# K-means Clustering

## K-means 描述

已知觀測集(x1，x2，...，xn)，其中每個觀測都是一個 d-維實向量，k-means 分類要把這 n 個觀測劃分到 k 個集合中(k≤n)，使得組內平方和（WCSS within-cluster sum of squares）最小。

換句話說，它的目標是滿足下列函式：

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x}\in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2$$

其中 μi 是 Si 中所有點的群中心。

已知初始的 k 個群中心$m_1^{(1)}$，...，$m_k^{(1)}$，k-means 演算法會按照下面兩個步驟來交替進行：

1. **分配(Assignment)：**將每個觀測值分配到 K 個群集中，使組內平方和(WCSS)達最小。

   因為這一平方和就是平方後的歐氏距離，所以很直觀地把觀測值分配到離它最近的群中心即可：

$$S_i^{(t)} = \left\{ x_p : \left\| x_p - m_i^{(t)} \right\|^2 \le \left\| x_p - m_j^{(t)} \right\|^2 \forall j, 1 \le j \le k \right\}$$

其中每個 xp 都只被分配到一個確定的群集 St 中，儘管在理論上它可能被分配到 2 個或者更多的群集。

2. **更新(Update)**：計算上步得到的群集中每一群集的群中心，作為新群中心

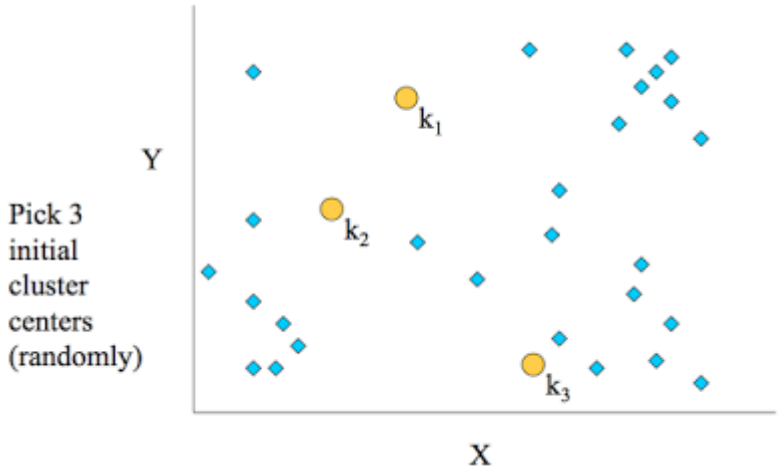
$$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j$$

因為算術平均是最小平方估計，所以這一步同樣減小了目標函式組內平方和（WCSS）的值。

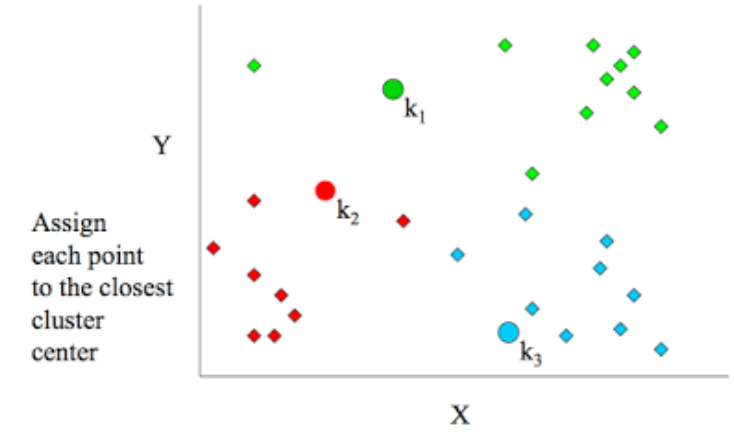
- 這一演算法將在對於觀測的分配不再變化時收斂。由於交替進行的兩個步驟都會減小目標函式 WCSS 的值，並且分配方案只有有限種，所以演算法一定會收斂於某一（局部）最優解。

