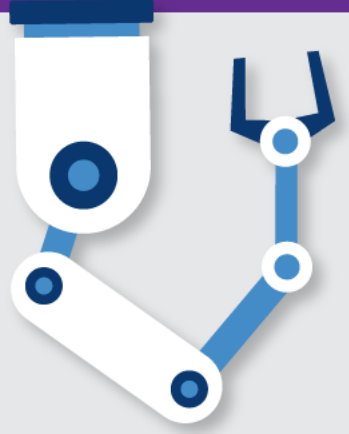
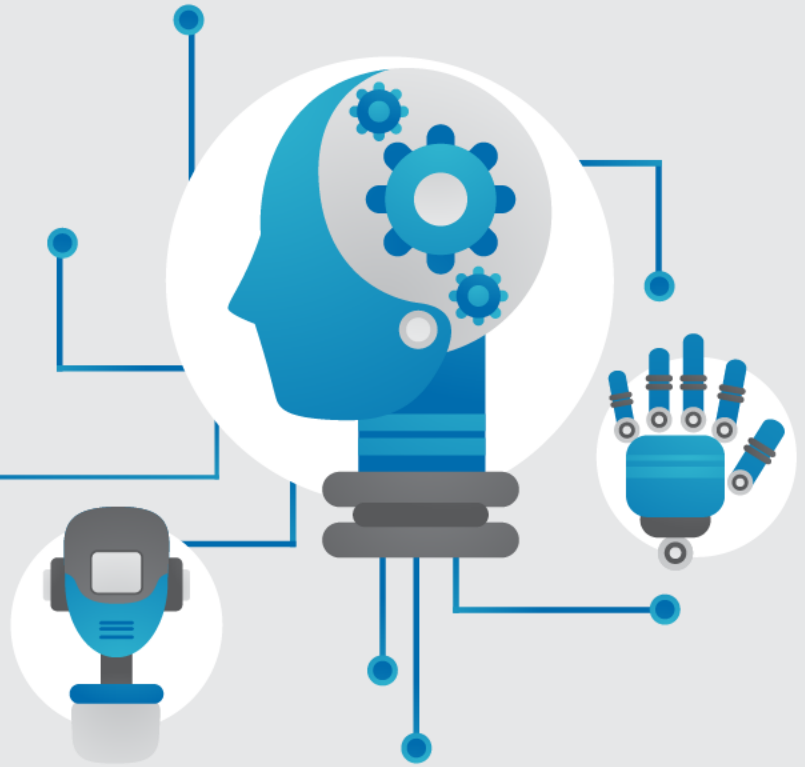
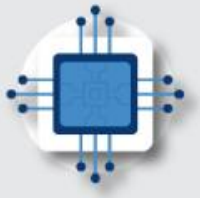


K-means 分群法



非監督式學習

- 非監督式學習是針對沒有事先標記過的資料，自動進行分類或分群。
- 非監督式學習的型態包含
  - 分群演算法（Clustering algorithm）：
    將資料分成不同群組，群組內的成員具有類似屬性。
    例如K-means演算法和DBSCAN 演算法。
  - 非監督式轉換（Unsupervised transformatio）：
    將原資料轉換為另一種表示方式，讓資料處理更為方便，例如降維（Dimension reduction）與特徵擷取（Feature extraction）。



