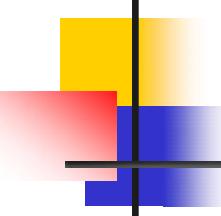


分群方法與軟體操作

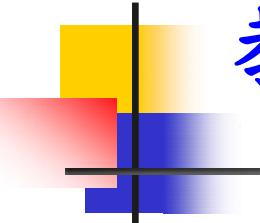


國立宜蘭大學資訊工程系
吳政璋助理教授
wucw@niu.edu.tw



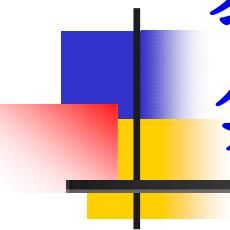
教學目標

- 了解集群分析的目的及應用
- 了解下列時序資料分群方法
 - 分割式分群法(Partitional Clustering)
 - 階層式分群法(Hierarchical Clustering)
 - 密度式分群法(Density-based Clustering)
 - 機率式分群法(Probability-based Clustering)
- 了解相似度計算方法
- 使用軟體進行集群分析



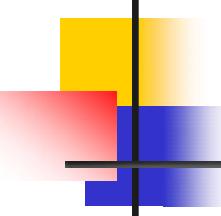
教學單元

- 集群分析的目的及應用
- 時序資料分群方法
- 相似度計算方法
- 使用軟體進行集群分析



教學單元

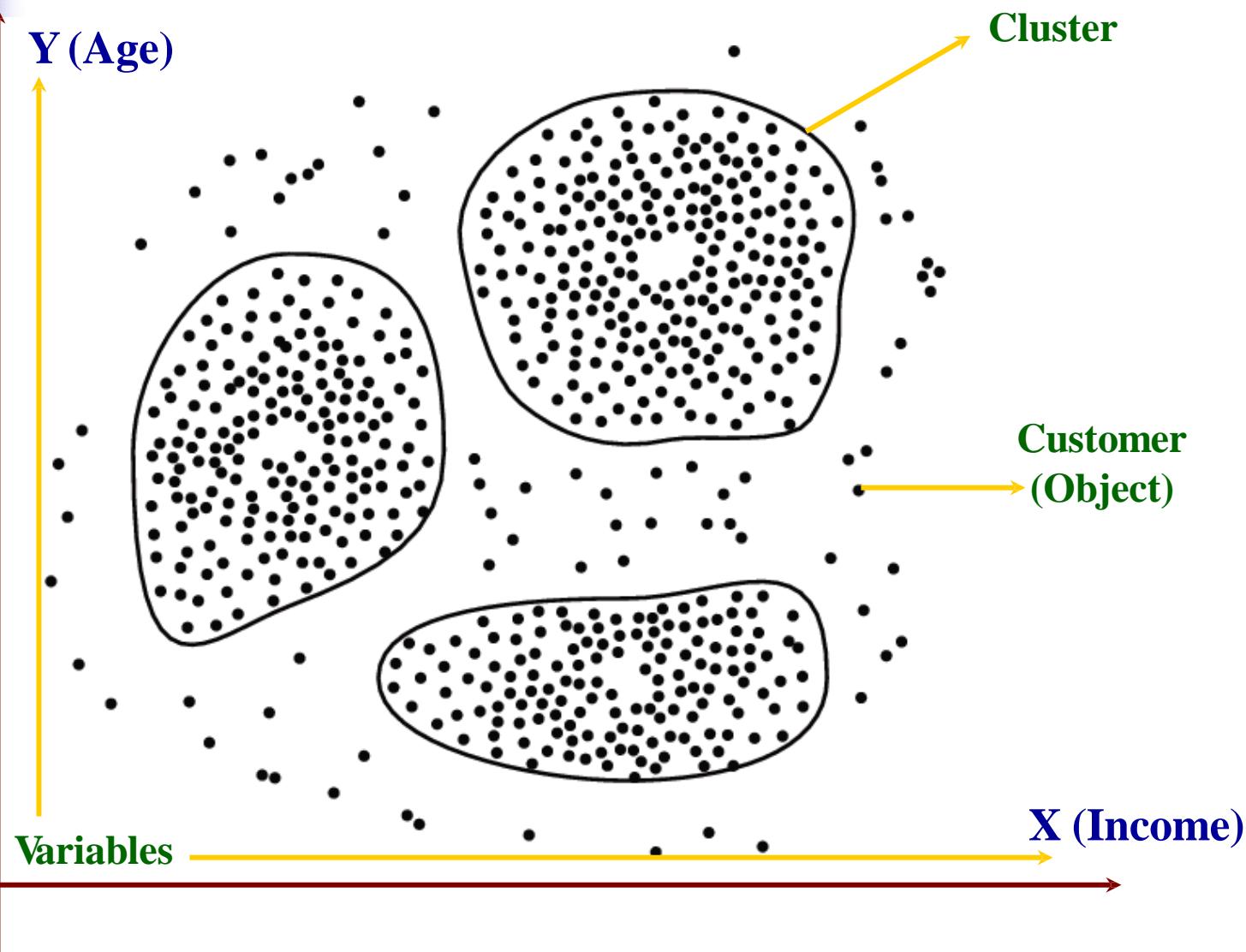
集群分析的目的及應用



分群演算法 (Clustering Algorithms)

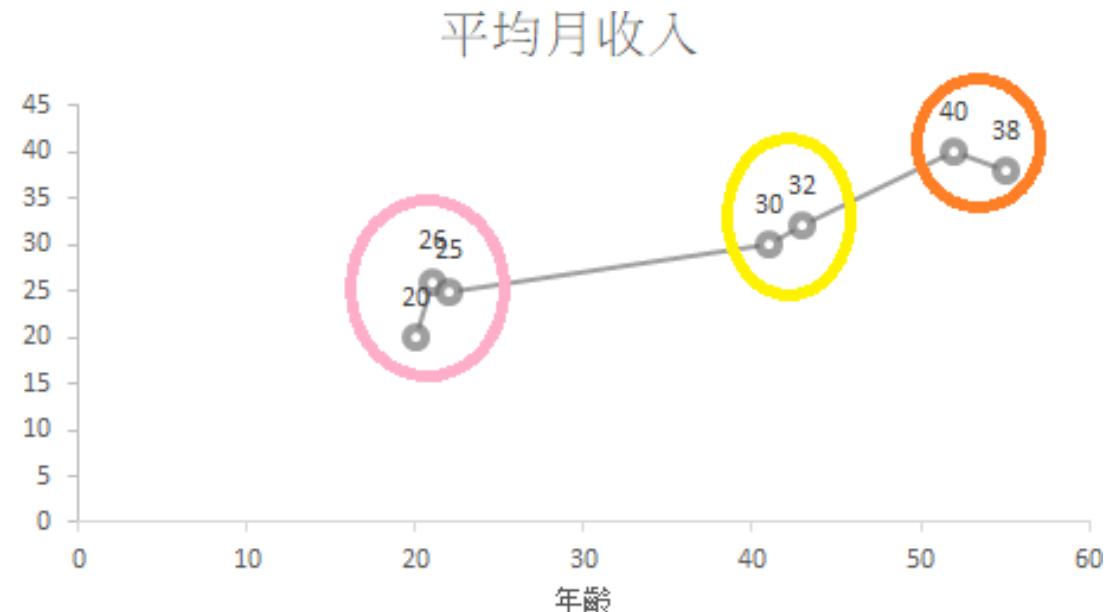
- 分群的目的
 - 以電腦自動化方式找出相似的物件(e.g., 圖片、音樂、文件、相似行為等)，並將之歸納成群。

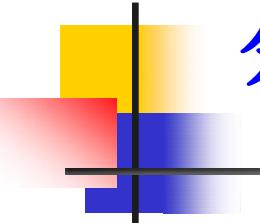
舉例說明：分群的目的



舉例說明：分群的目的

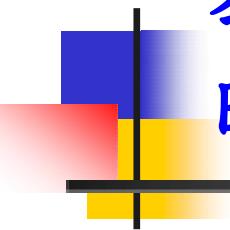
會員	年齡	平均月收入(千)
1	20	20
2	21	26
3	22	25
4	41	30
5	43	32
6	52	40
7	55	38





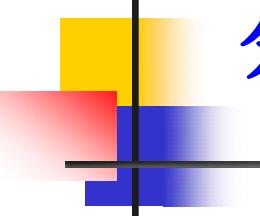
分群的相關應用

- 找出相似資料或物件
- 資料精簡化
- 找出具代表性資料
- 找出每群中隱含的特性或資訊
- 資訊檢索(如圖片/影像/音訊檢索)



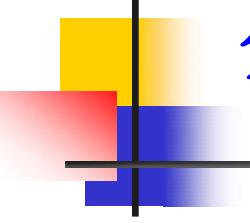
教學單元

時序資料分群方法



分群演算法的種類

- 分群方法常見的種類
 - 分割式分群法(Partitional Clustering)
 - 階層式分群法(Hierarchical Clustering)
 - 密度式分群法(Density-based Clustering)
 - 機率式分群法(Probability-based Clustering)



分割式分群法(Partitional Clustering)

- 分割式分群法
 - 試圖資料分配到一個群集中，群集彼此之間互不交集或重疊。
 - 每一群裡面的資料與該群的 **群中心(Clustering Center)**相似度高於其他群集中心。
- 常見的分割式分群法
 - K-Means (或稱K-平均法)
 - K-Medoid (或稱K-物件法)

