Day 55 非監督式機器學習

K-means 聚類算法

出題教練 周俊川

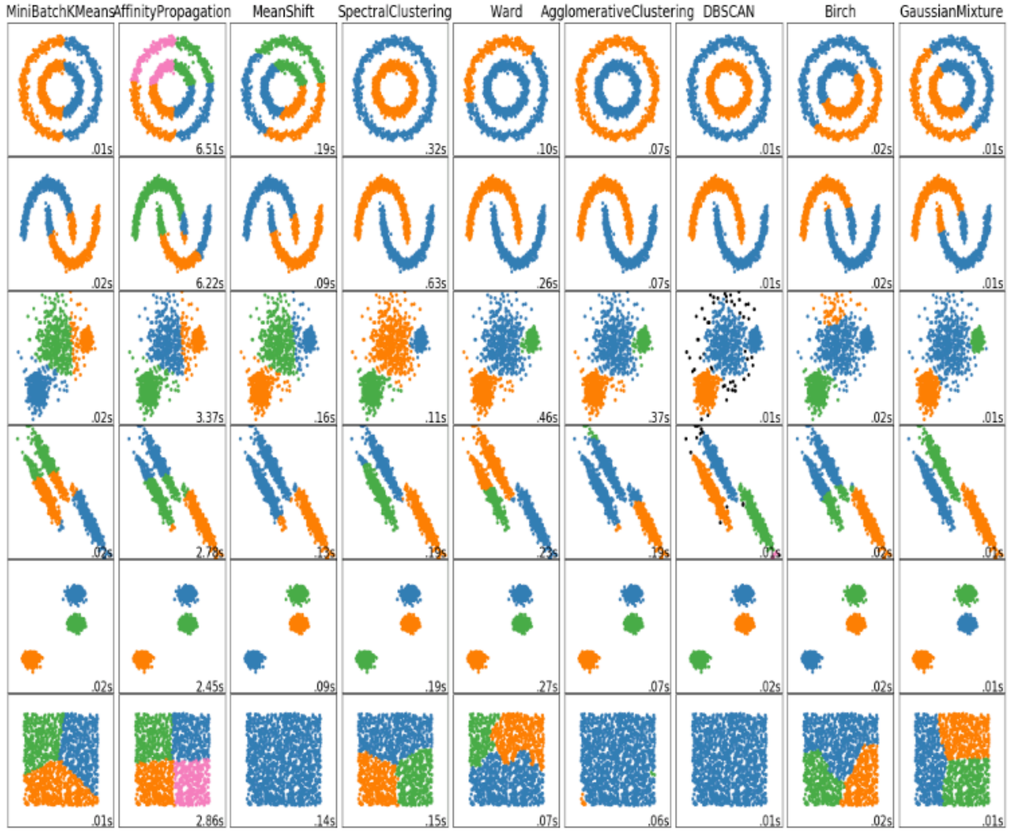
本日知識點目標

- 聚類算法與監督式學習的差異
- K-means 聚類算法簡介
- K-means 聚類算法的參數設計

聚類算法簡介

聚類算法用於把族群或資料點分隔成一系列的組合，使得相同 cluster 中的資料點比其他的組更相似。