非監督式學習

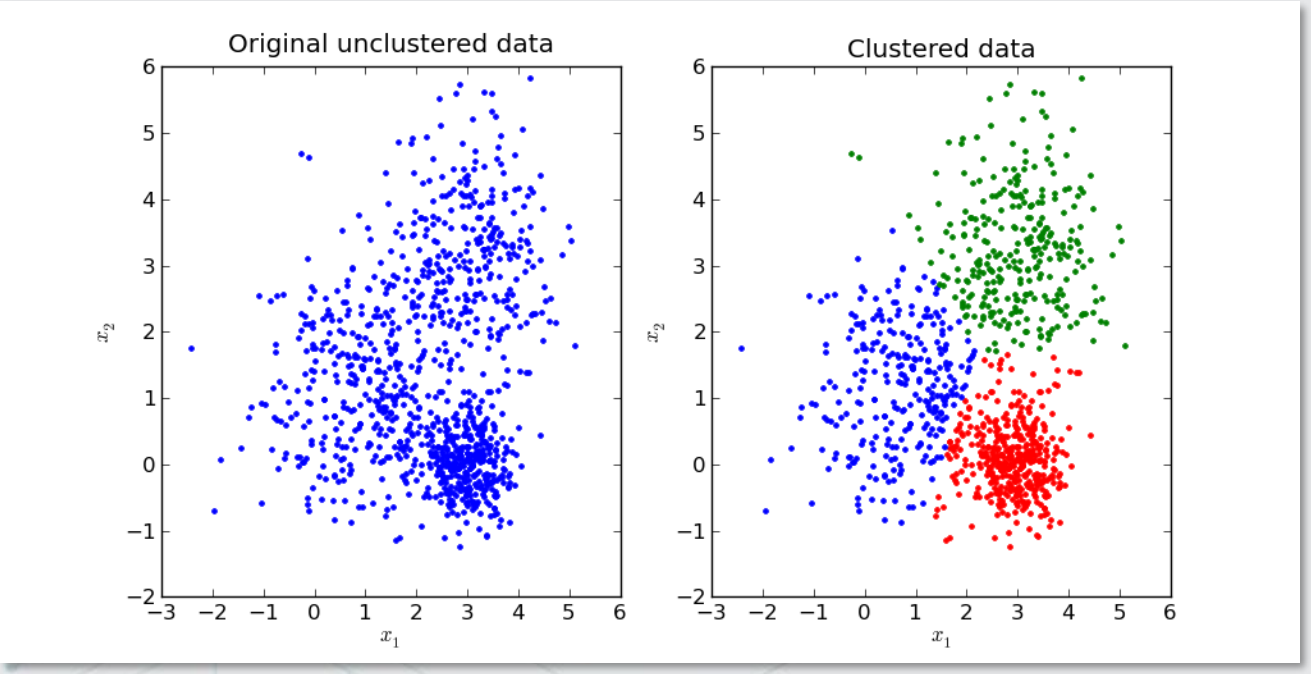
› 非監督式學習的特性

- 因為資料沒有輸出標籤，不易評估模型效率
- 非監督式學習常被用來了解資料集的特性，當作監督式學習的預先處理步驟，透過預先的資料分析，讓之後的監督式學習效率更好

K-means 演算法

- K-means演算法會將資料根據給定的K值進行分群，K代表群組數。

K = 3

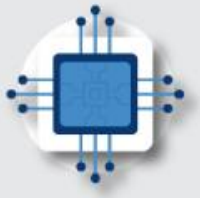


K-means 演算法

- K-means演算法的目標為最小化群組內資料和**群中心**的距離平方和：

  - x：S_i群內資料點
  - k：k群
  - S：群的分類
  - μ_i：S_i群中所有點的均值

$$\arg\min_{S} \sum_{i=1}^{k} \sum_{x \in S_i} \| x - \mu_i \|^2$$

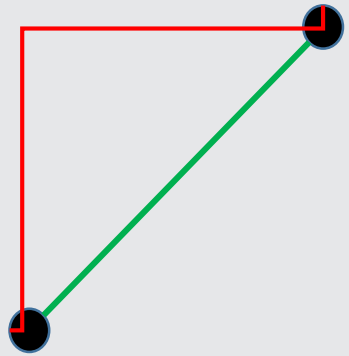


距離函數

- K-means演算法的距離函數計算
  - 歐幾里得距離（Euclidean distance）
  - 曼哈頓距離（Manhattan distance）
  - 餘弦相似（Cos similarity）

距離函數

- 假設 $x=(x_1,x_2,\ldots,x_n)$, $y=(y_1,y_2,\ldots y_n)$為兩個n維向量
- 歐幾里得距離（Euclidean Distance）
  - 函數：$\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$
  - 圖中**綠色斜線**為歐幾里得距離



