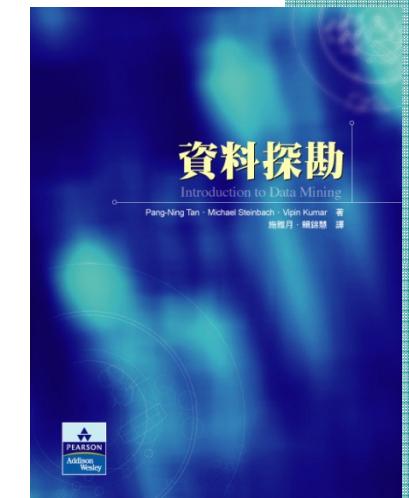


第 2 章

資料



什麼是資料？

- 資料**物件**和其**屬性**的集合

- 一個屬性是指物件的**特性**，而其特性可能會隨時間而變動

— 范例：眼睛的顏色、溫度 **物件**

屬性

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

屬性值

- 屬性值可以用**數值**或是**符號**來表示
- 屬性和屬性值的區別
 - 相同的屬性可以對映至不同的屬性值
 - ◆ 範例：測量出的長度可以尺或米來表示
 - 不同的屬性可以對映至相同的屬性值集合
 - ◆ 範例：身分證字號和年齡這兩種屬性，都可以用**整數**來表示

屬性的型態

- 不同的屬性型態

- 定性

- ◆ **名目 (Items)**: 為**類別屬性**或**定性屬性** (無法運算)

- 範例：員工編號、眼睛顏色、郵遞區號

- ◆ **順序**: 可以排序物件之間的**先後順序**

- 範例：成績(高低)、金屬硬度(程度)

- 定量

- ◆ **區間**: 區間裡的每個值都有意義

- 範例：日期、華氏或攝氏溫度

- ◆ **比例**: 對比例尺度而言，其差和比例是有意義的

- 範例：溫度、電子現金

可用以描述屬性的運算特性

- 差異性： $= \neq$
- 順序性： $< >$
- 加減： $+ -$
- 乘除： $* /$

- 名目屬性：差異性
- 順序屬性：差異性、順序性
- 區間屬性：差異性、順序性、加減
- 比例屬性：四種皆可

表 2.2 不同的屬性型態

屬性型態	描述	例子	運算
類別 (定性)	名目	每一個名目屬性值都有不同的名稱，而且物件之間也可以彼此區別 ($=, \neq$)	郵遞區號，員工編號，眼睛顏色，性別
	順序	其值足以排序物件間的先後順序 ($<, >$)	金屬硬度，{好, 較好, 最好}，成績，門牌號碼
數值 (定量)	區間	對區間值而言，每一個數值都有意義，可做 $(+, -)$	日期，華氏或攝氏溫度
	比例	對比例尺度而言，其差和比例是有意義的 $(*, /)$	溫度，電子現金

表 2.3 屬性型態間的轉換

屬性型態		轉換	建議
類別 (定性)	名目	1 對 1 的映射	如果員工編號重新給定，也不會造成任何差異
	順序	保留原有順序	如果其值為好、較好、最好，則可以用{1,2,3}來表示
數值 (定量)	區間	新值 = $a * \text{舊值} + b$, a 、 b 是常數	華氏與攝氏溫度的衡量單位不一樣，但可以互相轉換
	比例	新值 = $a * \text{舊值}$	長度可以用尺或腳來衡量

離散型和連續型屬性

- 離散型屬性

- 屬性是**有限的**或是**可數的**
- 範例：郵遞區號
- 通常以**整數值**表示
- 注意：二元屬性是離散型屬性的一個例外情形

- 連續型屬性

- 屬性值通常為**實數**
- 範例：氣溫
- 通常以**浮點數**來表示

資料集的型態

- 記錄型資料

- 資料矩陣
- 文件資料
- 交易資料

- 圖形資料

- 全球資訊網 (World Wide Web)
- 分子結構

- 順序資料

- 時序資料
- 序列資料
- 時間序列資料
- 空間資料

對資料探勘具重大影響的資料特性

- 維度

- 資料集的維度事實上就是物件的屬性，維度愈高的資料愈難分析，有時稱為維度的魔咒（curse of dimensionality）

- 稀疏性

- 對一些非對稱屬性資料而言，也許僅1%的資料是不為0；可是實際上，因為只有非0的數值需要被儲存和運算，因此節省很多時間和儲存空間，所以也算是稀疏資料的一項優點

- 解析度

- 不同解析度的資料其特性差異很大，例如在以公尺為單位的解析度上看地球的表面是很不平的，但是在以公里為單位的解析度上來看卻又相對平坦。所以如果解析度太大，那麼有些特性可能會因此消失

記錄資料

- 資料集包含很多固定欄位的記錄

<i>Tid</i>	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

資料矩陣

- 如果資料物件都有一些相同的屬性，那麼這些資料物件就可以視為一個多維空間中的一點或是向量，其中每個維度表示一個屬性
- 這些資料物件可以解釋成 $m \times n$ 的矩陣，其中m列表示**物件**；n行表示**屬性**

Projection of x Load	Projection of y Load	Distance	Load	Thickness
10.23	5.27	15.22	27	1.2
12.65	6.25	16.22	22	1.1
13.54	7.23	17.34	23	1.2
14.27	8.43	18.45	25	0.9

文件資料

- 如果文件中的某些字詞可被忽略，那麼其文件就可以形成一個字詞向量，也就是將字詞視為一個屬性，如此一來所形成的矩陣稱為文件—字詞矩陣（document-term matrix）

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

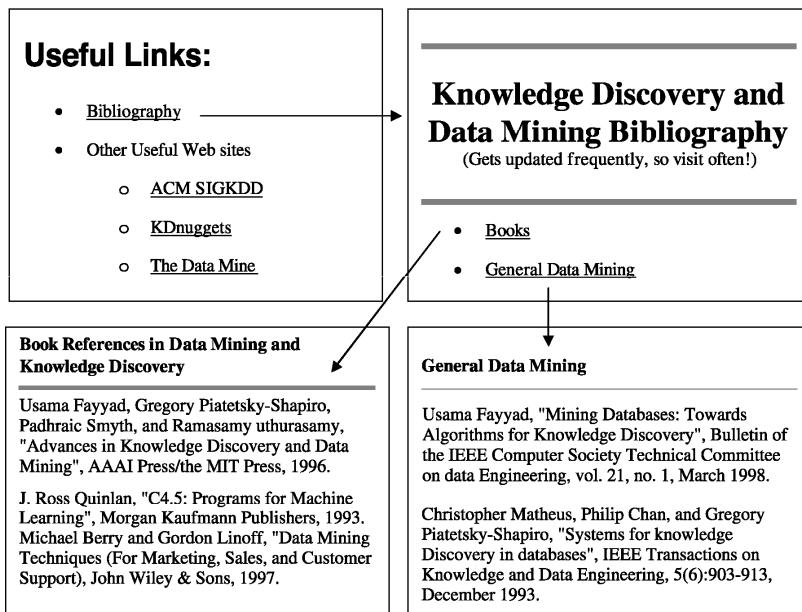
交易資料

- 是一種特殊的記錄資料類型
 - 每一筆記錄（或稱交易）都包含很多產品項目
 - 範例：顧客在超市的購買記錄。下圖的每一列，表示顧客在特定時間所購買的產品項目

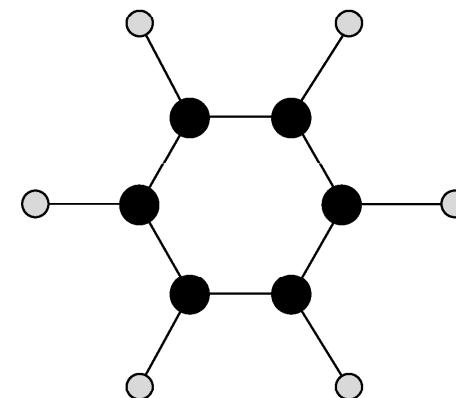
<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

圖形資料

● 範例：網頁連結和化學元素的結構



(a) 網頁連結



(b) 芳分子

順序資料

● 時序交易資料

時間	顧客	購買的項目
t1	C1	A, B
t2	C3	A, C
t2	C1	C, D
t3	C2	A, D
t4	C2	E
t5	C1	A, E

顧客	時間與購買的項目
C1	(t1: A,B) (t2:C,D) (t5:A,E)
C2	(t3: A, D) (t4: E)
C3	(t2: A, C)

