

4. CLUSTERING

客戶輪廓

假設你是影音串流平台的數據工程師

genre偏好(動作片, 動畫片, 劇情片……)
觀看時間偏好(夜貓族, 配飯吃族……)

ABSTRACT

- 分群基本概念
- Hierarchical clustering
- K-MEANS
- DBSCAN
- 商業問題定義

分群基本概念

分群的意義 – 同群體之間相似度較高



分群的意義 – 同群體之間相似度較高

較相似



湘北隊



陵南隊



翔陽隊



海南隊

較相似



新選組



十本刀



劍心



御庭番眾

較不相似

1 集群分析概念

- 將具有某些共同特性者予以整合在一起，然後分配到特定的群體，最後形成許多不同集群的一種分析方法。
- 將不同的觀察值依**相對距離**的遠近加以分類成不同集群，對不同集群所具有的**特性**加以歸納並**命名**
- 將某些具有共同特性的樣本予以整合，顯示出**內部同質性**與**外部異質性**，即達到群內差異最小、群間差異最大

2 變數與資料的篩選

集群分析在變數與資料的篩選上須考量：

- 1. 變數的共線性
- 2. 極端值的剔除
- 3. 資料的標準化

3 相似性衡量

相似性係指任何二個樣本，相對於其他樣本而言，如果在各種變數上有相近之處，即代表它們在很多方面具有共同的特性。

進行相似性衡量必須決定要採用何種設定基礎，不同的衡量方法，可能產生不同的分群結果

3 相似性衡量

在集群的概念中，距離最小的樣本會先集結，然後再找距離比較大者集結，至最後全部集結為止，其中較重要且較常使用到的，仍屬歐氏距離衡量法

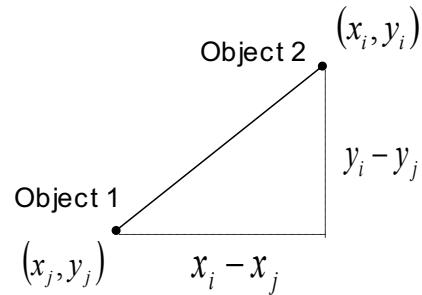
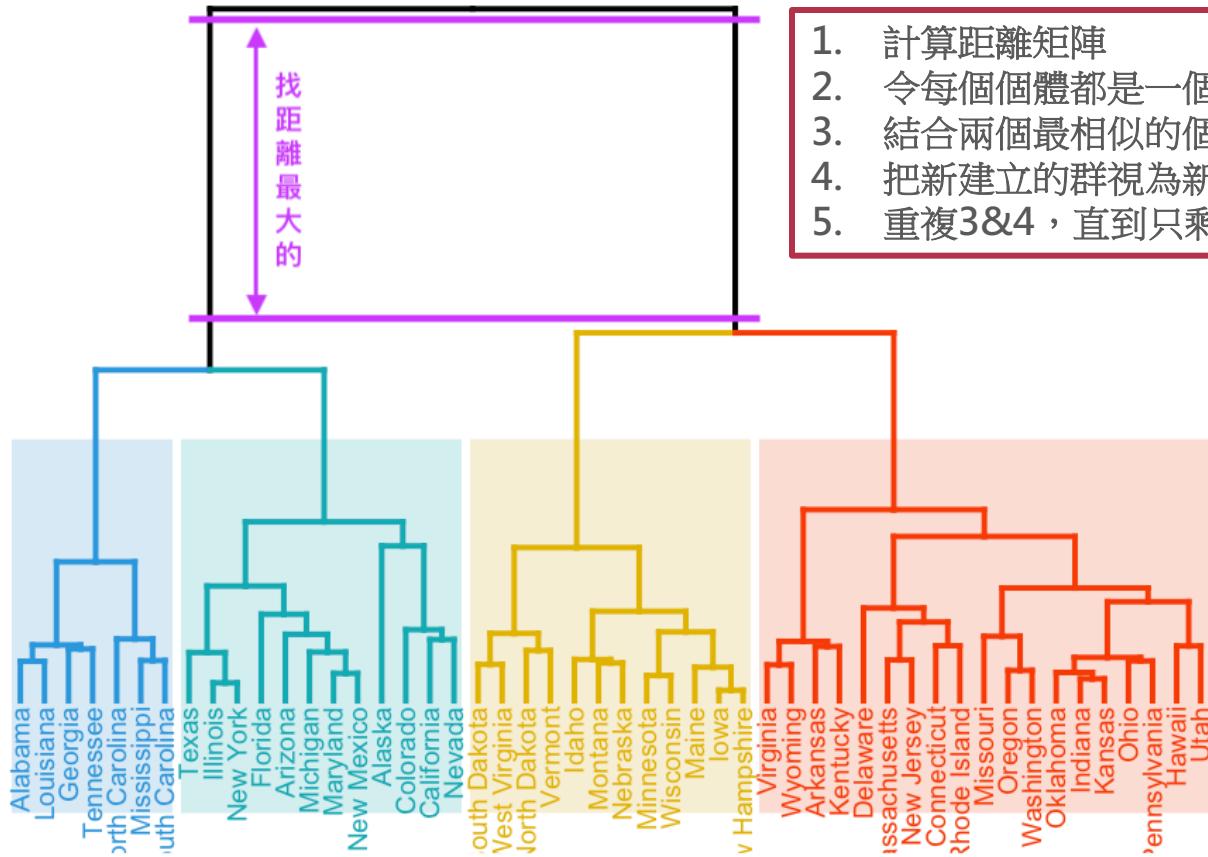


