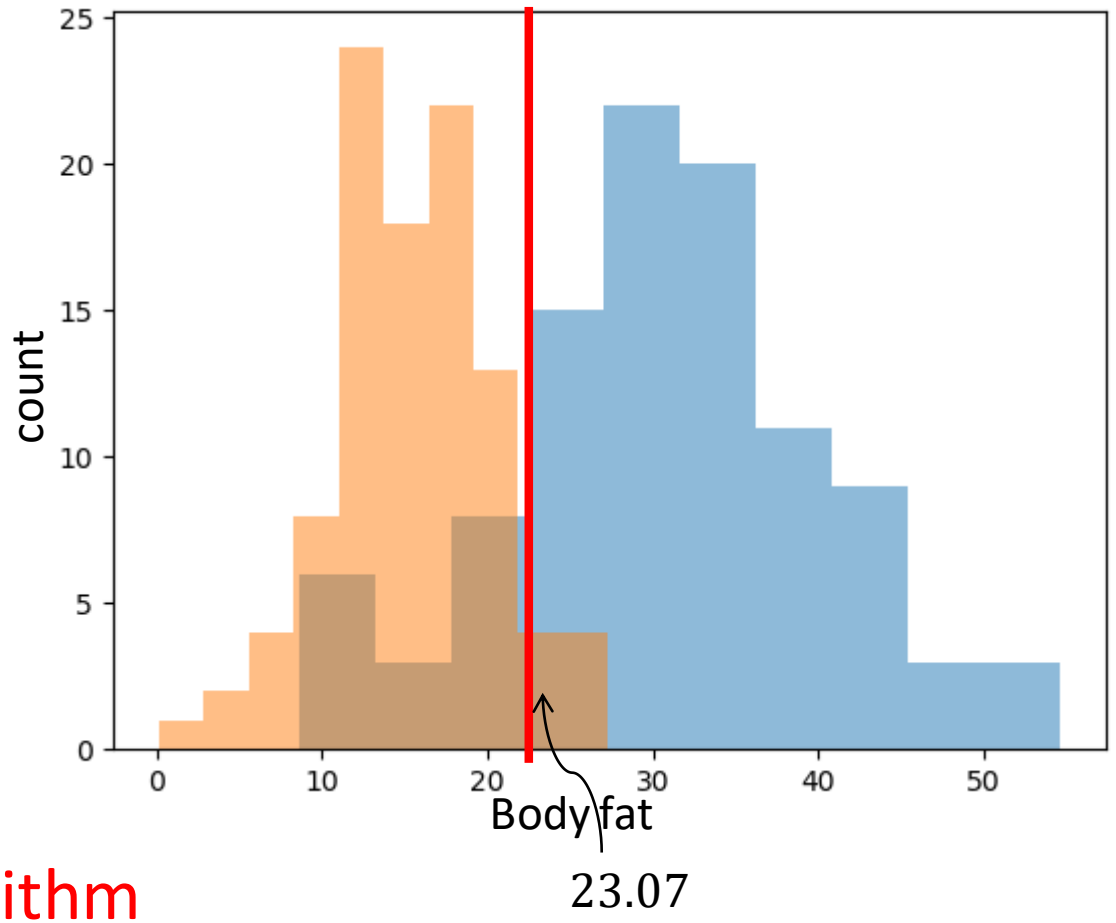
Visualization by histogram.

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)

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

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

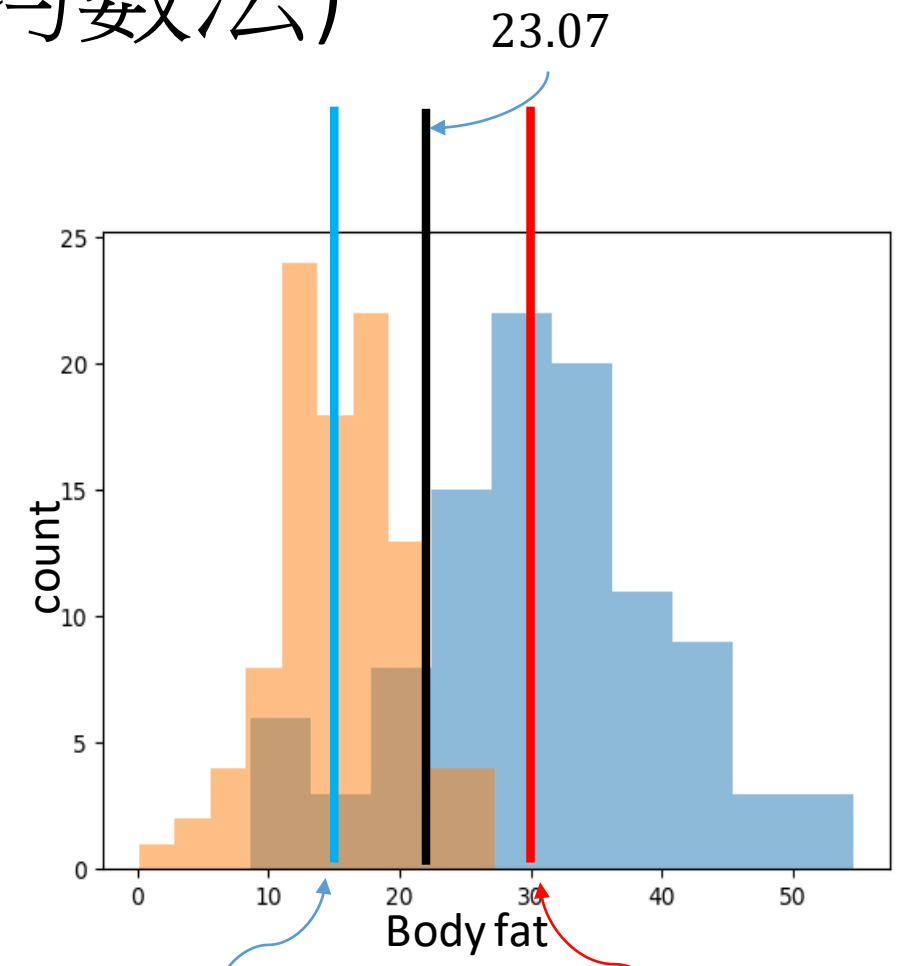
$$f_{male}(x) = x - \mu_{male}$$

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

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)

