分類技術

- 1 分類的概念
- 2 *K*最鄰近法
- 3 決策樹
- 4 分類模型的評估
- 5 結論

學習方向

- 瞭解分類技術的概念與分析流程
- 瞭解*K*最鄰近法的學習演算法與特性
- 瞭解決策樹的學習演算法與特性
- 瞭解分類模型的評估指標

分類技術

- 1 分類的概念
- 2 *K*最鄰近法
- 3 決策樹
- 4 分類模型的評估
- 5 結論

分類是什麼?

分類是什麼?

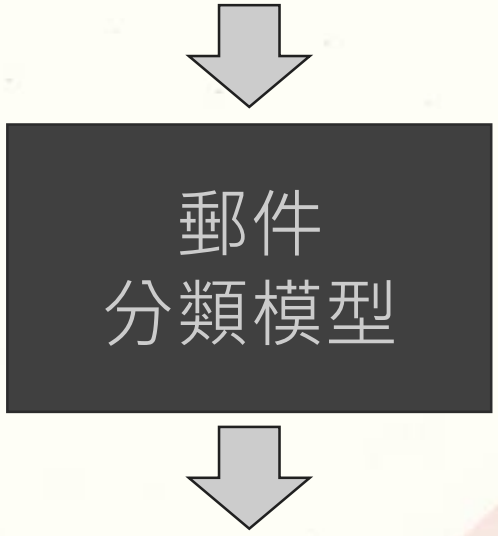
分類是什麼？ $y = f(x_1, x_2, \ldots, x_n)$

- 將一個物件指定預先定義的分類
- 貼標籤
  - 一般垃圾、資源回收
  - 垃圾郵件、一般信件、重要信件
  - 家人、摯友、同事

分類的定義 $y = f(x_1, x_2, \ldots, x_n)$

- 建立一個學習函數(分類模型)將每個屬性集$(x_1, x_2, \ldots, x_n)$對應到一個已定義的類別y

(標題x_1，內容x_2，寄件者x_3，寄件時間x_4)



郵件類別(y)

分類模型的功用

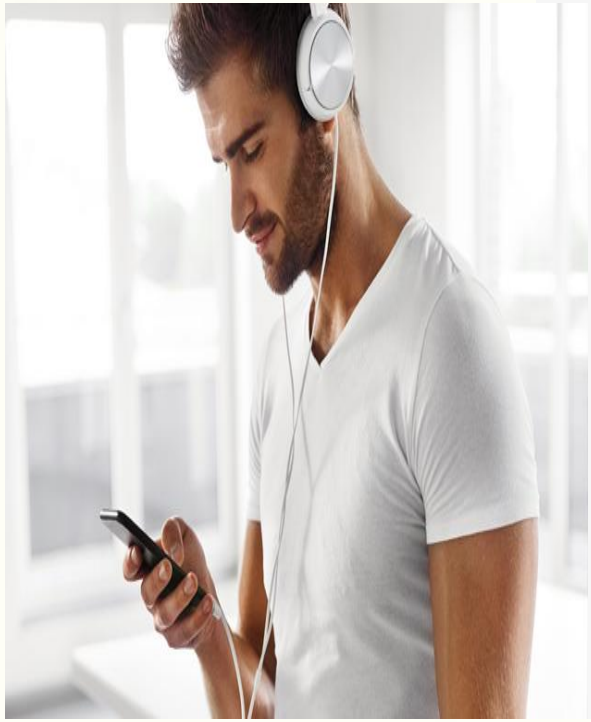
- 描述已知類別的特徵，用於解釋與區分不同類別
- 預測未知類別的物件，用於判斷物件所屬類別

如何建立分類模型？

- 請分辨此人是否為「奧客」？

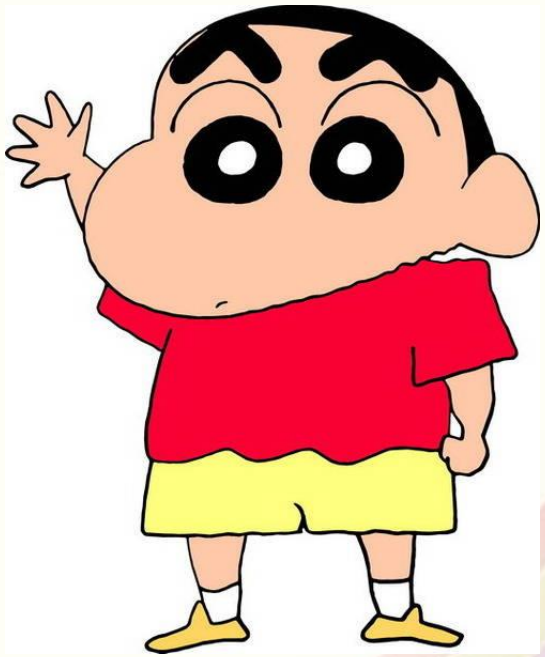
如何建立分類模型？

- 先觀察已經確定類別的顧客吧！
  - 好顧客



如何建立分類模型？

- 先觀察已經確定類別的顧客吧！
  - 奧客



如何建立分類模型？

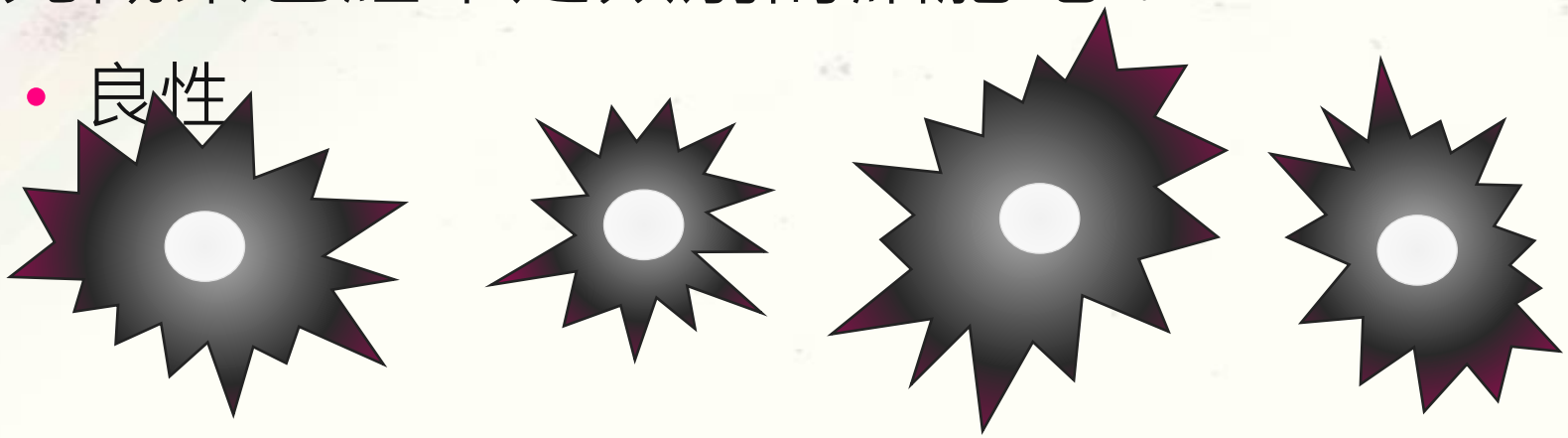
- 再判斷一次此人是否為「奧客」？

如何建立分類模型？(1)

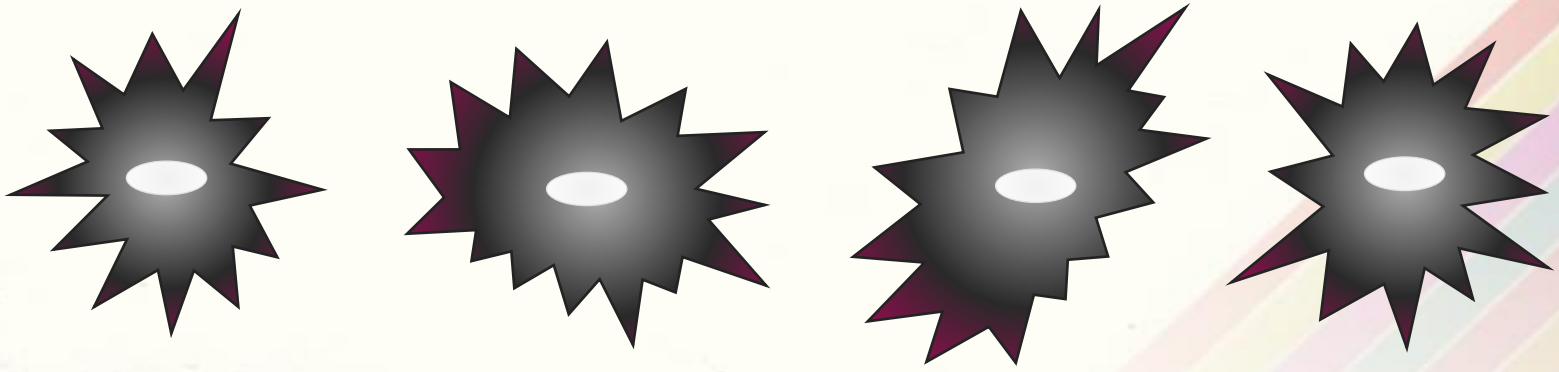
- 請分辨下列腫瘤細胞為良性還是惡性？

如何建立分類模型？(2)

- 先觀察已經確定類別的細胞吧！
  - 良性



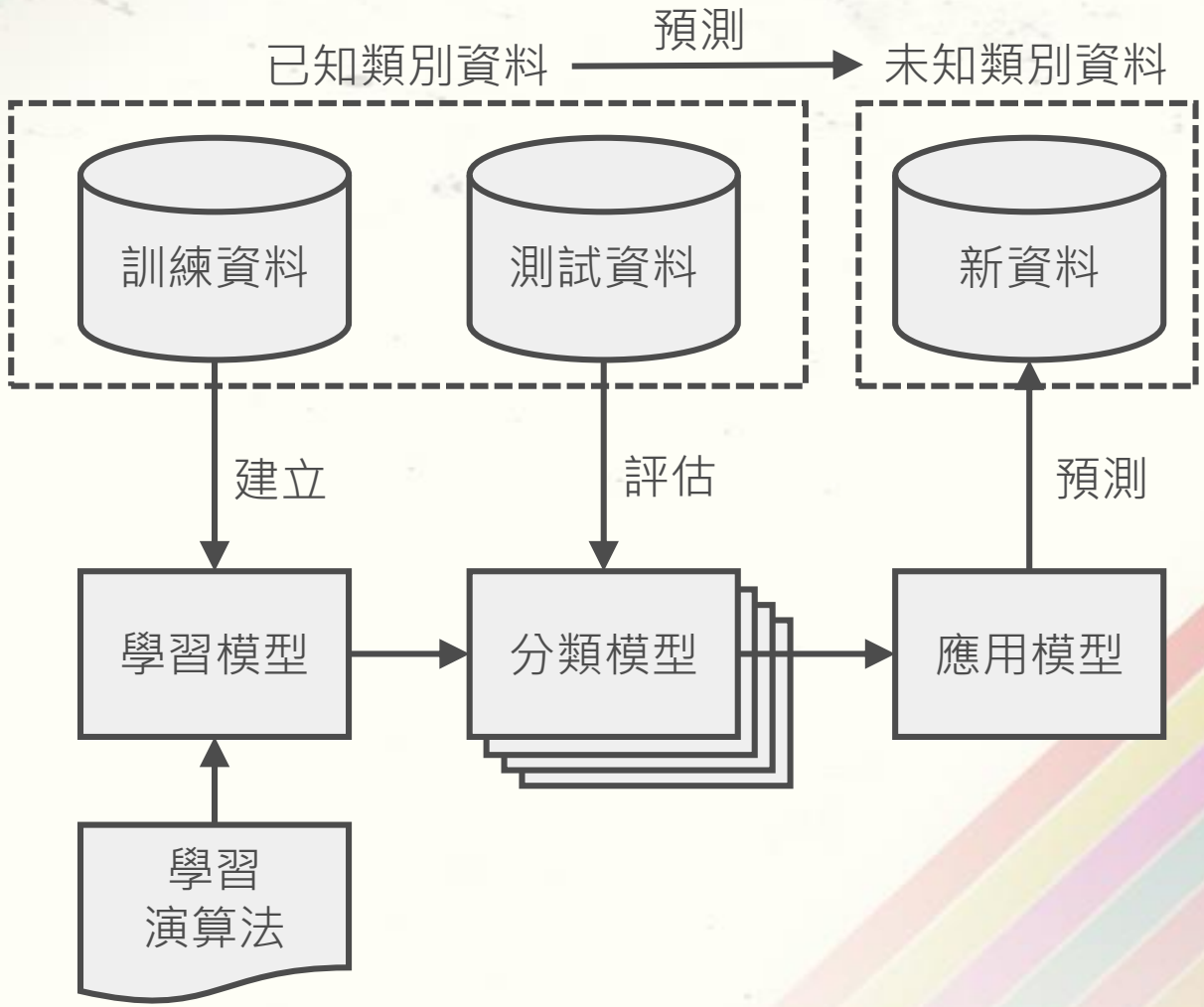
  - 惡性



如何建立分類模型？(3)