圖 集群分析歐氏距離計算方法

$$\text{歐氏距離之公式為 : } d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

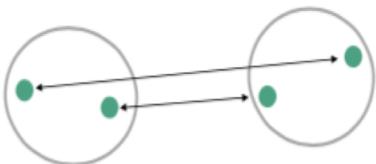
階層式分群 (HIERARCHICAL CLUSTERING)

算法流程

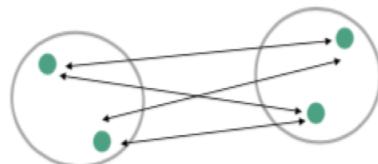


計算距離方法

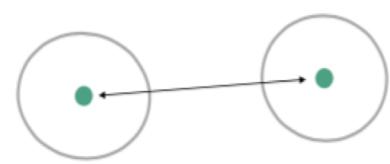
Single & Complete Linkage



Group Average



Dist. Between Centroid



衡量群體間最靠近的兩個點或最遠的兩個點

衡量兩群體的各點間，平均距離最大

計算兩群體間質心間的距離

計算距離方法

- 沃德法 (Ward's method)：群間的距離定義為兩群合併後，各點到合併群中心的距離平方和。

$$d(C_i, C_j) = \sum_{a \in C_i \cup C_j} \|a - \mu_{C_i \cup C_j}\|$$

- 假使有兩個cluster，為 C_i 和 C_j 。 μ 代表 C_i 和 C_j 裡面所有資料的平均。
- 找中心點，然後計算所有屬於 $C_i \cup C_j$ 的資料到中心距離的總和。
- 優勢：較single和complete的方法，更能夠全觀的考慮到整個cluster中的資料分布，相對average來說也能更有指標性的選擇適合的cluster。

非階層式集群分析法 (K-MEANS)