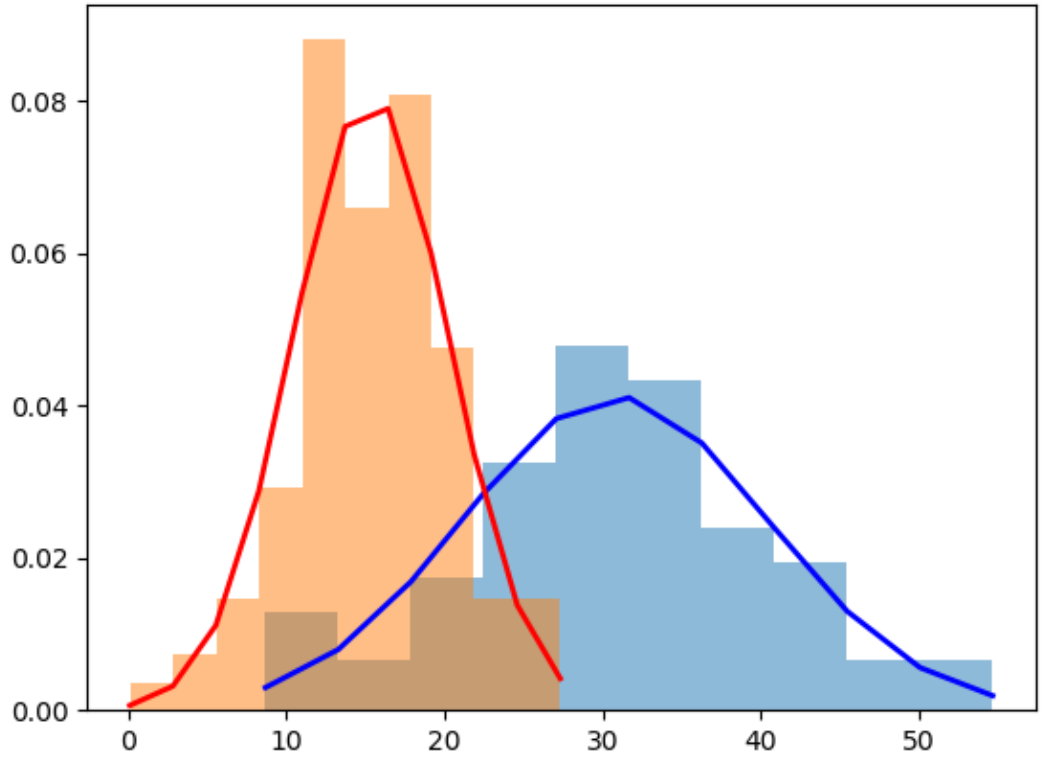
We can assume the histogram (density) is a **Gaussian** (normal)-like distribution.

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)

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

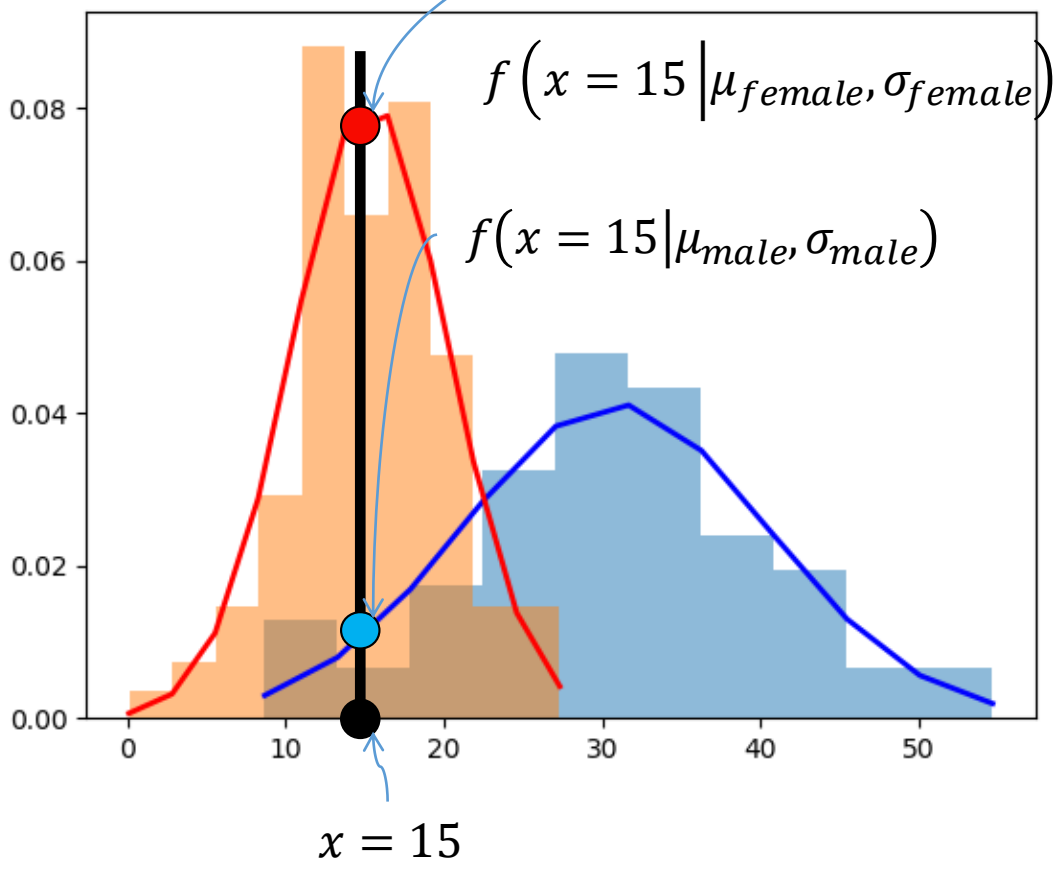
$$f(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

A unlabeled x: 15% body fat

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

So this unlabeled x would be classify to Female.

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



Classification (Multi-variables)(平均數法)

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

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

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

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

向量

We can't make decision with an array.

純量 純量

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

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): $\|\boldsymbol{x} - \boldsymbol{\mu}\|_{\mathrm{L2}} = \underset{1\times 2}{(\boldsymbol{x} - \boldsymbol{\mu})^{\mathrm{T}}} \underset{2\times 1}{(\boldsymbol{x} - \boldsymbol{\mu})}$
- 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

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

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

Likelihood function(Multi-variables)

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

純量

$|\Sigma|$: 共變異數的行列式值 → 純量

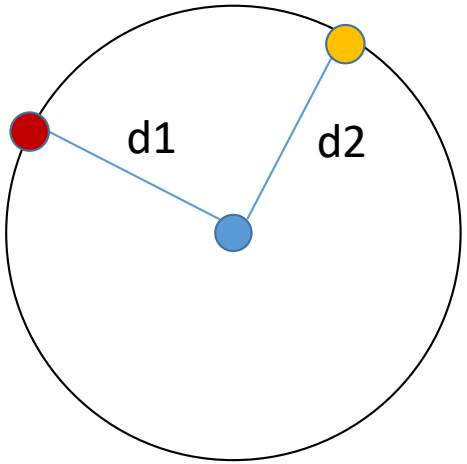
$$\underset{1\times d}{(\boldsymbol{x}-\boldsymbol{\mu})^T}\underset{d\times d}{\Sigma^{-1}}\underset{d\times 1}{(\boldsymbol{x}-\boldsymbol{\mu})}$$

輸出為 1×1→純量

Distance

Euclidean Distance

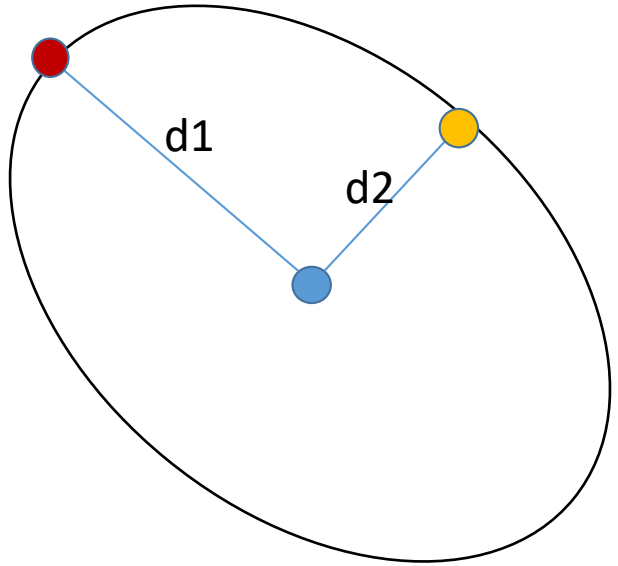
$$(\boldsymbol{x}-\boldsymbol{\mu})^T(\boldsymbol{x}-\boldsymbol{\mu})$$



d1 = d2

Mahalanobis Distance

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



d1 = d2

Maximum a posterior (MAP)

- In statistics, we hope to make decision by getting a maximum posterior probability for a given ***x***.