數值屬性(Numerical Attribute)

常用的不相似度計算方法

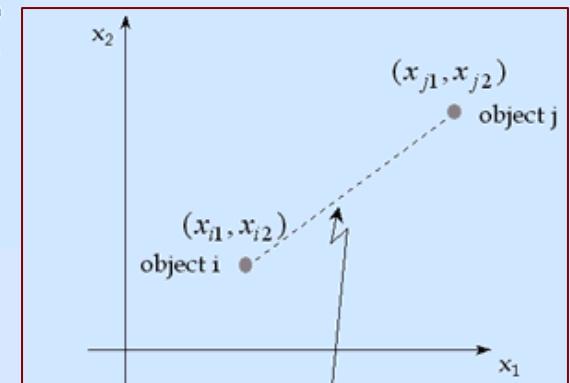
Euclidean Distance Measure

Suppose the data set contains p attributes. The Euclidean distance measure defines the dissimilarity $d(i, j)$ between objects i and j as:

$$d(i, j) = \sqrt{\left(\frac{|x_{i1} - x_{j1}|}{R_1}\right)^2 + \left(\frac{|x_{i2} - x_{j2}|}{R_2}\right)^2 + \dots + \left(\frac{|x_{ip} - x_{jp}|}{R_p}\right)^2}$$

where R_f is the range of attribute f, defined as:

$$R_f = \max_h x_{hf} - \min_h x_{hf}$$



$$d(i, j) = \sqrt{\left(\frac{|x_{i1} - x_{j1}|}{R_1}\right)^2 + \left(\frac{|x_{i2} - x_{j2}|}{R_2}\right)^2}$$

數值屬性(Numerical Attribute)

常用的不相似度計算方法

Example of Euclidean Distance Measure:

object	Age (20~70)	Income (20000~120000)
1	20	30000
2	30	50000

$$d(1,2) = \sqrt{\left(\frac{|20-30|}{50}\right)^2 + \left(\frac{|30000-50000|}{100000}\right)^2}$$
$$= \sqrt{\left(\frac{1}{5}\right)^2 + \left(\frac{1}{5}\right)^2} = 0.2828$$

類別屬性(Nominal Attribute)

常用的不相似度計算方法

General Dissimilarity Measure

Suppose the data set contains p attributes. The dissimilarity $d(i, j)$ between objects i and j is defined as:

$$d(i, j) = \frac{\sum_{f=1}^p d_{ij}^{(f)}}{p}$$

$d_{ij}^{(f)}$ is the contribution of the fth attribute to the dissimilarity between objects i and j, when the measurements x_{if} and x_{jf} are non-missing.

If attribute f is either binary or nominal, then $d_{ij}^{(f)}$ is defined as:

$$d_{ij}^{(f)} = 1 \text{ if } x_{if} \neq x_{jf}$$

$$= 0 \text{ if } x_{if} = x_{jf}$$

順序屬性(Ordinal Attribute) 常用的不相似度計算方法

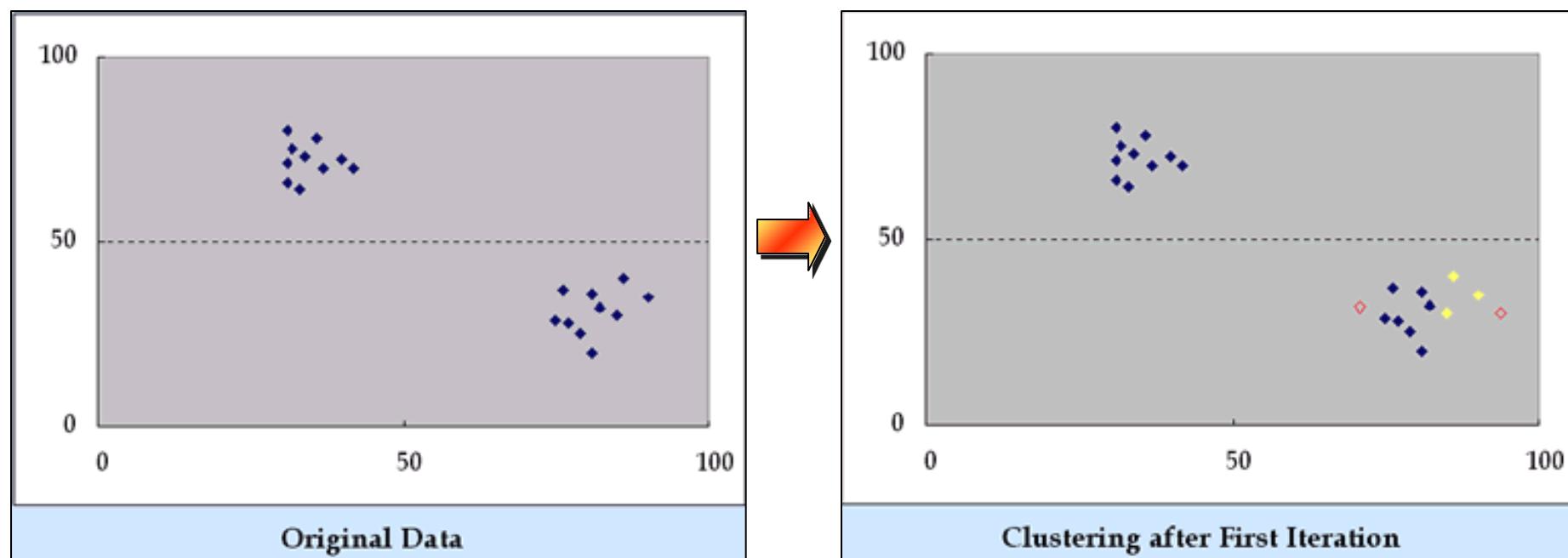
Example of General Dissimilarity Measure:

object	Age (20~70)	Occupation	Gender	Education Level
1	20	Banking	Male	High School
2	30	Education	Female	Master

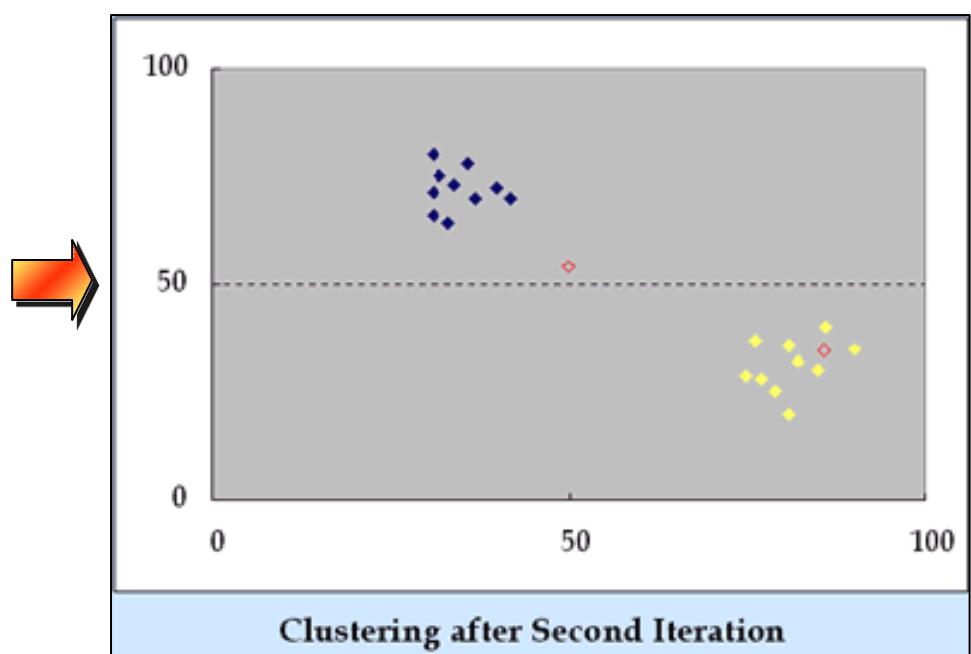
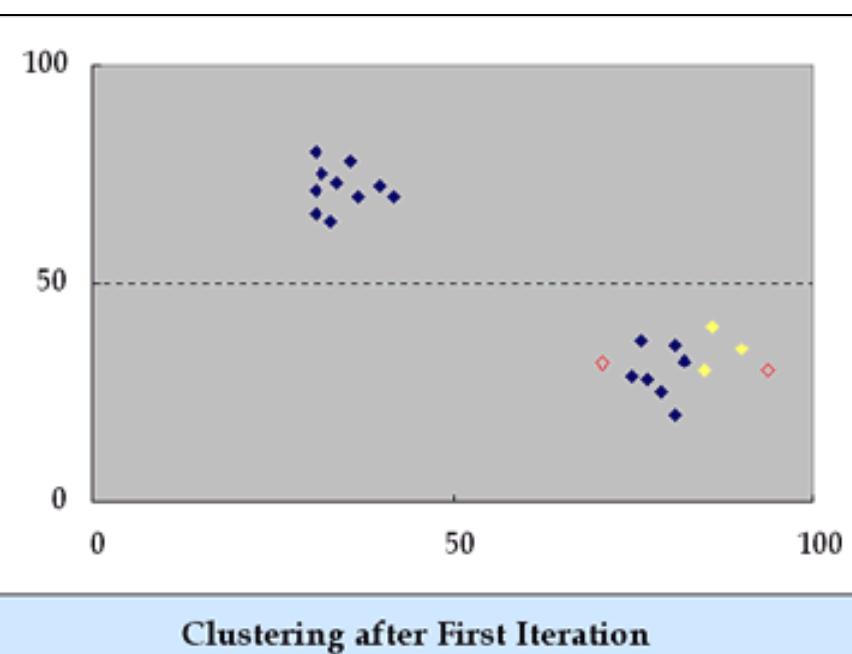
↓ Converting education level
1: High school, 2: Undergraduate
3: Master, 4: ph.D.