- 曼哈頓距離（Manhattan Distance）
  - 函數：$\sum_{i=1}^{n}|x_i-y_i|$
  - 圖中**紅線**為曼哈頓距離

距離函數

- 假設 $x=(x_1,x_2,\ldots,x_n)$, $y=(y_1,y_2,\ldots y_n)$為兩個n維向量
- 餘弦相似（Cos similarity）

$$\text{Similarity (distance)} = \frac{x \cdot y}{||x|| \times ||y||} = \frac{\sum_{i=1}^{n} x_i \times y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \times \sqrt{\sum_{i=1}^{n} y_i^2}}$$

by歐幾里得點積公式： $a \cdot b = ||a|| \times ||b|| \times cos\theta$

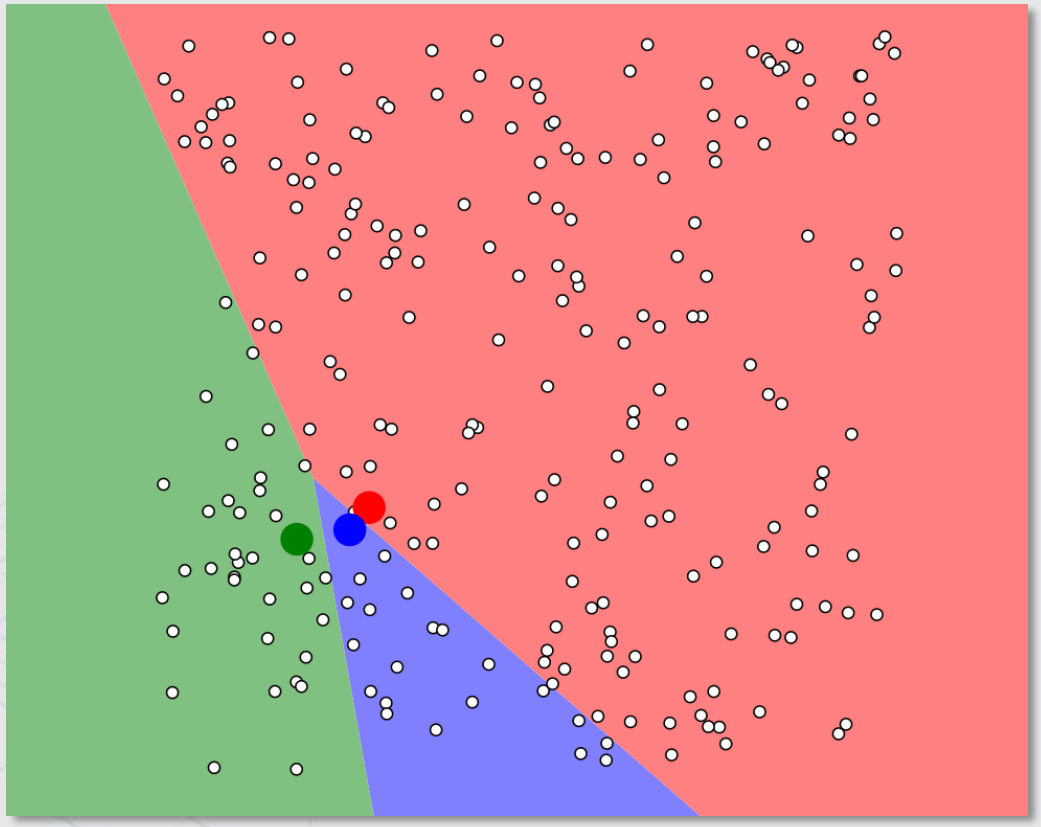
K-means 演算法步驟

› 演算法步驟

1. 決定k值，並隨意選取k個點當作群組中心
2. 將每一個點歸類於距離最近的群組中心
3. 重新計算分類後每個群組的中心
4. 重複步驟2、3直到群集不再變動或群組中心移動量小於設定的閾值