◆ **下面我們用散點圖來表示 K-means 演算法：**

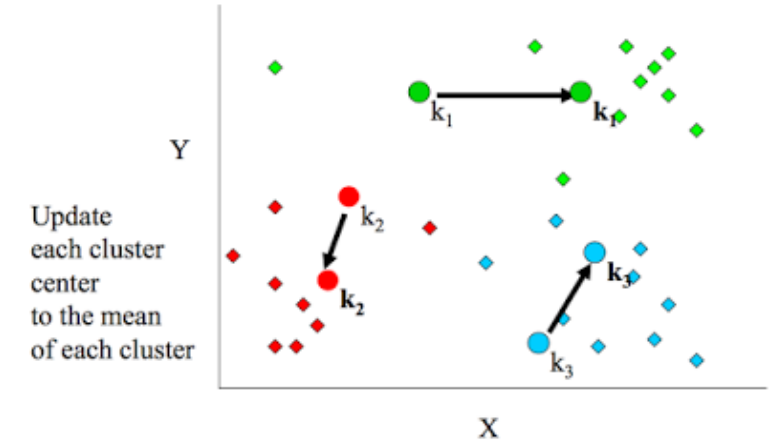
1. 假設欲分為 3 群，隨機選取 3 筆資料當作初始群中心($k_1$~$k_3$)