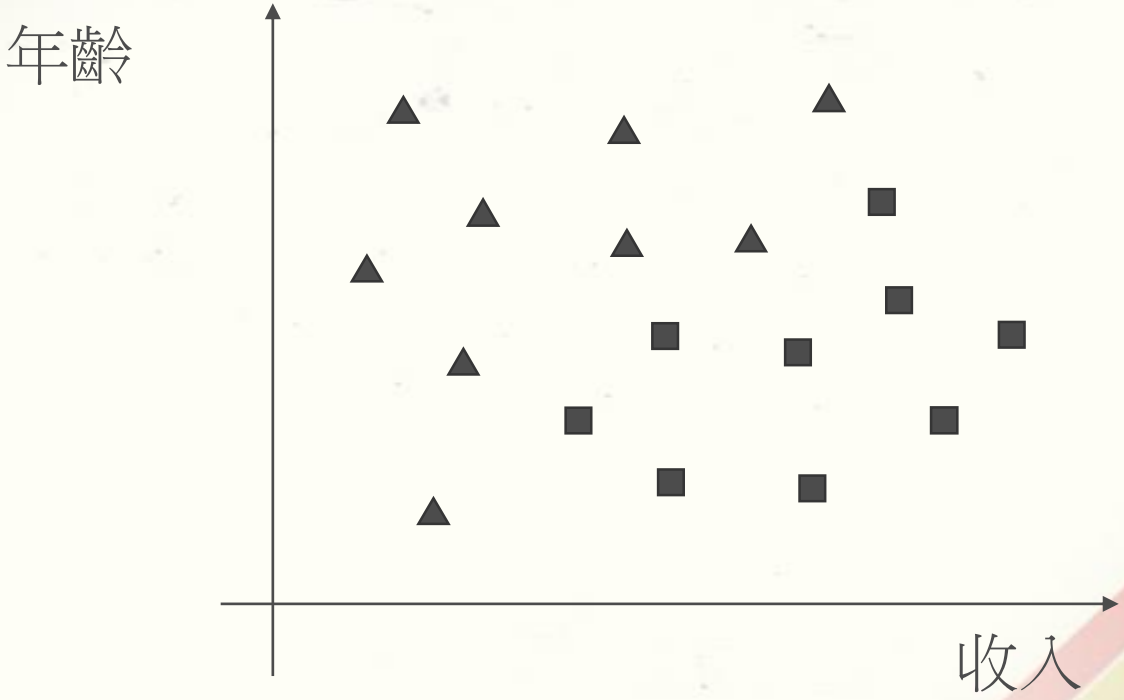
- 再分辨一次腫瘤細胞為良性還是惡性？

分類模型分析流程(1)



分類模型分析流程(2)

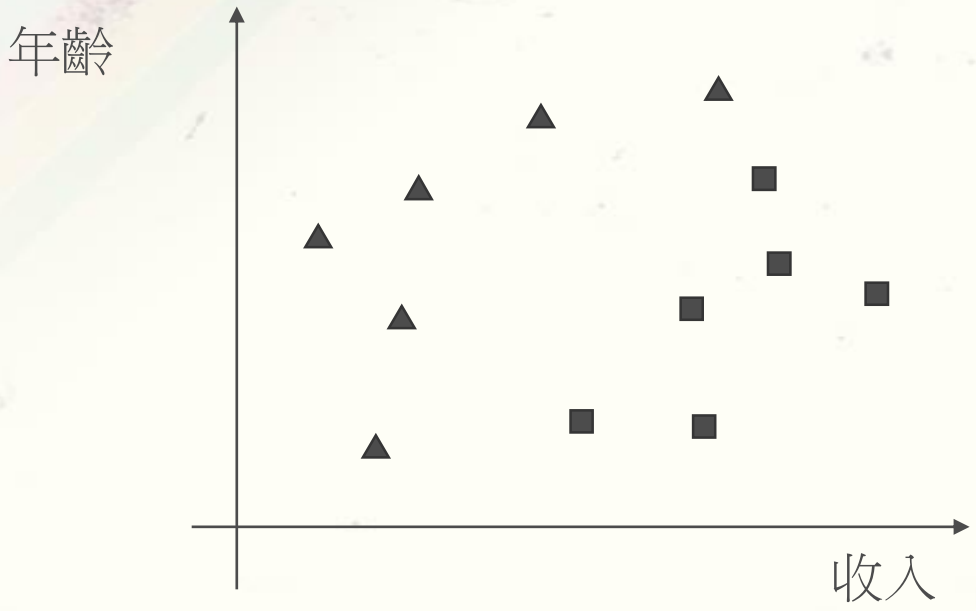
利用「收入」和「年齡」建立公司新客戶是否會加購加值服務的分類預測模型，蒐集現有18 筆資料繪製一個XY散佈圖



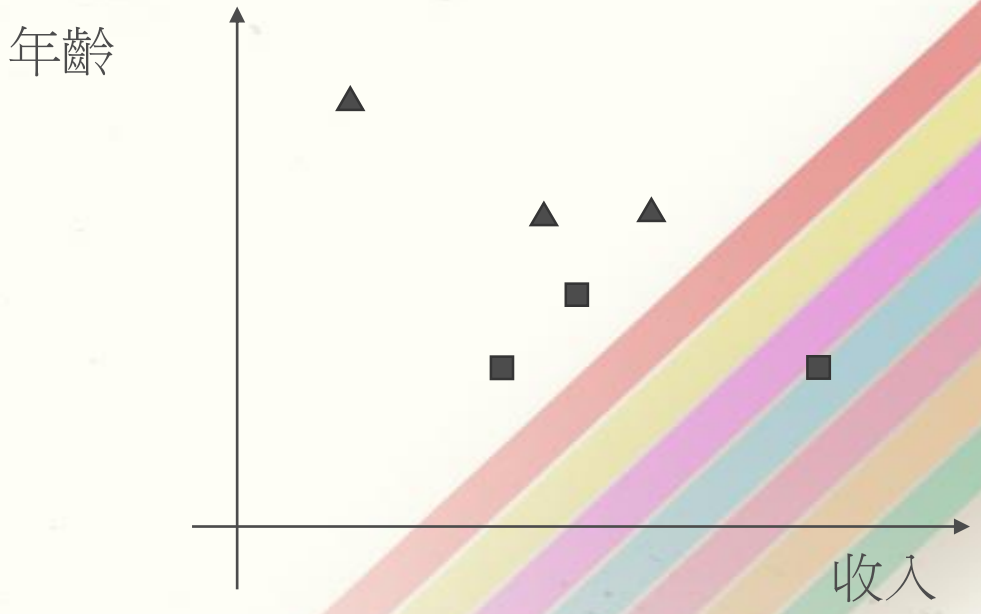
■ 表示購買加值服務的客戶

▲ 表示未購買加值服務的客戶

分類模型分析流程(3)



訓練資料
(training dataset)



測試資料
(testing dataset)

隨機分成訓練資料 12 筆和測試資料 6 筆

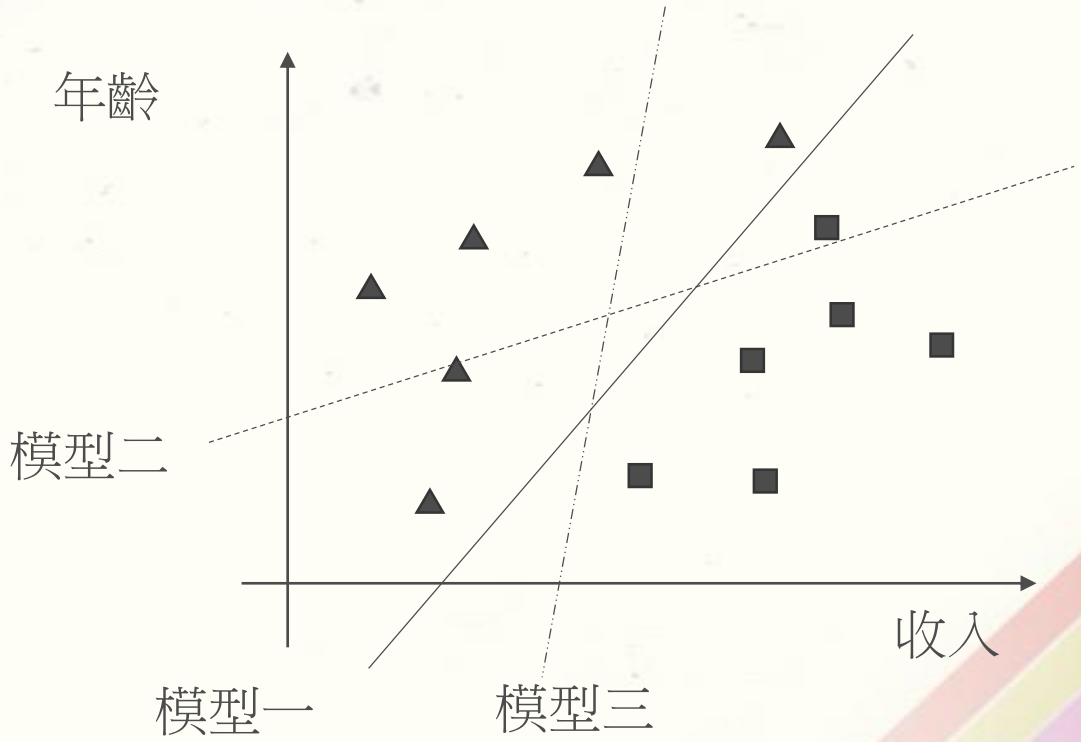
分類模型分析流程(4)

利用訓練資料來建立分類模型，至於要如何選擇學習演算法呢？可以想像在圖上畫一條直線，畫直線的原則為

“盡可能地”

讓直線兩邊都是一樣類型的客戶，

圖中三條不同的直線分別代表三個不同的分類模型



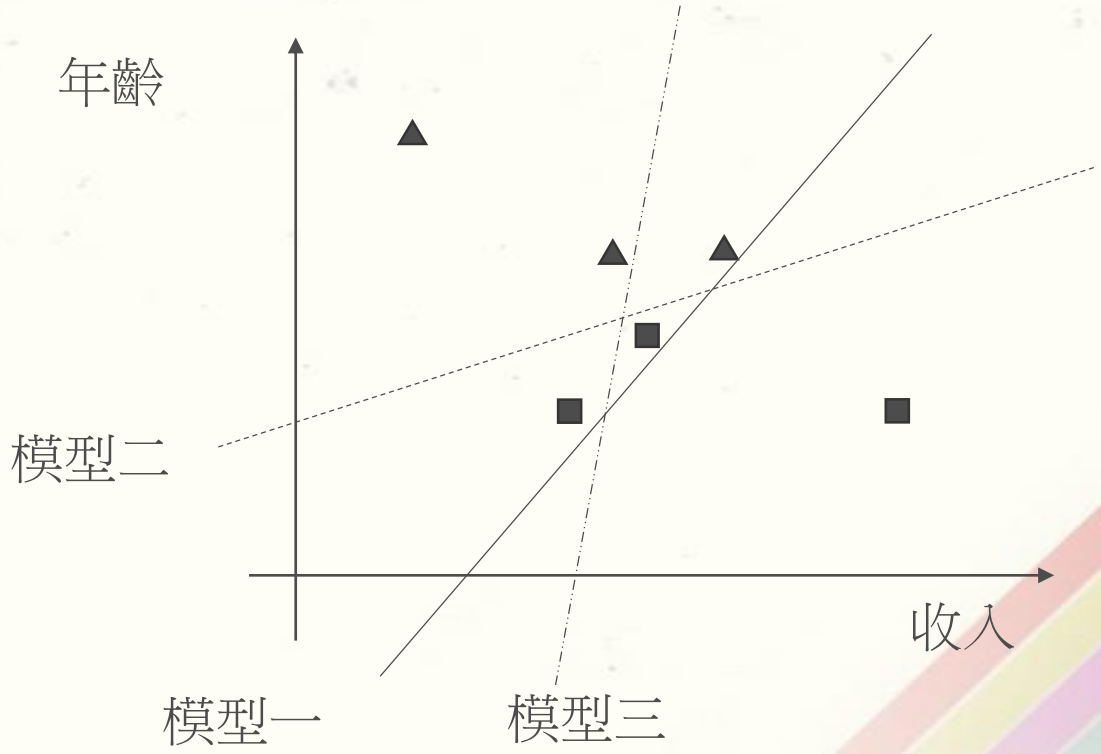
如果從訓練資料的結果評估三個模型，可以看出模型一的效果最好，因為在模型一的直線右邊都是會購買的客戶■，而直線左邊都是不會購買的客戶▲

而模型二和模型三因為容許一些些的失誤造成無法將客戶清楚地分到直線的兩邊，
但這樣的模型(指有失誤)是否就是效果不好呢？

分類模型分析流程(5)

從測試資料的結果可以發現反而是模型二的分類效果是較好的。

但是不管是哪一個模型都可以歸納描述出購買加值服務客戶的特色是高收入、年紀輕，也可以利用最後選擇的模型去預測新客戶是否有機會購買加值服務而進行推銷。



分類模型資料集(dataset)

	$(x_1,$	$x_2,$	$x_3,$	$x_4)$	y
	編號	是否負債	性別	婚姻狀態	收入	是否還款
訓練資料	1	是	男	單身	低	否
	2	否	女	單身	低	否
	3	是	男	單身	高	是
	4	否	女	結婚	低	是
	5	否	男	單身	高	是
	6	是	女	單身	高	否
	7	否	女	結婚	低	是
	8	是	男	結婚	高	否
	9	否	男	單身	低	是
測試資料	10	是	女	結婚	低	否
	11	否	女	結婚	高	是
	12	是	男	結婚	高	否