object	Age (20~70)	Occupation	Gender	Education Level
1	20	Banking	Male	1
2	30	Education	Female	3

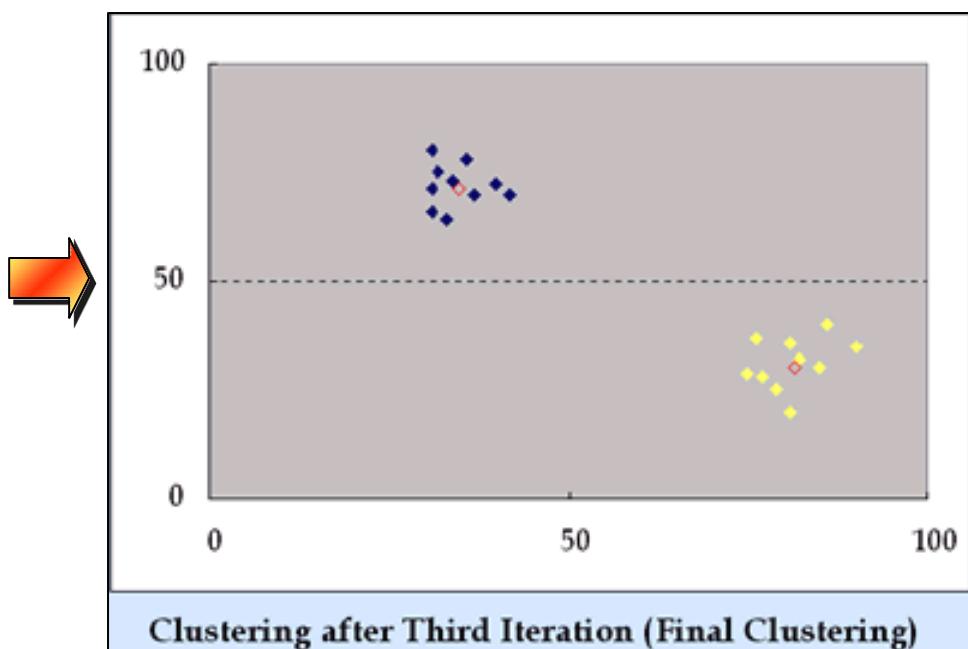
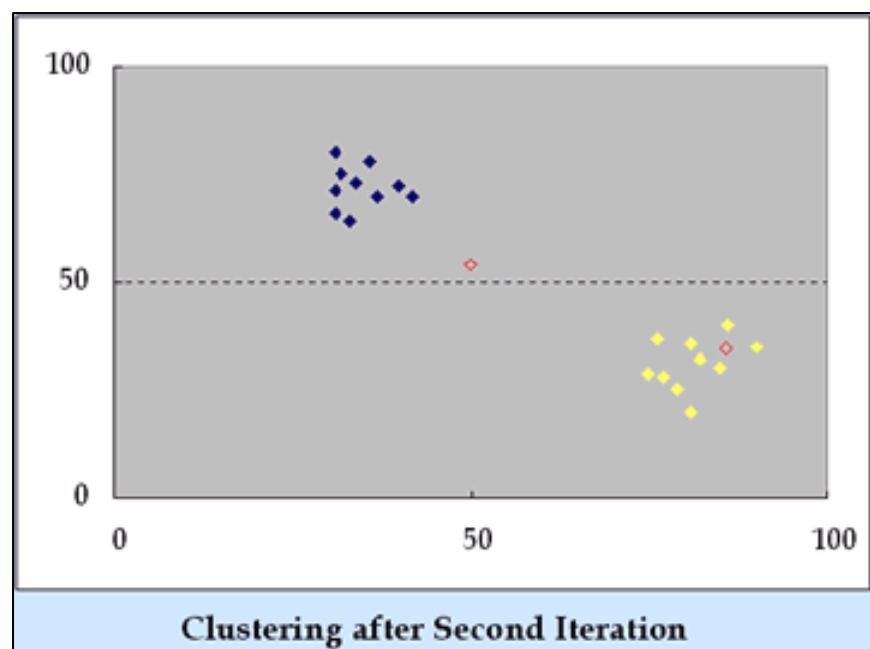
分割式分群法(Partitional Clustering) K-Means 舉例說明



分割式分群法(Partitional Clustering) K-Means 舉例說明



分割式分群法(Partitional Clustering) K-Means 舉例說明



分割式分群法(Partitional Clustering) K-Means 舉例說明

Algorithm: k -means. The k -means algorithm for partitioning based on the mean value of the objects in the cluster.

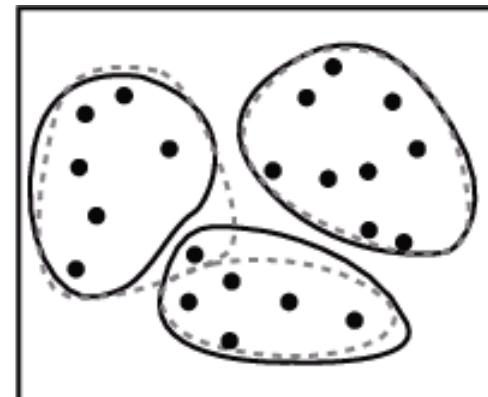
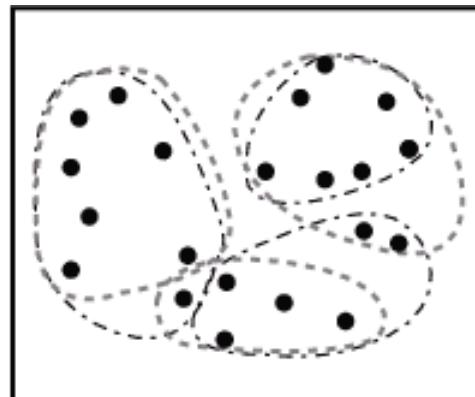
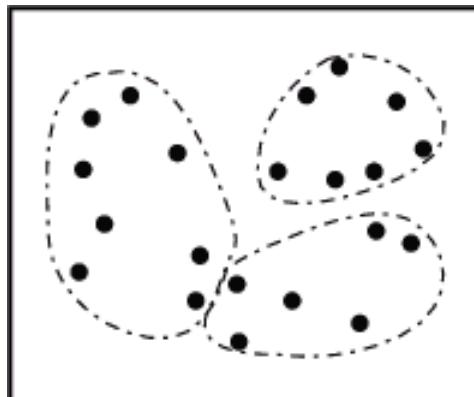
Input: The number of clusters k and a database containing n objects.

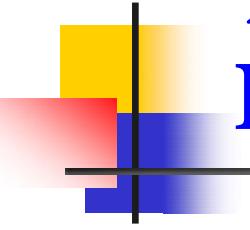
Output: A set of k clusters that minimizes the squared-error criterion.

Method:

**Sensitive to
Outlier!**

- (1) arbitrarily choose k objects as the initial cluster centers;
- (2) repeat
- (3) (re)assign each object to the cluster to which the object is the most similar,
based on the mean value of the objects in the cluster;
- (4) update the cluster means, i.e., calculate the mean value of the objects for each cluster;
- (5) until no change;



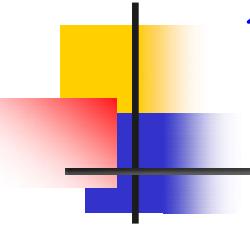


分割式分群法(Partitional Clustering)

K-Means Exercise

Number of clusters = 2

Object	X	Y
1	22	60
2	40	25
3	60	30
4	64	66
5	80	30
6	82	55



分割式分群法(Partitional Clustering) K-Medoid (或稱K-物件法)

- <https://www.youtube.com/watch?v=lAEPQz5tcFY>

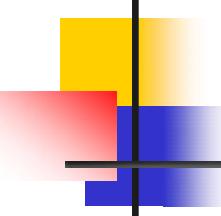
階層式分群法(Hierarchical Clustering)

■ 聚合法(由下往上)

- 一開始將每個資料點都視為是一個獨立的群集，然後依據群集間相似度計算公式，不斷地合併二個最相似的群集，直到最後所有的群集都合併成一個大群集或達到某個終止條件。

■ 分裂法(由上往下)

- 分裂法是採用由上而下的處理方式，一開始時將所有資料點視為一個大群集，同樣不斷地依據相似度計算公式將大群集分裂成較小的子群集，直到最後每個物件各自為一個獨立的群集或達到某個終止條件為止。



階層式分群法(Hierarchical Clustering)

聚合法的運作過程

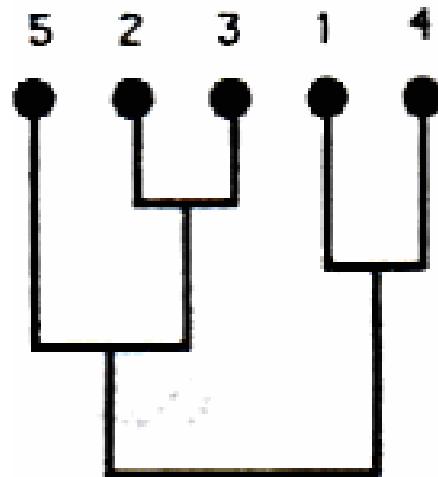
- 聚合法的運作過程
- 步驟 1
 - 將資料集合中每個資料點當作個別群集
- 步驟 2
 - 利用群集間相似度計算公式，將最相似的兩個群集加以合併，形成一新的群集，並以樹狀結構記錄此群集關係。
 - 重複執行步驟 2，直到所有的資料點都歸屬到同一群集或滿足使用者所設定之終止條件為止。