  $p(c|\boldsymbol{x})$: posterior probability of data ***x*** for class *c*.

- Likelihood : $f(\boldsymbol{x}|\mu,\sigma)$. Given on a parameter set, the function output for the data ***x***.

Maximum a posterior (MAP)

- $p(c|x)$: posterior probability of data $\boldsymbol{x}$ for class c.

$$p(c|\boldsymbol{x}) = \frac{p(c)p(\boldsymbol{x}|c)}{p(\boldsymbol{x})}$$

$p(c)$: prior probability for class c

$p(\boldsymbol{x}|c)$: likelihood function for class c

$p(\boldsymbol{x}) = \sum_{c=1}^{L} p(c)\, p(\boldsymbol{x}|c)$: normalizing constant

Maximum a posterior (MAP)

How to make decision?

By checking the posterior probability for all class.

$$\text{Decision}(\boldsymbol{x}) = \underset{c=\{1,2,\ldots,C\}}{arg\max} \{p(c|\boldsymbol{x})\}$$

$$\underset{c=\{1,2,\ldots,C\}}{arg\max} \{p(c)p(\boldsymbol{x}|c)\}$$

Prior probability

$p(c)$, in Chinese = 先驗機率 (在還沒有建模前，得到的先天訊息)

Male : Female = 25: 75

Sample a people.

Female
75%

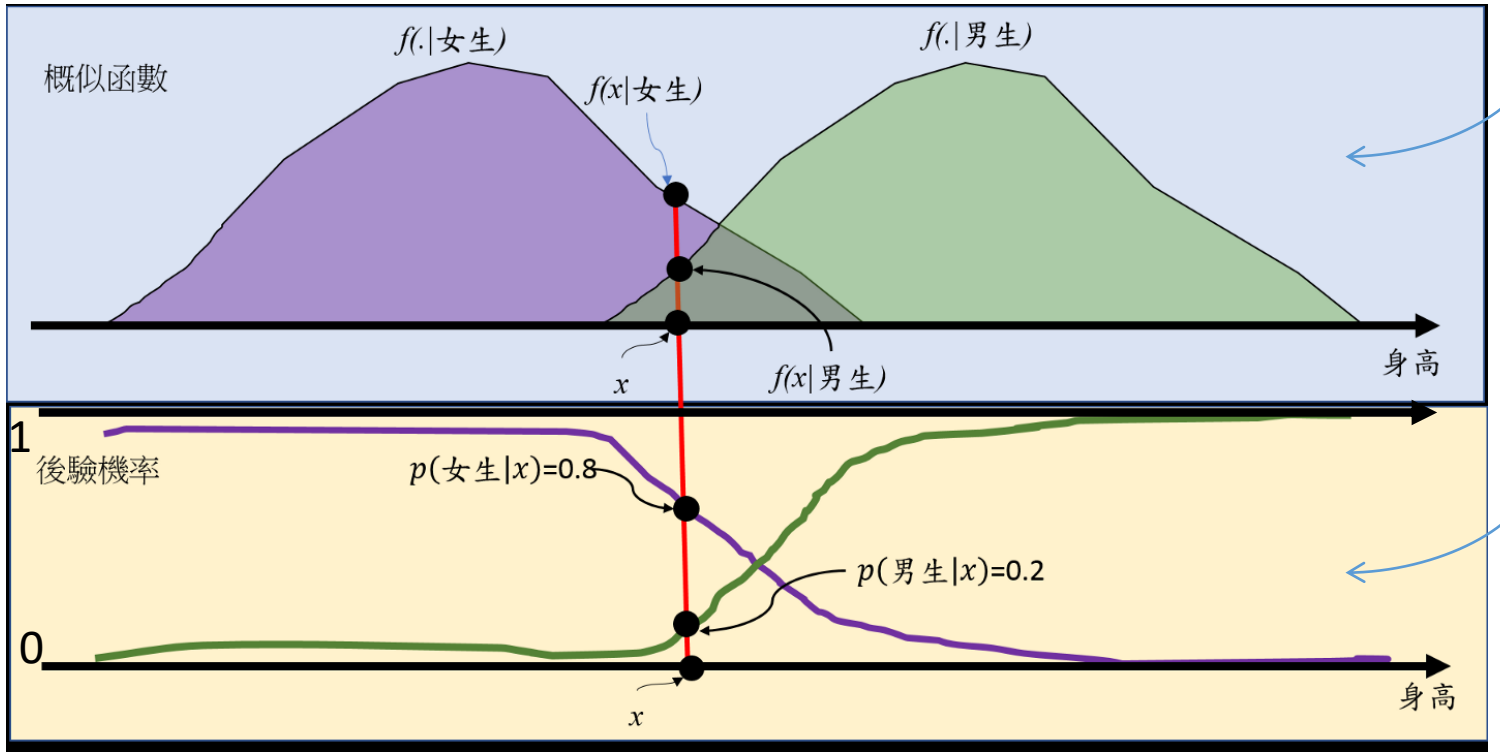
Gender?

Male
25%

Maximum a posterior (MAP)

Posterior probability: $p(c|\boldsymbol{x}) = \frac{p(c)p(\boldsymbol{x}|c)}{p(\boldsymbol{x})}$

Likelihood : $f(\boldsymbol{x}|\mu,\sigma)$



Likelihood function

Likelihood function是積分面積總和為1。

Posterior probability

Posterior probability是單點在命題的總類別數機率合為1。

Maximum a posterior (MAP)

Decision rule:

$$c_{MAP} = \underset{c=\{1,2,\ldots,L\}}{\arg\max}\{p(c|x)\}$$

$$= \underset{c=\{1,2,\ldots,L\}}{\arg\max}\left\{\frac{p(c)\,f(x|c)}{p(x)}\right\} = \underset{c=\{1,2,\ldots,L\}}{\arg\max}\left\{p(c)\,f(x|c)\right\}$$

MAP with Gaussian function

Gaussian function:

$$f(x|\mu_c, \Sigma_c) = (2\pi)^{-d/2} |\Sigma_c|^{-0.5} \exp\left\{-0.5(x-\mu_c)^T \Sigma^{-1}(x-\mu_c)\right\}$$

MAP:

$$c_{MAP} = \underset{c=\{1,2,...,L\}}{\arg\max}\{p(c|x)\} = \underset{c=\{1,2,...,L\}}{\arg\max}\left\{p(c)\, f(x|c)\right\}$$

MAP with Gaussian function
線性區別分析(linear discriminant classifier, LDC)

Gaussian function + MAP:

$$c_{MAP}^{GC} = \arg\max_{c=\{1,2,\ldots,L\}}\{p(c|x)\} = \arg\max_{c=\{1,2,\ldots,L\}}\left\{p(c)\, f(x|c)\right\}$$

$$= \arg\min_{c=\{1,2,\ldots,L\}}\{-2\ln(p(c)) + \ln(|\Sigma_c|) - 0.5(x-\mu_c)^T \Sigma^{-1}(x-\mu_c)\}$$