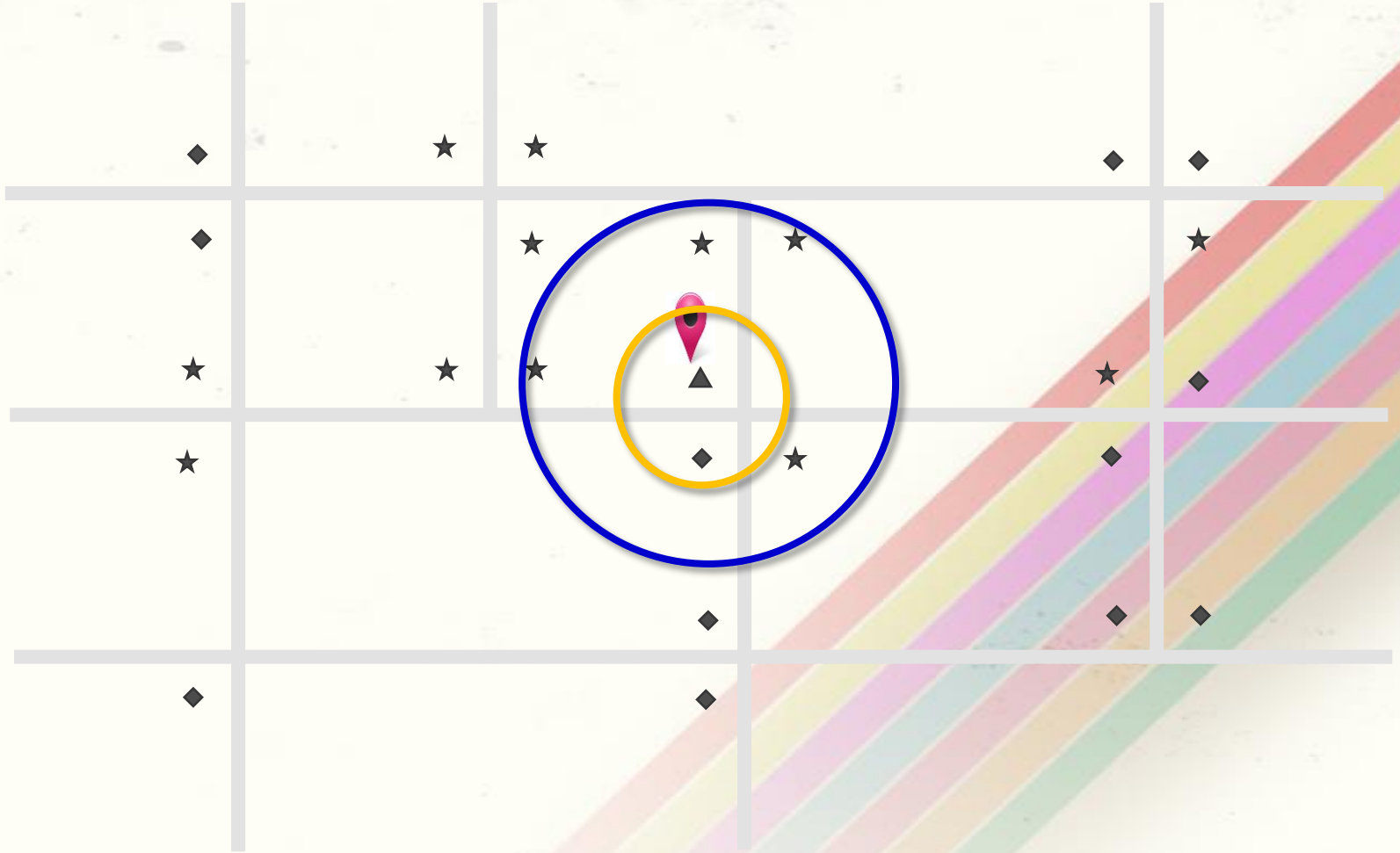
分類技術

- 1 分類的概念
- 2 *K*最鄰近法
- 3 決策樹
- 4 分類模型的評估
- 5 結論

城市的地圖與房價的判別

此為一張城市的地圖，地圖標示★為房價實價登錄高於此城市平均房價的地點，標示◆為房價實價登錄低於此城市平均房價的地點，請問標示▲的地點其房價會高於還是低於此城市平均房價？判斷的理由為何？

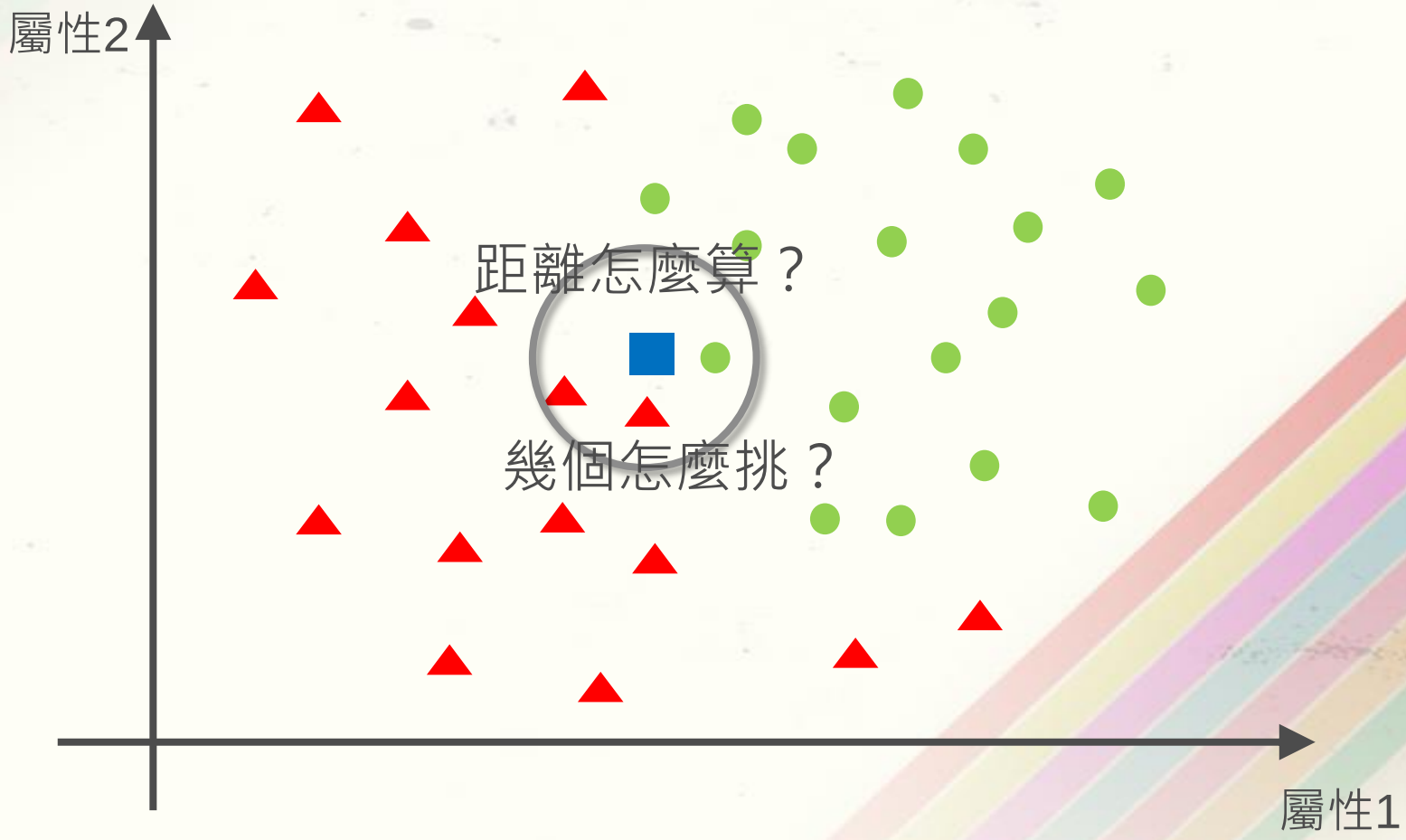
(1)假如 k=1？
(2)假如 k=5？



*K*最鄰近法的分類流程

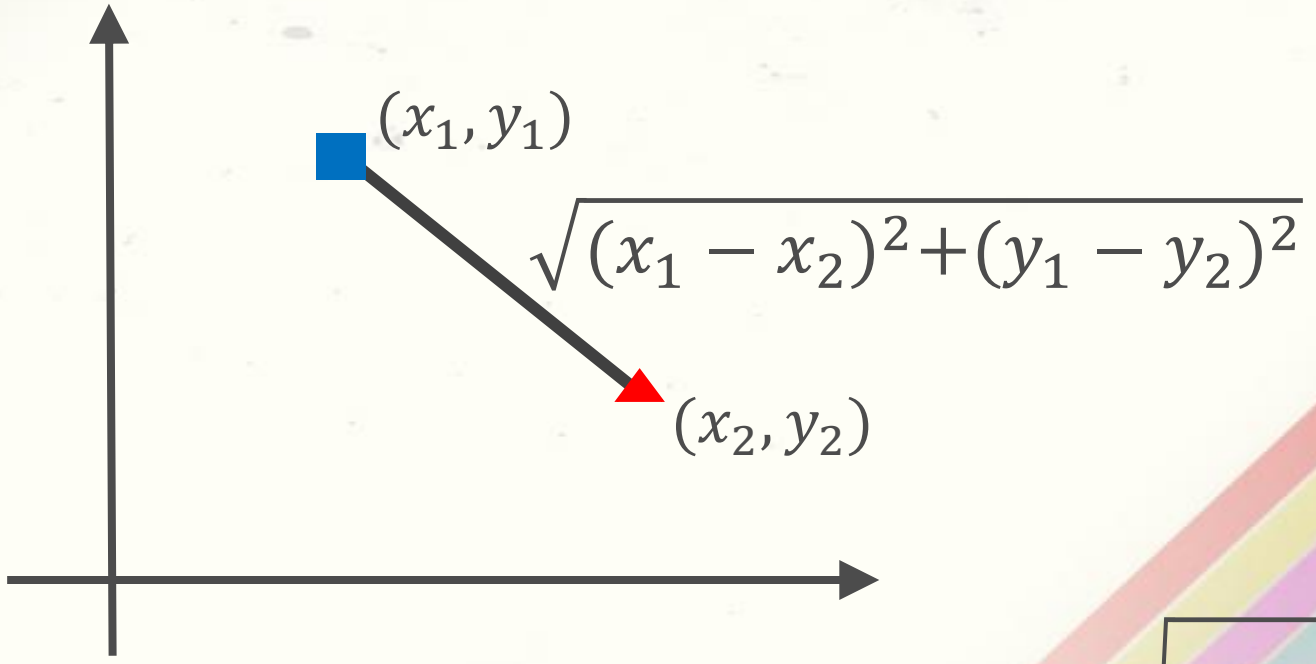
- 步驟一：決定參數*K*（最鄰近物件的個數）
- 步驟二：利用屬性資料計算未知類別物件與所有訓練資料的距離
- 步驟三：排序距離並選出最小的*K*個最鄰近物件
- 步驟四：根據最鄰近物件多數的類別決定待分類物件的類別

如何使用最鄰近法？



距離怎麼算？

- 歐氏距離



假如資料中的屬性是數值資料，計算距離方式可以使用歐氏距離

$$a = (a_1, a_2, \ldots, a_n)$$
$$b = (b_1, b_2, \ldots, b_n)$$

$$\Rightarrow \quad distance = \sqrt{\sum_{i=1}^{k} (a_i - b_i)^2}$$

數值資料歐式距離

會員編號	年齡	平均月收入(千)	…		
1	20	20			
2	21	26			
3	22	25			
4	41	30			
5	43	32			
6	52	40			
7	55	38			
…	…	…			

$$d_E(x_1, x_2) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

$$= \sqrt{(20 - 21)^2 + (20 - 26)^2}$$

$$d_M(x_1, x_2) = |20 - 21| + |20 - 26|$$

d_E歐氏距離(Euclidean distance)

d_M曼哈頓距離(Manhattan distance)

K最鄰近法的應用(1)

訓練資料

測試資料

編號	是否負債	性別	婚姻狀態	收入	是否還款
1	是	男	單身	低	否
2	否	女	單身	低	否
3	是	男	單身	高	是
4	否	女	結婚	低	是
5	否	男	單身	高	是
6	是	女	單身	高	否
7	否	女	結婚	低	是
8	是	男	結婚	高	否
9	否	男	單身	低	是
10	是	女	結婚	低	否
11	否	女	結婚	高	是
12	是	男	結婚	高	否

K最鄰近法的應用(2)

- 距離 = $\frac{\text{屬性值相異個數}}{\text{全部屬性個數}}$

因為資料中的屬性不是數值資料，而是類別資料，所以必須重新定義一個計算距離的公式如左

編號	是否負債	性別	婚姻狀態	收入	是否還款
1	是	男	單身	低	否
2	否	女	單身	低	否
10	是	女	結婚	低	否
11	否	女	結婚	高	是
12	是	男	結婚	高	否

K最鄰近法的應用(3)

舉例說明，編號 10 和編號 1 的四個屬性中，在是否負債和收入的值是一樣的，所以距離為 2/4 = 0.5

編號	是否負債	性別	婚姻狀態	收入	是否還款
1	是	男	單身	低	否
2	否	女	單身	低	否
3	是	男	單身	高	是
4	否	女	結婚	低	是
5	否	男	單身	高	是
6	是	女	單身	高	否
7	否	女	結婚	低	是
8	是	男	結婚	高	否
9	否	男	單身	低	是
10	是	女	結婚	低	否
11	否	女	結婚	高	是
12	是	男	結婚	高	否

距離	10 ○	11 ○	12 ✖	是否還款
1	0.5	1	0.5	否
2	0.5	0.5	1	否
3	0.75	0.75	0.25	是
4	0.25	0.25	0.75	是
5	1	0.5	0.5	是
6	0.5	0.5	0.5	否
7	0.25	0.25	0.75	是
8	0.5	0.5	0	否
9	0.75	0.75	0.75	是

說明編號 10、11、12

- 舉例說明，編號 10 和編號 1 的四個屬性中，在是否負債和收入的值是一樣的，所以距離為 2/4 = 0.5。

- 所以若決定參數 K = 3，編號 10 和編號 11 雖然在0.5 有多個可能結果，但因為距離最近的兩個都是會還款，所以將其分類為會還款

- 而編號 12 最近的兩個一個會還款一個不會還款，在第三近的三個0.5中，有兩個不會還款，於是判定編號 12 也不會還款，則採用 K = 3，只有編號 10 預測不正確。

K最鄰近法的討論(1)