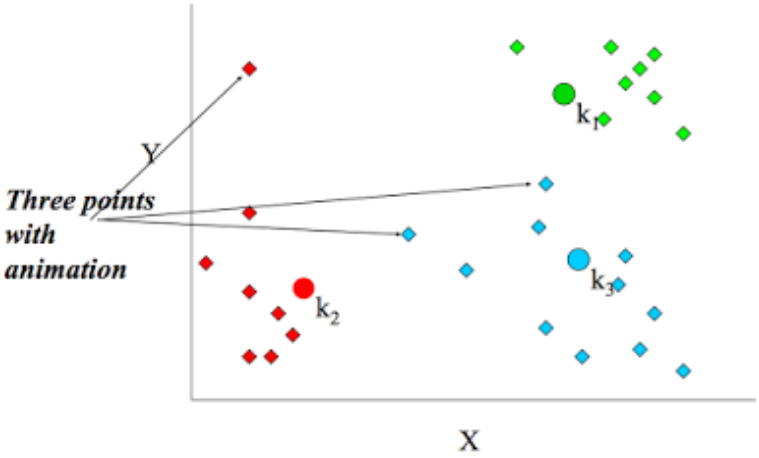
2. 計算每個資料 $x_i$ 對應到最短距離的群中心（分配）



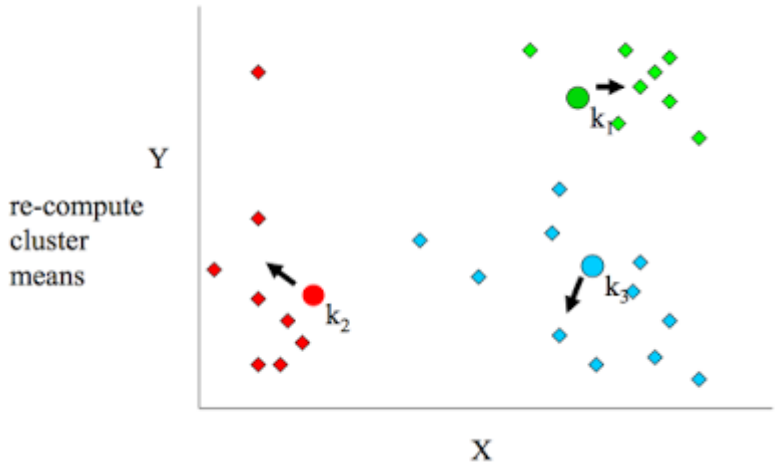

3. 利用目前得到的分類重新計算群中心（更新）



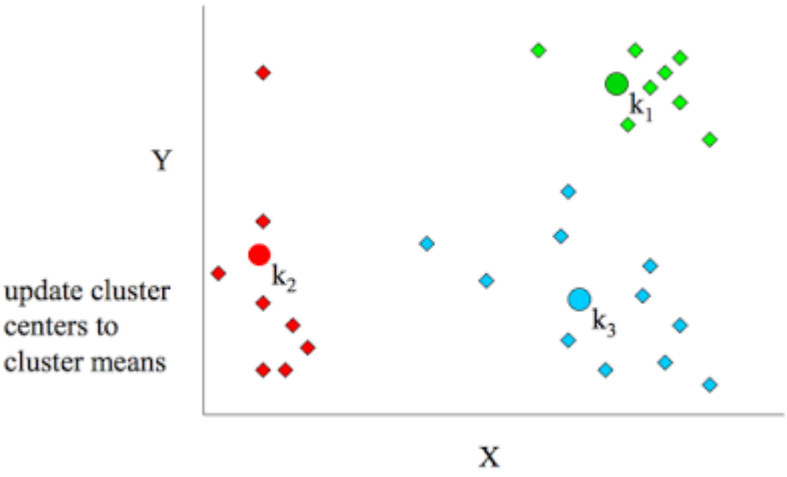

4. 再次計算每個資料 $x_i$ 對應到最短距離的群中心（分配）



5. 再次重新計算群中心（更新）




6. 不斷重複以上的動作（分配->更新->分配->更新），直到收斂至最佳解



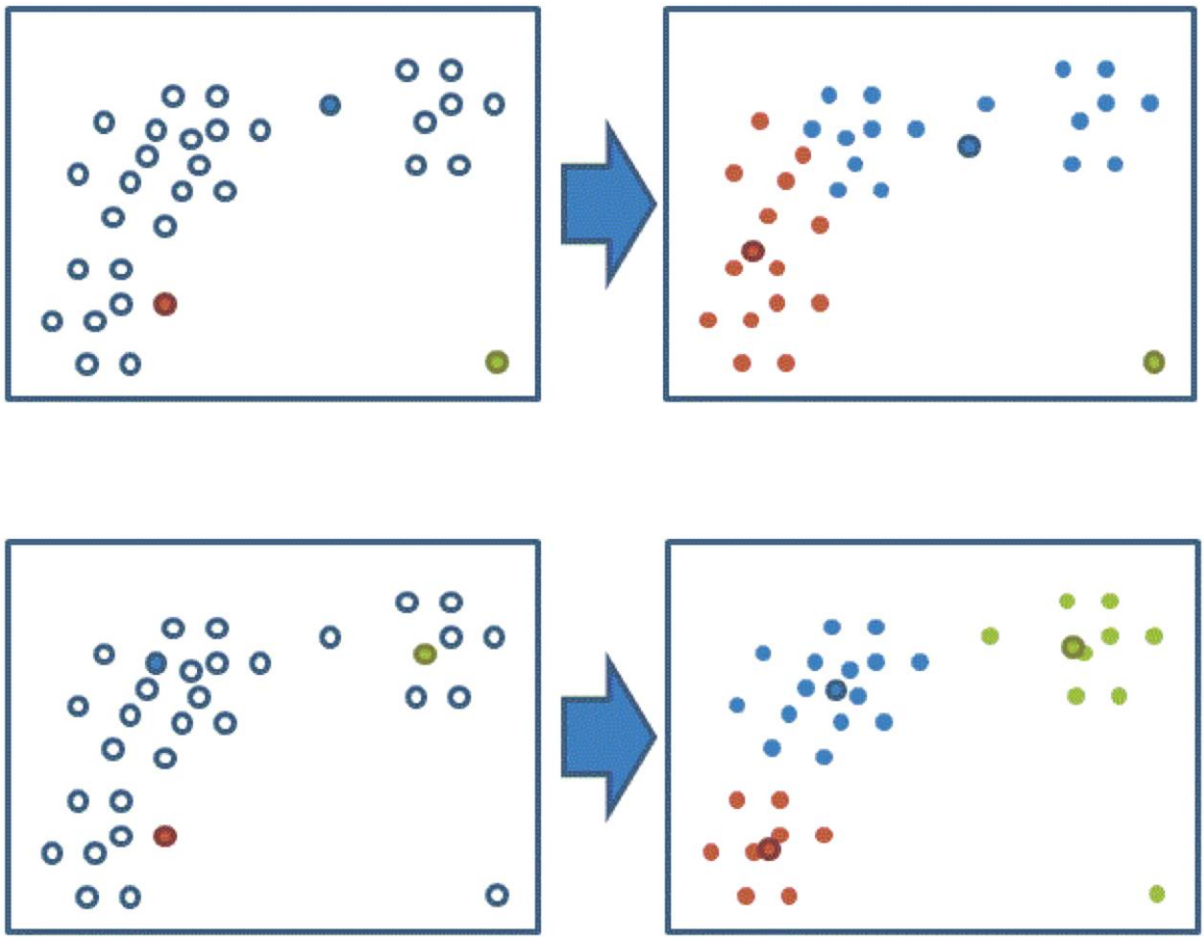
## 決定初始群中心 – Kmeans++

大概了解了「K means」的流程之後

我們可以發現一開始的群中心是隨機的

也就是說同一筆資料用「K means」跑 10 次，10 次的結果可能都不同