順序資料

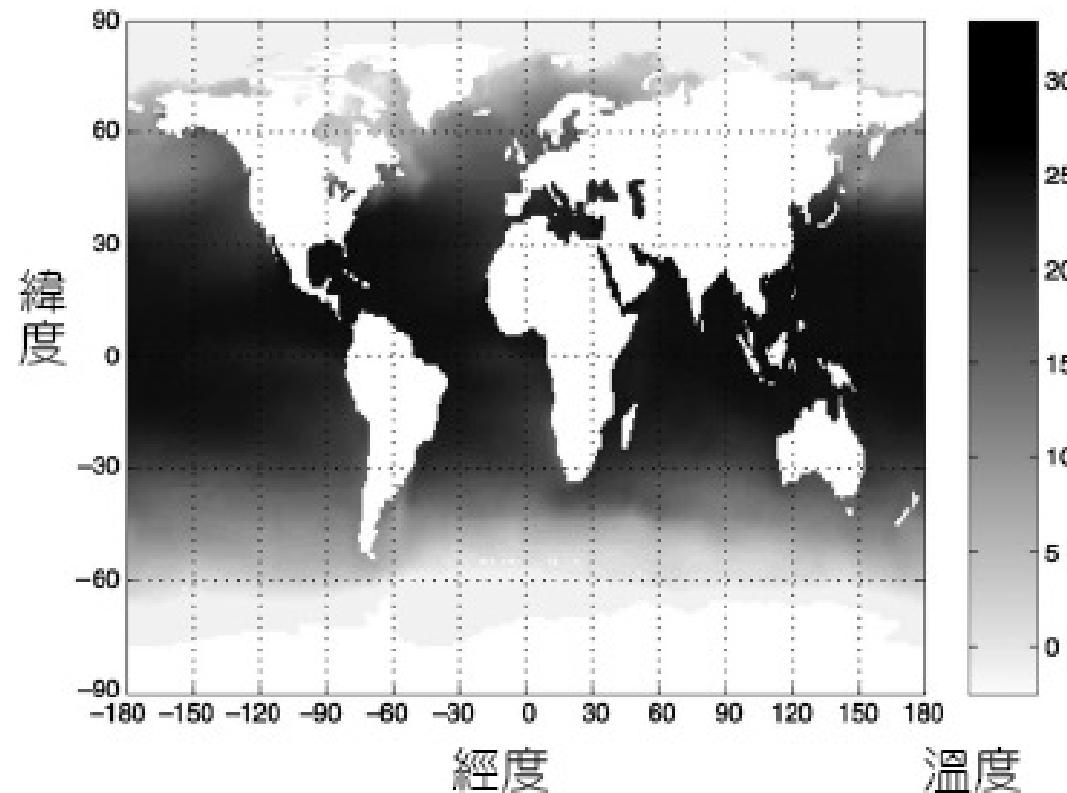
- 基因序列資料
 - 很多問題和基因序列類似，如預測結構的相似度

GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCgcCTGGCGGGCG
GGGGGAGGCAGGGGCCGCCCAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCAGCAGCGAACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG

人類基因編碼

順序資料

- 空間性暫時資料

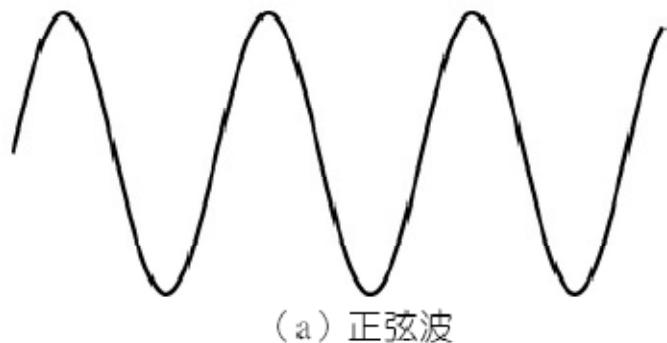


資料品質

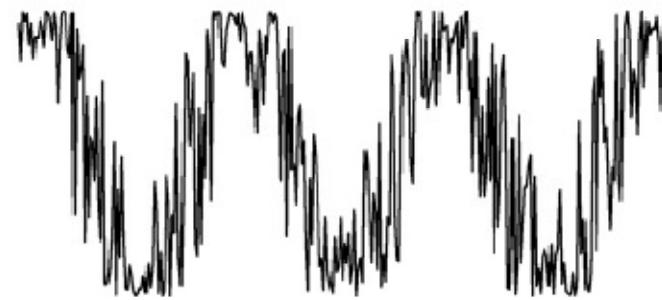
- 有哪些資料品質的問題？
 - 如何偵測資料的問題？
 - 我們如何處理這些問題？
-
- 資料品質的問題包括：
 - 雜訊和離群值
 - 遺漏值
 - 重複性資料

雜訊

- 雜訊值有可能來自於測量誤差，包含一些資料的扭曲或是不實
 - 下圖是一個刪除雜訊值前後的時間序列



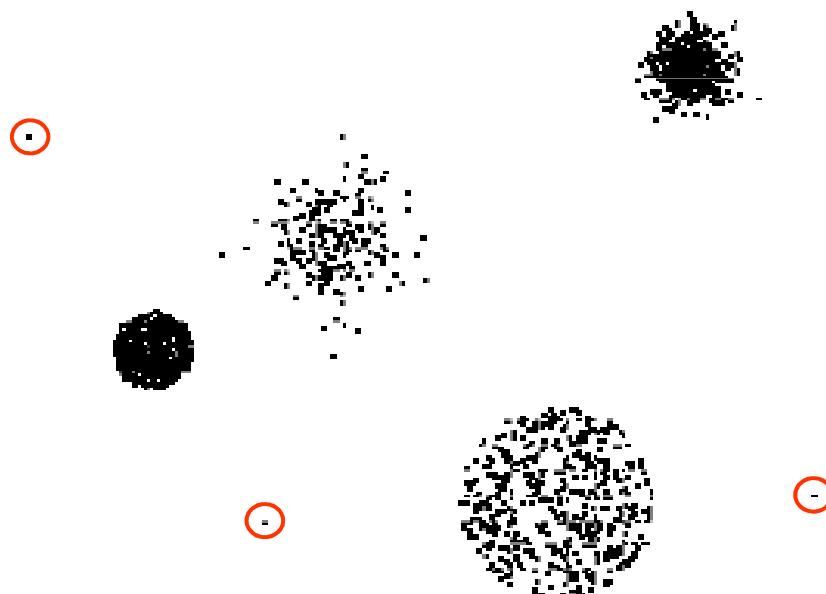
(a) 正弦波



(b) 具雜訊值的正弦波

離群值

- 離群值可能是因為資料物件的某些特性和其他物件不一樣，或者是其屬性值較不常出現在其他物件中



遺漏值

- 會有遺漏值的原因
 - 某些資料無法完整搜集
(例如，在問卷回答的過程中，很多人不想揭露年齡或是體重)
- 處理遺漏值的方法
 - 刪除資料物件
 - 估計遺漏值
 - 在分析過程中忽略遺漏值

重複性資料

- 資料有可能包含重複的物件或者是幾乎都是重複的物件
 - 整合來自不同來源的資料時
- 範例：一個人有多個不同的電子郵件帳號
- 資料清理（data cleaning）
 - 處理重複性資料的過程
 - 要避免不小心結合一些相似但非重複的物件

資料前處理

- 聚合 (Aggregation)
- 抽樣
- 維度縮減
- 特徵選取
- 特徵的產生
- 離散化及二元化
- 變數的轉換

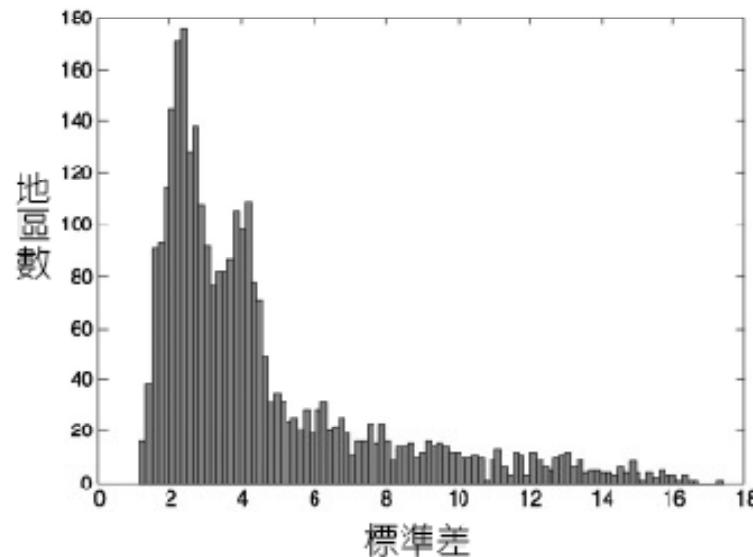
聚合

- 聚少成多
- 假設有一個記錄產品在各分店每日交易的資料，我們可以用聚合的觀念將每個分店的每日銷售額彙總出來，如此一來資料量就可以大幅降低
- 運算時可節省記憶體
- 缺點是看不到細節變化

