

非監督式學習(Unsupervised Learning)

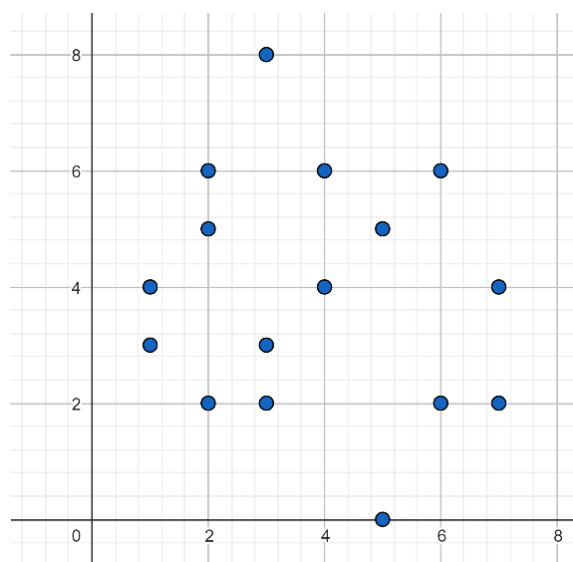
非監督式學習的訓練資料沒有標準答案，機器只能自行摸索、找出潛在的規則進行分類，因此這種學習方式通常用來處理「分群(Clustering)」問題。利用現有的資料特徵分成不同群體，每個群體之間的特徵相似。為了讓資料間的關聯性更加接近，會將特徵值數值化計算資料間的相近程度。

K-平均演算法(K-means Clustering)是先設定要分群的數量，將相近資料彼此分在同一群體，其概念就像是國中數學裡的「重心」，透過公式求得群體間的距離關係，再將資料逐漸分群。

K-means 執行步驟

- 步驟一：設定 K 值，代表接下來要將資料分 K 群
- 步驟二：任意指定座標平面上的 K 個點，作為初始分群中心點
- 步驟三：用「歐幾里得距離」計算座標上各點與初始分群中心點距離
- 步驟四：經由距離關係決定座標點歸屬於哪一個群體中
- 步驟五：根據分群結果以「算術平均」來求得新的分群中心點
- 步驟六：重複步驟三到步驟五，直到分群結果與分群中心點不再變動

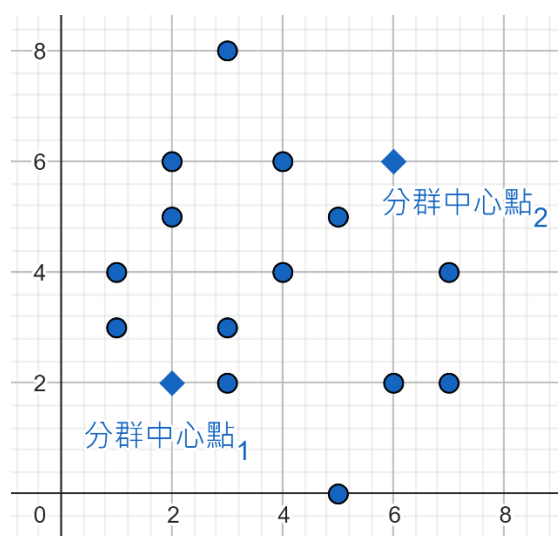
座標上16點分別(1, 3)、(1, 4)、(2, 2)、(2, 5)、(2, 6)、(3, 2)、(3, 3)、(3, 8)、(4, 4)、(4, 6)、(5, 0)、(5, 5)、(6, 2)、(6, 6)、(7, 2)、(7, 4)。