2.3.1. Overview of clustering methods

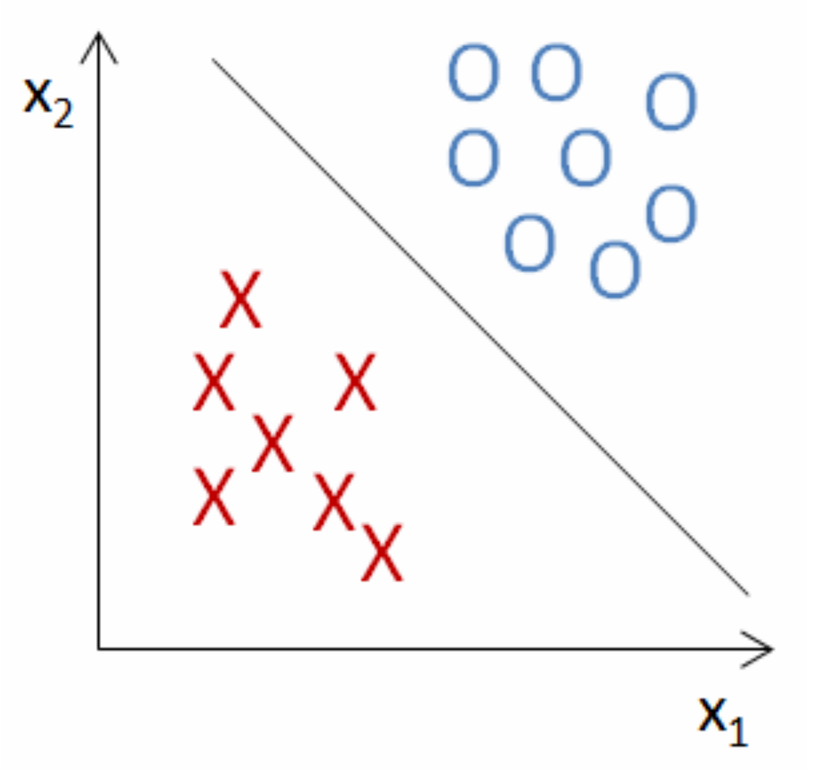


圖片來源：scikit-learn

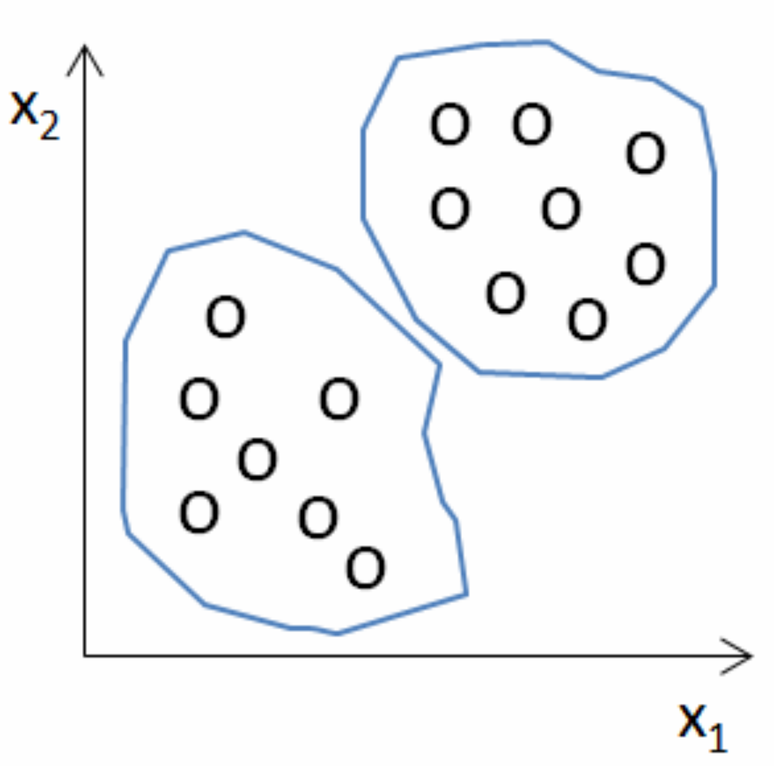
Supervised learning vs. clustering

CUPOY

監督式學習目標在於找出決策邊界 (decision boundary)