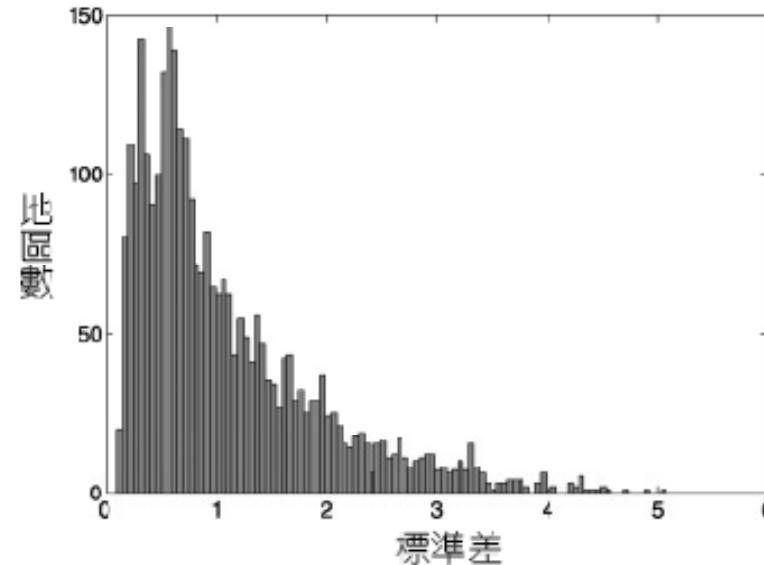
聚合的範例

澳洲降雨量

變化較小



(a) 平均月降雨量標準差



(b) 平均年降雨量標準差

圖 2.8 ▶ 1982–1993 年澳洲降雨量的資料

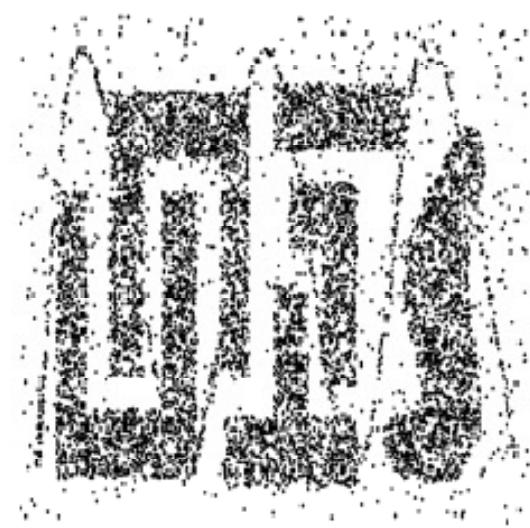
抽樣

- 抽樣是用來選取欲分析資料的主要技術
 - 通常用在**資料調查及資料分析**上
- 統計學上的抽樣主要在於要得到所有資料太過耗時
資料探勘的抽樣主要在於計算的時間太過耗時
- 有效的抽樣原則在於樣本必須是具有**代表性**：
 - 抽樣的樣本所得到的結果會和整個原始資料的結果很接近
 - 如果某一個資料的平均數很接近整體資料的平均數，那麼就具有**代表性**

抽樣的方法：隨機抽樣

- 隨機抽樣
 - 每個項目被選取的機率是一樣的
- 抽樣後不放回
 - 被抽中的樣本就不再繼續抽
- 抽樣後放回
 - 被抽中的樣本有可能繼續被抽中
- 分層抽樣法 (Stratified sampling)
 - 第一種分層抽樣法，是假設不同類型中的資料不管數量多少都有相同的抽樣機率；
 - 第二種分層抽樣法是依照類型中的資料所佔比例來決定其抽樣的個數

樣本的大小



(a) 8000 樣本點



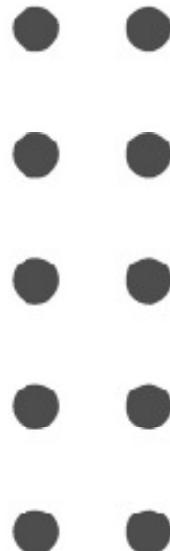
(b) 2000 樣本點



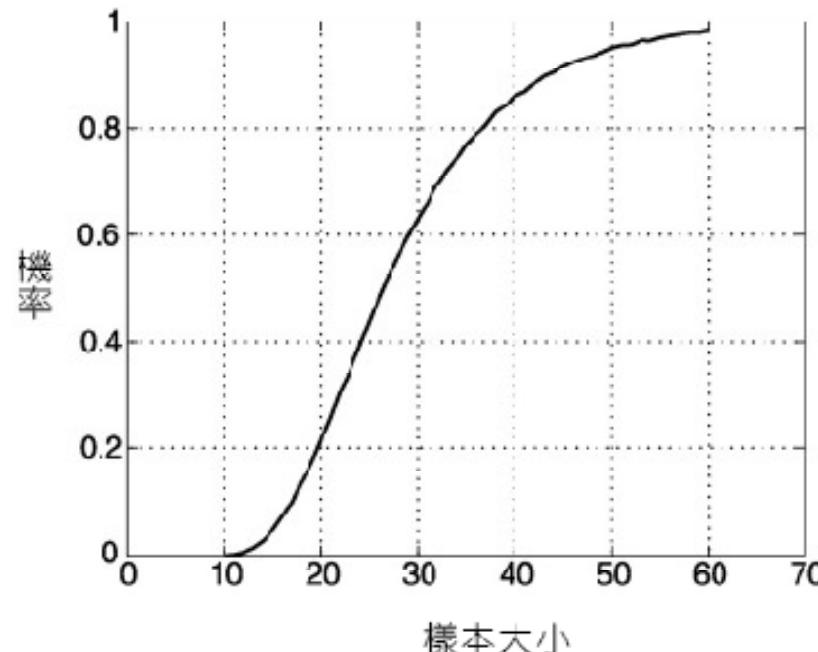
(c) 500 樣本點

樣本大小

- 圖 (a) 假設群體數很少，只有10個。圖 (b) 顯示從10個群體中各抽出一個物件的機率，其樣本大小從10到60



(a) 10 群樣本點



(b) 樣本點包含所有 10 個群體的機率

維度的問題

- 當維度增加時，資料分析的工作會變得很困難，因為它可能會增加空間中的稀疏性
- 對**分類**和**分群**問題而言，**資料密度**以及**樣本點間的距離**是很重要的，可是卻因為維度太多而變得沒有意義

縮減維度

- 目的

- 避免維度的問題
- 降低資料探勘演算法所需的時間和記憶體
- 讓資料更易於用視覺化方式呈現出來
- 協助刪除掉一些無關的特徵或是雜訊值

- 技術

- 主成份分析 (PCA)
- 奇異值分解法 (SVD)

特徵選取

- 另一個縮減資料維度的方法
- 重複的特徵
 - 是指大部分資訊包含一個或多個其他屬性
- 無關的特徵
 - 是指大部分的資訊是不可用的情形。例如學生的學號就和成績的表現無關

特徵選取

- 技術

- 嵌入法

- ◆ 在資料探勘演算的過程中，可以自己決定所要用的屬性或是必須忽略的屬性

- 過濾法

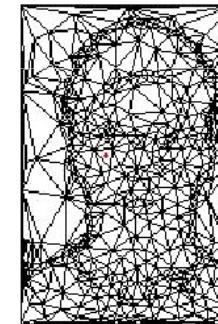
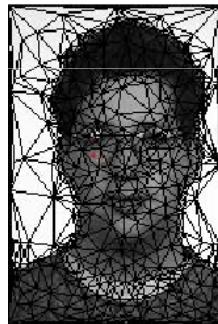
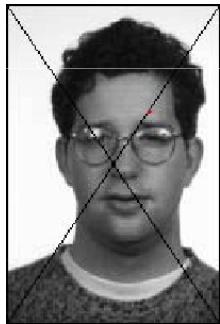
- ◆ 在進行資料探勘之前，可以先選擇一些相關性較低的屬性

- 包裝法

- ◆ 是將資料探勘視為一個黑盒子，利用此黑盒子來找到最好的屬性，但是不會處理所有可能的特徵組合

特徵的產生

- 新的屬性通常是從原始屬性中建立出來的，而且新的屬性個數一定會比原始屬性個數要來的少
- 三種常見的方法
 - 特徵的萃取 (不同應用領域，萃取方法各有差異)