MAP with Gaussian function

$$c_{MAP}^{GC} = \underset{c=\{1,2,\ldots,L\}}{\arg\min}\{-2\ln(p(c)) + \ln(|\Sigma_c|) - 0.5(x-\mu_c)^T \Sigma^{-1}(x-\mu_c)\}$$

The most important term of this formula is measuring the distance between x and center of distribution

Quadratic discriminant classifier

$$\text{QDC: } w_{MAP} = \underset{i=\{1,2,\ldots,L\}}{\arg\max}\{lnp(w_i) - 0.5ln|\boldsymbol{\Sigma}_i| - 0.5(\boldsymbol{x}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)\}$$

linear discriminant classifier

$$\text{LDC: } w_{MAP} = \underset{i=\{1,2,\ldots,L\}}{\arg\max}\{lnp(w_i) - 0.5(\boldsymbol{x}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)\}$$

minimum Euclidean classifier

$$\text{MEC: } w_{MAP} = \underset{i=\{1,2,\ldots,L\}}{\arg\max}\{lnp(w_i) - 0.5(\boldsymbol{x}-\boldsymbol{\mu}_i)^T I_d(\boldsymbol{x}-\boldsymbol{\mu}_i)\}$$

I_d: Identity matrix $(d \times d)$

範例: 假設男生和女生的先驗機率都是0.5

- 收集的訓練資料中，估計出男生和女生的身高與體重的平均數向量，身高和體重的共變異數矩陣

$$\mathbf{x} \sim N\left(\boldsymbol{\mu}_{男生} = \begin{bmatrix} 175 \\ 80 \end{bmatrix}, \boldsymbol{\Sigma}_{男生} = \begin{bmatrix} 10^2 & 30 \\ 30 & 5^2 \end{bmatrix}\right)$$

$$\mathbf{x} \sim N\left(\boldsymbol{\mu}_{女生} = \begin{bmatrix} 160 \\ 50 \end{bmatrix}, \boldsymbol{\Sigma}_{女生} = \begin{bmatrix} 8^2 & 50 \\ 50 & 9^2 \end{bmatrix}\right)$$

範例: 假設男生和女生的先驗機率都是0.5

- LDA 男生的隸屬函數為：

$$\mathbf{x} \sim N\left(\boldsymbol{\mu}_{\text{男生}} = \begin{bmatrix} 175 \\ 80 \end{bmatrix}, \boldsymbol{\Sigma}_{\text{男生}} = \begin{bmatrix} 10^2 & 30 \\ 30 & 5^2 \end{bmatrix}\right)$$

$$-0.5\ln\left(\left|\boldsymbol{\Sigma}_{C_{\text{男生}}}\right|\right) - \frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}_{C_{\text{男生}}}\right)^T \boldsymbol{\Sigma}_C^{-1} \left(\mathbf{x}-\boldsymbol{\mu}_{C_{\text{男生}}}\right)$$

$$= -0.5\ln(1600) - \frac{1}{2}\left(\mathbf{x} - \begin{bmatrix} 175 \\ 80 \end{bmatrix}\right)^T \begin{bmatrix} 10^2 & 30 \\ 30 & 5^2 \end{bmatrix}^{-1} \left(\mathbf{x} - \begin{bmatrix} 175 \\ 80 \end{bmatrix}\right)$$

$$\mathbf{x} \sim N\left(\boldsymbol{\mu}_{\text{女生}} = \begin{bmatrix} 160 \\ 50 \end{bmatrix}, \boldsymbol{\Sigma}_{\text{女生}} = \begin{bmatrix} 8^2 & 50 \\ 50 & 9^2 \end{bmatrix}\right)$$

$$-0.5\ln\left(\left|\boldsymbol{\Sigma}_{C_{\text{女生}}}\right|\right) - \frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}_{C_{\text{女生}}}\right)^T \boldsymbol{\Sigma}_C^{-1} \left(\mathbf{x}-\boldsymbol{\mu}_{C_{\text{女生}}}\right)$$

$$= -0.5\ln(2684) - \frac{1}{2}\left(\mathbf{x} - \begin{bmatrix} 160 \\ 50 \end{bmatrix}\right)^T \begin{bmatrix} 8^2 & 50 \\ 50 & 9^2 \end{bmatrix}^{-1} \left(\mathbf{x} - \begin{bmatrix} 160 \\ 50 \end{bmatrix}\right)$$

範例: 假設男生和女生的先驗機率都是0.5

$$\mathbf{x}^* = \begin{bmatrix} 身高 \\ 體重 \end{bmatrix} = \begin{bmatrix} 170 \\ 75 \end{bmatrix}$$

男生的隸屬函數值為

$$-0.5\ln(1600) - \frac{1}{2}\left(\begin{bmatrix} 170 \\ 75 \end{bmatrix} - \begin{bmatrix} 175 \\ 80 \end{bmatrix}\right)^T \begin{bmatrix} 10^2 & 30 \\ 30 & 5^2 \end{bmatrix}^{-1} \left(\begin{bmatrix} 170 \\ 75 \end{bmatrix} - \begin{bmatrix} 175 \\ 80 \end{bmatrix}\right)$$

$$= -0.5\ln(1600) - \frac{1}{2}\frac{1}{1600}\begin{bmatrix} -5 \\ -5 \end{bmatrix}^T \begin{bmatrix} 25 & -30 \\ -30 & 100 \end{bmatrix}\begin{bmatrix} -5 \\ -5 \end{bmatrix} \approx -4.1966915$$