聚合法的代表例子 Single Linkage、Complete Linkage

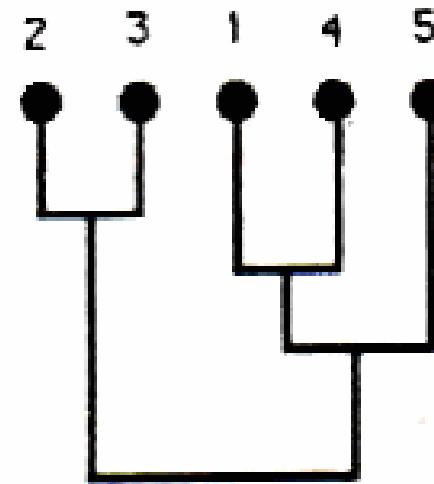
Dissimilarity
Matrix (5×5)

	x_1	x_2	x_3	x_4	x_5
x_1	0	6	8	2	7
x_2	6	0	1	5	3
x_3	8	1	0	10	9
x_4	2	5	10	0	4
x_5	7	3	9	4	0

樹狀圖(Dendrogram)



Single Link



Complete Link

群與群的相似度

(A)最近法(單一聯結法 Single Linkage)：

$$d_{AB} = \min_{i \in A, j \in B} d_{ij}$$

(B)最遠法(完全聯結法 Complete Linkage)：

$$d_{AB} = \max_{i \in A, j \in B} d_{ij}$$

(C)平均法(Average Linkage)：

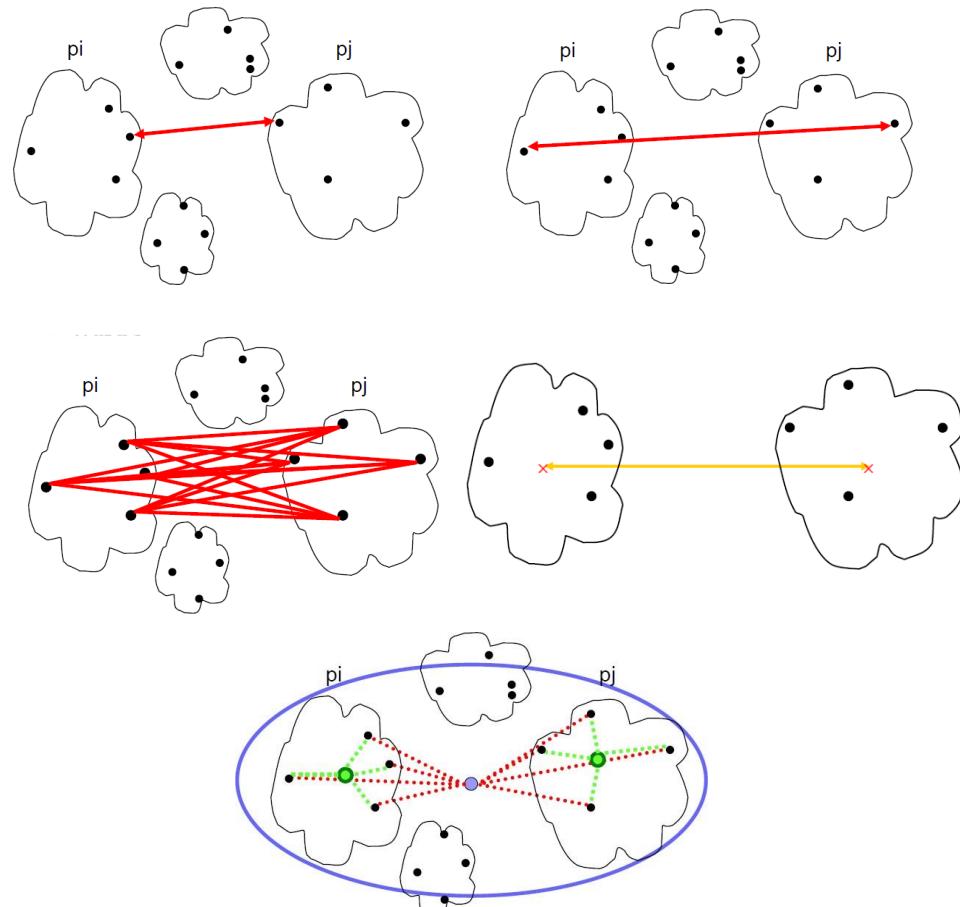
$$d_{AB} = \frac{\sum \sum d_{ij}}{n}, n \text{ 為全部距離的個數}$$

(D)中心法(Centroid Method)：

$$d_{AB} = d(\bar{x}_A, \bar{x}_B) = \|\bar{x}_A - \bar{x}_B\|^2$$

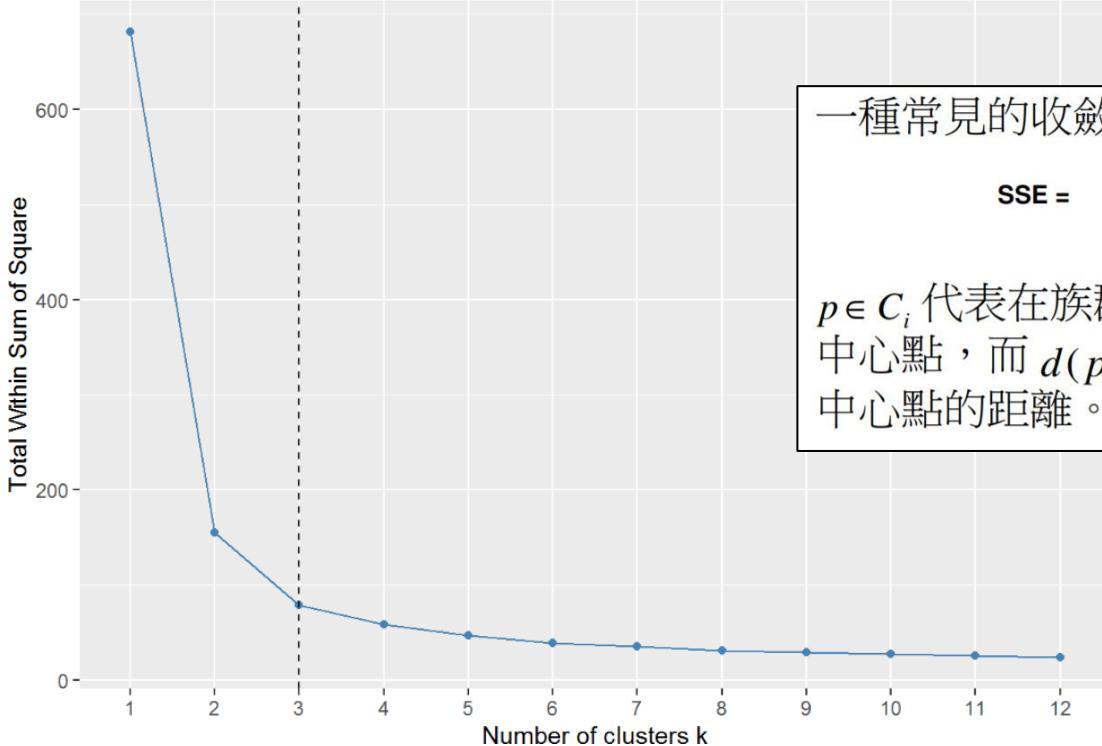
(E)華德法(Wards Method 華德最小變異法)：

$$d_{AB} = n_A \|\bar{x}_A - \bar{x}\|^2 + n_B \|\bar{x}_B - \bar{x}\|^2$$



Elbow Method

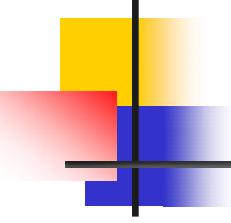
Elbow Method for HC



一種常見的收斂標準是Sum of Squared Errors (SSE)：

$$\text{SSE} = \sum_{i=1}^k \sum_{p \in C_i} d(p, m_i)^2$$

$p \in C_i$ 代表在族群*i*中的每一個資料點， m_i 是族群 *i* 的中心點，而 $d(p, m_i)$ 代表每一個資料點和它所屬族群中心點的距離。