讓我們來看一個極端點的例子：



由上圖可以發現，如果初始群中心設定的不好可能導致不會的結果。

為了解決上述問題，改進的 K-menas 算法，即 **K-means++**算法被提出

K-means++算法主要是為了能夠在初始決定群中心時選擇較優的群中心

- **K-means++算法的基本原則是使得初始群中心之間的相互距離盡可能遠**

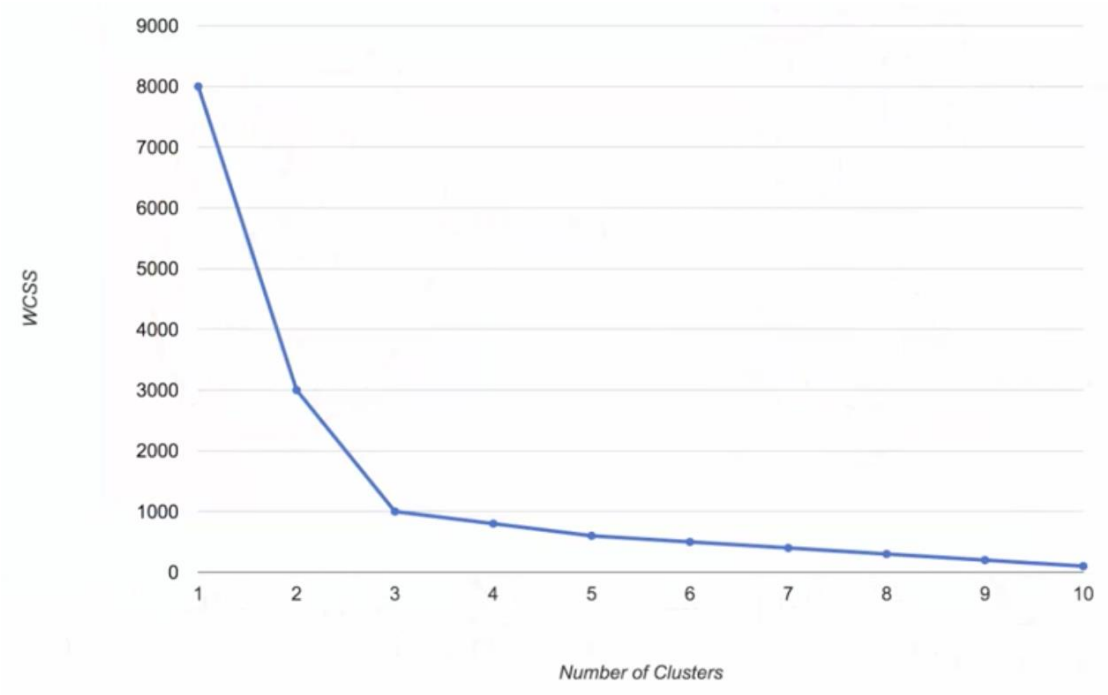
◆ K-means++算法的初始化過程如下：

1. 從觀測值中隨機選擇一個點作為第一個群中心
2. 計算每個觀測值與已選擇的群中心距離 D
3. 選擇一個新的觀測值作為新的群中心，選擇的原則是：D 較大的點，被選取作為群中心的機率較大
4. 重複 2 和 3 直到 k 個群中心被選出來
5. 利用這 k 個初始的群中心來運行標準的 k-means 計算