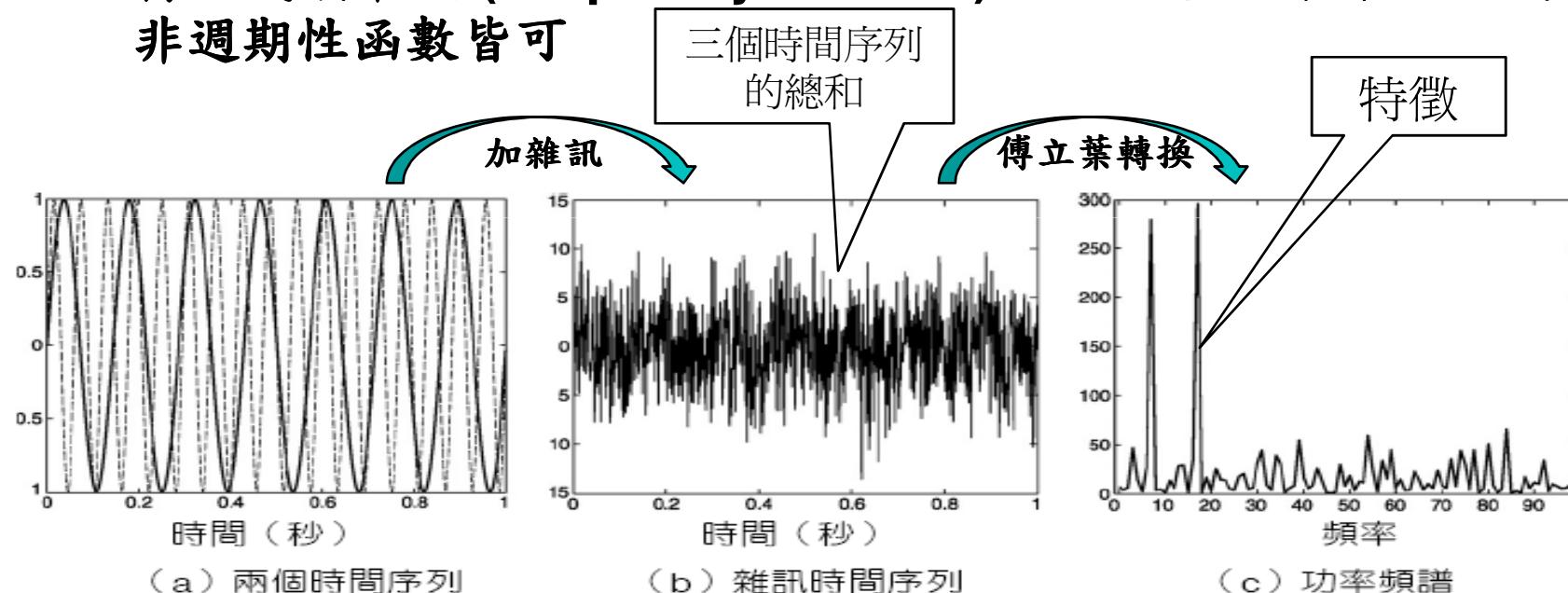
人臉偵測
(Face Detection)
影像正規化
(Normalization)
特徵偵測
(Feature Detection)
活體偵測
(Factuality Detection)
遮沒物偵測
(Occlusion Detection)
光線自動補償
(Lighting Compensation)
可辨識性分析
(Recognizability Analysis)

- 將資料映射到新的空間
- 特徵的建構

將資料映射到新的空間

●傅立葉轉換

- Fourier 轉換能把任意的時域 (time domain) 函數以數學方法轉換成頻率域 (frequency domain) 函數，包括周期性函數與非週期性函數皆可



● 波轉換 (Wavelet)

離散小波轉換

1. 離散小波轉換(Discrete Wavelet Transformation, DWT)與DCT相似，都是一種將**空間域影像**轉換成**頻率域影像**的技術。
2. 低頻特性：
 1. 肉眼對**低頻敏**感度較高
 2. 像素與像素之間的**變化小**
 3. 影像較**平滑**、細緻且清楚
 4. 低頻部分的值稍有改變，人的眼睛一看就知道

離散小波轉換 (Cont.)

3. 高頻特性：
 1. 肉眼對**高頻敏**感度較低
 2. 像素與像素之間的**變化大**
 3. 影像較**粗糙**、模糊
 4. 高頻部分的值稍有改變，肉眼是無法清楚的看出的

Haar的離散小波轉換

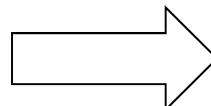
- Harr函數離散小波轉換的運算大致上有兩個步驟：
 - 一為**水平分割**，另一為**垂直分割**，水平分割是讀取**係數**的順序是依照**水平方向**由左至右來取；儲存時是水平方向儲存。
 - **垂直分割**是讀取係數的順序是垂直方向由上至下來取；儲存時是垂直方向儲存。

Haar的離散小波轉換 (Cont.)

Harr函數離散小波轉換過程：

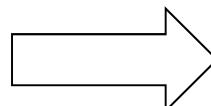
步驟一：第一次水平分割

A	B	C	D



A+B	C+D	A-B	C-D
L		H	

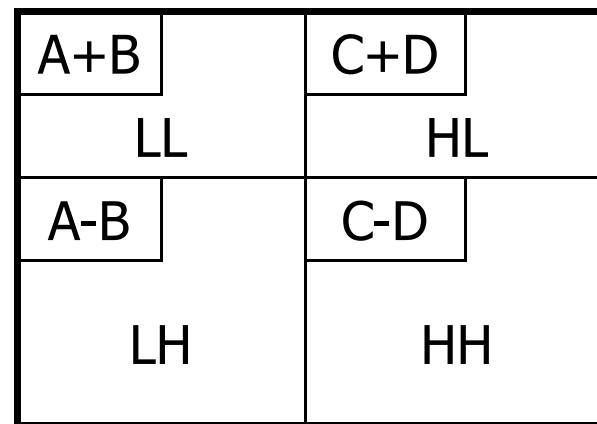
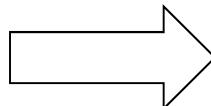
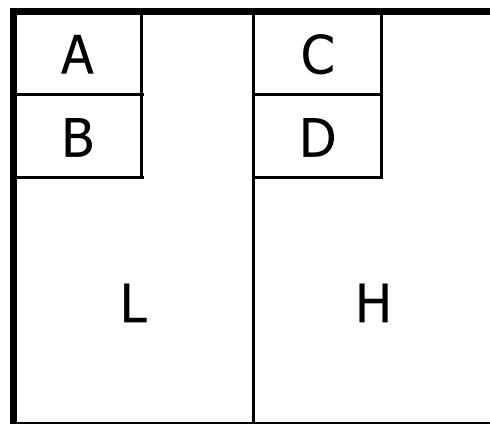
(0,0)	(0,1)	(0,2)	(0,3)
(1,0)	(1,1)	(1,2)	(1,3)
(2,0)	(2,1)	(2,2)	(2,3)
(3,0)	(3,1)	(3,2)	(3,3)



(0,0)	(0,1)	(0,2)	(0,3)
(1,0)	(1,1)	(1,2)	(1,3)
(2,0)	(2,1)	(2,2)	(2,3)
(3,0)	(3,1)	(3,2)	(3,3)

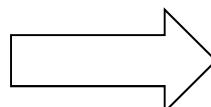
Haar的離散小波轉換 (Cont.)

步驟二：第一次垂直分割



A 4x4 matrix representing the input image:

(0,0)	(0,1)	(0,2)	(0,3)
(1,0)	(1,1)	(1,2)	(1,3)
(2,0)	(2,1)	(2,2)	(2,3)
(3,0)	(3,1)	(3,2)	(3,3)



L H

原始影像 O

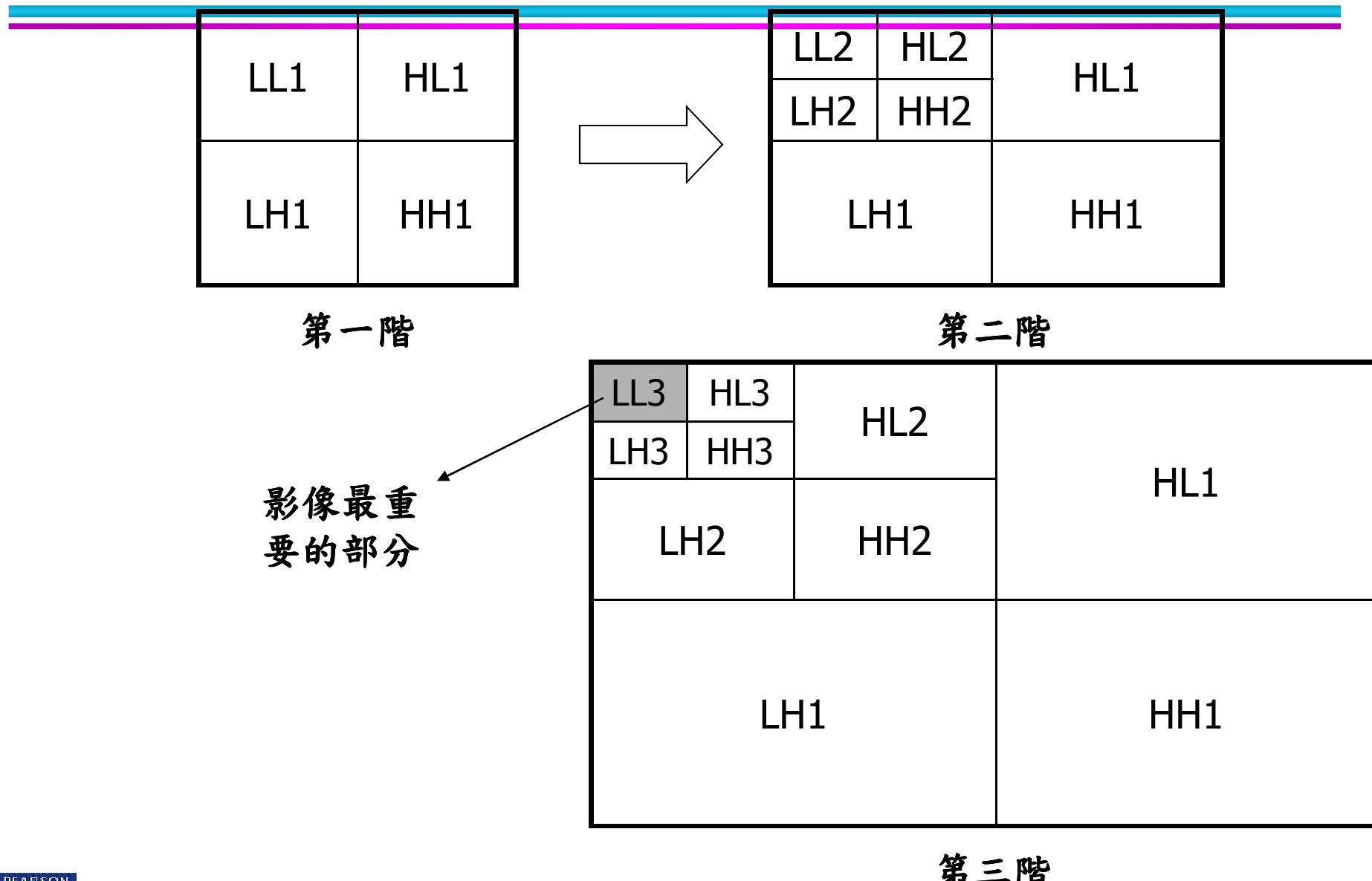
The LL submatrix from the first vertical Haar transform:

(0,0)	(0,1)	(0,2)	(0,3)
(1,0)	(1,1)	(1,2)	(1,3)
(2,0)	(2,1)	(2,2)	(2,3)
(3,0)	(3,1)	(3,2)	(3,3)

LL HL

Harr函數離散小波轉換第一次
水平分割後結果 O'

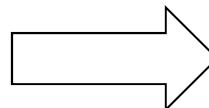
Haar的離散小波轉換 (Cont.)



Haar的離散小波轉換 (Cont.)

20	15	30	20
17	16	31	22
15	18	17	25
21	22	19	18

原始影像 O

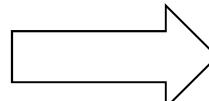


35	50	5	10
33	53	1	9
33	42	-3	-8
43	37	-1	1

第一次水平分割的結果 O'

68	103	6	19
76	79	-4	-7
2	-3	4	1
-10	5	-2	-9

第一次垂直分割的結果 O''



326	-38	6	19
16	-32	2	-7
2	-3	4	1
-10	5	-2	-9

第二階離散小波轉換的結果
(即第一階離散小波轉換的結果)

特徵的架構

- 原始資料中的**特徵**有時也是**重要資訊**，但不一定適合資料探勘演算法使用
- 例：歷史手工藝資料分類
 - 手工藝由各式各樣的材料(e.g., 木頭、黏土)組成
 - 可採 特徵密度 = 質量 / 總容積
 - 需要**領域專家**參與

離散化與二元化

- 連續資料不利於某些分類演算法，或資料類別很多，但部份類別不常發生，可將不常發生的類別予以合併以降低類別個數

- 二元化

類別值	總數	x_1	x_2	x_3
Awful	0	0	0	0
Poor	1	0	0	1
OK	2	0	1	0
Good	3	0	1	1
Great	4	1	0	0

- 離散化

- ◆ 決定多少類別 (亦即決定多少分割點)
- ◆ 將所有值對映到類別中
- ◆ 離散化後的表示法: $\{(x_0, x_1], (x_1, x_2], \dots, (x_{n-1}, x_n)\}$
或 $x_0 < x \leq x_1, \dots, x_{n-1} < x < x_n$

離散化

- 監督式

- 類別已知

- ◆ 先將資料分成多個區間，再將相似的區間合併起來

- 非監督式

- 類別未知

- 方法

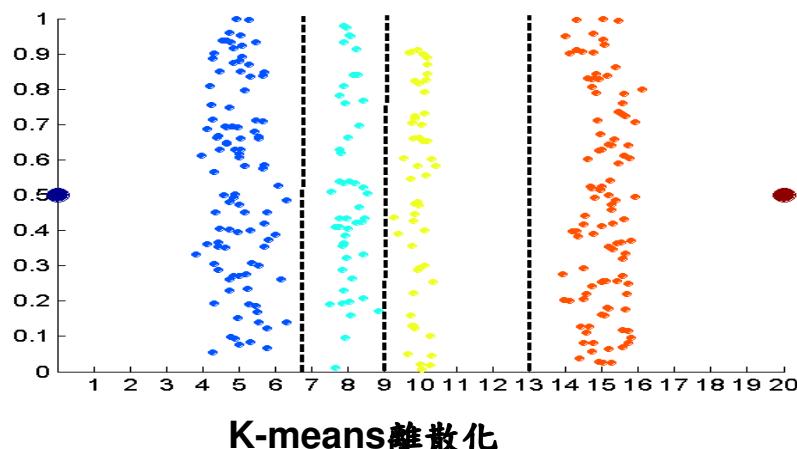
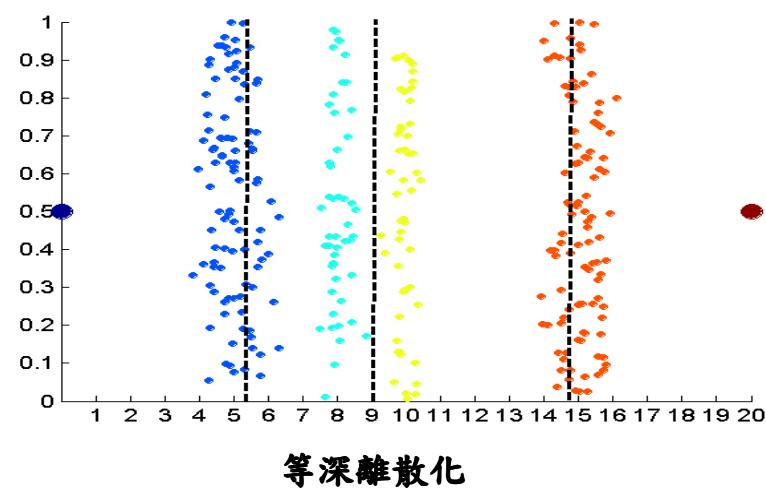
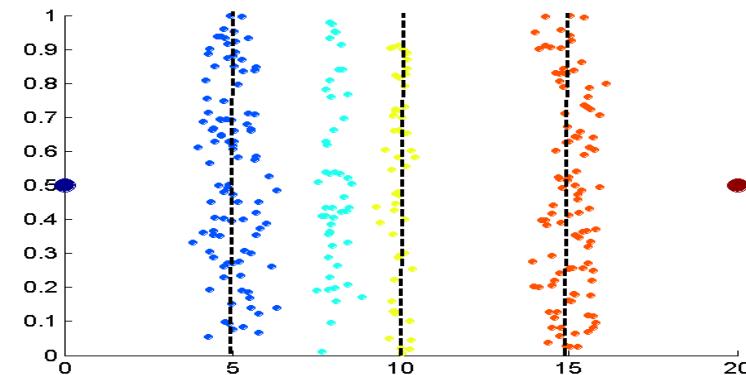
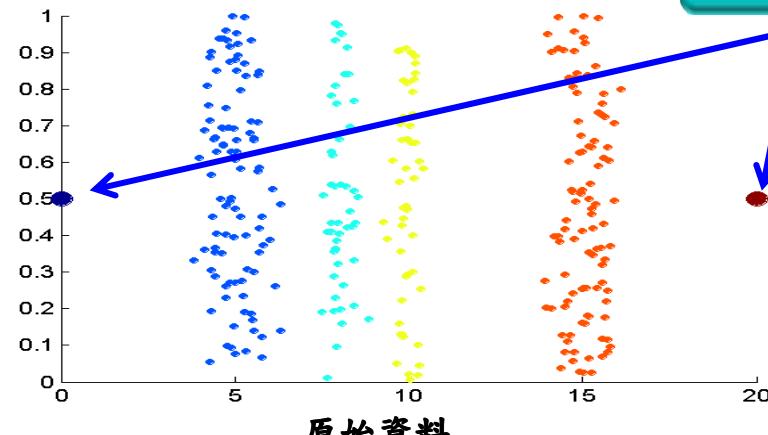
- ◆ 等距區間 (等寬)，此法易受離散值影響