核心目標

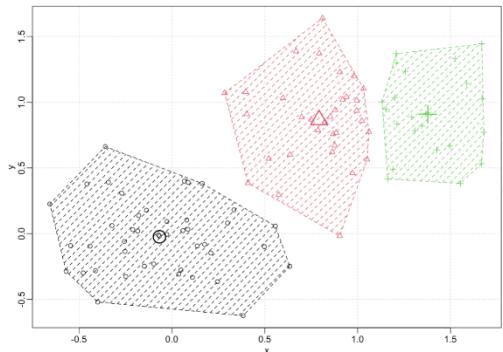
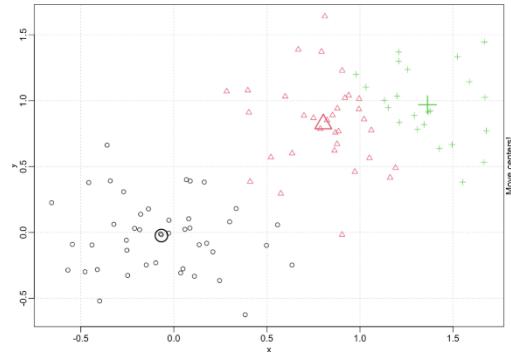
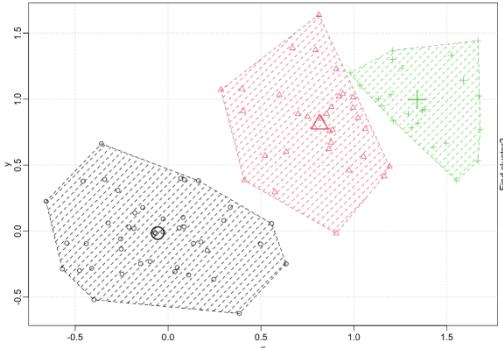
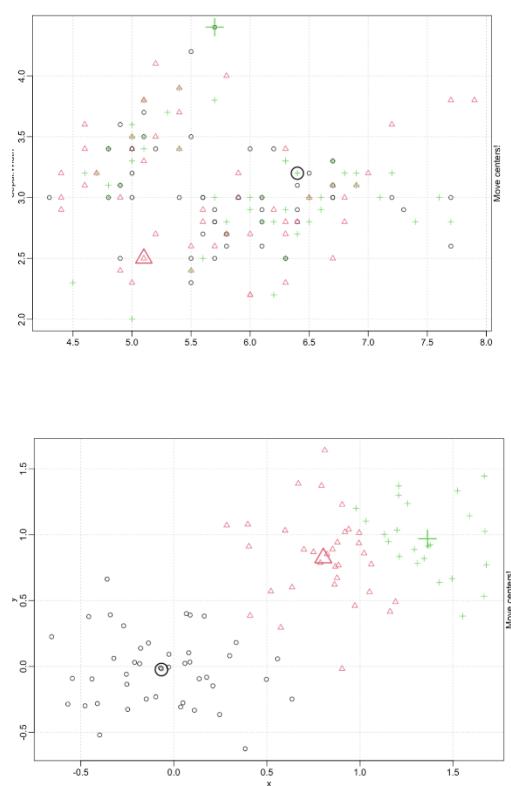
- 同一群裡的資料同質性高，不同群的資料同質性低

同質性以「距離」作為指標

- 同一群裡內的距離近，不同群間的距離遠

1. 非層次集群分析概念

- 由觀察值中指定K群的中心值(cluster seed)，以K個中心值為中心
- 開始將觀察樣本值中與中心值較接近者納入各群中
- 依各觀察值到各群中心值之距離遠近重新計算出各集群之集結係數，再試著移動中心值之位置，並重新計算集結係數
- 直到中心點及各群之樣本不能再移動為止



Find cluster?

Find cluster?

2. 怎麼選 K ?

要決定分為多少群才有意義：

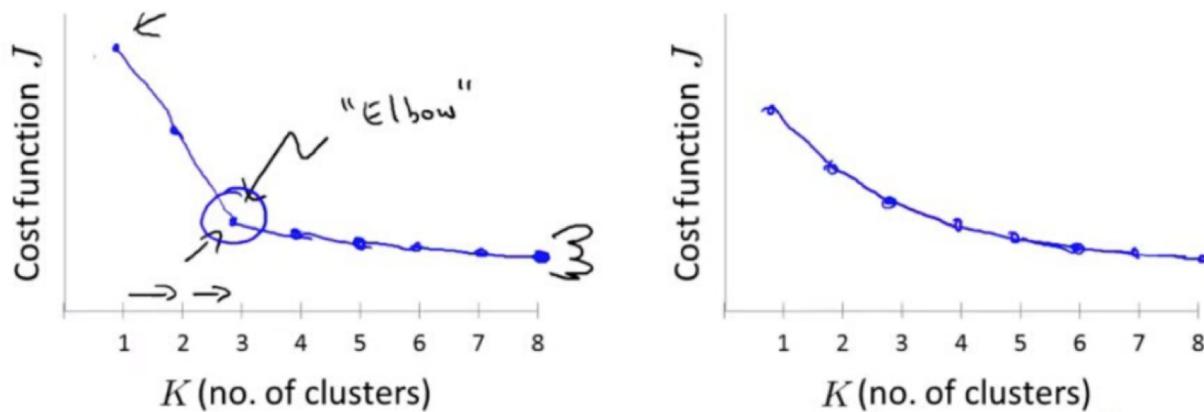
- 集群之群數以在2~6群為宜，超過 6 群則其後續分析將變得相當瑣碎。
- 集群完成後，各群數量應盡量接近。(如，若第一群有100個觀察值，第二群只有5個觀察值即非常不適當。)
- 當集群數減少，集群內各觀察值的同質性便會降低。應權衡集群數與同質性兩者，儘可能找到較少的集群，但仍滿足同質性的必要水準。
- 盡量依照公司策略方向來決定集群數目。