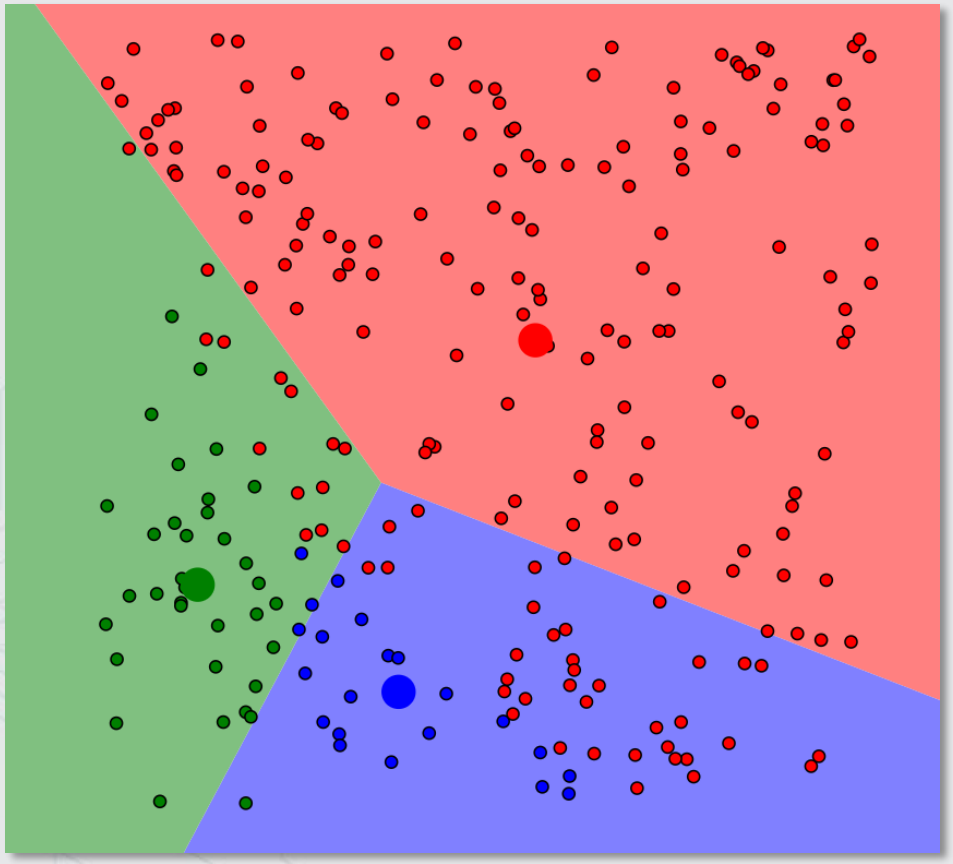
K-means 步驟說明

- 選定初始群中心，將每一個點歸類於距離最近的群集中心



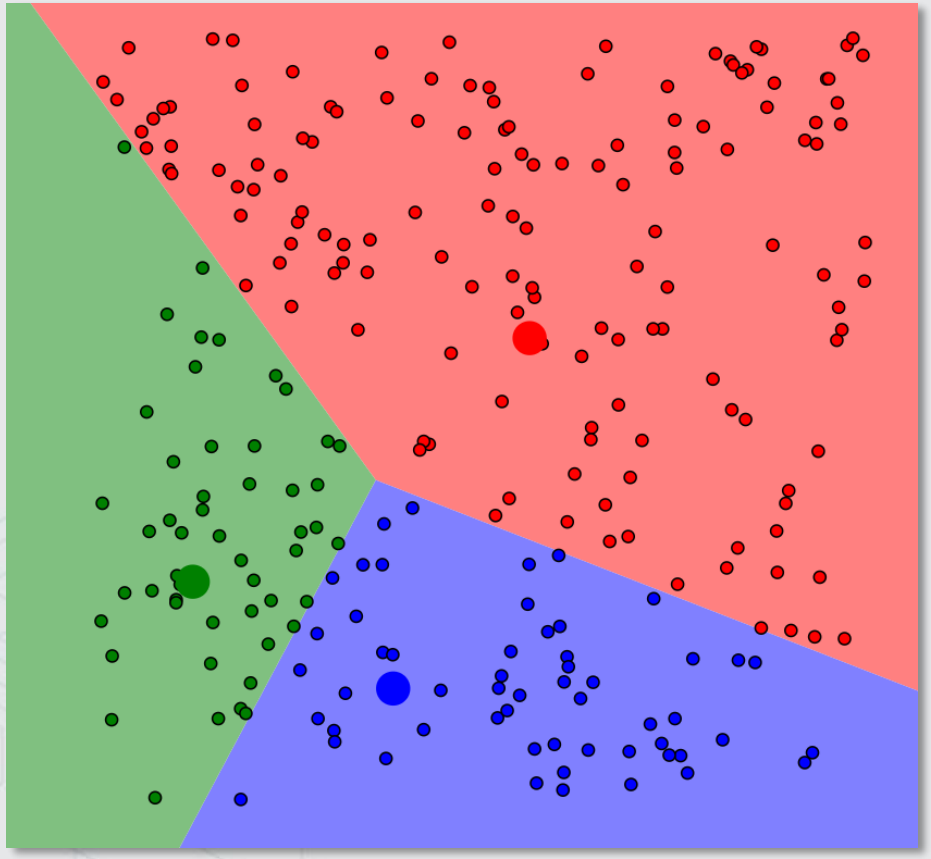
