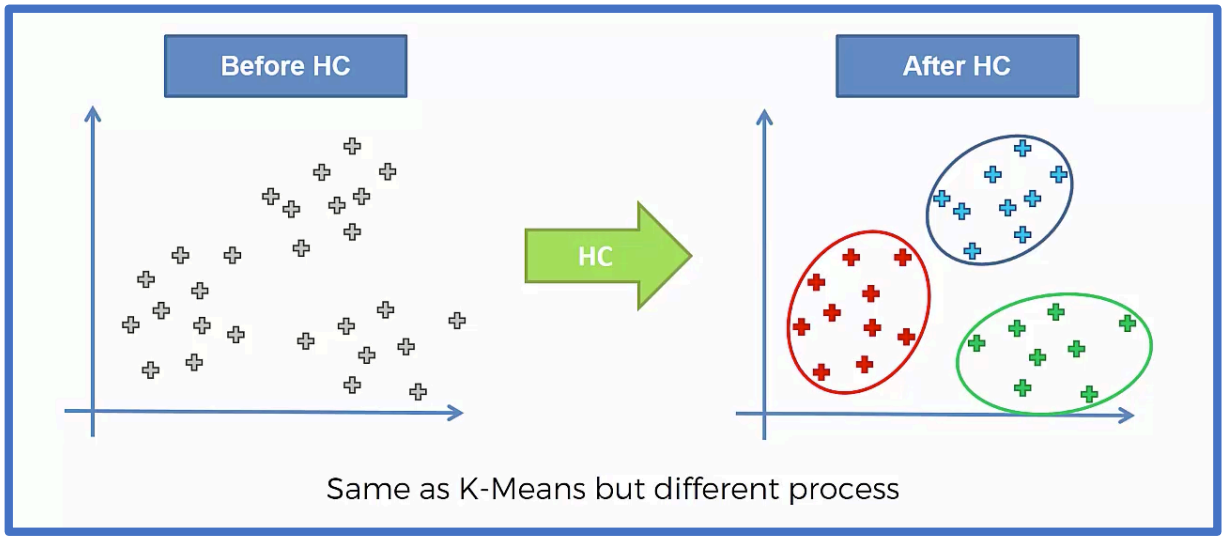
Hierarchical Clustering



一、說明

階層式分群法透過一種階層架構的方式，將資料層層反覆地進行分裂或聚合，以產生最後的樹狀結構，常見的方式有兩種：

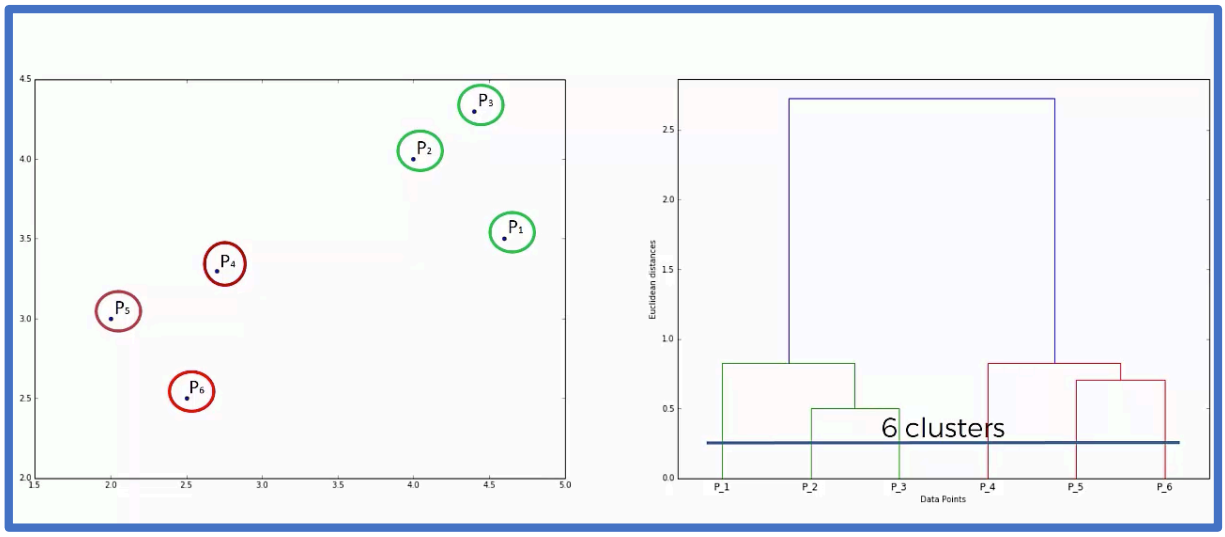
- 聚合式階層分群法 (Bottom-up, agglomerative) : 如果採用聚合的方式，階層式分群法可由樹狀結構的底部開始將資料或群聚逐次合併。
- 分裂式階層分群法 (Top-down, divisible) : 如果採用分裂的方式，則由樹狀結構的頂端開始，將群聚逐次分裂。