- 尺度變異的控制：正規化($normalization$)
  - 考慮三筆資料：A(25, 3000)、B(75, 3000)、C(25, 3100)，分別代表 A、B、C 的 (年齡、收入) 資料

$$\mathrm{A}(25{,}3000)\ vs.\ \mathrm{B}(75{,}3000)$$

$$\mathrm{A}(25{,}3000)\ vs.\ \mathrm{C}(25{,}3100)$$

$$min - max\ normalization$$

$$\boldsymbol{v' = \frac{v - min}{max - min}}$$

說明

- **其中 A 和 B 年齡差距 50 歲，收入一樣 3000元，A 和 C 收入差距 100 元，年齡一樣是 25 歲，概念上 A 和 B 年齡上的差異遠大於 A 和 C 收入上的差異，可是若利用歐氏距離計算，得到 A 和 B 的歐氏距離為 50，A 和 C 的歐氏距離為 100，和概念不符合，究其原因為未考慮屬性資料值的變異，為解決這樣的問題，可以將屬性資料進行正規化**

- 假設年齡屬性中最小值min為 25，最大值max為 75，而收入屬性中最小值min為 3000，最大值max為 5000，經過最小值 - 最大值正規化法轉換後可以得到A(0, 0)、B(1, 0)、C(0, 0.05)，這樣再次計算歐氏距離可以得到與概念上較一致的結果。

K最鄰近法的討論(2)

- 類別個數一樣，當甲兩票，乙也兩票這時候該怎麼辦?
  - 加上權重的概念，**距離越近的影響力可能較大**
  - 甲（2）、丙（2）、乙（2）、乙（3）、甲（5）
  - 使用距離的倒數當作權重
  - 甲的權重得分為$\frac{1}{2}+\frac{1}{5}=\frac{7}{10}$
  - 乙的權重得分為$\frac{1}{2}+\frac{1}{3}=\frac{5}{6}$
  - 歸類為類別乙

當決定 K 值為 5，而最近 5 個物件的類別分別為甲 (2)、丙 (2)、乙 (2)、乙 (3)、甲 (5)，括號內的值為距離，則該分類為甲或是乙？

K最鄰近法的特點

- 不需要事先建立分類模型，直接使用已知類別的資料即可對未知類別的資料進行預測
- 運算成本過高，每一個待分類的資料點都要計算其到全部訓練資料的距離，為解決這個問題，可以事先將訓練資料進行集群，先計算距離哪一個群最近，再計算與最近的群內的資料距離

*K*最鄰近法的特點

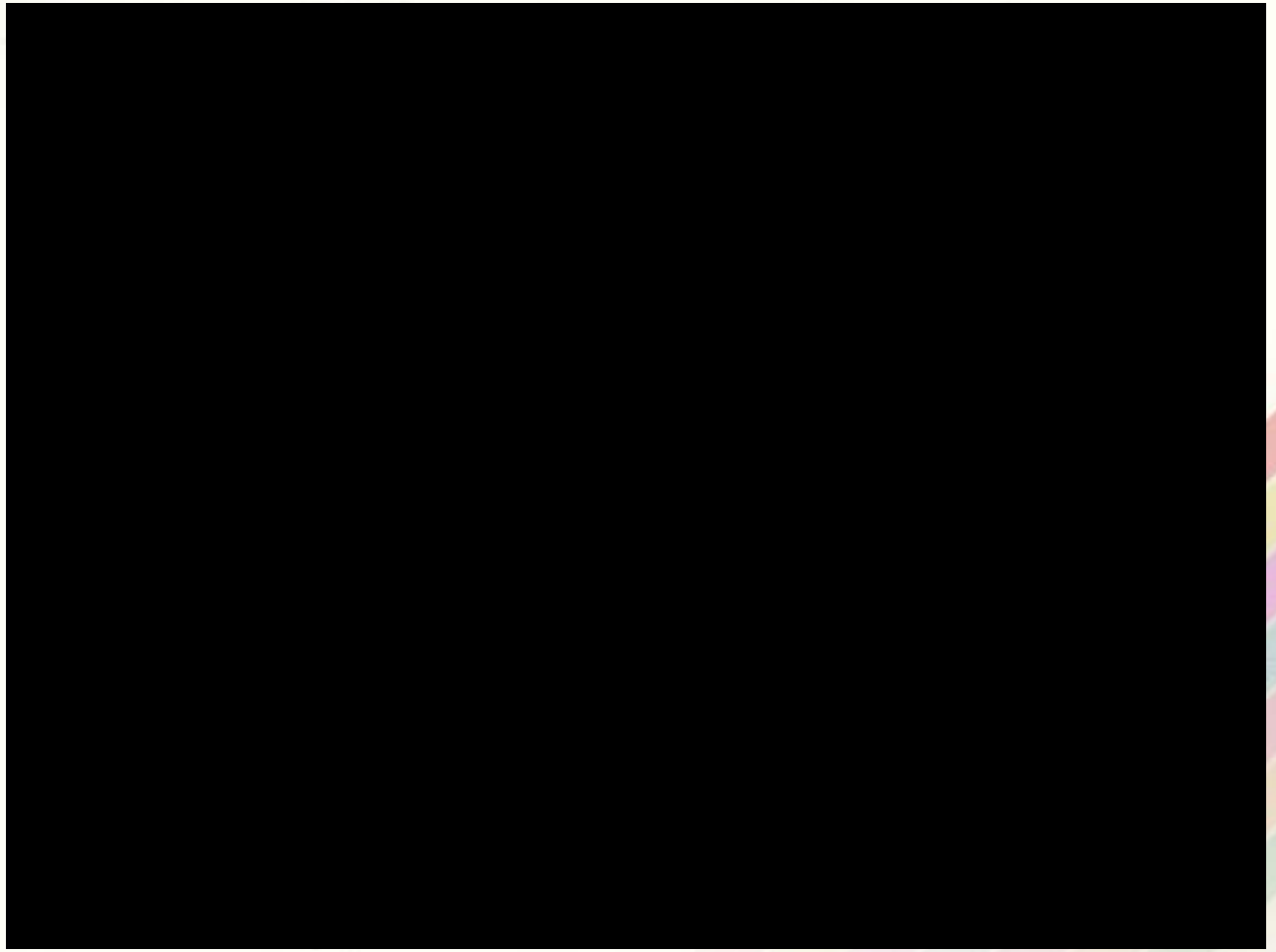
- 類別的決定只需鄰近的*K*個樣本，可以避免類別分布不平衡的問題，對於不同類別的資料數差異過大或類別重疊過多的資料，*K*最鄰近法的效果良好
- *K*值的決定會影響結果，故需多嘗試不同的*K*值以決定分類效果最好的*K*值

分類技術

- 1 分類的概念
- 2 *K*最鄰近法
- 3 決策樹
- 4 分類模型的評估
- 5 結論

什麼是決策樹？

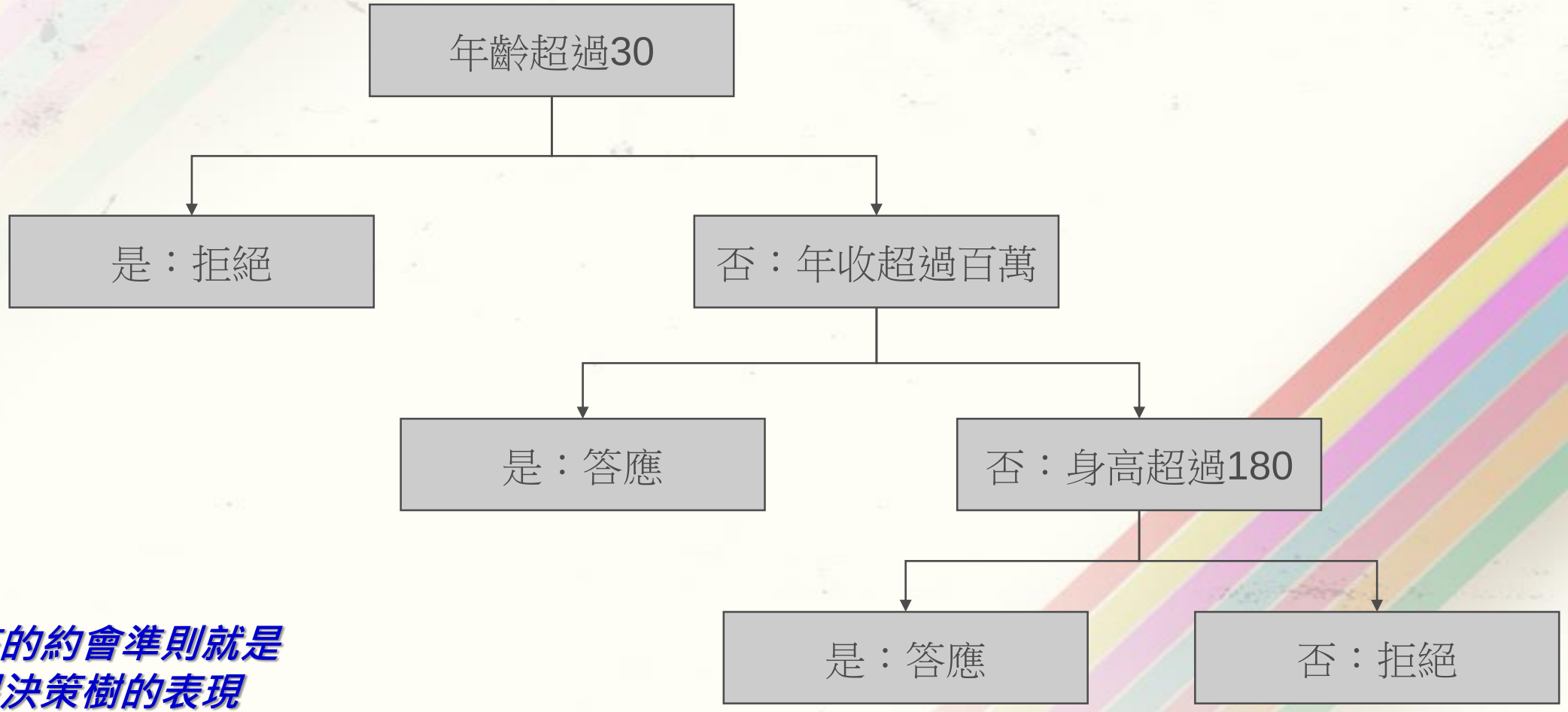
- 心理測驗



小花的約會準則

- 小花是一個受歡迎的女生，每天都有一堆男生想約她出去玩，小花總是在答應與不答應之間煩惱，為了解決這樣的困擾，小花決定依照自己的經驗與偏好用以下的準則決定是否答應。
- 小花的約會準則就是一棵決策樹的表現，雖然這棵決策樹的建構過於主觀，但仍清楚地描繪了小花在意的屬性，並且小花可以透過這棵決策樹快速的依照男生的屬性做決策決定是否答應，生活中有許多決策過程都可以繪製出這樣的決策過程

小花的約會準則



小花的約會準則就是
一棵決策樹的表現

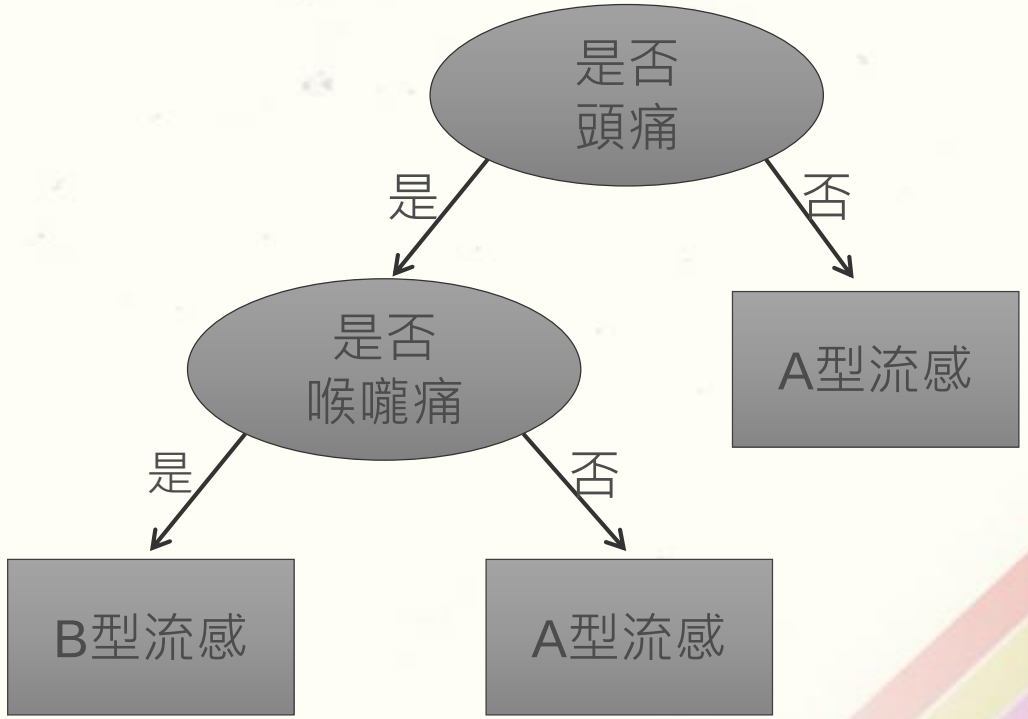
醫生問診準則

- 醫生透過不斷地詢問病患的病徵，最後決定病患的疾病名稱並施以治療，這樣的過程可以描述如下
  - **(1)** 若頭痛或是頭不痛但喉嚨痛，則為 A 型流感；
  - **(2)** 若頭不痛且喉嚨也不痛，則為 B 型流感。