嘗試用 Single Linkage 、Complete Linkage 進行分群

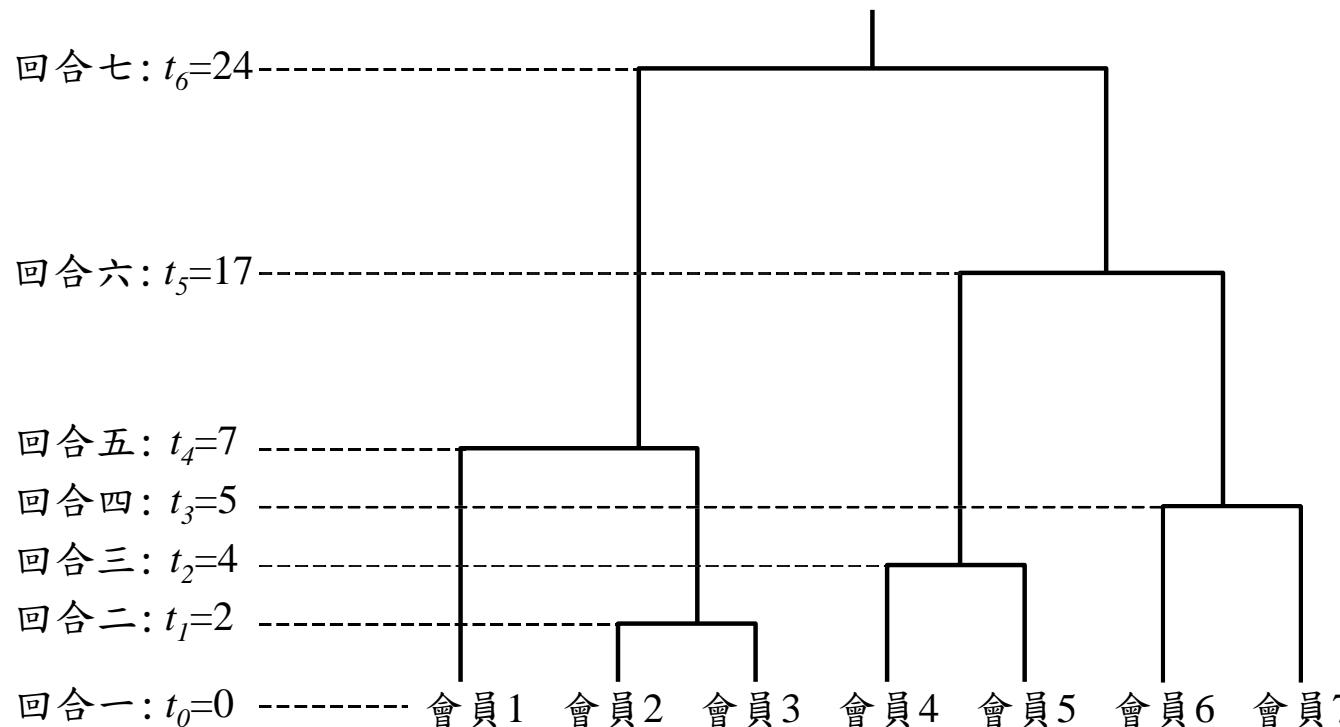
Number of clusters = 2

Object	X	Y
1	22	60
2	40	25
3	60	30
4	64	66
5	80	30
6	82	55

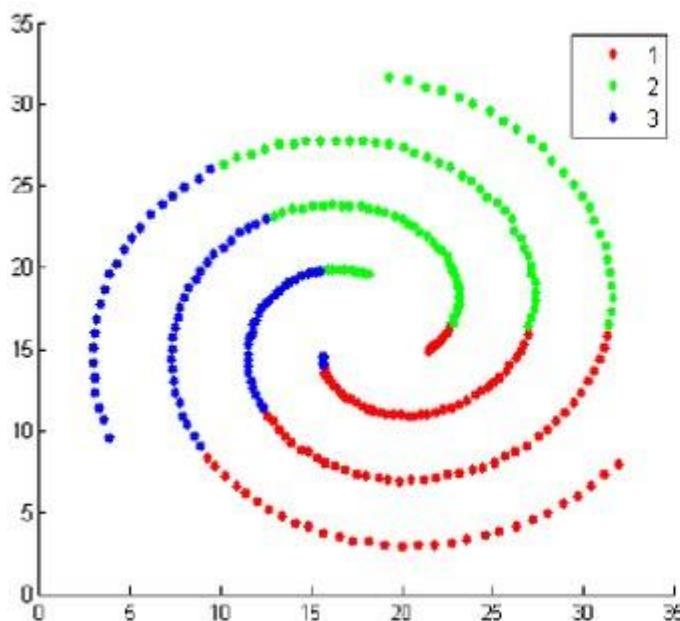
嘗試用 Single Linkage 、Complete Linkage 進行分群

- 右圖為會員之間的距離
下圖為聚合法產生的樹狀結構圖

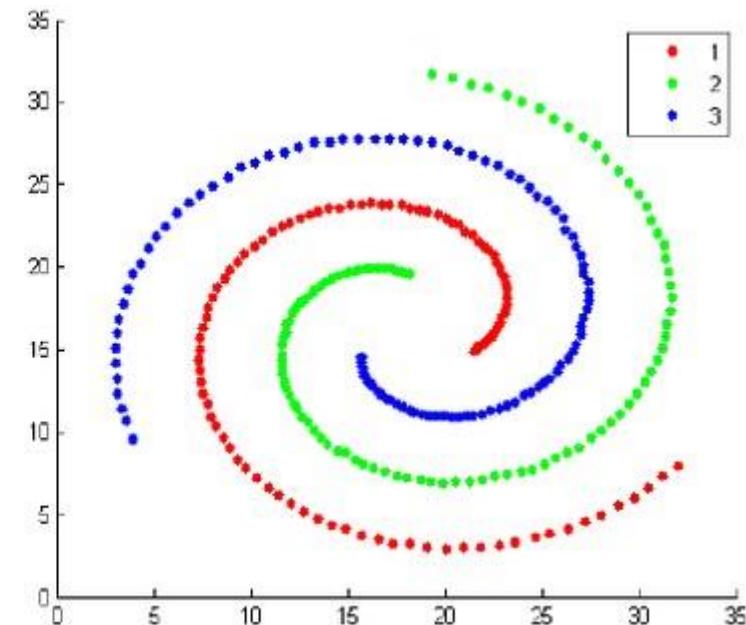
	會員1	會員2	會員3	會員4	會員5	會員6	會員7
會員1	0	7	7	31	35	52	53
會員2		0	2	24	28	45	46
會員3			0	24	28	45	46
會員4				0	4	21	22
會員5					0	17	18
會員6						0	5
會員7							0



密度式分群法(Density-based Clustering) DBSCAN

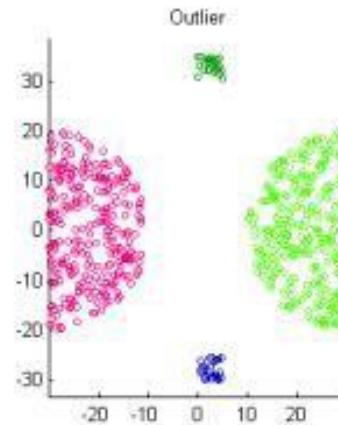
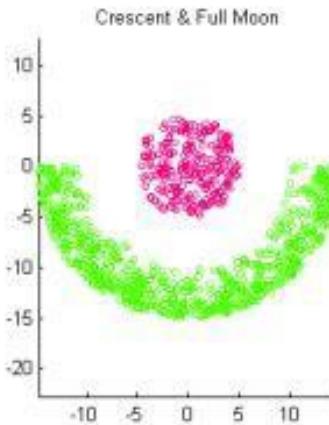
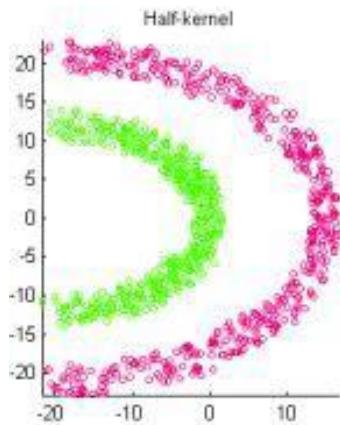
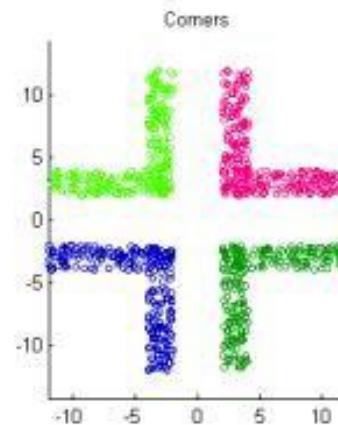
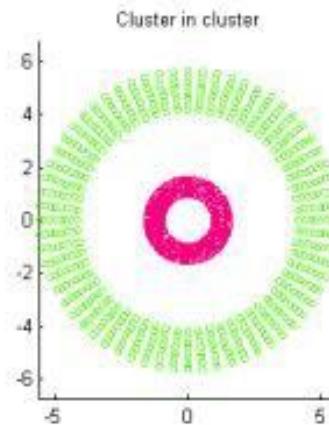
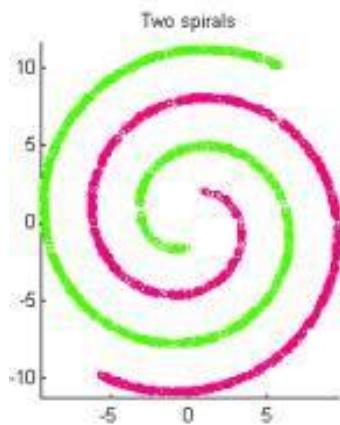


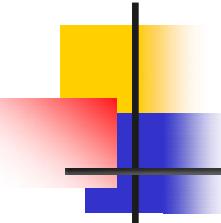
K-Means 分群結果



DBSCAN 分群結果

DBSCAN 分群結果



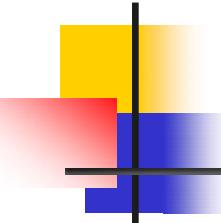


高密度關連區域分群法(DBSCAN)

名詞介紹(1/3)