2.1 手肘法 (ELBOW METHOD)

- 以SSE (sum of the squared errors, 誤差平方和) 為指標，計算每群中的每一個點，到群中心的距離。共 K 個群， C_i 代表其中一個群， m_i 表示該群中心點

$$SSE = \sum_{i=1}^k \sum_{p \in C_i} |p - m_i|^2$$

- 根據 K 與 SSE 作圖，可以從中觀察到使 SSE 的下降幅度由「快速轉為平緩」的點，一般稱這個點為拐點 (Inflection point)，會將它挑選為 K 。
- 該點可以確保 K 值由小逐漸遞增時的一個集群效益，因此適合作為分群的標準。



2. 2 輪廓係數法 (Silhouette Coefficient)

- 「找出相同群凝聚度越小、不同群分離度越高」的值
- 其凝聚度 (a) 是指與相同群內的其他點的平均距離；分離度 (b) 是指與不同群的其他點的平均距離。 S 是指以一個點作為計算的值，輪廓係數法則是將所有的點都計算 S 後再總和。 S 值越大，表示效果越好，適合作為 K 。

$$S = \frac{b - a}{\max(a, b)}$$

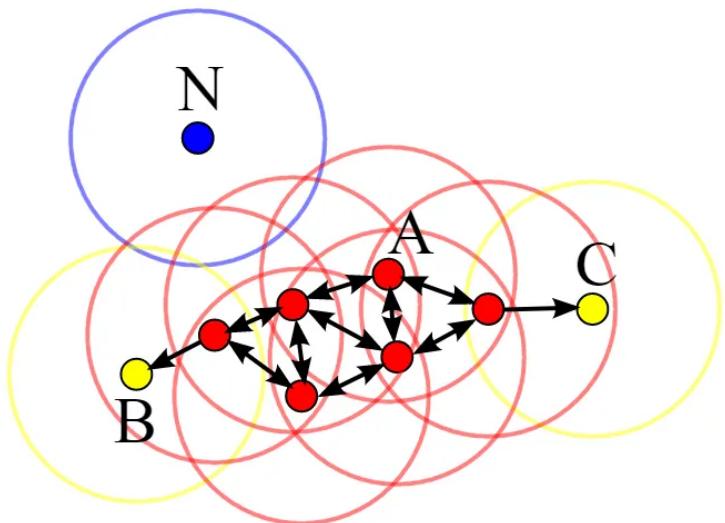
DBSCAN

(DENSITY-BASED SPATIAL CLUSTERING OF APPLICATIONS WITH NOISE)

核心目標

- 發現任意形狀的群：
基於密度的方式將數據點聚合成形狀任意的群。
即所謂Density-Based的方法，重視的是data的密度。
- 識別雜訊點：
區分屬於群的數據點與孤立的雜訊點。
能夠自動處理noise。
- 不依賴預定義群數：
根據數據分佈特徵，自動決定群的數量。

核心概念



- eps: neighborhood radius
- min_samples: 4
- A: Core 核心點
- B, C: not core 邊界點
- N: noise 雜訊點