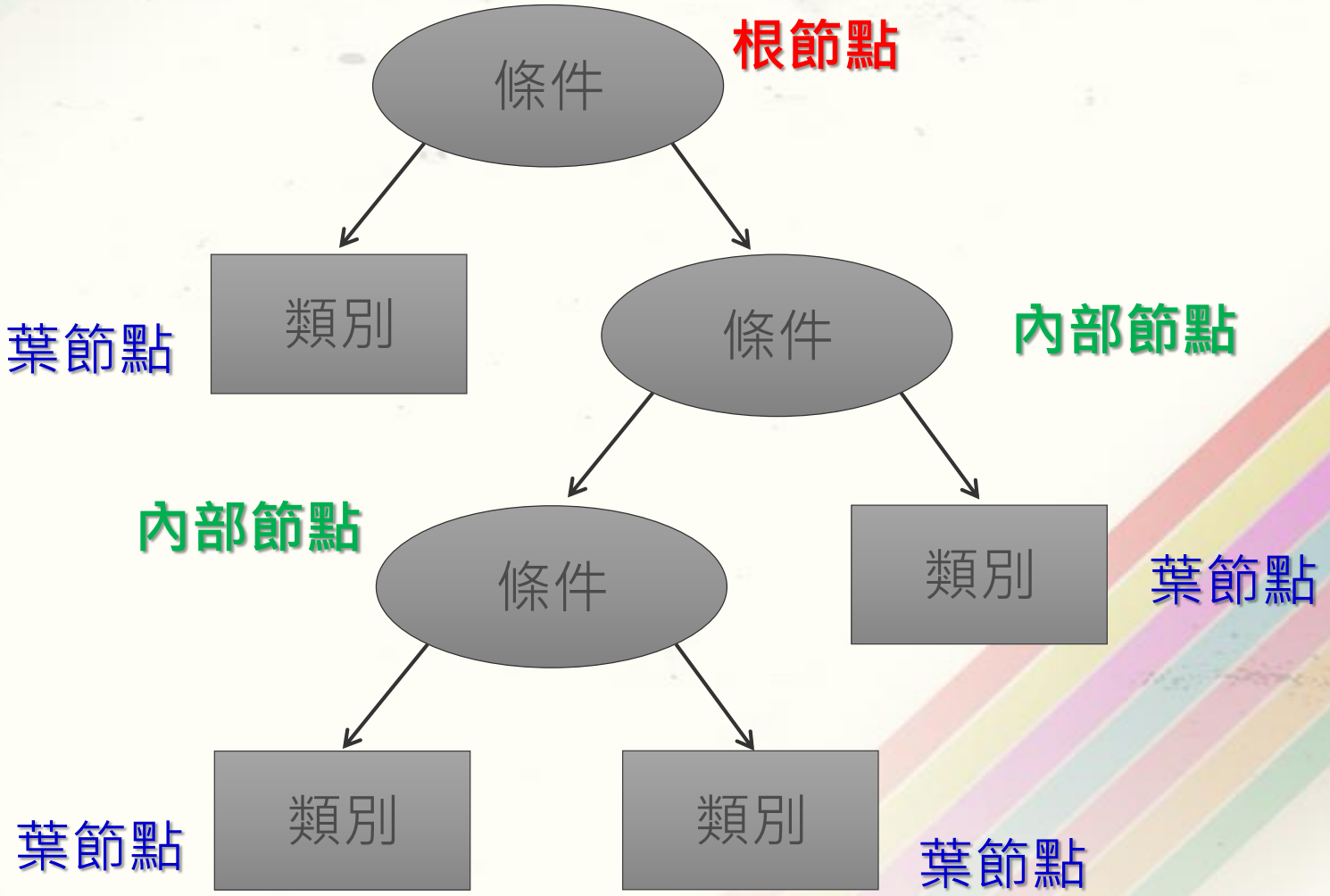
醫生問診準則也是一棵決策樹的表現

- 決策樹是一個處理分類問題的樹狀結構，最上層的節點稱為根節點，沒有分支的節點稱為葉節點，其他的節點稱為內部節點，根節點與內部節點表示一個評估條件，每個分支表示評估條件的可能結果，而葉節點代表是分類結果的類別

醫生問診



究竟什麼是決策樹？

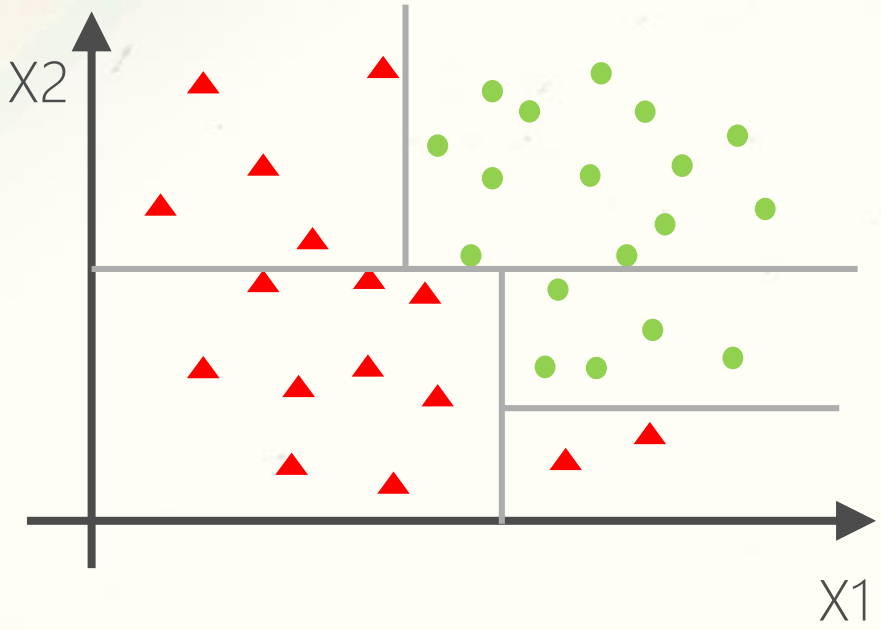


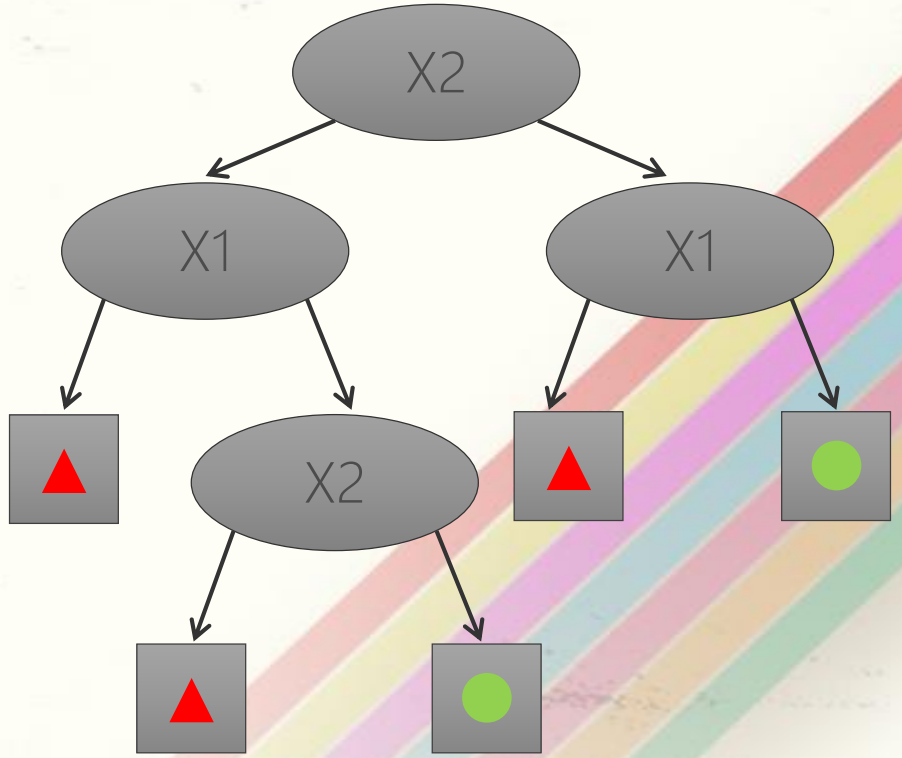
如何建立決策樹？

- Hunt's演算法
  - 概念
    - 遞迴分割
  - 步驟
    - 如果節點內的資料都是同個類別，則此節點變成此類別的葉節點
    - 如果節點內的資料包含多個類別，則選擇適當地屬性測試條件將資料繼續分割

建立決策樹的學習演算法很多，譬如 ID3、C4.5、C5.0、CART、……等演算法，但這些演算法都依循著 Hunt's 演算法的想法，透過不斷地遞迴分割訓練資料，使分割完的資料集中的資料類別都相同

如何建立決策樹？(1)



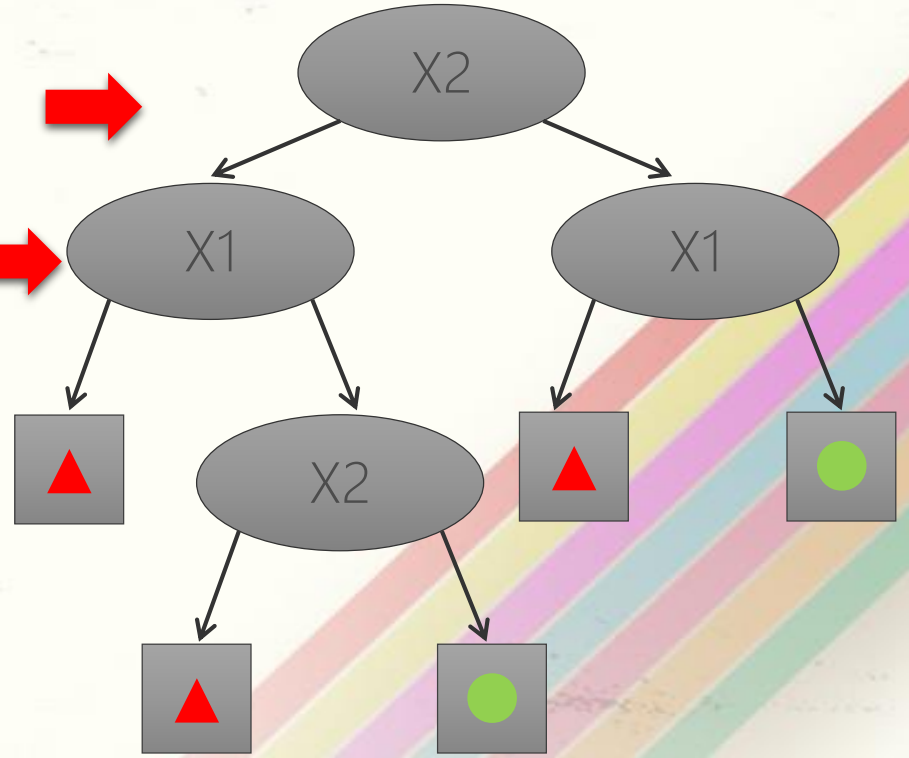


如何建立決策樹？(2)

屬性的條件可以有幾個分支？

選哪個屬性來分割比較好？

何時該停止分割？

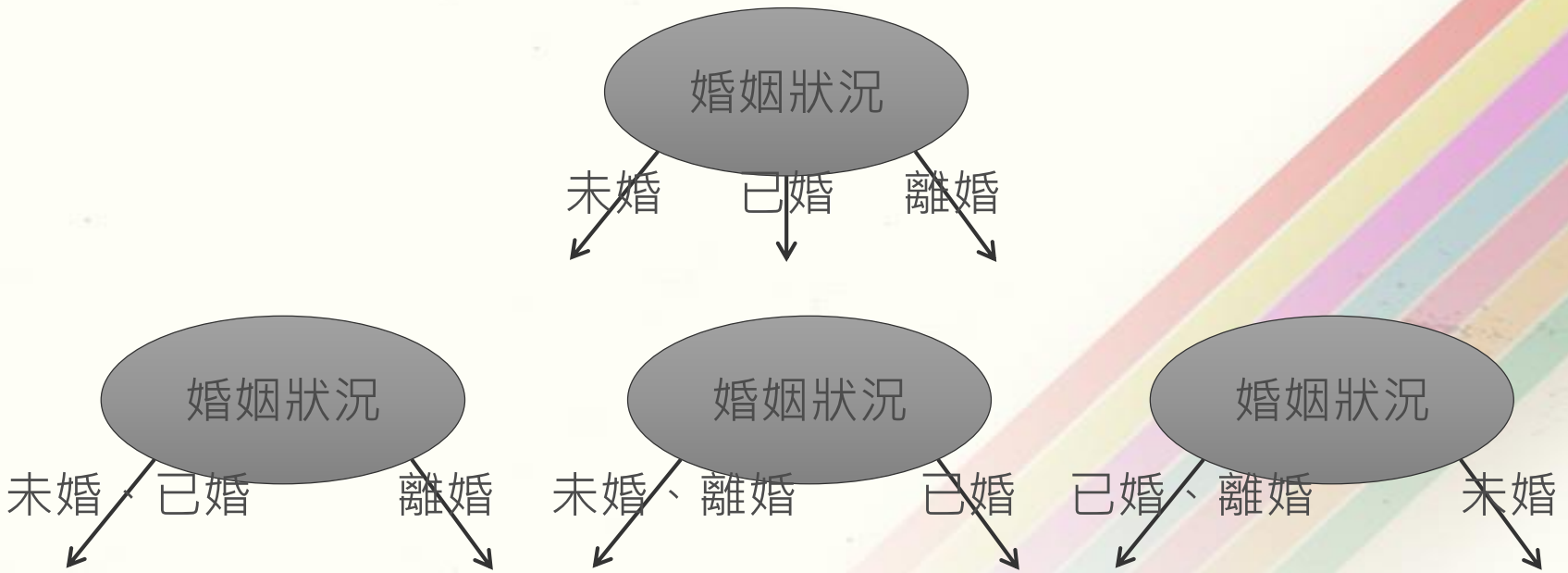


屬性的條件可以有幾個分支？(1)