- DBSCAN之相關定義

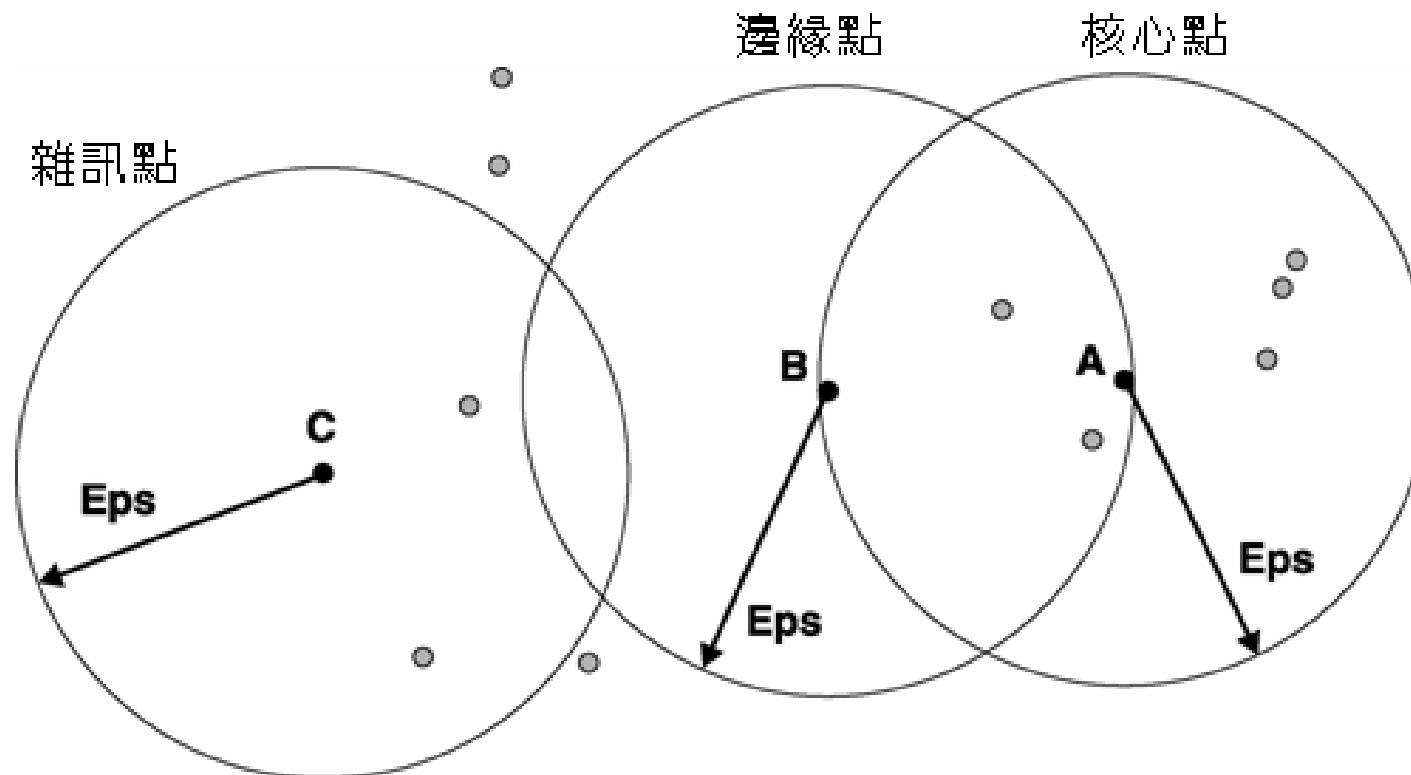
- 距離資料點半徑長度*Eps*以內的鄰近區域，則為該資料點的*Eps*-鄰近區域
- 資料點的*Eps*-鄰近區域中包含了至少*Minpts*個資料點，則該資料點為**核心物件(Core Point)**
- Core Point 的意義，周圍足夠密集



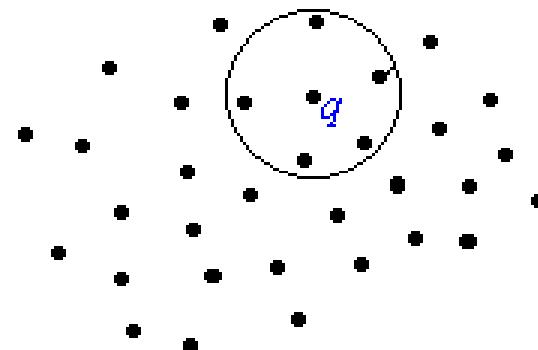
高密度關連區域分群法(DBSCAN) 名詞介紹(1/3)

- 資料點 p 的位置是在某核心物件 q 的 Eps -鄰近區域內，則資料點 p 可被稱為“可由 q 直接密度可達(directly density-reachable)”的物件
- 假如資料點 p 可由 q_1 直接密度可達、而 q_1 可由 q_2 直接密度可達、.....、而 q_{i-1} 是可由 q_i 直接密度可達，則資料點 p 可以被稱為“可由 q_i 密度可達(density-reachable)”的物件

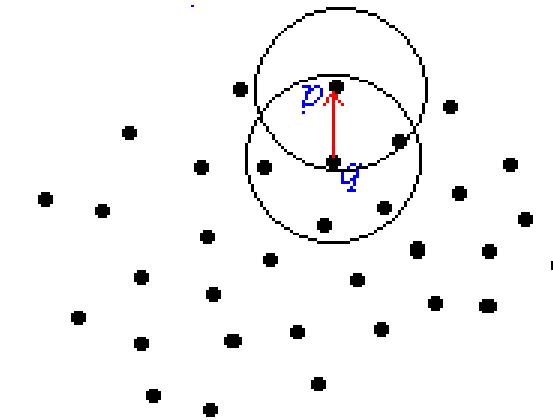
高密度關連區域分群法(DBSCAN) 名詞介紹(3/3)



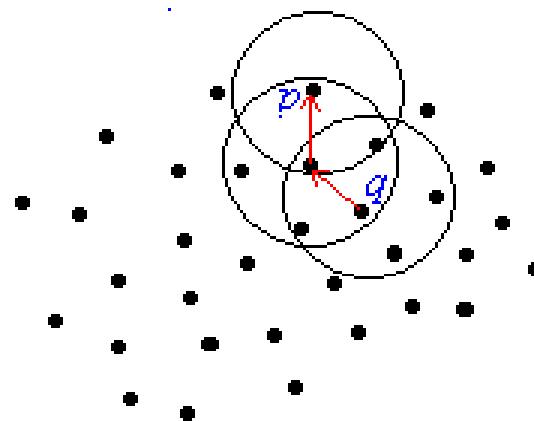
高密度關連區域分群法(DBSCAN) 舉例說明



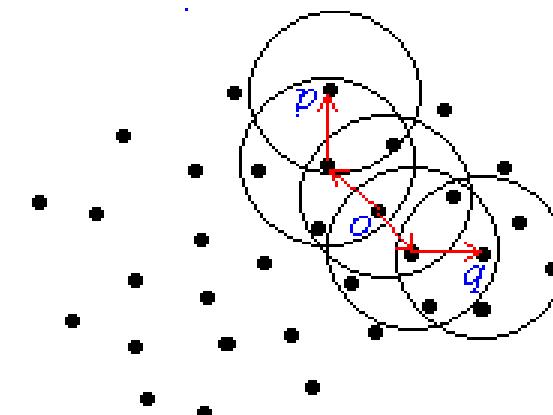
(a)



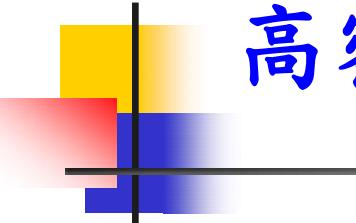
(b)



(c)

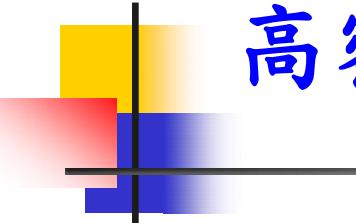


(d)



高密度關連區域分群法(DBSCAN)