## 選擇分類群數 K - Elbow Method(肘部法)

如何決定分類群數 K 非常重要，使用 Elbow Method 來估計最佳分類群數 K

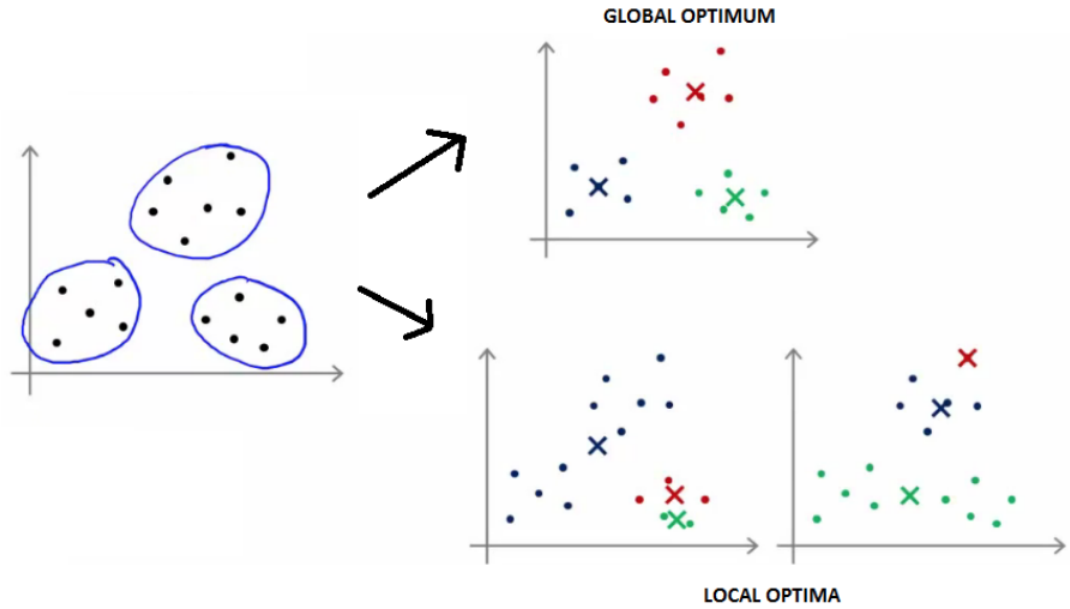
Elbow Method 是將 K 分別設置為 2、3、4...，讓組內平方和(WCSS)逐漸收斂，如下圖所示：




- 從圖中可以看出，K 值從 1 到 3 時，平均畸變程度變化最大。超過 3 以後，平均畸變程度變化顯著降低，因此肘部就是 3，選擇 K=3 作為分類群數來進行 K-means 運算

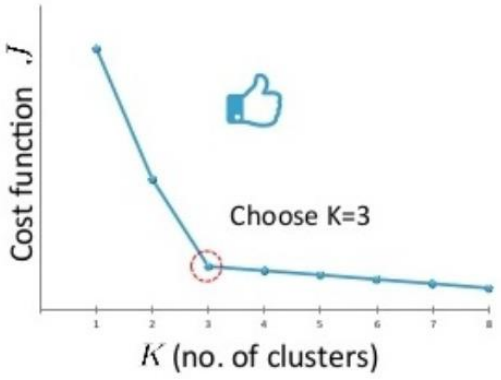
## 需注意的問題

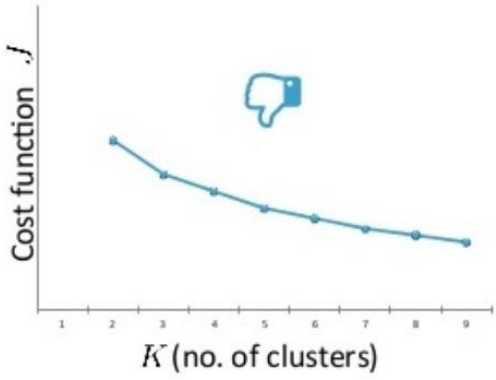
1. 局部最佳解(LOCAL OPTIMA)




➢ 解決方法：重新選擇初始群中心

2. 肘部不明顯




➢ 解決方法：可根據 K-means 算法後續的目標進行選擇