- 二元屬性
  - 是/否、男/女、...
  - 兩個分支條件恰好可以表示

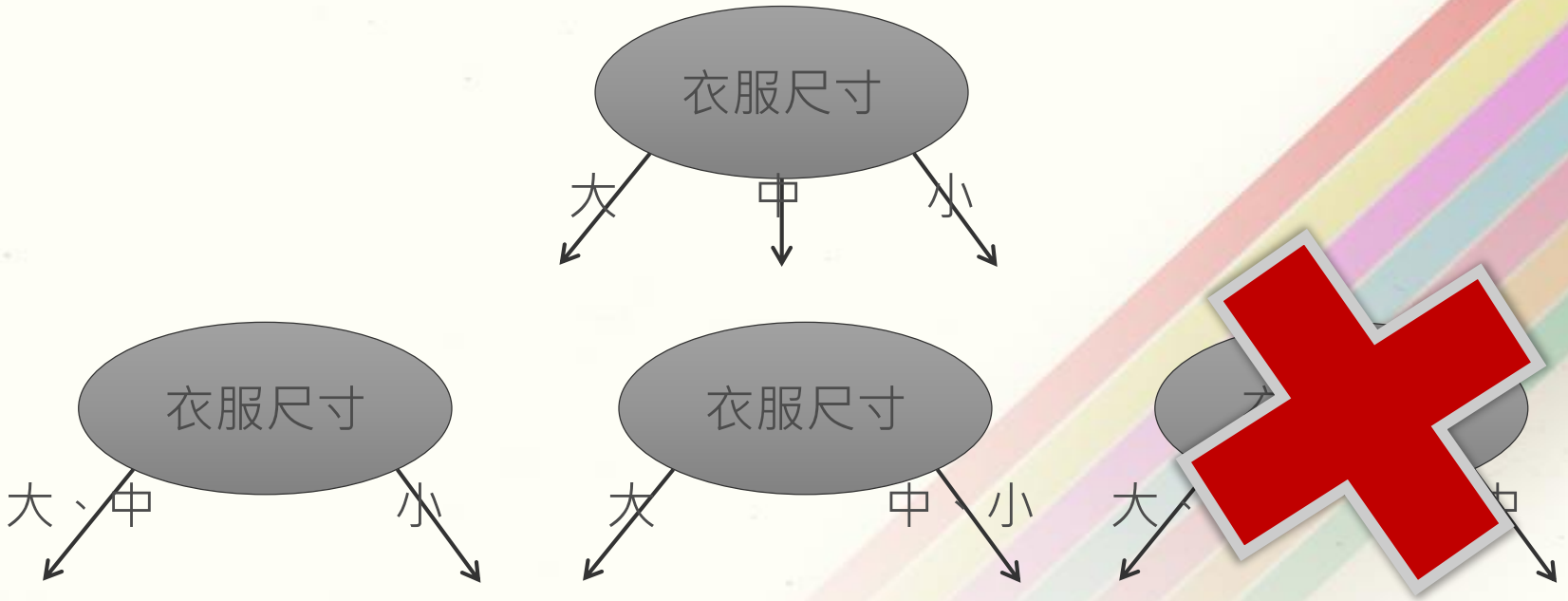
屬性的條件可以有幾個分支？(2)

- 類別屬性
  - 血型、隊名、...
  - 可多個分支條件亦可轉成兩個分支條件



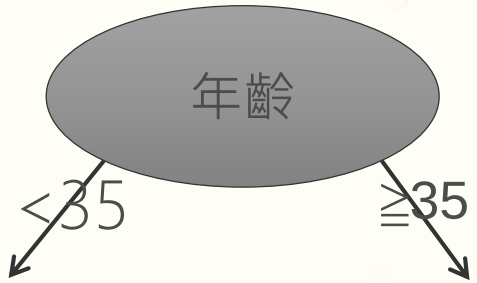
屬性的條件可以有幾個分支？(3)

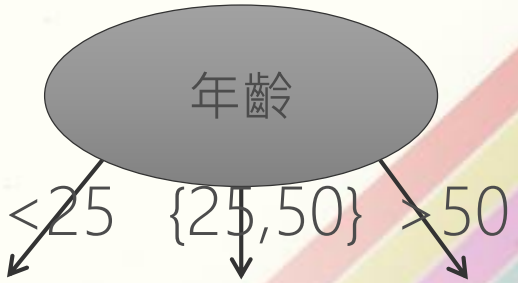
- 順序屬性
  - 學歷、等級、...
  - 可多個分支條件亦可轉成兩個分支條件



屬性的條件可以有幾個分支？(4)

- 連續變數
  - 溫度、身高、薪資、...
  - 使用多個區間條件或選一個值做分割

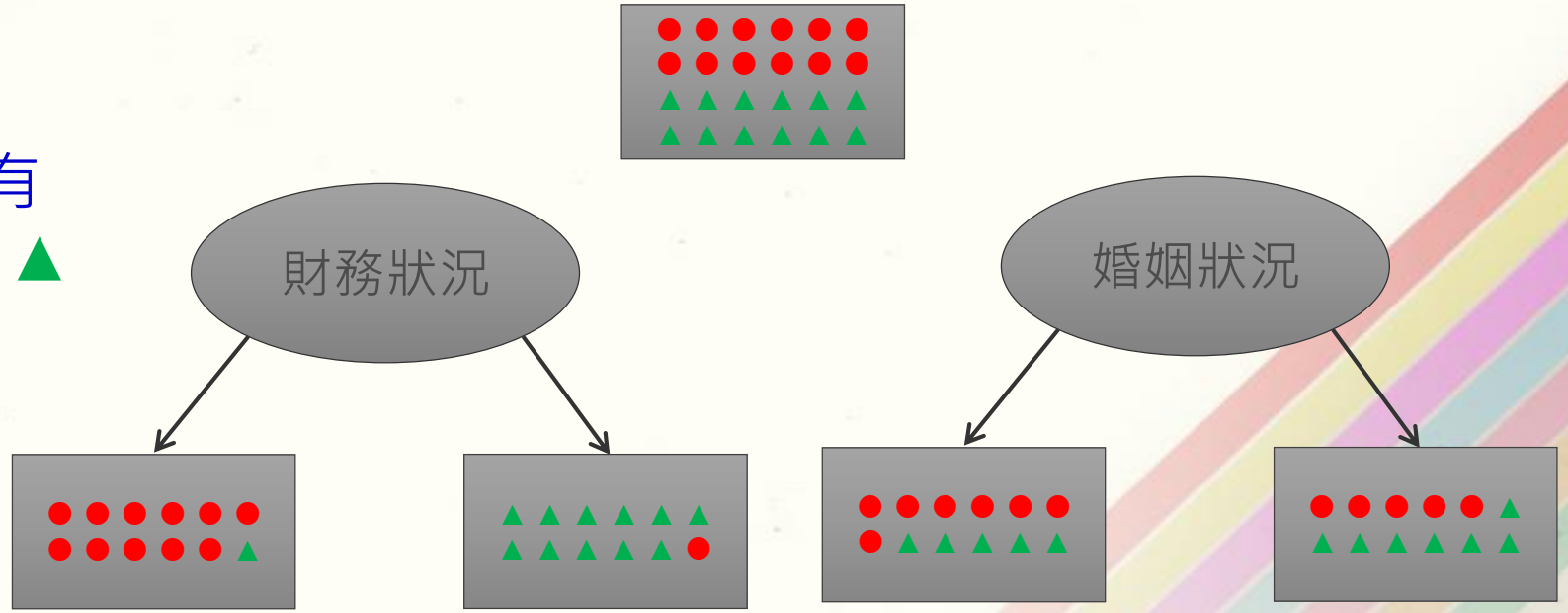




選哪個屬性來分割比較好？

- 思考一下，你會選擇財務狀況抑或婚姻狀況當做分割的屬性條件？

原始的資料分布有
12 個 ● 和 12 個 ▲

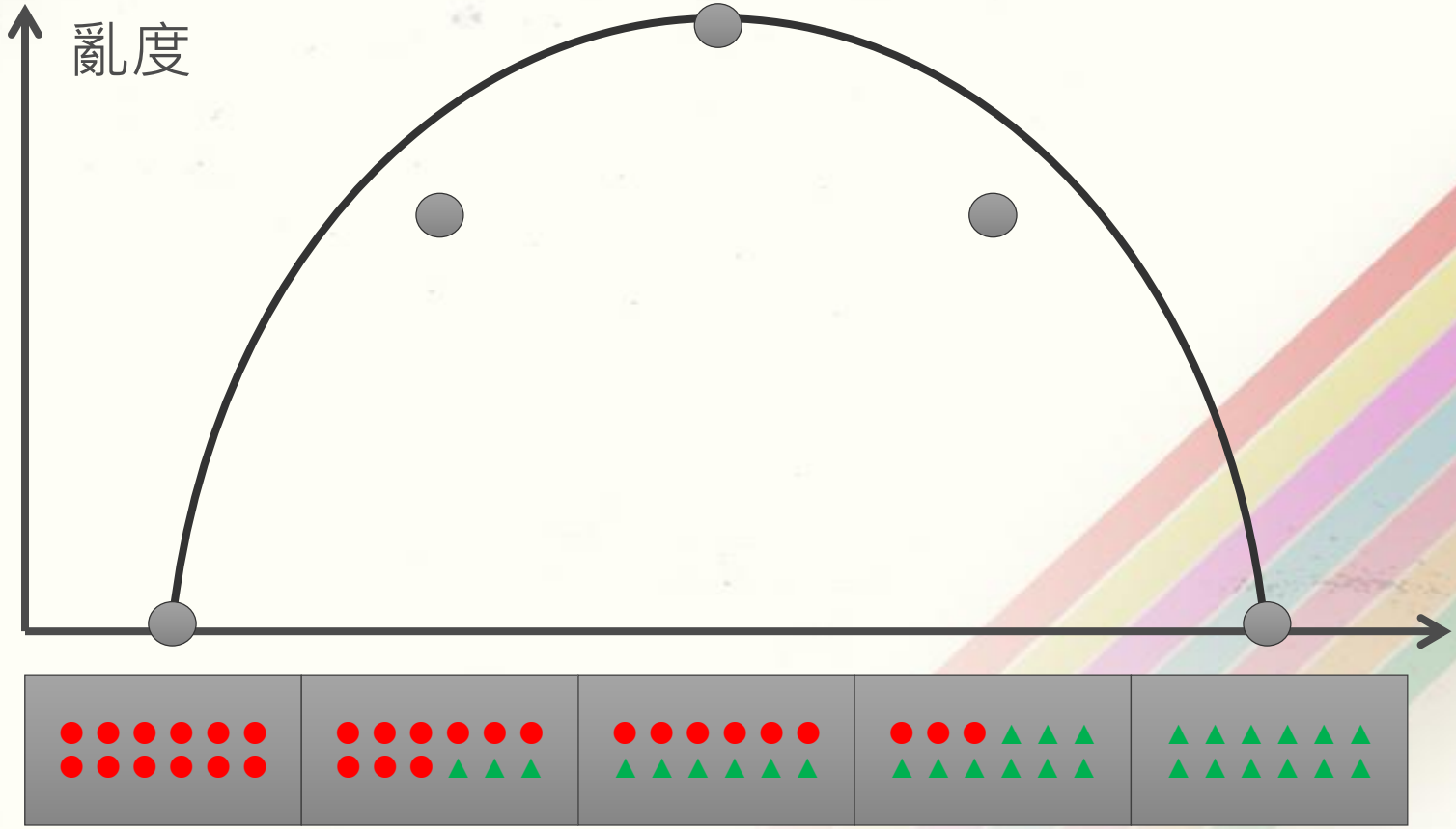


Key point
只要比較分割前和分割後哪一個屬性可以使資料的亂度降低最多，就選擇該屬性作為條件節點

分類的結果越乾淨越好

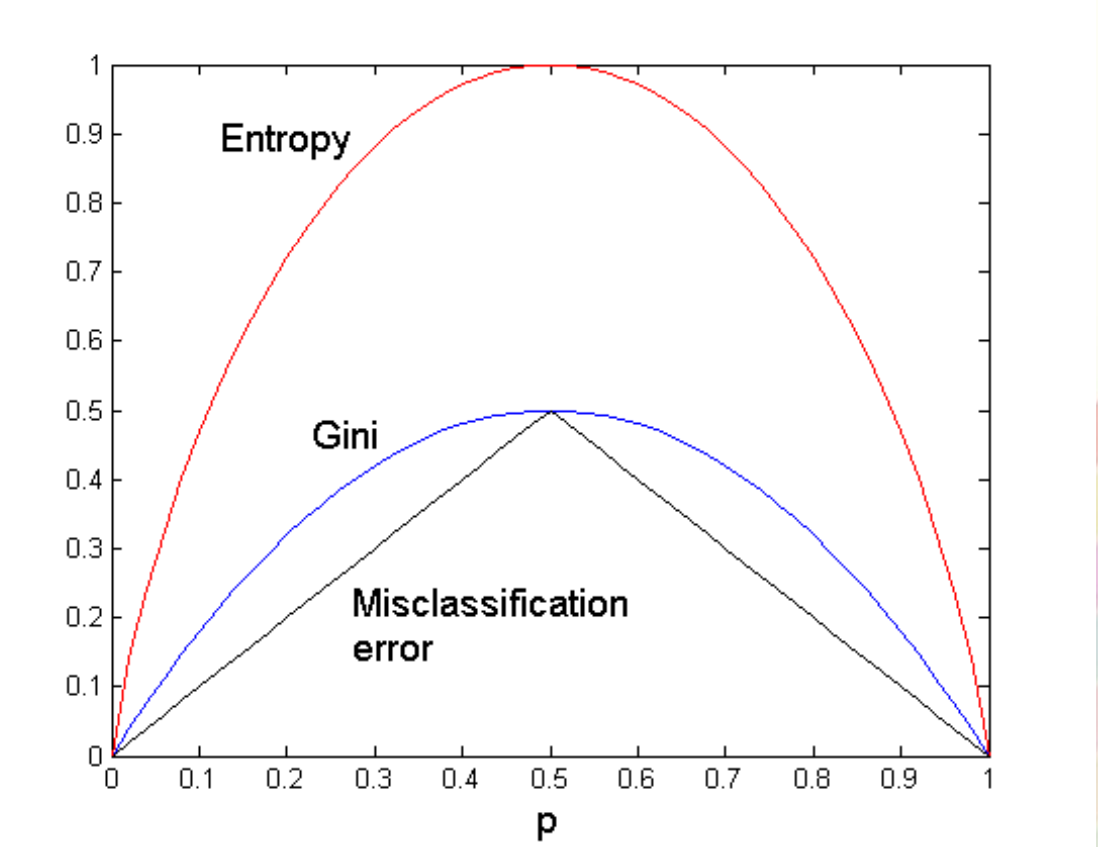
選哪個屬性來分割比較好？

- 如何描述乾淨(混亂)的程度？



選哪個屬性來分割比較好？

- 是否有這樣的亂度衡量指標？



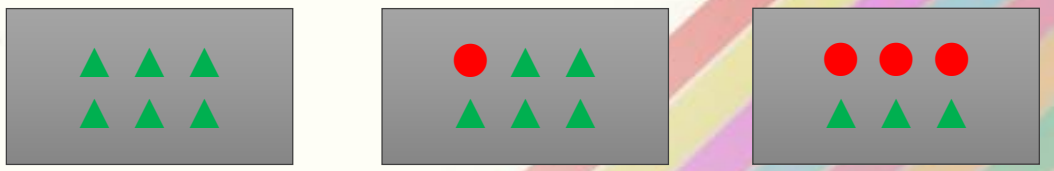
選哪個屬性來分割比較好？

- 亂度的衡量通常使用熵 (Entropy) 和 Gini 係數兩個指標，亂度衡量指標的定義

$$Entropy = -\sum_{i=1}^{k} p_i \log_2 p_i$$

$$Gini = 1 - \sum_{i=1}^{k} p_i^2$$

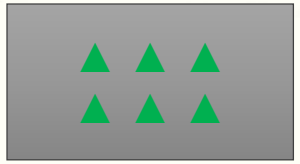
其中 k 表示分割中的類別數，而 pi 表示第 i 個類別所占的比率



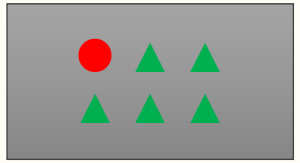
狀況一：分割內類別 ● 有 0 個，類別 ▲ 有 6 個
狀況二：分割內類別 ● 有 1 個，類別 ▲ 有 5 個
狀況三：分割內類別 ● 有 3 個，類別 ▲ 有 3 個