Clustering 目標在於找出資料結構

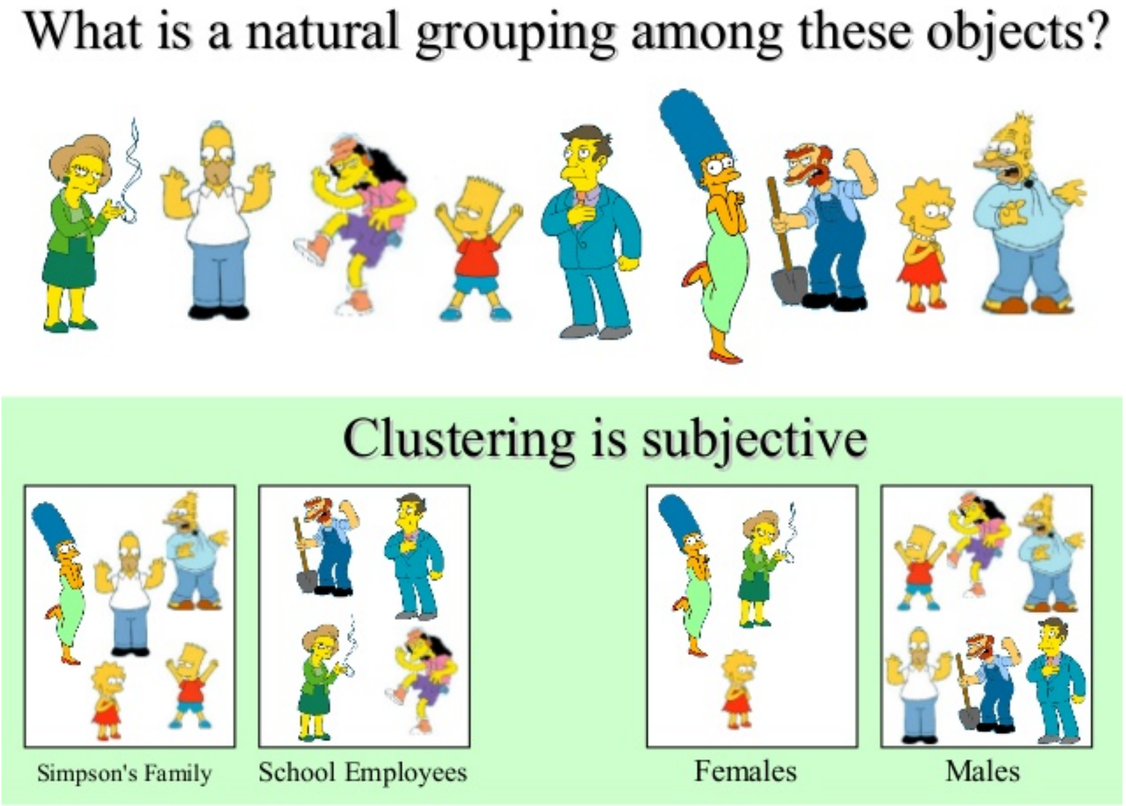


圖片來源：murphymind.blogspot

Why clustering?