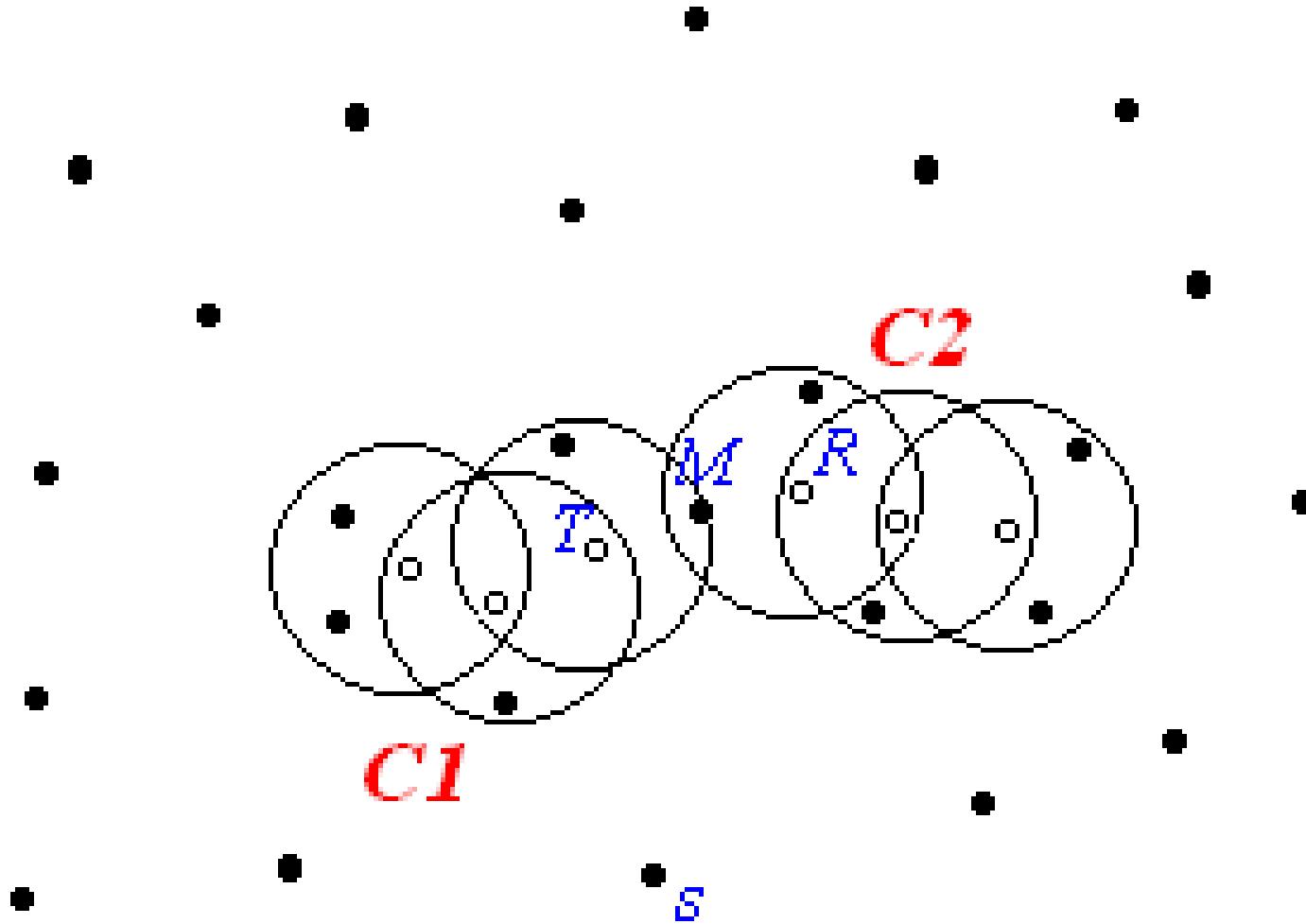
- DBSCAN對於群集和偏移值的定義
- 對資料點 p 和 q 而言，假如 q 歸屬於群集 A ，且 p 可由 q 密度可達，則 p 也將歸屬於群集 A ；對於歸屬於相同群集 A 的 p 和 q 而言， p 和 q 必為密度連接
- 對無法歸屬到任何群集之資料點，將被視為雜訊、偏移值

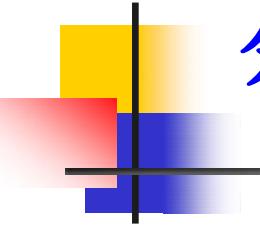


高密度關連區域分群法(DBSCAN)

- DBSCAN的運作過程
- 步驟 1
 - 找出所有 Core Points
- 步驟 2
 - 依序選擇核心物件 x ，依序探索其鄰近區域內的其它 Core Point y ，合併 x 與 y 鄰近區域內的所有資料點
 - 重複步驟 2，直到所有 Core Point 皆被處理完畢。

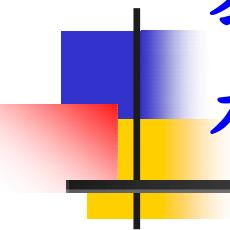
高密度關連區域分群法(DBSCAN)





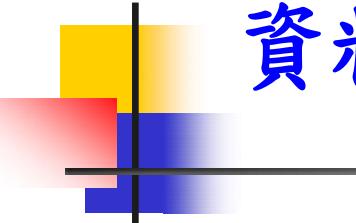
分群的相關應用

- 找出相似資料或物件
- 資料精簡化
- 找出具代表性資料
- 找出每群中隱含的特性或資訊
- 資訊檢索(如圖片/影像/音訊檢索)



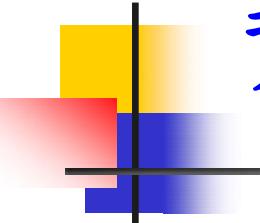
教學單元

相似度計算方法



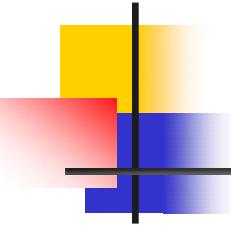
資料相似度

- 衡量資料紀錄間的相似度將決定資料紀錄所歸屬的群集，並影響整個分群的結果，因此相似度測量法是群集分析中最根本的課題。
- 相似度依據不同的資料型態、應用範圍、資料集合離散程度、資料複雜性將有不同的測量方法，因此根據這些考量選擇一些適合的測量方法，對於群集分析將有決定性的影響。



資料型態的考量

- 兩資料點之間的相似度是由其相對應資料維度中資料數值間的差異性來決定。
- 若所描述之資料維度都是連續性資料維度，則可由距離公式簡單地計算彼此在空間上的接近程度。
- 但若涉及類別資料維度，如何決定二資料點相似度的量測標準就更為複雜。



資料型態的考量

- 以連續性資料維度而言，測量資料點間的相似度通常利用簡單的空間距離計算公式，透過衡量資料點間距離的遠近來判斷彼此間的相似程度。
- 以下是兩種較常用的距離計算公式：
 1. 尤拉距離(Euclidean distance)
 2. 曼哈頓距離(Manhattan distance)

尤拉距離

- 計算資料紀錄間的尤拉距離來評估彼此相似度是其中最常見的測量方法。假設資料點 $x = (x_1, \dots, x_n)$ 和 $y = (y_1, \dots, y_n)$ 為兩筆資料，則可用下列公式計算：