女生的隸屬函數值為

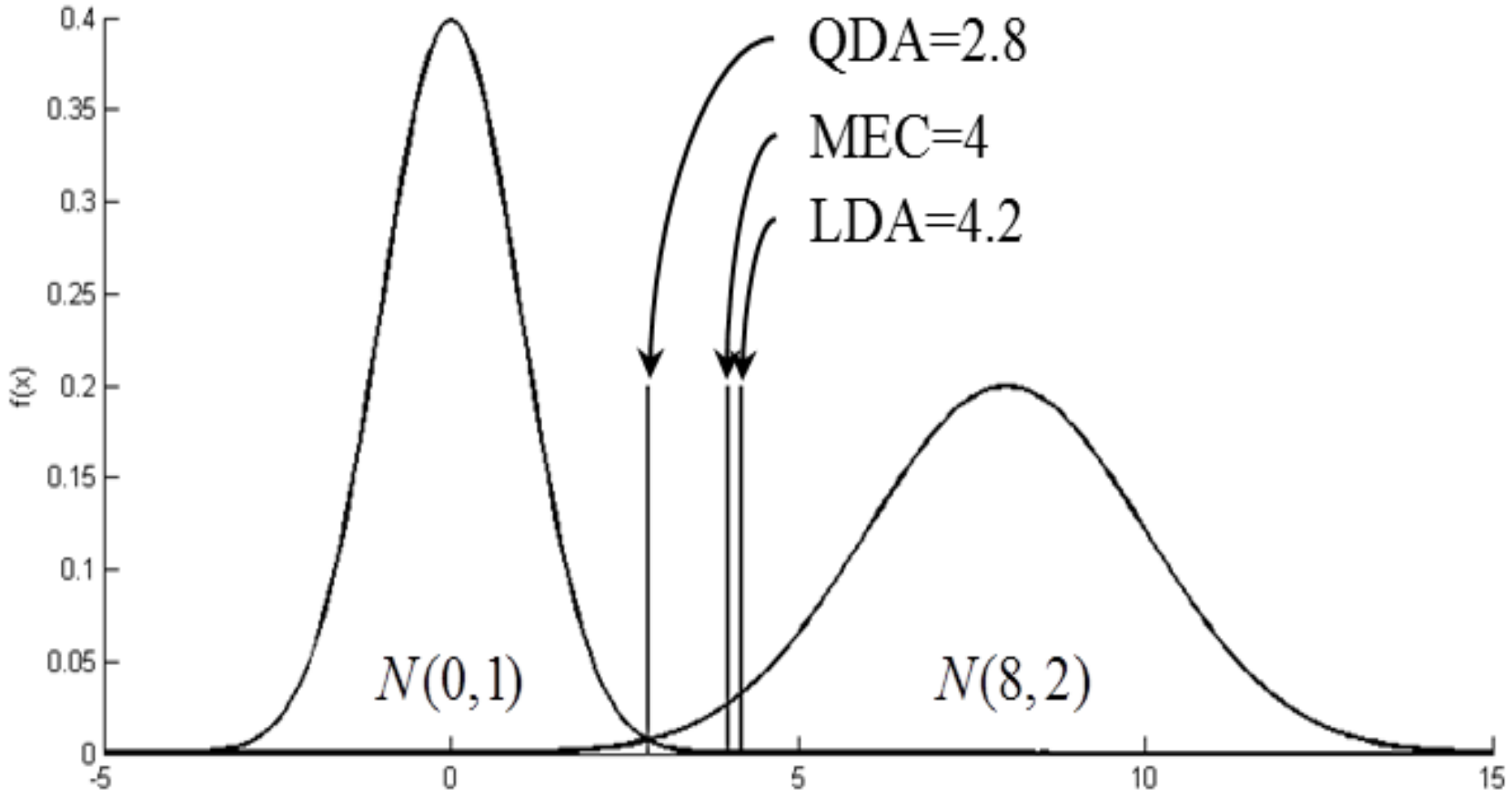
$$-0.5\ln(2684) - \frac{1}{2}\left(\begin{bmatrix} 170 \\ 75 \end{bmatrix} - \begin{bmatrix} 160 \\ 50 \end{bmatrix}\right)^T \begin{bmatrix} 8^2 & 50 \\ 50 & 9^2 \end{bmatrix}^{-1} \left(\begin{bmatrix} 170 \\ 75 \end{bmatrix} - \begin{bmatrix} 160 \\ 50 \end{bmatrix}\right)$$

$$= -0.5\ln(2684) - \frac{1}{2}\frac{1}{2684}\begin{bmatrix} 10 \\ 25 \end{bmatrix}^T \begin{bmatrix} 81 & -50 \\ -50 & 64 \end{bmatrix}\begin{bmatrix} 10 \\ 25 \end{bmatrix} \approx -8.250811$$

$$C_{QDA}\left(\boldsymbol{x}^* = \begin{bmatrix} 170 \\ 75 \end{bmatrix}\right) = \text{argmax}\left\{C_{男生} = -4.1966915,\ C_{女生} = -8.250811\right\}$$

$$= C_{男生}$$

範例：一維度的線性區別分析

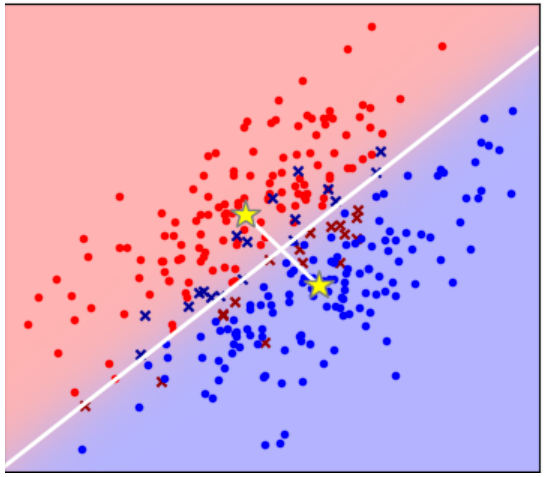


範例: 二維度的線性區別分析

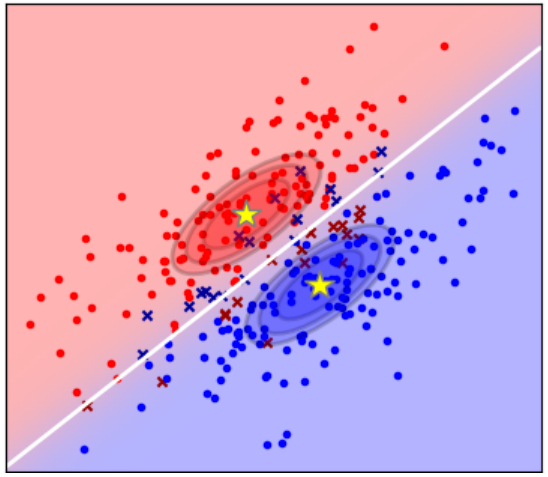
類別的共變異數矩陣一致，但平均數不同

$$\text{類別 1}: N\left(\boldsymbol{\mu}_{C_1}=\begin{bmatrix}0\\1\end{bmatrix}, \boldsymbol{\Sigma}=\begin{bmatrix}2 & 1\\1 & 1\end{bmatrix}\right) \qquad \text{類別 2}: N\left(\boldsymbol{\mu}_{C_2}=\begin{bmatrix}1\\0\end{bmatrix}, \boldsymbol{\Sigma}=\begin{bmatrix}2 & 1\\1 & 1\end{bmatrix}\right)$$