二、實作流程

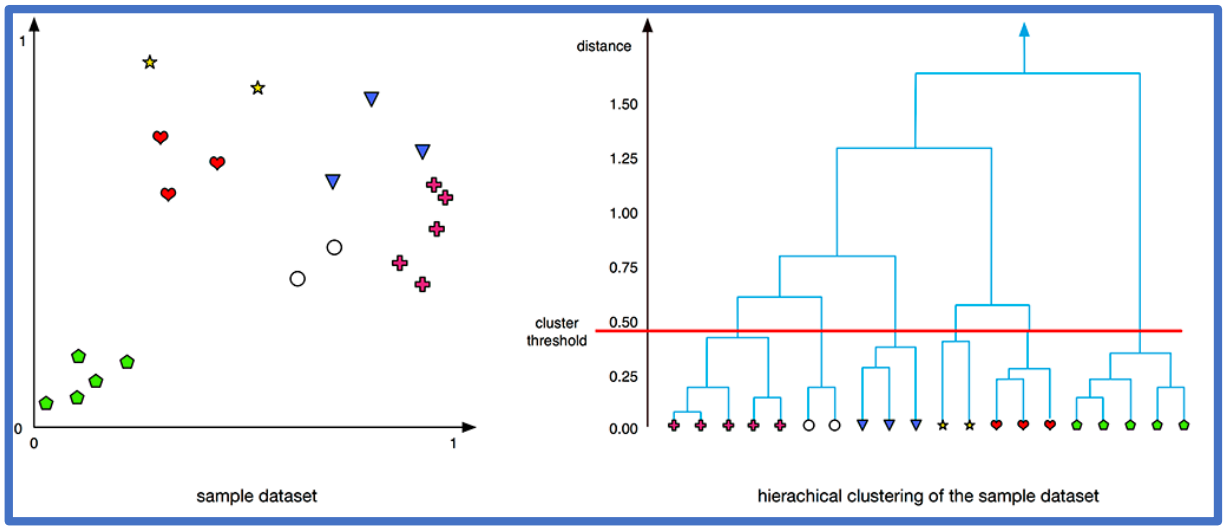
聚合式階層分群法（Hierarchical Agglomerative Clustering, HAC）
由樹狀結構的底部開始層層聚合，一開始我們將每一筆資料視為一個群聚（cluster），假設我們現在擁有 n 筆資料，則將這 n 筆資料視為 n 個群聚，亦即每個群聚包含一筆資料：

1. 將每筆資料視為一個群聚 $C_i, i = 1\ 1\ to\ n$
2. 找出所有群聚間，距離最接近的兩個群聚 C_i, C_j
3. 合併 C_i, C_j 成為一個新的群聚
4. 假如目前的群聚數目多於我們預期的群聚數目，則反覆重複步驟二至四，直到群聚數目已將降到我們所要求的數目。