選哪個屬性來分割比較好？

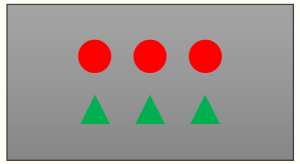
$$Gini = 1 - \sum_{i=1}^{k} p_i^2$$



$$Gini = 1 - \left(\frac{0}{6}\right)^2 - \left(\frac{6}{6}\right)^2 = 0$$



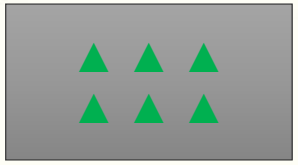
$$Gini = 1 - \left(\frac{1}{6}\right)^2 - \left(\frac{5}{6}\right)^2 = 0.278$$



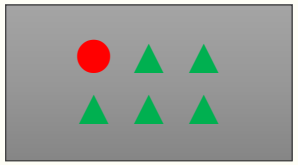
$$Gini = 1 - \left(\frac{3}{6}\right)^2 - \left(\frac{3}{6}\right)^2 = 0.5$$

選哪個屬性來分割比較好？

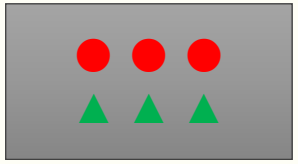
$$Entropy = -\sum_{i=1}^{k} p_i \log_2 p_i$$



$$Entropy = -\left(\frac{0}{6}\right)\log_2\left(\frac{0}{6}\right) - \left(\frac{6}{6}\right)\log_2\left(\frac{6}{6}\right) = 0$$



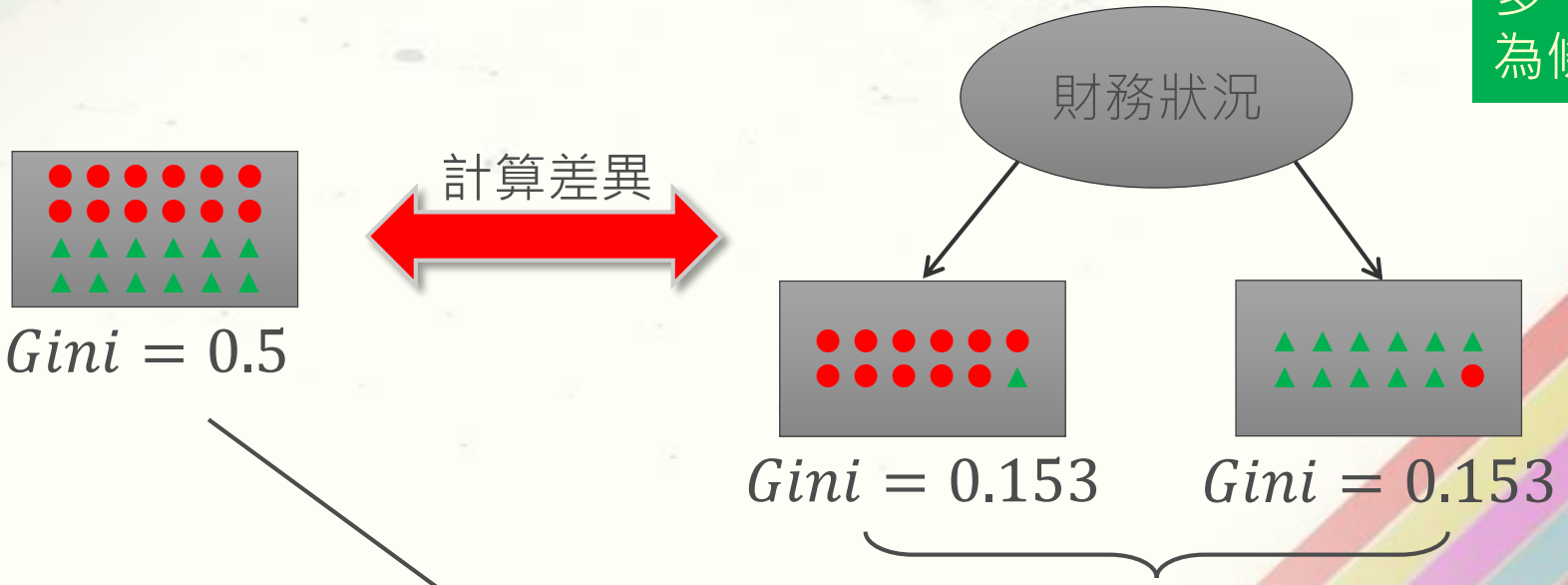
$$Entropy = -\left(\frac{1}{6}\right)\log_2\left(\frac{1}{6}\right) - \left(\frac{5}{6}\right)\log_2\left(\frac{5}{6}\right) = 0.650$$



$$Entropy = -\left(\frac{3}{6}\right)\log_2\left(\frac{3}{6}\right) - \left(\frac{3}{6}\right)\log_2\left(\frac{3}{6}\right) = 1$$

選哪個屬性來分割比較好？

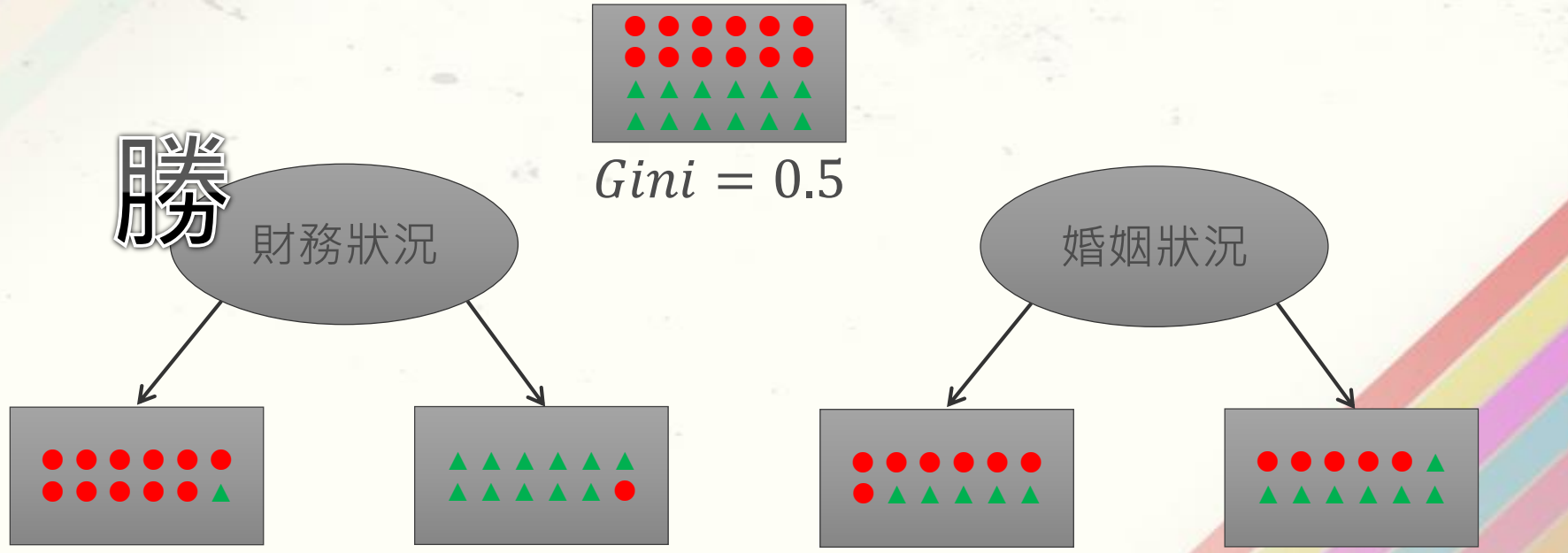
Key point
只要比較分割前和分割後哪一個屬性可以使資料的亂度降低最多，就選擇該屬性作為條件節點



$$定義Gain\Delta = 0.5 - \left[\left(\frac{12}{24}\right) \times 0.153 + \left(\frac{12}{24}\right) \times 0.153\right]$$

$$Gain(A) = I(D) - I_A(D)$$

選哪個屬性來分割比較好？



$Gini = 0.5$

$Gini = 0.153$ $Gini = 0.153$ $Gini = 0.486$ $Gini = 0.486$

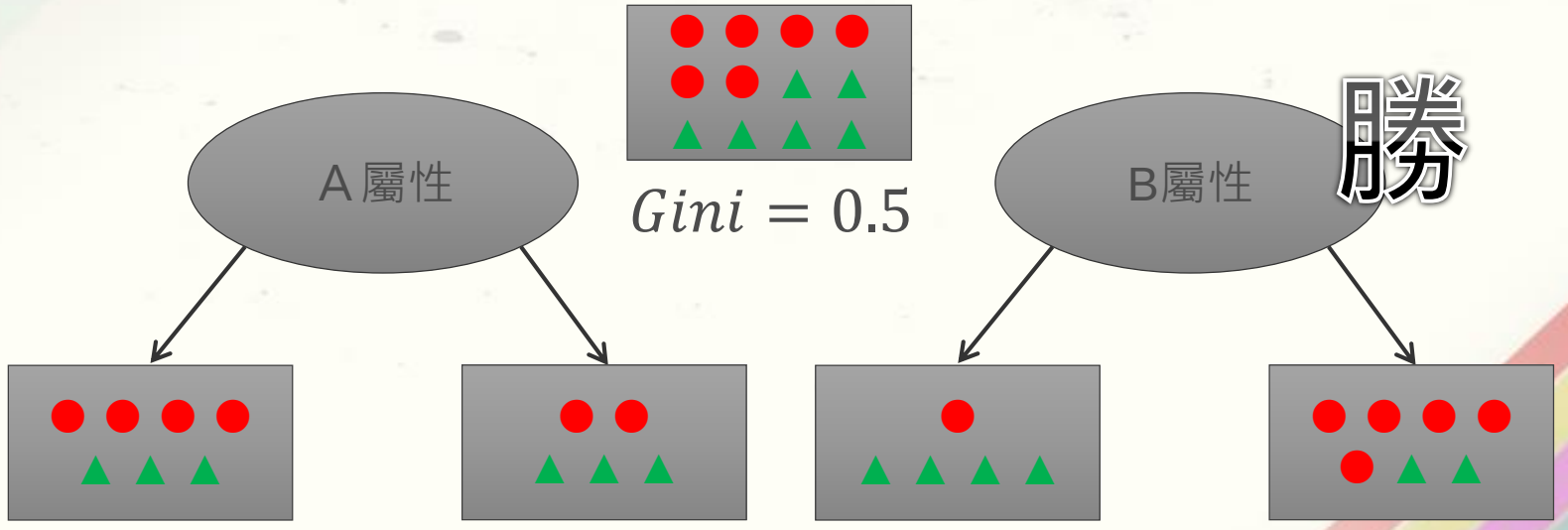
$$\left(\frac{12}{24}\right) \times 0.153 + \left(\frac{12}{24}\right) \times 0.153 = 0.153$$

$$\left(\frac{12}{24}\right) \times 0.486 + \left(\frac{12}{24}\right) \times 0.486 = 0.486$$

$$\Delta = 0.5 - 0.153 = 0.347$$

$$\Delta = 0.5 - 0.486 = 0.014$$

選哪個屬性來分割比較好？



$$\left(\frac{7}{12}\right) \times 0.4898 + \left(\frac{5}{12}\right) \times 0.48 = 0.486$$