$$d(x, y) := \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

- 尤拉距離是純數值，以二維空間來看，則為兩點之間最短距離直線。

尤拉距離

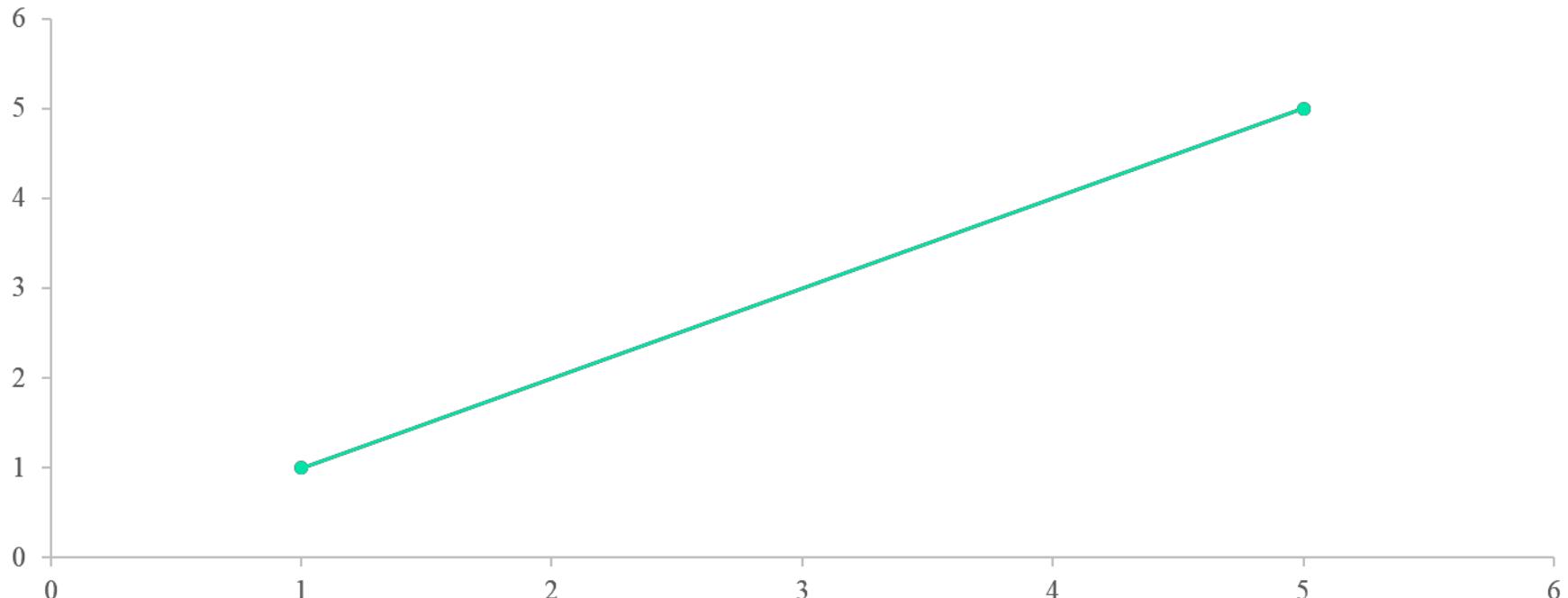


圖 1-1 尤拉距離在二維空間上的物理意義

曼哈頓距離

- 在平面上，坐標 (x_1, y_1) 的點 P_1 與坐標 (x_2, y_2) 的點 P_2 的曼哈頓距離為：

$$|x_1 - x_2| + |y_1 - y_2|$$

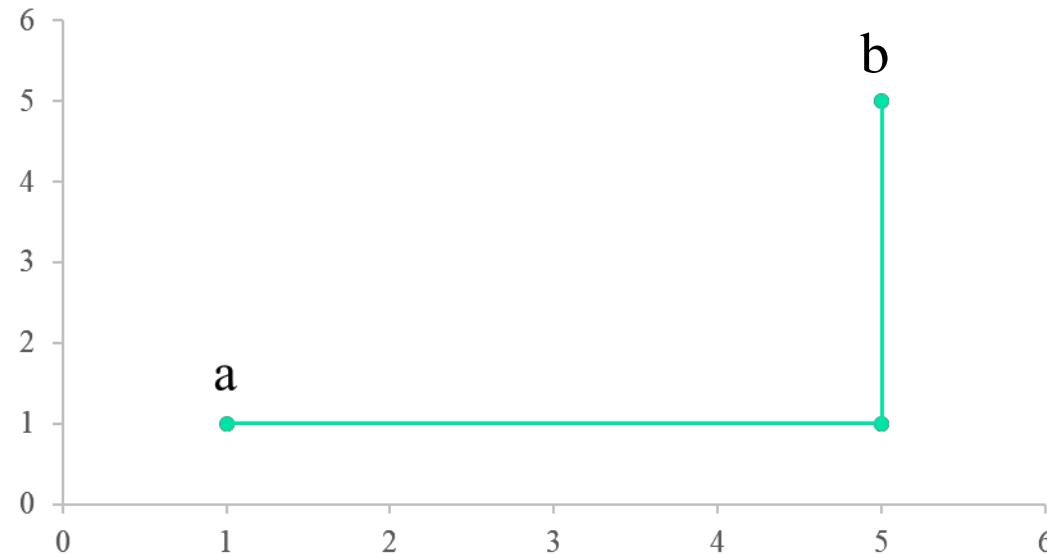
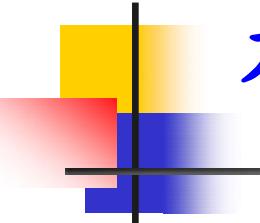
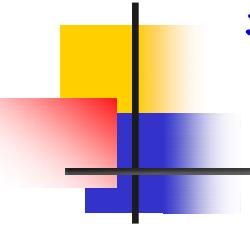


圖 1-2 a,b 點的曼哈頓距離



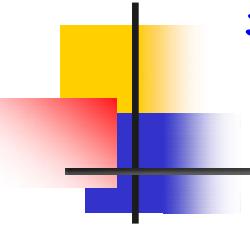
尤拉距離與曼哈頓距離

- 以資料點 $a<20,20>$ 與 $b<21,26>$ 分別計算尤拉距離和
曼哈頓距離：
- 尤拉距離： $\sqrt{[(21-20)*(21-20) + (26-20)*(26-20)]} = 6$
- 曼哈頓距離： $|21 - 20| + |26 - 20| = 7$



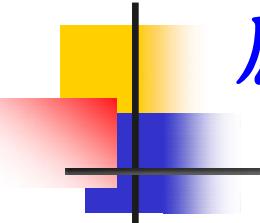
資料型態的考量

- 然而在實際應用上，資料紀錄通常需要用類別型態 資料維度來描述，例如會員所居住之縣市、會員的 婚姻狀態、會員的教育程度等等。
- 因此非連續性資料維度的考量是必須的，一般來講 對於非連續性資料維度在相似度計算有以下作法：
- 1.利用字串比對的方式，對於資料數值完全相同時 相似度以1表示，否則以0表示。



資料型態的考量

- 2.針對個別不同之非連續性資料維度，透過專家
事先訂定資料數值間的相似度與輔助之計算公式。
 -
- 3.將類別型態的資料數值先轉換或對應成連續性的
資料數值，再套用距離計算公式來計算其相似度。
 -



應用範圍的考量

- 在群集分析過程中，除了需要衡量資料點之間的相似程度之外，衡量群集間的相似程度在某些分群法與應用上是必要且不可或缺的。
- 在階層式分群法中，**聚合法(AGNES)**其運作方式是透過彼此相似度高的小群集合併成較大的群集；而**分裂法(DIANA)**則是將較大的群集進行分離的動作；在每一回合聚合或分裂過程，都必須衡量群集間的相似程度來做決策。

應用範圍的考量

- 衡量群集間彼此之相似程度可用下列公式

:

$$d_{min}(C_i, C_j) = \min_{p \in C_i, p' \in C_j} d(p - p'),$$

$$d_{mean}(C_i, C_j) = d(m_i - m_j),$$

$$d_{avg}(C_i, C_j) = \frac{1}{n_i n_j} \sum_{p \in C_i} \sum_{p' \in C_j} d(p - p'),$$

$$d_{max}(C_i, C_j) = \max_{p \in C_i, p' \in C_j} d(p - p'),$$

- 其中 m_i 表示群集 C_i 的平均值， n_i 表示屬於 C_i 的資料點數量，而 $d(p-p')$ 表示資料點 p 和 p' 的距離，可為尤拉距離或曼哈頓距離。

應用範圍的考量

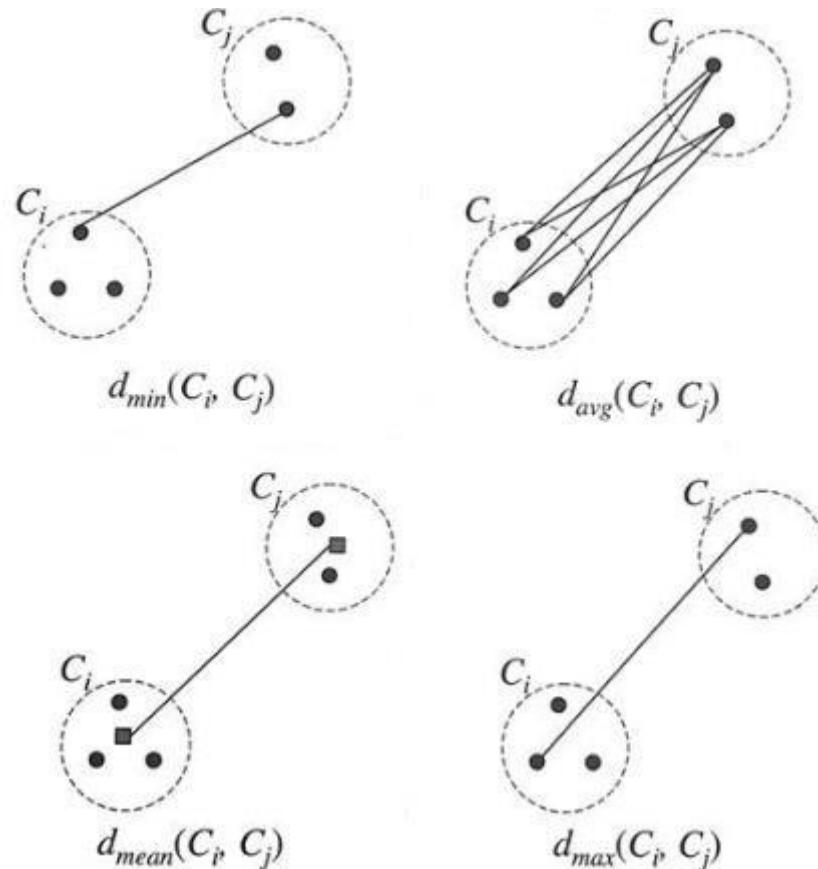
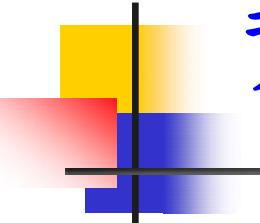
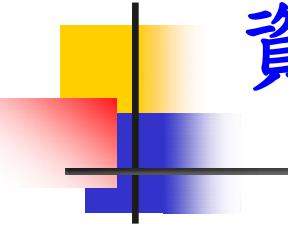


圖1-3 四種群集間相似度計算公式其空間上的物理意義



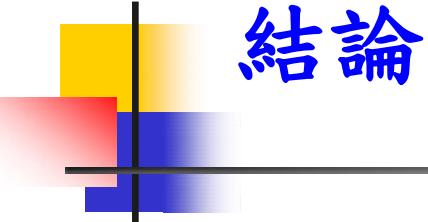
資料離散程度與複雜性的考量

- 一般相似度計算公式通常會對資料點中各資料維度給予相同的重要性，然而這將造成值域(**domain**)較大的資料維度將左右分群的結果。
- 例如以年齡與平均月收入所描述的會員資料而言，假設A<20,20>、B<21,22>、C<40,21>，以尤拉距離來計算，則A與C相似度較高；用年齡來判斷，則A與B較可能屬於同一個族群。



資料離散程度與複雜性的考量

- 為了避免這樣的情形發生，在計算相似度與資料分群之前，可以利用在前置處理章節所提到的轉換與標準化技術，先對資料數值進行前置處理。
- 同樣地，由於非連續性資料維度之計算方式與連續性資料維度之計算方式可能不同，因此在相似度計算上也必須注意到所可能產生的偏差與影響。



結論

- 集群分析的目的及應用
- 時序資料分群方法
 - 分割式分群法(Partitional Clustering)
 - 階層式分群法(Hierarchical Clustering)
 - 密度式分群法(Density-based Clustering)
 - 機率式分群法(Probability-based Clustering)
- 相似度計算方法
- 使用軟體進行集群分析