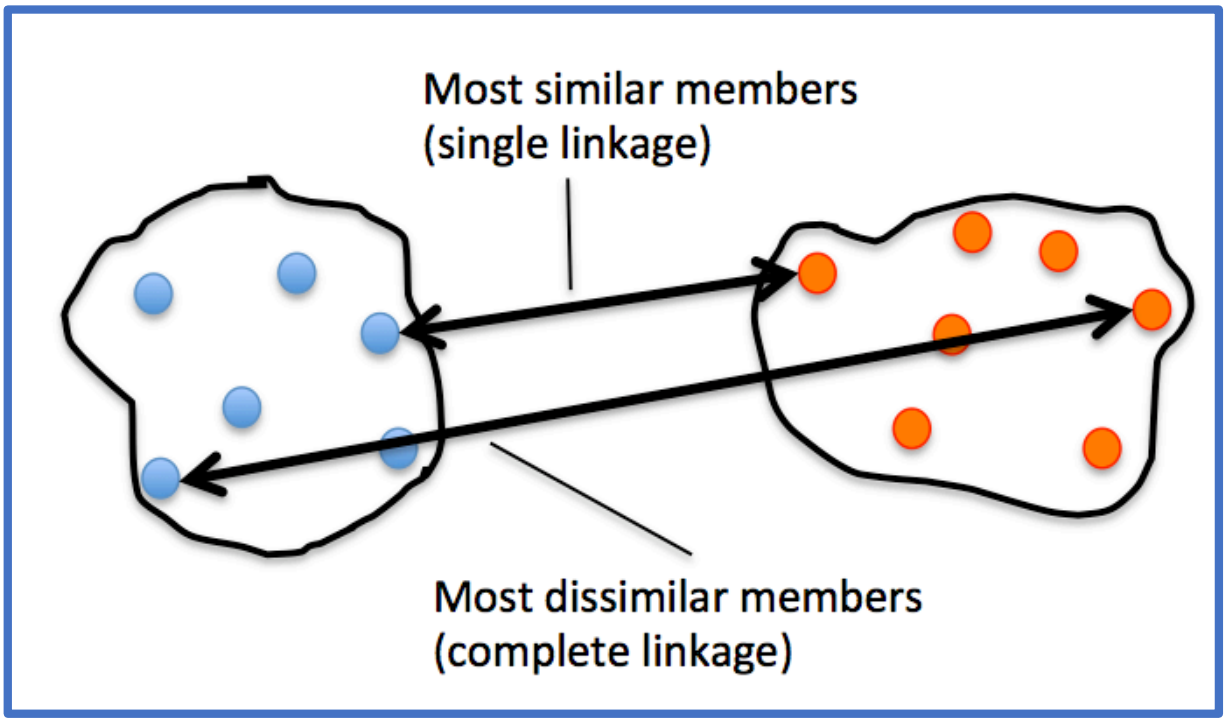
圖一



圖二



三、定義兩個群聚之間的距離



- 單一連結聚合演算法(single-linkage agglomerative algorithm)：群聚與群聚間的距離可以定義為不同群聚中最接近兩點間的距離。

$$d(C_i, C_j) = \min_{a \in C_i, b \in C_j} d(a, b)$$

- 完整連結聚合演算法(complete-linkage agglomerative algorithm)：群聚間的距離定義為不同群聚中最遠兩點間的距離，這樣可以保證這兩個集合合併後, 任何一對的距離不會大於 d。

$$d(C_i, C_j) = \max_{a \in C_i, b \in C_j} d(a, b)$$

- 平均連結聚合演算法(average-linkage agglomerative algorithm)：群聚間的距離定義為不同群聚間各點與各點間距離總和的平均。

$$d(C_i, C_j) = \sum_{a \in C_i, b \in C_j} \frac{d(a, b)}{|C_i||C_j|}$$

where $|C_i|$ and $|C_j|$ are the sizes for $|C_i|$ and $|C_j|$, respectively.

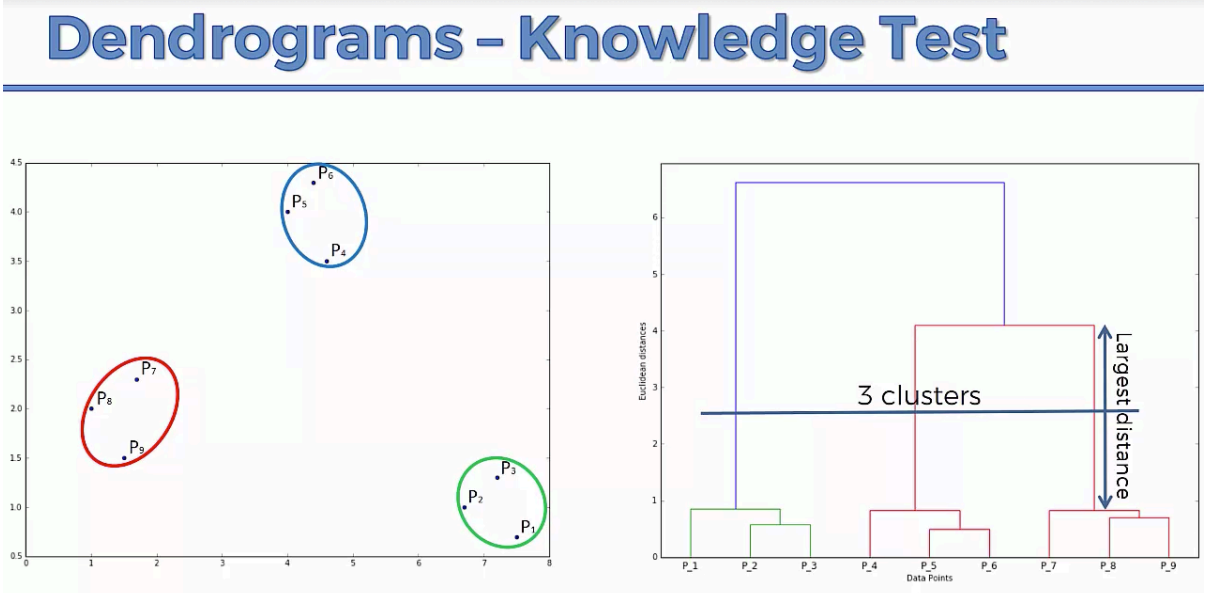
- 沃德法（Ward's method）：群聚間的距離定義為在將兩群合併後，各點到合併後的群中心的距離平方和。

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

where μ is the mean vector of $C_i \cup C_j$.

四、最佳化

方法：尋找最長的垂直線，並且不能有水平線跨過。



參考資料

1. http://mropengate.blogspot.tw/2015/06/ai-ch17-6-clustering-hierarchical.html
2. http://mirlab.org/jang/books/dcpr/dcHierClustering.asp?title=3-2%20Hierarchical%20Clustering%20(%B6%A5%BCh%A6%A1%A4%C0%B8s%AAk)&language=chinese
3. http://yourgene.pixnet.net/blog/post/117264518-%E6%B7%BA%E8%AB%87%E8%81%9A%E5%90%88%E5%BC%8F%E9%9A%8E%E5%B1%A4%E5%88%86%E7%BE%A4%E6%B3%95%E8%88%87%E7%86%B1%E5%9C%96