CUPOY

在資料還沒有標記、問題還沒定義清楚時，聚類算法可以幫助我們理解資料特性，評估機器學習問題方向等，也是一種呈現資料的方式。



圖片來源：slideshare.com

K-means 聚類算法

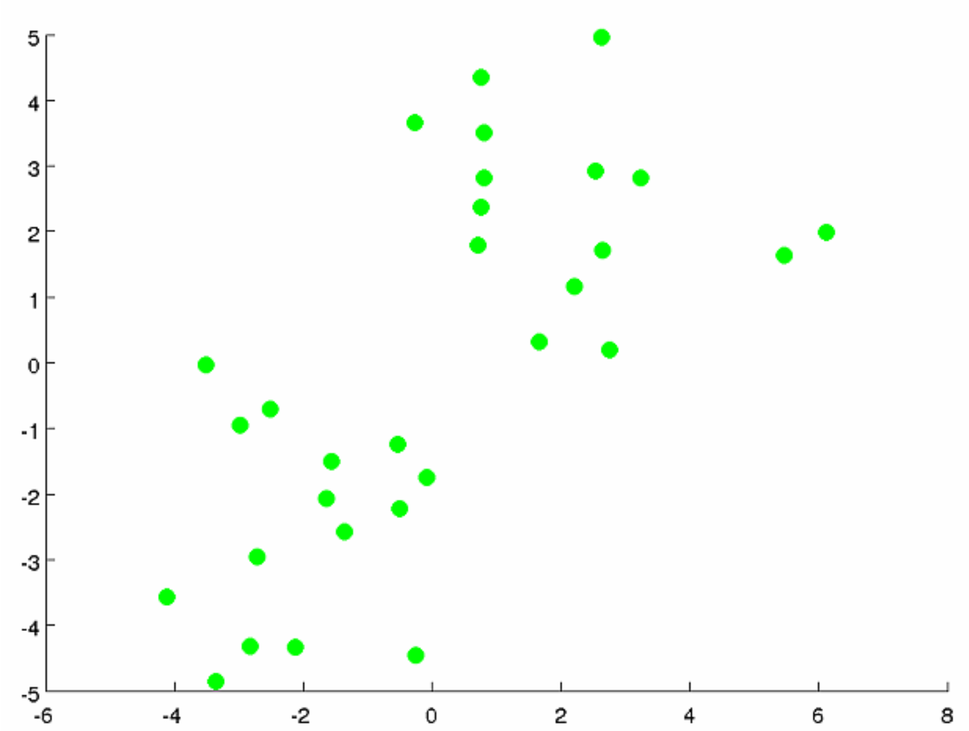
CUPOY

- 把所有資料點分成 k 個 cluster，使得相同 cluster 中的所有資料點彼此儘量相似，而不同 cluster 的資料點儘量不同。
- 距離測量（e.g. 歐氏距離）用於計算資料點的相似度和相異度。每個 cluster 有一個中心點。中心點可理解為最能代表 cluster 的點。

K-means 算法流程（一）

CUPOY

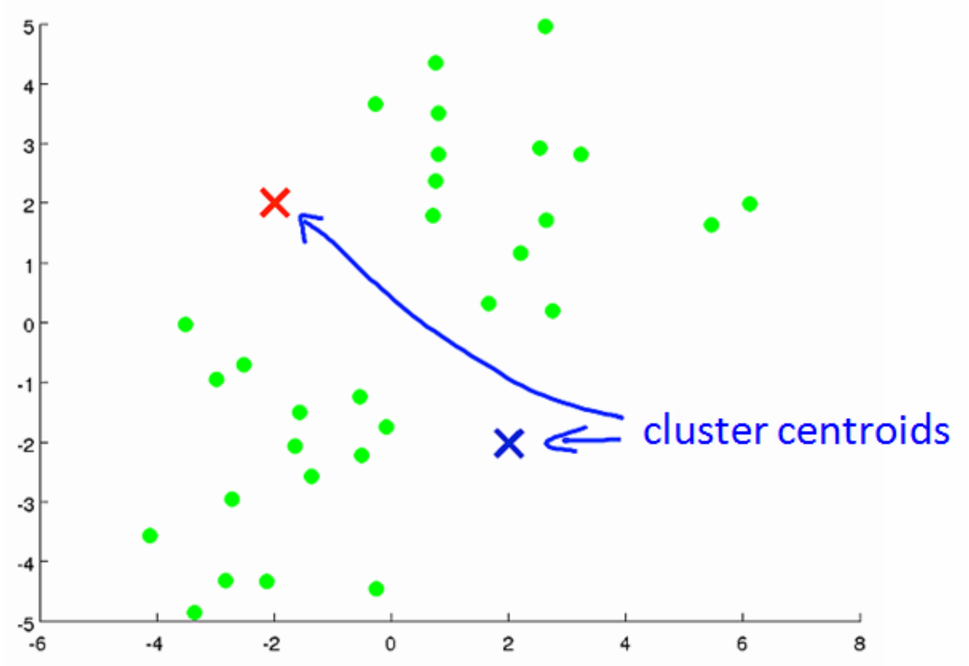
假設下圖是我們的 training set，我們目標是將資料分成 2 群



圖片來源：murphymind.blogspot

K-means 算法流程（二）

隨機選取 2 個點，稱爲 cluster centroid.