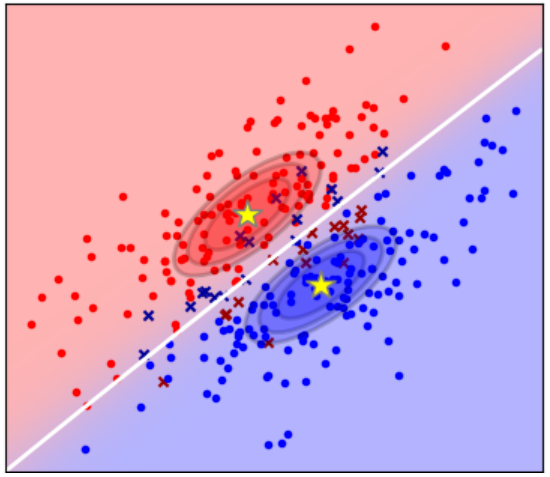
最小歐式距離分類器
minimum Euclidean classifier



線性區別分析
linear discriminant analysis



二次式區別分析
quadratic discriminant analysis



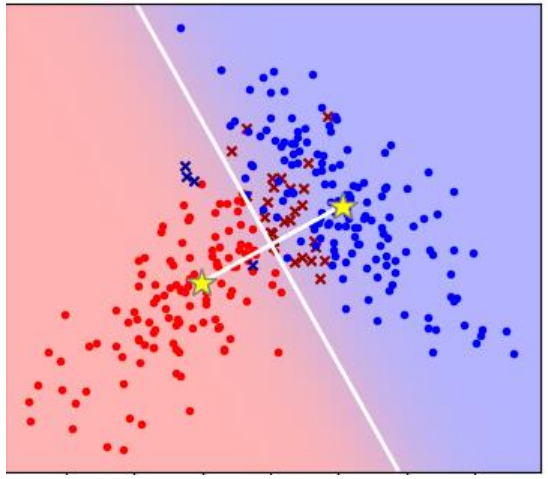
範例 、二維度的線性區別分析

每個類別的共變異數矩陣和平均數皆不同

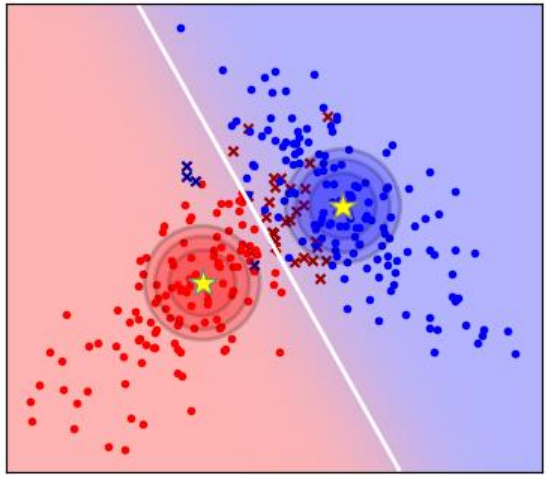
$$\text{類別 } 1：N\left(\boldsymbol{\mu}_{C_1}=\begin{bmatrix}0\\1\end{bmatrix},\boldsymbol{\Sigma}_1=\begin{bmatrix}1 & 1\\1 & 2\end{bmatrix}\right)$$

$$\text{類別 } 2：N\left(\boldsymbol{\mu}_{C_2}=\begin{bmatrix}2\\2\end{bmatrix},\boldsymbol{\Sigma}=\begin{bmatrix}1 & -1\\-1 & 2\end{bmatrix}\right)$$