- ◆ 等量法 (等深)，讓區間中的資料個數相等

- ◆ K-means

非監督式離散化

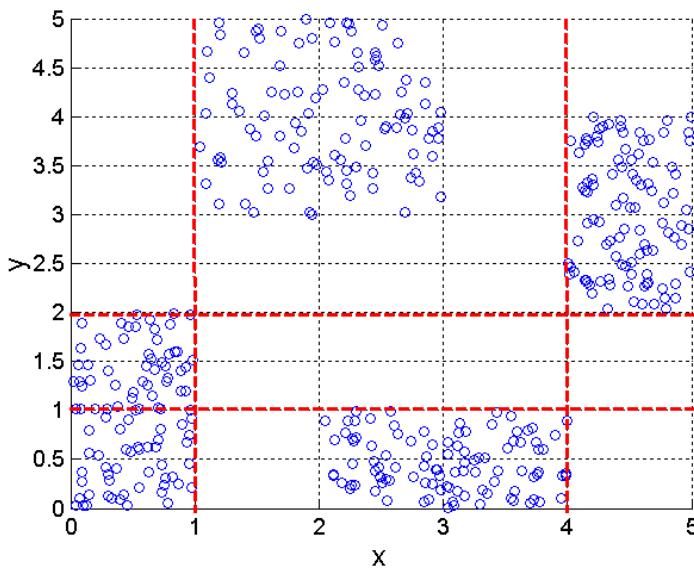
離群值



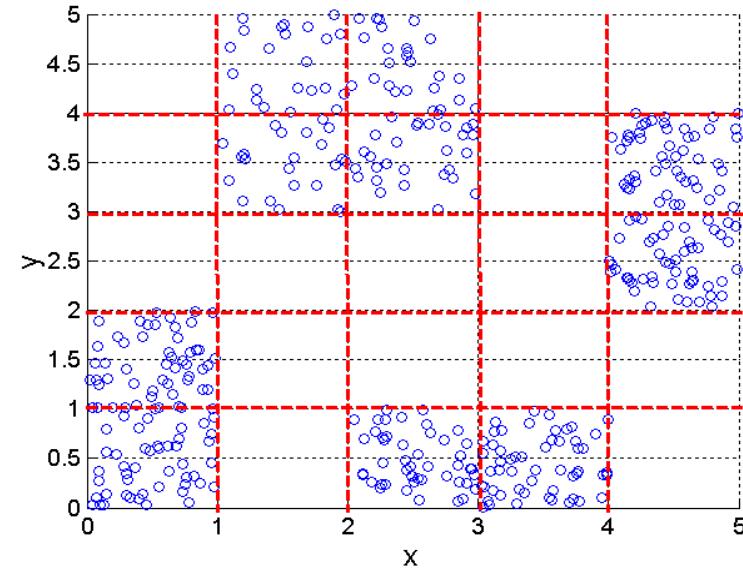
監督式離散化

● 使用亂度 (Entropy) 方法

- 一個區間內只包含一個類別，是很單純的情形，則其亂度為0，如果區間中包含很多類別，那表示亂度值會很高
- 假設區間包含順序性的資料值，其分割程序是不斷將資料歸類至其他區間，通常是選擇亂度最高的區間，直到使用者所定義的區間個數或滿足停止條件為止



三個區間



五個區間

變數的轉換

- 變數的轉換可以應用到所有變數上（**變數**其實就是**屬性**），換句話說，例如只有一個變數很重要，但有很大的值，那麼其值就可以用**絕對值**進行轉換
 - 簡單函數轉換： x^k , $\log(x)$, e^x , $|x|$
 - ◆ 統計學上，若變數不符合高斯(常態)分佈時，可將變數取sqrt, log或 $1/x$
 - ◆ 資料轉換時要考量資料特性 例如原為{1, 2, 3} 以 $1/x$ 轉換則為{1, $1/2$, $1/3$ }，調整後順序變了
 - 正規化
 - ◆ 目的：是將**變數**轉換成**常態分配**，目的在使整個值的單位能一致
 - ◆ 例：年齡與收入，有兩個人，其收入(幾百或幾千美金)，和年紀(不超過 150)的差異很大，在未考慮差異性時比較這兩個人的相似度時會有問題

變數的轉換 (Cont.)

- 正規化(Normalization)：

- 假設有三組資料，將每組資料的數值調整成介於0~1之間，以取得**單位的一致性**與**可比較性**，常用於**數量研究方法**。

- 標準化(Standardization)：

- 假設有三組資料，將每組資料減去該組之**平均數**再除以該組之**標準差**的過程就叫**標準化**，每組資料此時資料的中心值會變成0、標準差會變成1，即成為標準常態分配，值會介於 $-X \sim X$ 之間，以取得單位的一致性與可比較性，常用於**統計分析**。

相似度與不相似度

- 相似度

- 相似度表示物件間**相同的程度**
- 物件之間的相似度愈高，其物件愈相像
- 其值僅會介於0~1之間

- 不相似度

- 不相似度表示**兩個物件間**差異的程度
- **不相似度**和**距離**其實是同義字，距離愈大，不相似度愈高
- 其值介於0—1之間，但是有時其範圍可以是0到 ∞
- 鄰近值（proximity）來表示相似度與不相似度

簡單屬性間的相似度與不相似度

- 順序屬性值 {p1 = 很好, p2=好, p3 = OK, p4 = 劣, p5 = 差}，如何比較 p1 與 p2 相似度?
 - 給定 {很好 = 4, 好 = 3, OK = 2, 劣 = 1, 差 = 0} 則
 - 不相似度為 $d(p1, p2) = (4-3)/4 = 0.25$ ，相似度為 $1-0.25 = 0.75$

下表是各種屬性型態的不相似度及相似度之計算方法，其中兩個物件x與y，各有一個屬性，而 $d(x,y)$ 與 $s(x,y)$ 分別表示不相似度及相似度

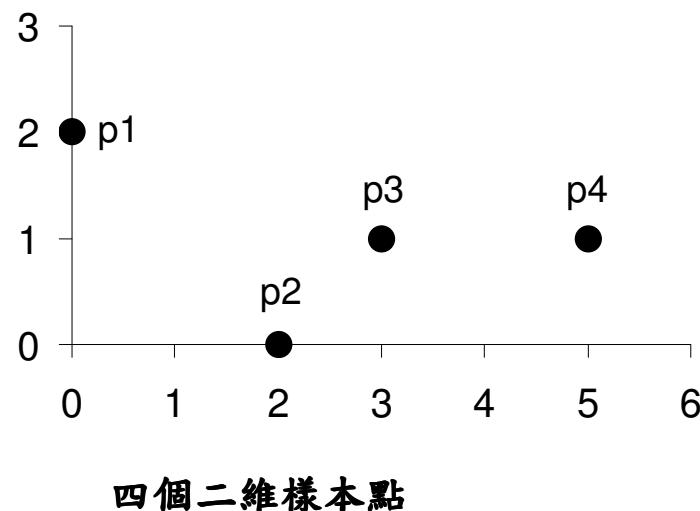
屬性型態	不相似度	相似度
名目	$d = \begin{cases} 0 & \text{若 } x = y \\ 1 & \text{若 } x \neq y \end{cases}$	$s = \begin{cases} 1 & \text{若 } x = y \\ 0 & \text{若 } x \neq y \end{cases}$
順序	$d = x - y /(n - 1)$ (將值對映至整數 0~n-1 之值，其中 n 為數值)	$s = 1 - d$
區間或比例	$d = x - y $	$s = -d$, $s = \frac{1}{1+d}$, $s = e^{-d}$, $s = 1 - \frac{d-min_d}{max_d-min_d}$

歐幾里德距離

- 歐幾里德距離

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{k=1}^n (x_k - y_k)^2}$$

其中 n 是指維度個數，而 x_k 及 y_k 分別表示 \mathbf{x} 與 \mathbf{y} 的第 k 個屬性



Point	x 軸	y 軸
p1	0	2
p2	2	0
p3	3	1
p4	5	1

\mathbf{x} 與 \mathbf{y} 座標軸上的四個點

	p1	p2	p3	p4
p1	0.0	2.8	3.2	5.1
p2	2.8	0.0	1.4	3.2
p3	3.2	1.4	0.0	2.0
p4	5.1	3.2	2.0	0.0

距離矩陣

Minkowski (閔可夫斯基)距離

- Minkowski 距離是由歐幾里德距離衍生而來

$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_{k=1}^n |x_k - y_k|^r \right)^{1/r}$$

其中r是一個參數，n是指維度個數，而 x_k 及 y_k 分別表示x與y的第k個屬性

Minkowski 距離：範例

- 當 $r = 1$ 時 (L_1 範數(Norm))，常見的例子是漢明(Hamming)距離，它是用來計算物件間有多少個不同的位元個數
- $r = 2$ ，則是用歐幾里德距離 (L_2 範數)。
- $r = \infty$ ，是指物件間任何屬性的最大距離 (L_{\max} 或 L_∞ 範數)，其的公式如下：

$$\lim_{n \rightarrow \infty} \left(\sum_{k=1}^n |x_k - y_k|^r \right)^{1/r} = \max_{k=1}^n |x_k - y_k|$$

- r 的參數和 n 維的個數不一樣，所以從以上情形中，我們可以瞭解不同的屬性個數所用的公式不一樣。.