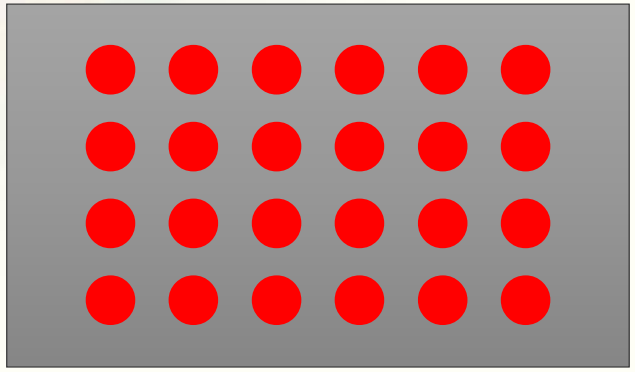
$$\Delta = 0.5 - 0.486 = 0.014$$

$$\left(\frac{5}{12}\right) \times 0.32 + \left(\frac{7}{12}\right) \times 0.2778 = 0.375$$

$$\Delta = 0.5 - 0.375 = 0.125$$

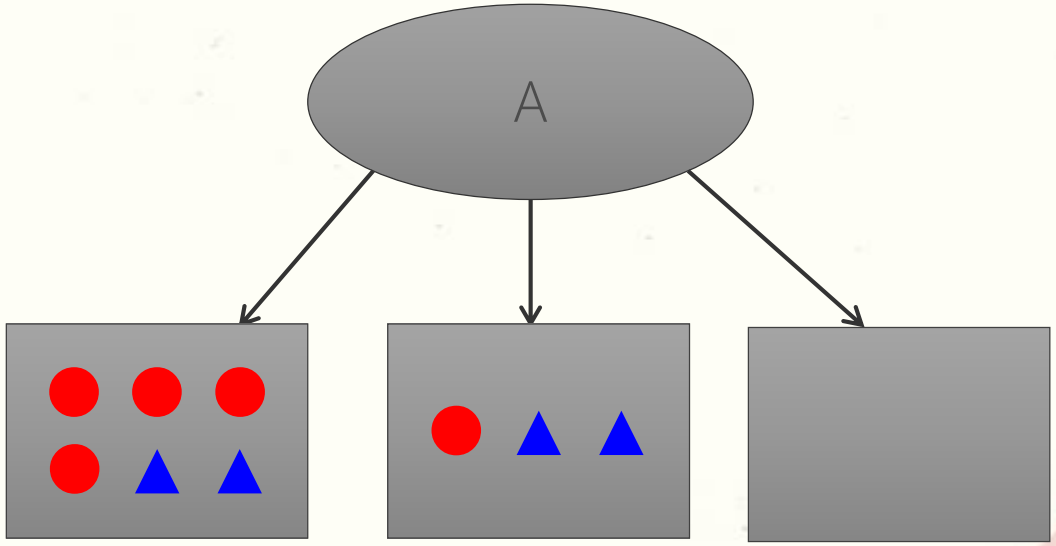
何時該停止分割？

- Hunt’s 演算法本身的停止條件，分割內都是同一個類別



何時該停止分割？

- 分割的子樹是空的，空節點用父結點多數決



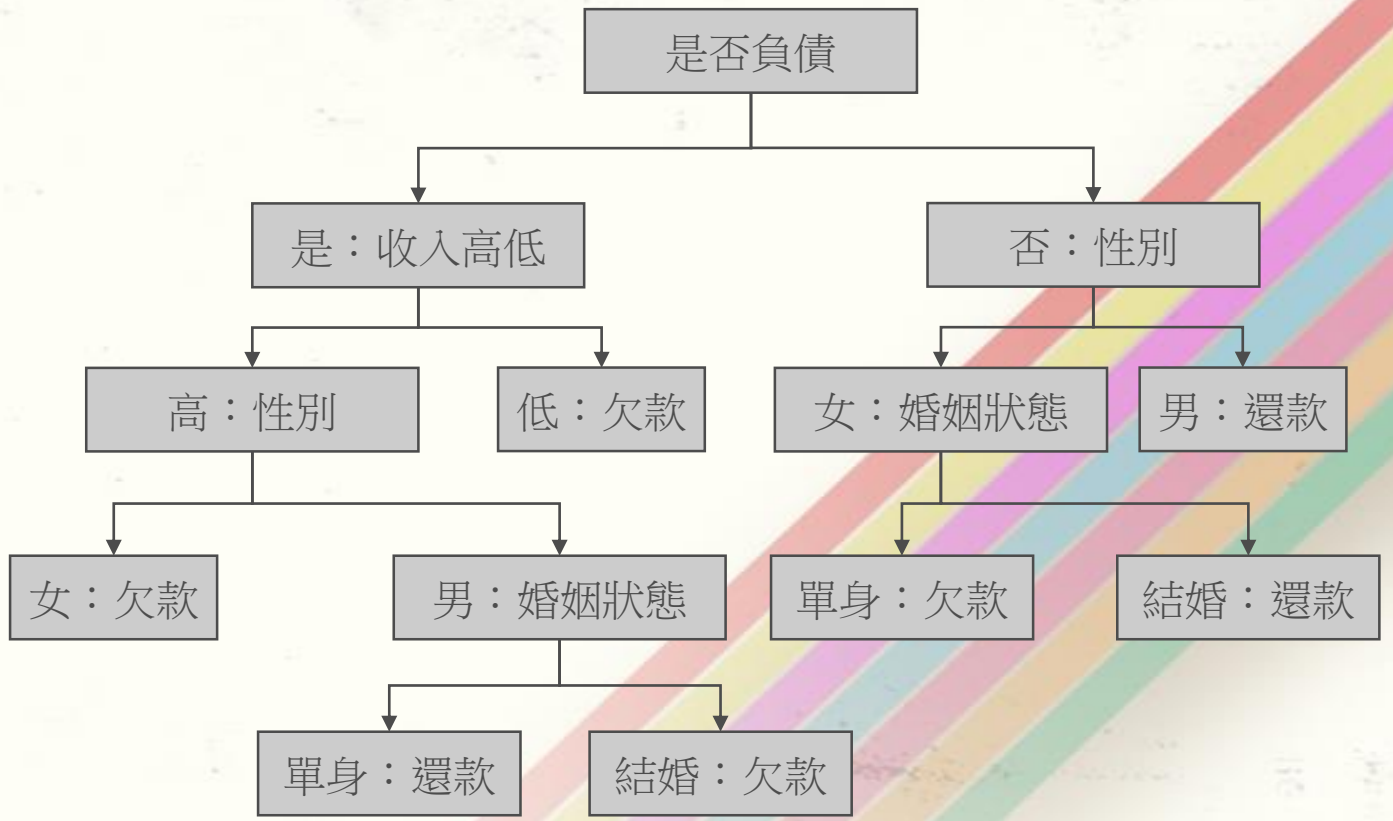
何時該停止分割？

- 屬性的值皆一樣，多數決

ID	已婚	收入	規格	類別
1	是	22K	大	是
2	是	22K	大	否
3	是	22K	大	是

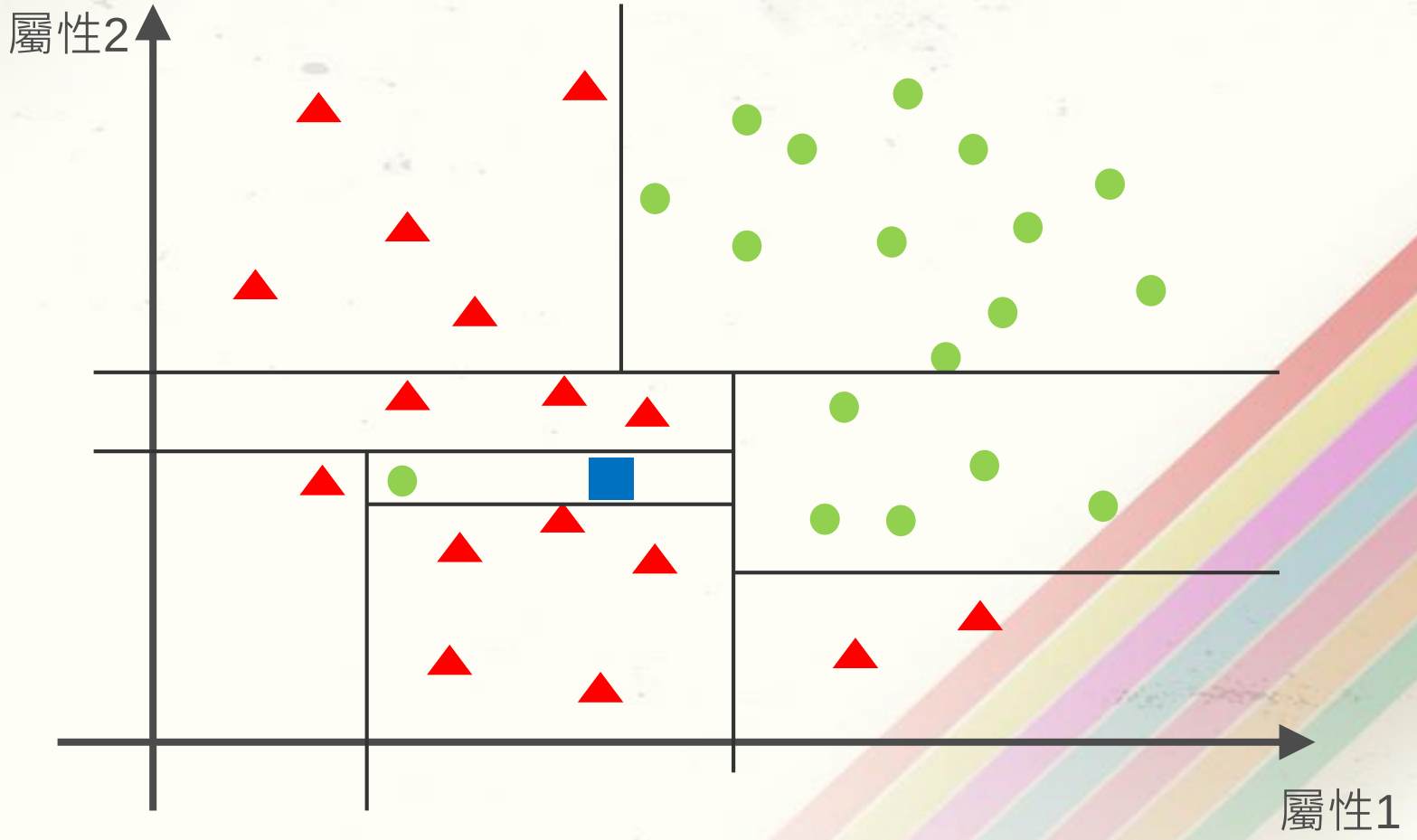
決策樹的應用

編號	是否負債	性別	婚姻狀態	收入	是否還款
10	是	女	結婚	低	否
11	否	女	結婚	高	是
12	是	男	結婚	高	否



使用此決策樹判別測試資料可以得到
編號 10 為不會還款，
編號 11 會還款，
編號 12 不會還款，
全部的測試資料都正確預測。

決策樹的討論



為了少數的資料製造過度複雜的模型

決策樹的討論

- 預先修剪(Pre-Pruning)
  - 定義門檻值阻止後續子樹的分割
    - 類別數、類別比率、Gain獲取率、…
  - 優點
    - 避免產生複雜的子樹
  - 缺點
    - 門檻值不易選取
      - 低則過度配適，高則配適不足
    - 現在沒有好的分割方式不代表之後沒有

決策樹的討論

- 事後修剪(Post-Pruning)
  - 完整的決策樹產生後再進行評估
  - 方法
    - 常被使用→提昇子樹的位置
    - 效果不好→刪除子樹
  - 優點
    - 修剪後的決策樹模型較佳
  - 缺點
    - 需要較多的額外計算成本

決策樹的特點

- 不要求資料的分配
- 建置成本低、速度快、表示方法容易
- 對未知類別的紀錄，分類速度快
- 小型的決策樹容易對模型做解釋
- 與其他分類技術相比精確度不差
- 子樹可能重複出現
- 亂度測量方法的選擇對分類結果影響不大

分類技術

- 1 分類的概念
- 2 *K*最鄰近法
- 3 決策樹
- 4 分類模型的評估
- 5 結論

混淆矩陣(Confusion Matrix)

		預測類別	
		C_1	C_0
實際類別	C_1	$TP(a)$	$FN(b)$
	C_0	$FP(c)$	$TN(d)$

$$Accuracy = \frac{a+d}{a+b+c+d}$$

$$Error\ rate = \frac{b+c}{a+b+c+d}$$

$$Precision(p) = \frac{a}{a+c}$$

$$Recall(r) = \frac{a}{a+b}$$

$$F_1\ mesure = \frac{2rp}{r+p}$$

混淆矩陣(Confusion Matrix)

		預測類別	
		C_1 真	C_0 假
實際類別	C_1 真	$TP(a)$	$FN(b)$
	C_0 假	$FP(c)$	$TN(d)$

一般來說，Precision就是檢索出來的條目(比如：文檔、網頁等)有多少是準確的，Recall就是所有準確的條目有多少被檢索出來了

$$Accuracy = \frac{a+d}{a+b+c+d}$$

$$Error\ rate = \frac{b+c}{a+b+c+d}$$

$$Precision(p) = \frac{a}{a+c}$$

$$Recall(r) = \frac{a}{a+b}$$

$$F_1\ mesure = \frac{2rp}{r+p}$$

利用評估指標評估模型

- 利用正確率(Accuracy)來判別建構的二個分類模型哪個比較好，在三筆測試資料中，二個分類模型的正確率如下，所以最後決定使用決策樹當作預測是否如期還款的分類模型。
  - K最鄰近法：66.7%
  - 決策樹：100%

分類技術

- 1 分類的概念
- 2 *K*最鄰近法
- 3 決策樹
- .4 分類模型的評估
- 5 結論

結論

- 分類是大數據分析與資料探勘的重要技術之一，此單元介紹了分類的重要核心概念和KNN、Decision Tree兩個常用的分類技術，每個分類技術都有其較適合的應用場合，並沒有絕對的好壞，而每個方法深入研究也都有更複雜的延伸和變化。

結論

- 分類技術透過特徵找出每一個資料的類別，並可以利用這些特徵去辨識未知類別的資料，實務上對於分類的應用更是廣泛，舉凡醫學診斷、貸款風險評估、手寫 / 語音 / 人臉辨識、廣告推薦等都有分類的技術元素存在存在，除此之外，在模型的選擇上，除了根據正確率、錯誤率這些指標的選擇之外，分類模型建置的成本也是常需考慮的。