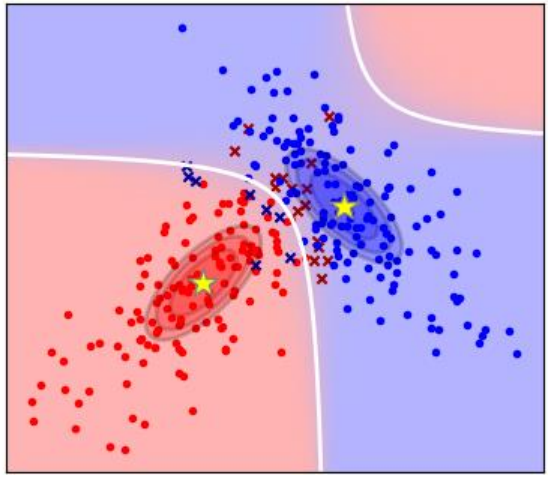
最小歐式距離分類器
minimum Euclidean classifier



線性區別分析
linear discriminant analysis

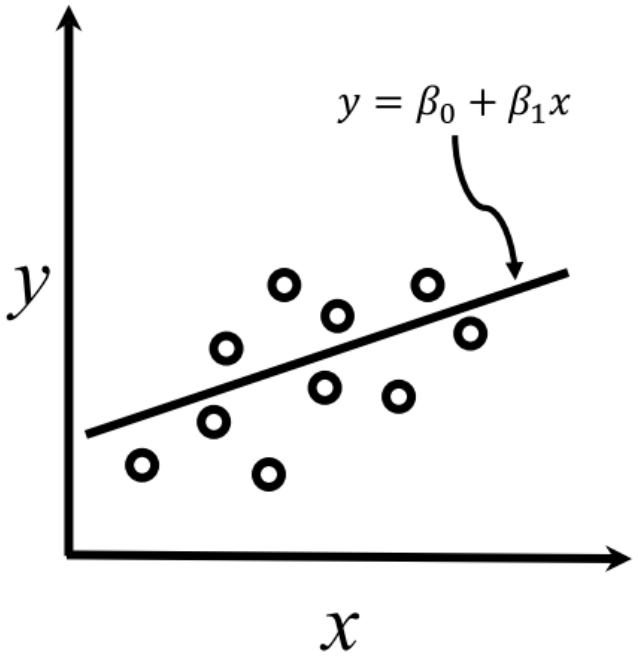


二次式區別分析
quadratic discriminant analysis

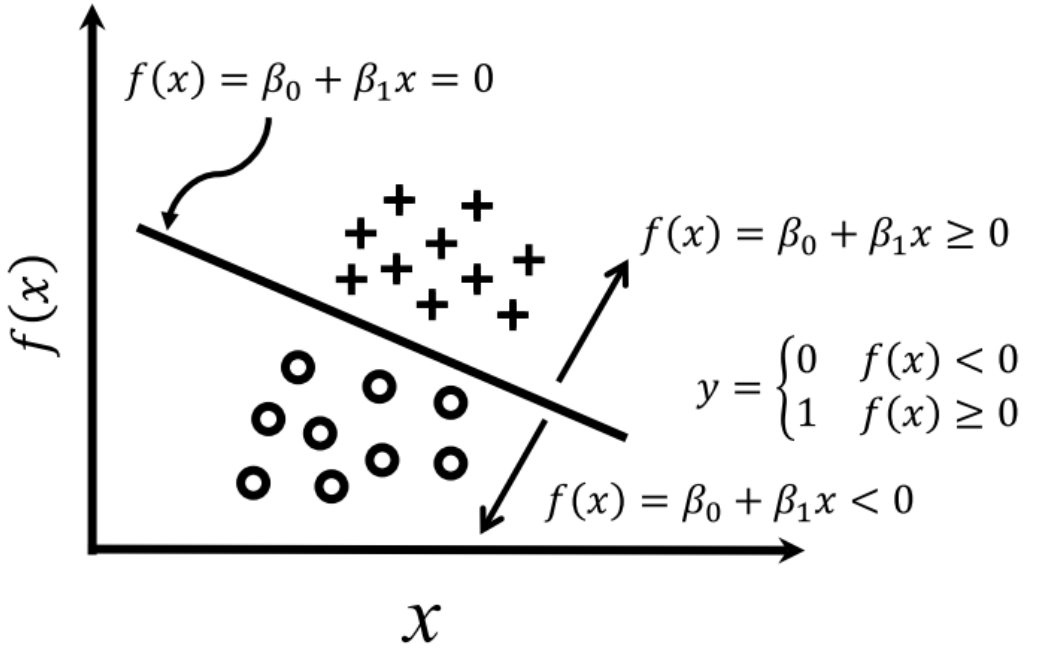


羅吉斯迴歸（Logistic Regression）

迴歸

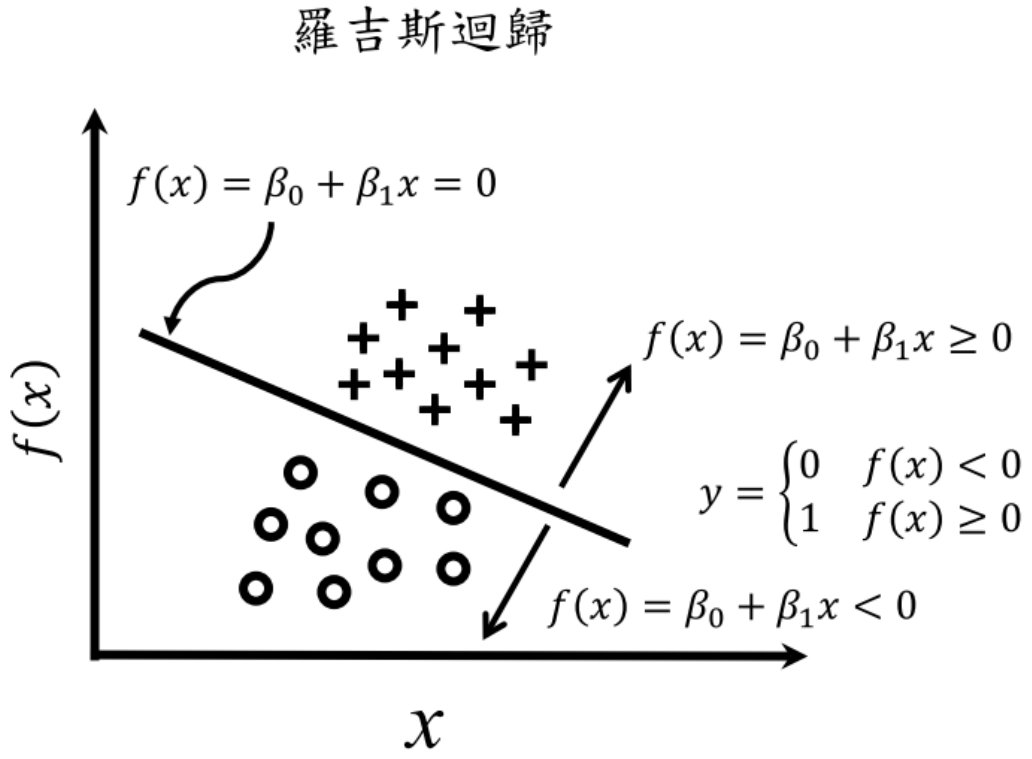


羅吉斯迴歸



- 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)

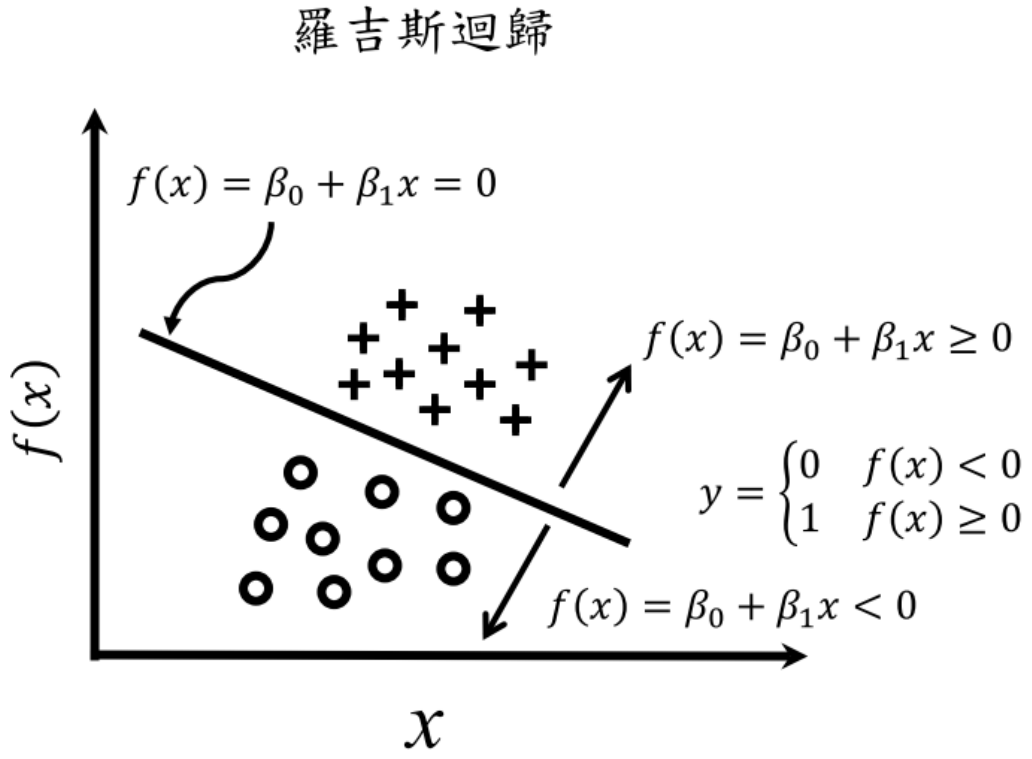
羅吉斯迴歸（Logistic Regression）



- 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)

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

羅吉斯迴歸（Logistic Regression）



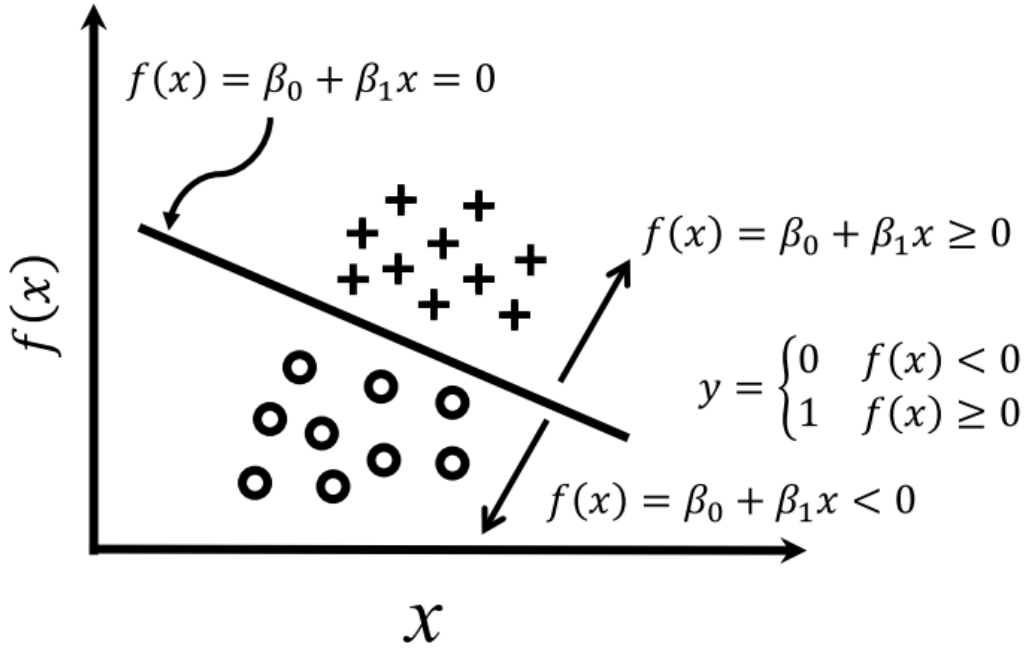
- σ(.)在機器學習上稱為單位階梯函數(unit step function)，大於一個閾值(threshold)是一類，反之為另一類，此例的閾值為0。