1. Eps (ε, 半徑) :

定義一個數據點的鄰域範圍，只有在這個範圍內的點才被認為是相鄰的。

2. MinPts (最小點數量) :

一個數據點需要在其Eps鄰域內至少包含MinPts個點才能成為核心點 (Core Point) 。

3. 核心點 (Core Point) :

鄰域內有至少MinPts個點的數據點。

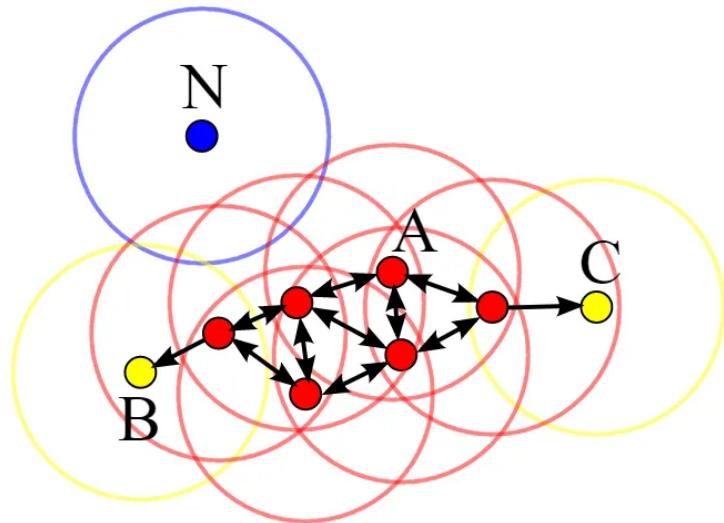
4. 邊界點 (Border Point) :

不滿足成為核心點條件，但在核心點的鄰域範圍內。

5. 雜訊點 (Noise Point) :

既不是核心點也不是邊界點的點。

算法流程

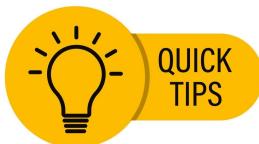


- eps : neighborhood radius
- min_samples : 4
- A: Core
- B, C: not core
- N: noise

1. DBSCAN會自行從任意一個點出發。
2. 假設從A出發，然後搜尋A周圍 eps 範圍以內的「資料數量」，當前的 eps 範圍裡有超過 min_samples 個資料時，我們就認為A是一個Core。
3. 然後開始去對A的 eps 範圍內的其他資料做一樣的事情，直到現在某一個點的 eps 範圍內不具備 min_samples 數量的點就停止。
4. 如果今天出發的點是N，則在最一開始周圍就找不到足夠數量的點，N就會被判斷為Noise。