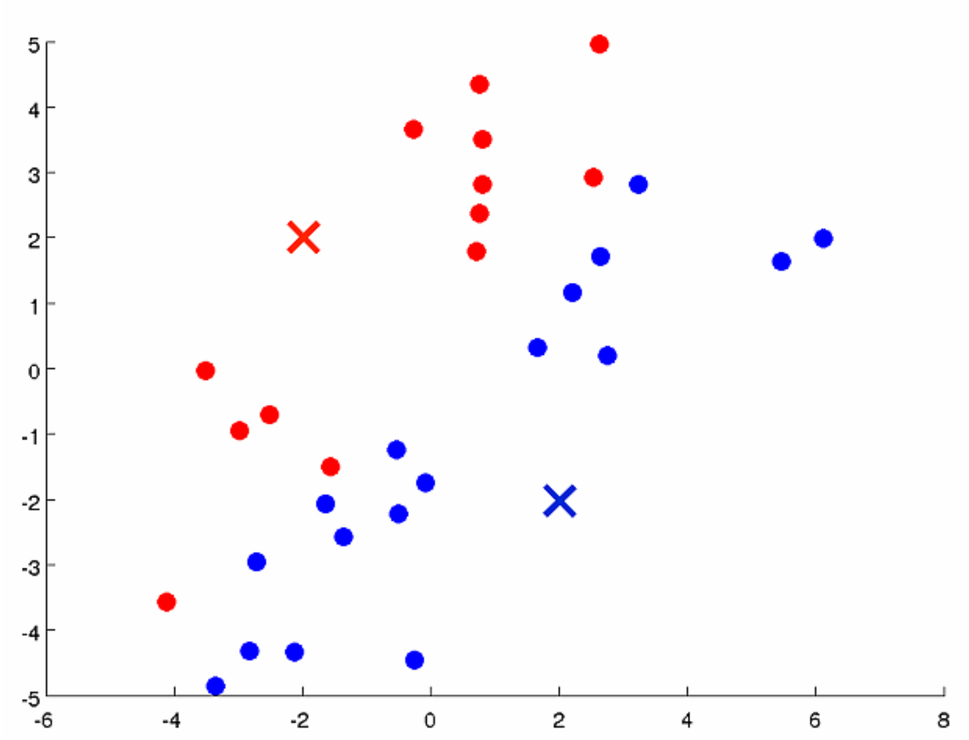


圖片來源：murphymind.blogspot

K-means 算法流程（三）

CUPOY

對每一個 training example 根據它距離哪一個 cluster centroid 較近 ， 標記爲其中之一 (cluster assignment)

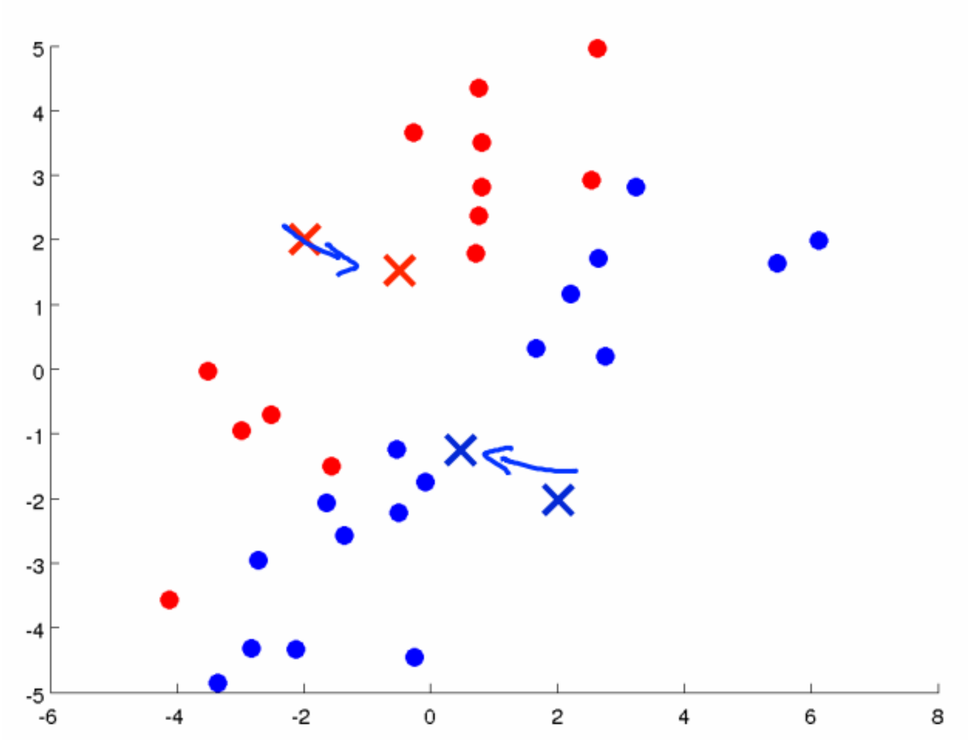


圖片來源：murphymind.blogspot

K-means 算法流程（四）

CUPOY

- 然後把 centroid 移到同一群 training examples 的中心點 (update centroid)
- 反覆進行 cluster assignment 及 update centroid, 直到 cluster assignment 不再導致 training example 被 assign 爲不同的標記 (算法收斂)