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

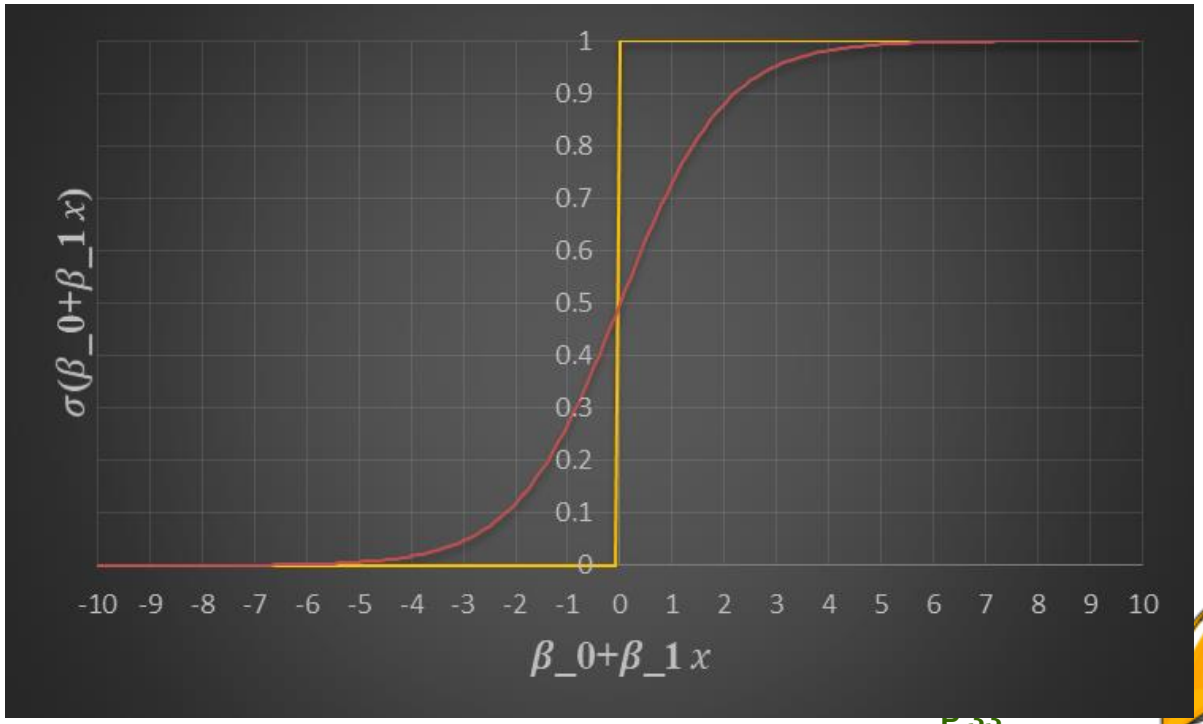
羅吉斯迴歸用Sigmoid函數限制值域

羅吉斯迴歸

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



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



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

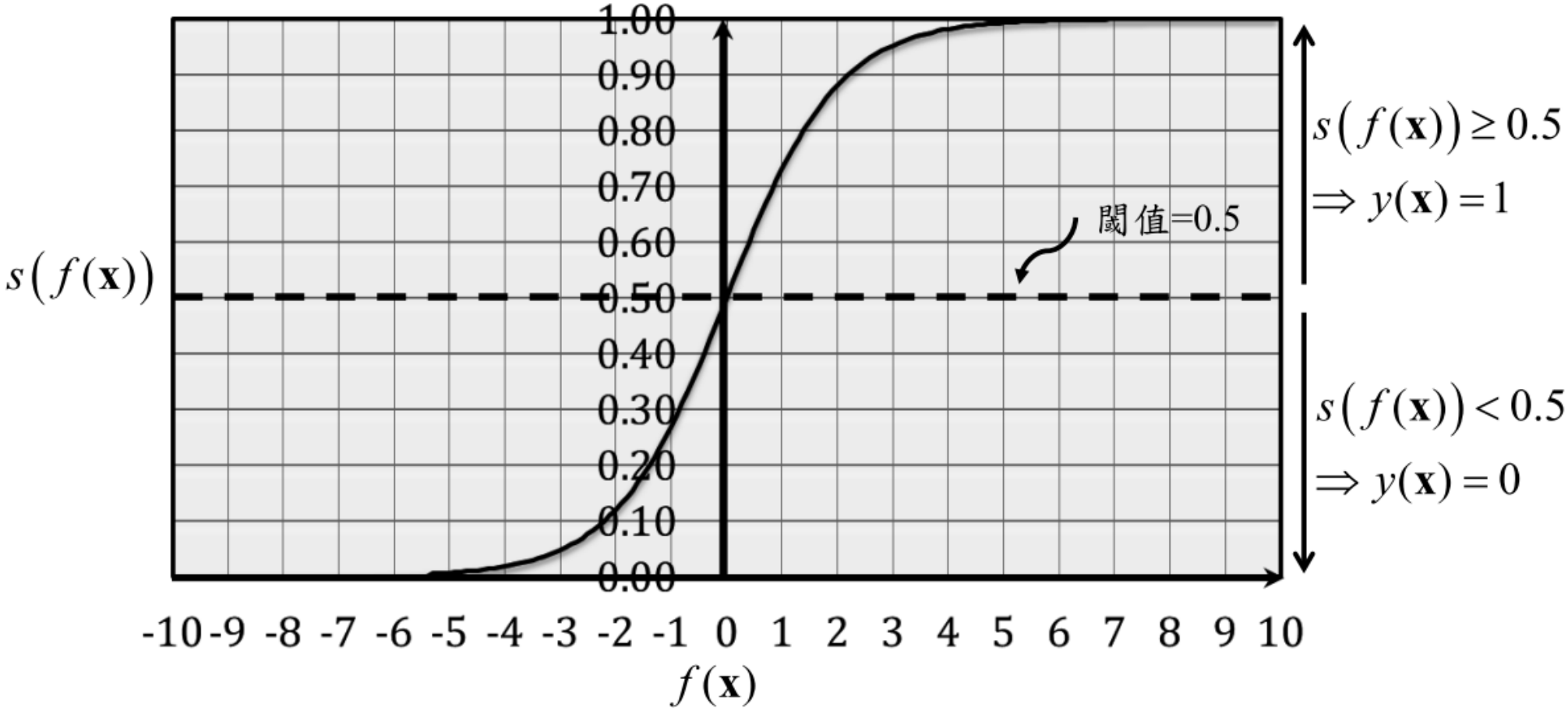
$$s\left(f(\mathbf{x})\right)=\frac{e^{f(\mathbf{x})}}{1+e^{f(\mathbf{x})}}=\frac{e^{\mathbf{x}^T\boldsymbol{\beta}}}{1+e^{\mathbf{x}^T\boldsymbol{\beta}}}$$

羅吉斯迴歸是二分類演算法

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

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

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

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

概似函數最大化→不好做

我們把問題轉成-log，找最小化的問題。

$$\mathcal{L}(\boldsymbol{\beta}) = -logL(\boldsymbol{\beta}) = -log\left(\prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}\right)$$

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

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

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

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

$$= -\sum_{i=1}^{n} \left(y_i log\left(\frac{p_i}{1-p_i}\right) + log(1-p_i)\right)$$

p_i是什麼

就是交叉熵的公式

要怎麼找(求解)羅吉斯回歸參數(β)?

p_i: 羅吉斯回歸的輸出

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

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

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

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

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

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

α 為學習率

- 牛頓法（Newton's Method）求羅吉斯迴歸參數

Recall

線性區別分析: (有closed-form solution)

- Quadratic discriminant classifier
  (距離: 馬氏距離，不同類別用各自的共變異數矩陣)
- linear discriminant classifier
  (距離: 馬氏距離，不同類別用共同的共變異數矩陣)
- minimum Euclidean classifier
  (平均數法: 歐式距離)

羅吉斯回歸: (沒有closed-form solution)

- 梯度下降法找解
- 牛頓法找解

Classification

We just learned model-based algorithm.

Model-based: data is assumed as normal distribution. (parameters: mean vector and covariance matrix)

Can we learn without model?

ANS:
Yes.
Nearest neighbors, SVM, neural network.