參數選擇

- ϵ (eps)：
由這個參數值為半徑劃出的圓型區域稱為 ϵ -鄰域。
- minPts：
構成高密度區域需要最少有幾個點。



noise太多 => 加大eps，縮小min_sample。反之亦然。
注意不要讓eps跟min_sample變成太極端的值

優劣勢

優點：

- 能夠發現任意形狀的簇。
- 自動處理雜訊數據。（檢測任務）
- 不需要預先指定群的數量。
- 只需決定兩個參數。

缺點：

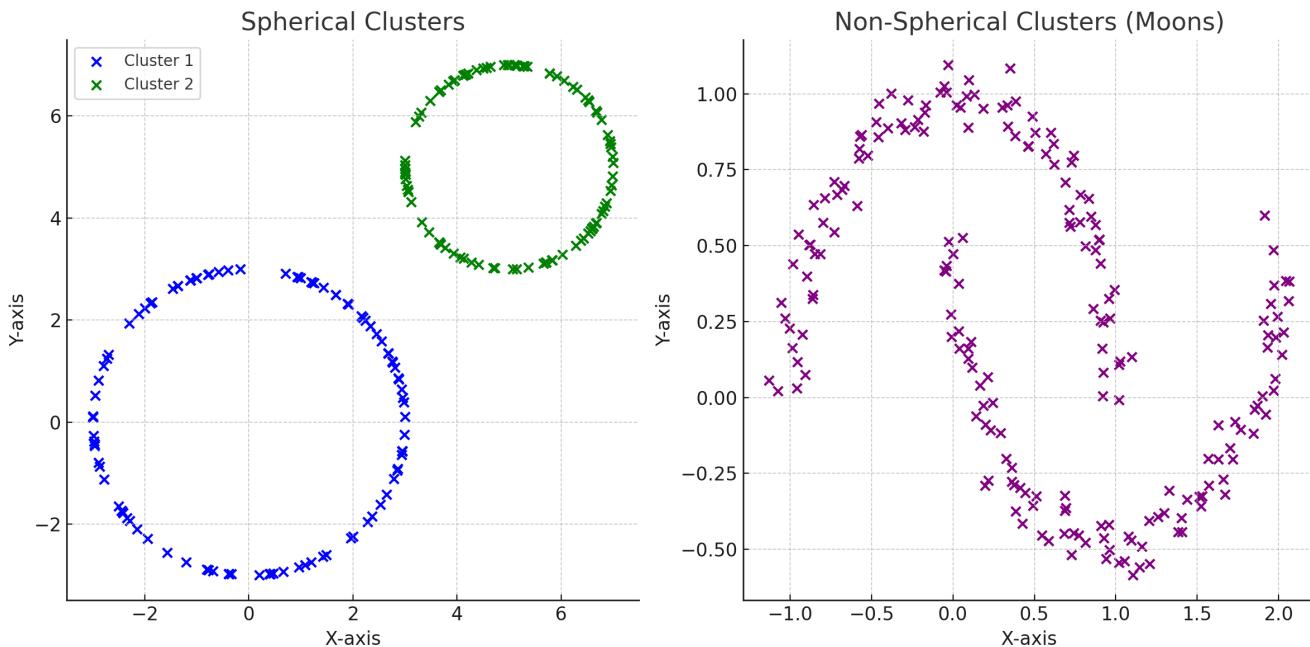
- 需要正確設定Eps和MinPts，否則可能導致結果不佳。
- 對於高維數據，密度計算可能會受到「維度詛咒」的影響，導致效果下降。（可以做降維）
- 無法有效處理密度差異過大的簇。

方法比較

比較總結

特性	Hierarchical Clustering	K-Means	DBSCAN
簇形狀	通常為球形	球形	任意形狀
簇數需要指定	不需要	必須指定 k	不需要
對噪聲的處理	不佳	不佳	良好
計算效率	低（適用於小數據集）	高（適合大數據集）	中等
適用數據類型	小型數據集 低噪聲	大型數據集 球形均勻分佈簇	任意密度的數據集
應用場景	分析數據的結構 樹狀可視化	快速分群 性能要求高	密度驅動分群 噪聲數據分析

球形 vs 非球形



左側（球形簇）：

- 數據點分佈在兩個圓形範圍內，形成明顯的球形簇。
- 適合使用 K-Means 等基於距離的分群算法。

右側（非球形簇）：

- 數據呈現「月牙形」分佈，典型的非球形結構。
- 需要像 DBSCAN 這樣的算法來捕捉密度連續的結構。

商業問題定義

請問 應該選擇哪種問題來做分群？

1

根據訂單金額區分為高中低，預測訂單金額高的客戶，進行精準行銷，賺取最大利潤

2

區分不同行為模式的客群，並搭配相應的行銷活動，以符合消費者需求，賺取最大利潤

請問 應該選擇哪種問題來做分群？

1

根據訂單金額區分為高中低，預測訂單金額高的客戶，進行精準行銷，賺取最大利潤

2

區分不同行為模式的客群，並搭配相應的行銷活動，以符合消費者需求，賺取最大利潤

探討客群分析中的分類和分群問題

分析目的

分類

分群

應用情境

精準行銷

客戶輪廓

分析指標

針對該名單進行廣告投放，或促銷活動等，有效提高收益

分類模型的準確度等

針對各群體特性在各種時間段，或各種通路，推出不同方案

清楚的客群特徵區別及描述

探討客群分析中的分類和分群問題

分析目的

分類

分群

應用情境

精準行銷

客戶輪廓

分析指標

針對該名單進行廣告投放，或促銷活動等，有效提高收益

分類模型的準確度等

針對各群體特性在各種時間段，或各種通路，推出不同方案

清楚的客群特徵區別及描述

THANKS