圖：座標點分布圖

步驟一與步驟二：

設定K值為2，初始分群中心點為(2, 2)和(6, 6)。



圖：分群中心點分布圖

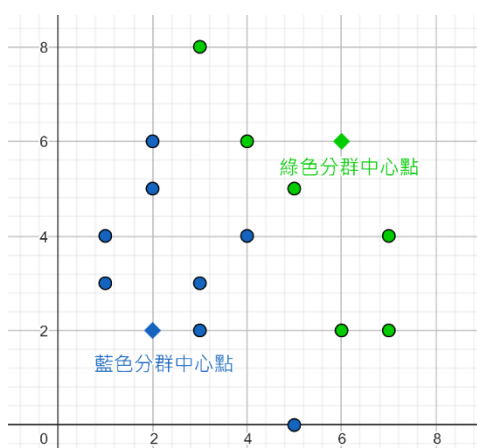
步驟三：

計算座標上各點與初始分群中心點(2, 2)和(6, 6)距離。

座標(x, y)	與(2, 2)距離	與(6, 6)距離	座標所在歸屬
(1, 3)	1.14	5.83	(2, 2)
(1, 4)	2.24	5.39	(2, 2)
(2, 5)	3	4.12	(2, 2)
(2, 6)	4	4	(2, 2) or (6, 6)
(3, 2)	1	5	(2, 2)
(3, 3)	1.41	4.24	(2, 2)
(3, 8)	6.08	3.61	(6, 6)
(4, 4)	2.83	2.83	(2, 2) or (6, 6)
(4, 6)	4.47	2	(6, 6)
(5, 0)	3.61	6.08	(2, 2)
(5, 5)	4.24	1.41	(6, 6)
(6, 2)	4	4	(2, 2) or (6, 6)
(7, 2)	5	4.12	(6, 6)
(7, 4)	5.39	2.24	(6, 6)

步驟四：

經由距離關係決定座標點歸屬於哪一個群體中，分為藍綠兩群。



圖：步驟四座標點分布圖

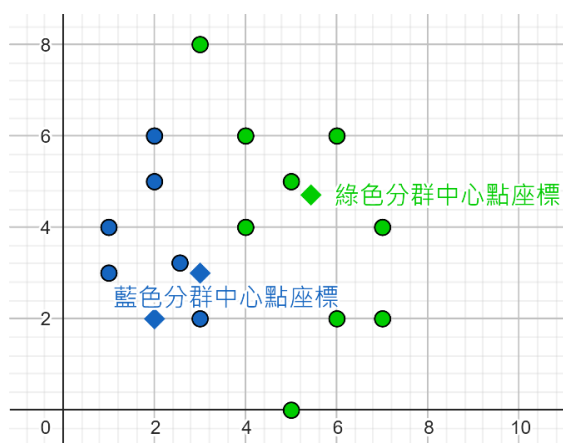
步驟五：

根據分群結果以「算術平均」來求得新的分群中心點(2.56, 3.22)和(5.43, 4.71)。

	(1, 3)、(1, 4)、(2, 2)、 (2, 5)、(2, 6)、(3, 2)、 (3, 3)、(4, 4)、(5, 0)	(3, 8)、(4, 6)、(5, 5)、 (6, 2)、(6, 6)、(7, 2)、 (7, 4)
新分群中心X	2.56	5.43
新分群中心Y	3.22	4.71

步驟六：

重複步驟三到步驟五，直到分群結果與分群中心點不再變動。



圖：步驟六座標點分布圖

階層分群法(Hierarchical Clustering)可以動態決定要分群的數量，這裡的「階層」代表分群數量階層，其又有分為兩種方法「聚合法(Bottom-up Clustering)」和「分裂法(Top-down Clustering)」。

「聚合法」概念如同「異中求同」，將所有資料先視作為不同的群，再找最相似的兩群，將其結合為一個新群。重複上述行為，直到聚合後的群數為目標群數。

「分裂法」概念如同「同中求異」，將所有資料先視作為相同的群，再依據群內的相異性，將其拆分為兩個群。重複上述行為，直到分裂後的群數為目標群數。

由於資料數值化的關係，資料會在座標上呈現不均勻分布。如果想要得知點和點、點和群、群和群間的距離關係，就要用到不同的計算距離方式。

中心連結(Centroid-linkage)

$$d(G_1, G_2) = d(\bar{a}, \bar{b})$$

$$\bullet G_1 \in a, G_2 \in b$$

在不同的兩群中，選擇各群的中心點，即為兩群距離

單一連結(Single-linkage)

$$d(G_1, G_2) = \min(a, b)$$

$$\bullet G_1 \in a, G_2 \in b$$

在不同的兩群中，選擇最短距離兩點，即為兩群距離

平均連結(Average-linkage)

$$d(G_1, G_2) = \frac{\sum d(a, b)}{|G_1||G_2|}$$

$$\bullet G_1 \in a, G_2 \in b$$

在不同的兩群中，各點之距離的平均，即為兩群距離