圖片來源：murphymind.blogspot

Optimization Objective

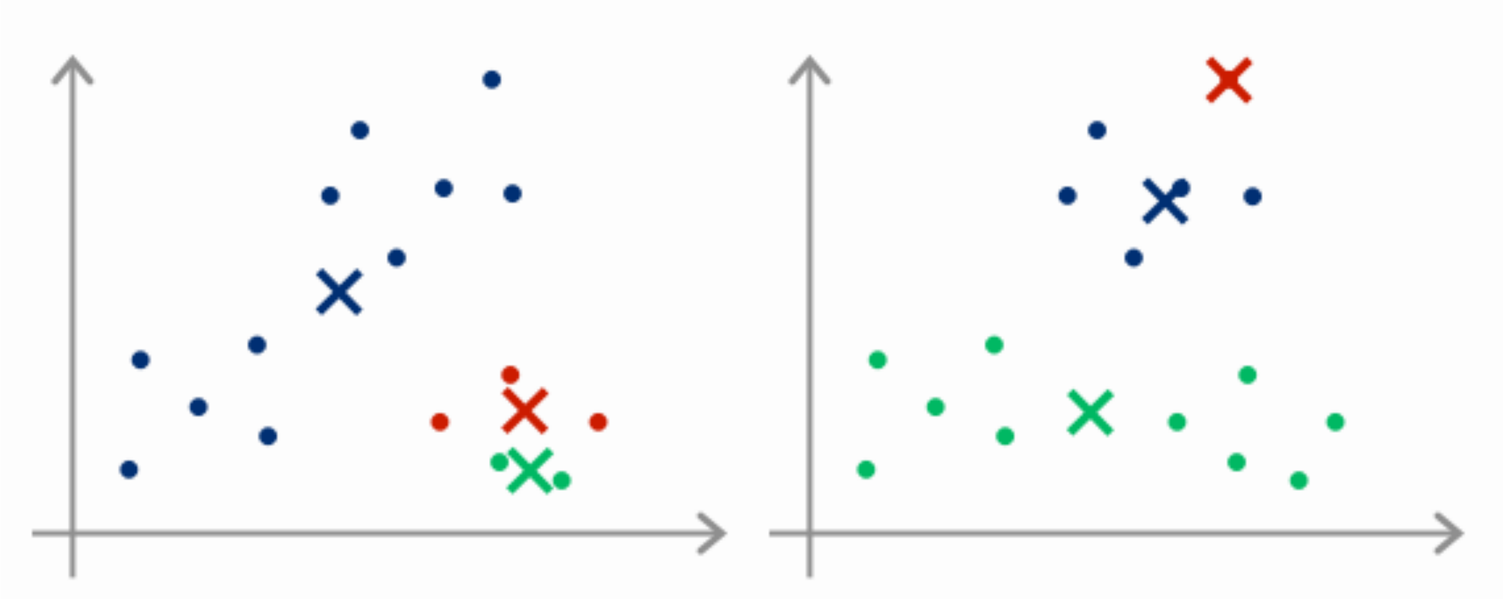
CUPOY

K-means 目標是使總體群內平方誤差最小

$$\sum_{i=0}^{n} \min_{\mu_j \epsilon C} (||x_i - \mu_j||^2)$$

K-means 注意事項 （一）

Random initialization: initial 設定的不同，會導致得到不同 clustering 的結果，可能導致 local optima，而非 global optima。



圖片來源：murphymind.blogspot

K-means 注意事項 （二）

CUPOY

因爲沒有預先的標記，對於 cluster 數量多少才是最佳解，沒有標準答案，得靠手動測試觀察。

重要知識點複習

CUPOY

- 當問題不清楚或是資料未有標註的情況下，可以嘗試用分群算法幫助瞭解資料結構，而其中一個方法是運用 K-means 聚類算法幫助分群資料
- 分群算法需要事先定義群數，因此效果評估只能藉由人爲觀察。

解題時間
It’s Your Turn

請跳出PDF至官網Sample Code＆作業
開始解題