Minkowski 距離

Point	x 軸	y 軸
p1	0	2
p2	2	0
p3	3	1
p4	5	1

L ₁	p1	p2	p3	p4
p1	0.0	4.0	4.0	6.0
p2	4.0	0.0	2.0	4.0
p3	4.0	2.0	0.0	2.0
p4	6.0	4.0	2.0	0.0

L _∞	p1	p2	p3	p4
p1	0.0	2.0	3.0	5.0
p2	2.0	0.0	1.0	3.0
p3	3.0	1.0	0.0	2.0
p4	5.0	3.0	2.0	0.0

距離矩陣

距離公式常見的衡量方法

- 歐幾里德距離公式是用常見的衡量方法，如果有兩個點 x 與 y ，那麼其 $d(x, y)$ 具有以下特性：
 1. 正向性
 - (a) $d(x, y) \geq 0$ ，對所有 x 與 y 而言
 - (b) $d(x, y) = 0$ ，只有當 $x = y$ 時
 2. 對稱性
 - $d(x, y) = d(y, x)$ ，對所有的 x 與 y 而言
 3. 三角不等式
 - $d(x, z) \leq d(x, y) + d(y, z)$ ，對所有的 x, y 與 z 的點而言
- 滿足以上三個條件稱為metrics（度量）

距離公式常見的衡量方法 (Cont.)

- 非度量性的不相似度—時間

$$d(t_1, t_2) = \begin{cases} t_2 - t_1 & \text{if } t_1 \leq t_2 \\ 24 + (t_2 - t_1) & \text{if } t_1 > t_2 \end{cases}$$

- $d(2\text{PM}, 1\text{PM}) = ?$
- $d(1\text{PM}, 2\text{PM}) = ?$

相似度的特性

- 對相似度而言，三角不等式通常很難成立，但是一定會具有對稱性以及正向性，例如 $s(x, y)$ 是 x 與 y 樣本點間的相似度，如下：
 1. $s(x, y) = 1$ ，只有當 $x = y$ 時 ($0 \leq s \leq 1$)
 2. $s(x, y) = s(y, x)$ ，對所有 x 與 y 而言（對稱）

二元資料間的相似度

- 二元資料間的相似度值稱為**相似係數**，通常介於0–1之間，其值為1表示物件間具有**完全相關**的特性，為0表示**完全不相關**
- 假設 x 與 y 是具有 n 個二位元屬性的物件，而所產生的二個向量有以下四種關係

f_{00} : $x = 0$ 且 $y = 0$ 的屬性個數

f_{01} : $x = 0$ 且 $y = 1$ 的屬性個數

f_{10} : $x = 1$ 且 $y = 0$ 的屬性個數

f_{11} : $x = 1$ 且 $y = 1$ 的屬性個數

二元資料間的相似度

- 簡單配對係數 (simple matching coefficient, SMC)

$$SMC = \frac{\text{同時為 1 或同時為 0 的屬性個數}}{\text{屬性個數}} = \frac{f_{11} + f_{00}}{f_{01} + f_{10} + f_{11} + f_{00}}$$

是計算所有向量中同時為 1 或同時為 0 的元素個數比例。

- Jaccard (杰卡德) 係數

$$J = \frac{\text{同時為 1 的屬性個數}}{\text{不同時為 0 的屬性個數}} = \frac{f_{11}}{f_{01} + f_{10} + f_{11}}$$

SMC 和 Jaccard：範例

分析x, y 兩筆交易之相似度，其中 0 表未購買, 1 表示購買

$$x = (1, 0, 0, 0, 0, 0, 0, 0, 0, 0)$$

$$y = (0, 0, 0, 0, 0, 0, 1, 0, 0, 1)$$

$f_{01} = 2$: x = 0 且 y = 1 的屬性個數

$f_{10} = 1$: x = 1 且 y = 0 的屬性個數

$f_{00} = 7$: x = 0 且 y = 0 的屬性個數

$f_{11} = 0$: x = 1 且 y = 1 的屬性個數

$$SMC = \frac{f_{11} + f_{00}}{f_{01} + f_{10} + f_{11} + f_{00}} = \frac{0+7}{2+1+0+7} = 0.7$$

$$J = \frac{f_{11}}{f_{01} + f_{10} + f_{11}} = \frac{0}{2+1+0} = 0$$

Cosine 相似度

●文件通常用向量來表示，屬性是文件中某個字詞出現的次數

- 假設有兩個文件向量x與y

$$\cos(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}$$

其中 \bullet 是指向量的乘積，而 $\|\mathbf{x}\|$ 是向量的長度

- 範例：

$$\mathbf{x} = (3, 2, 0, 5, 0, 0, 0, 2, 0, 0)$$

$$\mathbf{y} = (1, 0, 0, 0, 0, 0, 0, 1, 0, 2)$$

$$\sqrt{\sum_{i=1}^n x^2}$$

$$\mathbf{x} \cdot \mathbf{y} = 3 * 1 + 2 * 0 + 0 * 0 + 5 * 0 + 0 * 0 + 0 * 0 + 0 * 0 + 2 * 1 + 0 * 0 + 0 * 2 = 5$$

$$\|\mathbf{x}\| = \sqrt{3 * 3 + 2 * 2 + 0 * 0 + 5 * 5 + 0 * 0 + 0 * 0 + 0 * 0 + 2 * 2 + 0 * 0 + 0 * 0} = 6.48$$

$$\|\mathbf{y}\| = \sqrt{1 * 1 + 0 * 0 + 0 * 0 + 0 * 0 + 0 * 0 + 0 * 0 + 0 * 0 + 1 * 1 + 0 * 0 + 2 * 2} = 2.24$$

$$\cos(\mathbf{x}, \mathbf{y}) = 0.31$$

當 cosine 夾角為0 時，其值為1，表兩文件相同，如果其值為0 表兩文件並沒有出現相同的字詞

Extended Jaccard 系數 (Tanimoto (谷本)系數)

- Extended Jaccard係數可以用在文件資料上，而且也鬆綁
Jaccard係數僅能處理二元屬性的限制，其方法也稱為Tanimoto
係數；但是有一些其他的係數也稱為Tanimoto係數。這個係數
我們將用EJ來表示，定義如下：

$$EJ(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - \mathbf{x} \cdot \mathbf{y}}$$

相關性

- 具有二元或是連續屬性的二個物件，可以用線性函數來計算其相關性，一般也稱為相似度
- 其中皮爾森相關係數的定義如下：

$$\text{corr}(\mathbf{x}, \mathbf{y}) = \frac{\text{共變異數}(\mathbf{x}, \mathbf{y})}{\text{標準差}(\mathbf{x}) * \text{標準差}(\mathbf{y})} = \frac{s_{xy}}{s_x s_y}$$

平均數：反映這些數字的中心位置



$$\text{共變異數}(\mathbf{x}, \mathbf{y}) = s_{xy} = \frac{1}{n - 1} \sum_{k=1}^n (x_k - \bar{x})(y_k - \bar{y})$$

- 通常兩個隨機變數之間是有關係的，亦即當知道其中之一的值之後，隱約可以猜測另一隨機變數的值，「身高」與「體重」就是一個明顯的例子。
- 例如，一個高個子的體重很可能也是比較重的。一般所稱的「高個子」指的是比「平均身高」高，而「比較重」意謂比「平均體重」重，所以我們定義一個數量來量化兩個隨機變數 X 與 Y 之間的相關程度，這個數量稱為 X 與 Y 的 「共變異數」(COVARIANCE)。

變異數：則反映這些數字變化的程度



相關性 (Cont.)

$$\text{標準差}(x) = s_x = \sqrt{\frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x})^2}$$

$$\text{標準差}(y) = s_y = \sqrt{\frac{1}{n-1} \sum_{k=1}^n (y_k - \bar{y})^2}$$

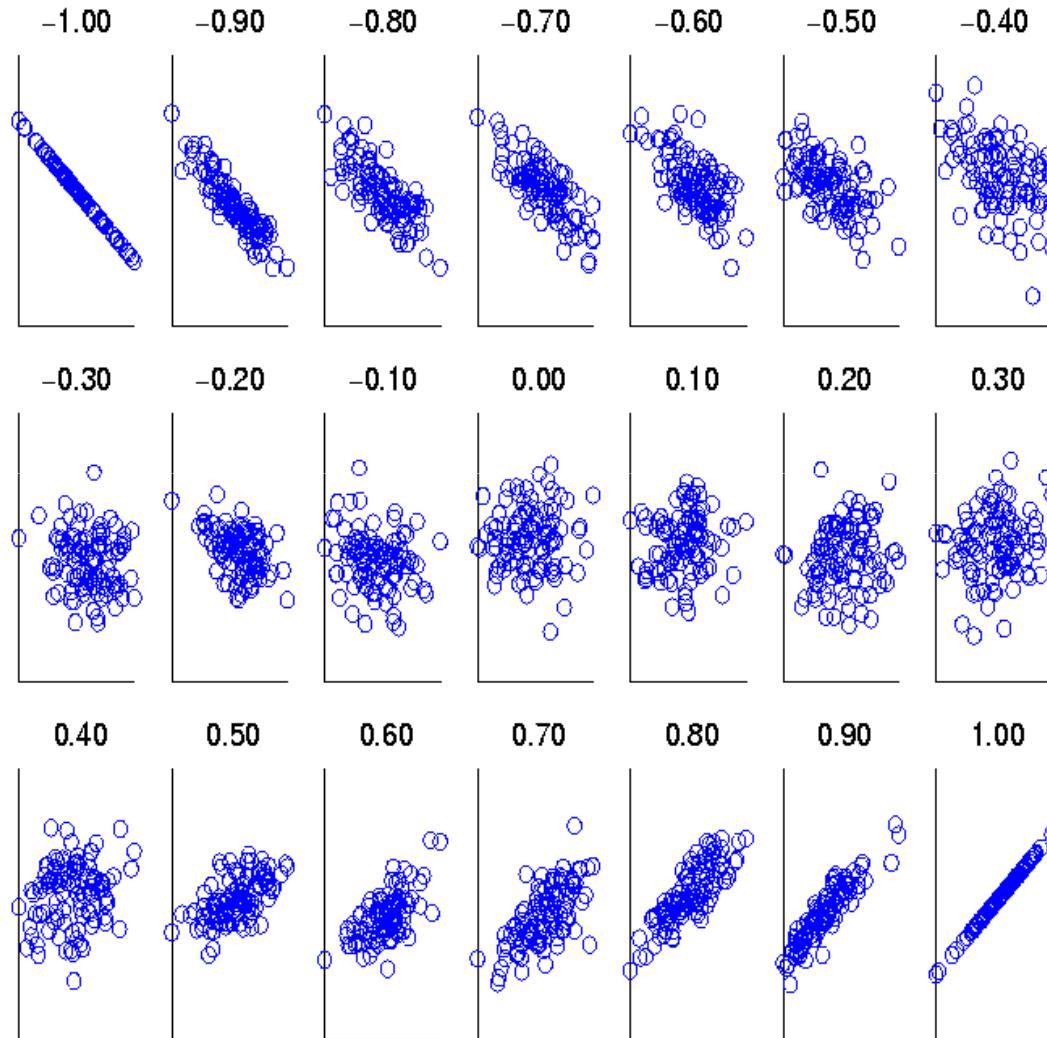


標準差是反映組內個體間的離散程度。簡單來說，標準差是一組數值自平均值分散開來的程度的一種測量觀念。一個較大的標準差，代表大部分的數值和其平均值之間差異較大；一個較小的標準差，代表這些數值較接近平均值。

相關性 (Cont.)

- 關係係數值介於 -1 到 1
- 完全相關
 - 係數值為 1 表**完全正相關**
 - 係數值為 -1 表**完全負相關**
- 非線性關係
 - 相關係數為 0 表**不具線性關係**，但仍**可能存在非線性關係**
- 視覺化(Visualizing)關係

視覺化關係



相關係數 -1 到
1 之間的散佈圖

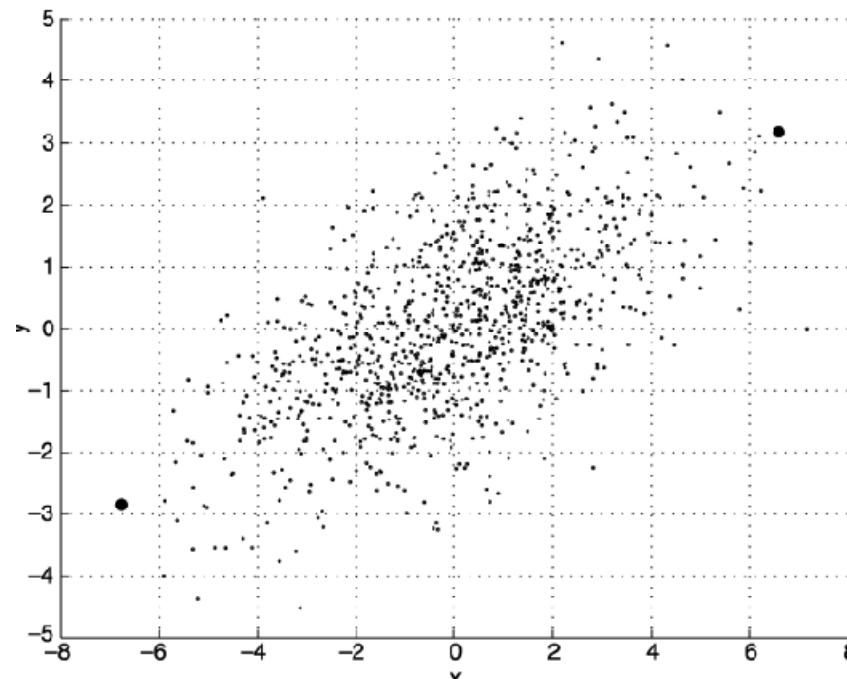
馬氏 (Mahalanobis) 距離

- 可用來處理屬性間具有相關性的問題
- 資料分配須接近於常態分配

反矩陣

$$\text{mahalanobis}(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})\Sigma^{-1}(\mathbf{x} - \mathbf{y})^T \quad \text{其中 } \Sigma^{-1} \text{ 是共變異矩陣的反矩陣}$$

計算較複雜，但適用於屬性間具高度相關的情形



$$A = \begin{bmatrix} 2 & 7 & 1 \\ 1 & 4 & -1 \\ 1 & 3 & 0 \end{bmatrix}$$

$$A^{-1} = \begin{bmatrix} \frac{-3}{2} & \frac{-1}{2} & \frac{11}{2} \\ \frac{1}{2} & \frac{1}{2} & \frac{-3}{2} \\ \frac{1}{2} & \frac{-1}{2} & \frac{-1}{2} \end{bmatrix}$$

$$AA^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

相關性為 0.6 馬氏距離是 6；而歐幾里德距離為 14.7

反矩陣

異質屬性的相似度

- 當屬性型態不一樣時，較為簡單的方法是分別計算每個屬性的相似度，然後將其結果整合成介於0–1之間的相似度，其作法通常是取其平均數

演算法 2.1 異質屬性的相似度

- 1: For the k^{th} attribute, compute a similarity, $s_k(\mathbf{x}, \mathbf{y})$, in the range [0,1].
- 2: Define an indicator variable, δ_k , for the k^{th} attribute as follows:
$$\delta_k = \begin{cases} 0 & \text{if the } k^{th} \text{ attribute is an asymmetric attribute and both objects have} \\ & \text{a value of 0, or if one of the objects has a missing value for the } k^{th} \\ & \text{attribute otherwise} \\ 1 & \end{cases}$$
3. Compute the overall similarity between the two objects using the following formula:

$$\text{相似度}(\mathbf{x}, \mathbf{y}) = \frac{\sum_{k=1}^n \delta_k s_k(\mathbf{x}, \mathbf{y})}{\sum_{k=1}^n \delta_k} \quad (2.15)$$

使用權重值

- 當權重不相同時
 - 如果權重和為 1：

$$\text{相似度}(p,q) = \frac{\sum_{k=1}^n w_k \delta_k s_k(\mathbf{x}, \mathbf{y})}{\sum_{k=1}^n \delta_k}$$

Mikowski距離可修改如下：

$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_{k=1}^n w_k |x_k - y_k|^r \right)^{1/r}$$

選擇適當的鄰近值衡量公式

- 對於一些**連續型**的資料，可以使用**歐幾里德距離**公式
- 對於**稀疏資料**，餘弦、Jaccard及extended Jaccard都可以處理這類型的問題
- 對於**時間序列**，若時間序列的**長度**很重要，可以使用**歐幾里德距離**公式；如果時間序列表示不同的值（像是血壓及氧氣消耗），那麼我們可以決定是否其時間序列具有相同的形狀，而不是相同的長度。而**相關係數**則可以用來量不同**長度**或是**等級**的問題

練習

- 令 $p_1 = (2, 3)$, $p_2 = (5, 1)$, $p_3 = (1, 4)$, $p_4 = (0, 1)$ ，請分別算出
 - 歐基理得距離矩陣
 - 閔可夫斯基之 L_1 , 與 L_∞ 之距離矩陣
- 令 $x = \{0, 1, 1, 0, 0, 0, 1, 0, 1, 1\}$, $y = \{1, 0, 1, 0, 1, 0, 0, 1, 0, 1\}$ ，請算出
 - SMC
 - Jaccard
- 令 $x = \{3, 2, 1, 0, 1, 0, 1, 0, 1, 1\}$, $y = \{5, 3, 1, 0, 1, 2, 0, 4, 0, 1\}$
 - $\cos(x, y)$
 - EJ