K-means 步驟說明

- 每群計算所有資料的平均值，更新群組中心



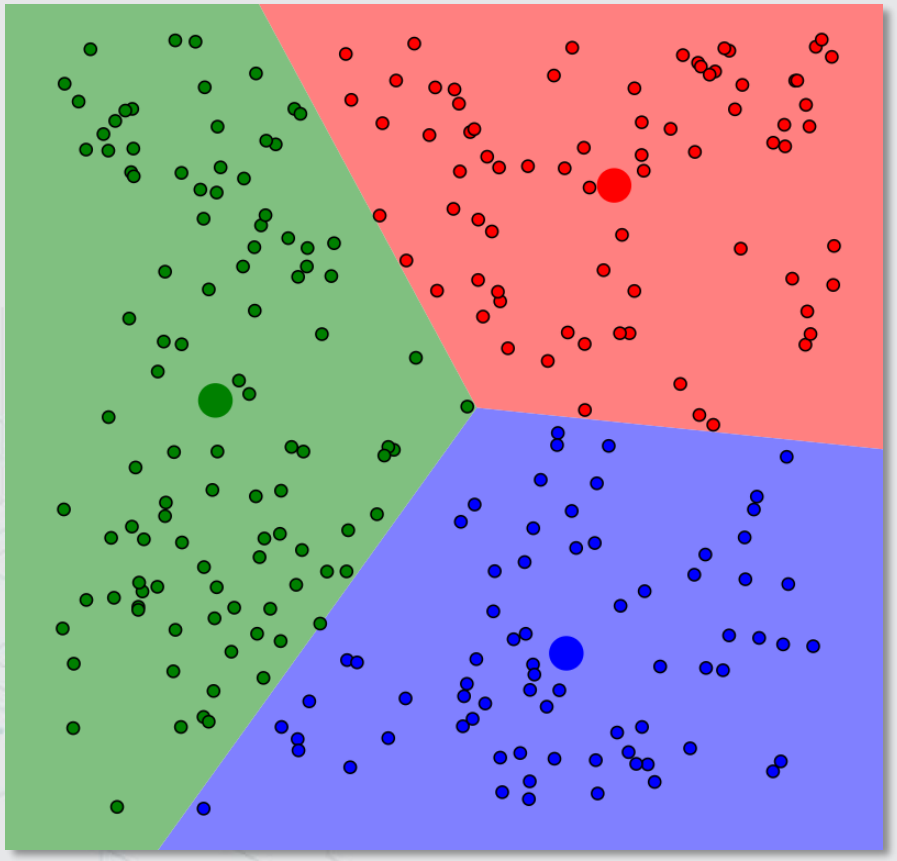
K-means 步驟說明

- 資料重新歸類所屬群組



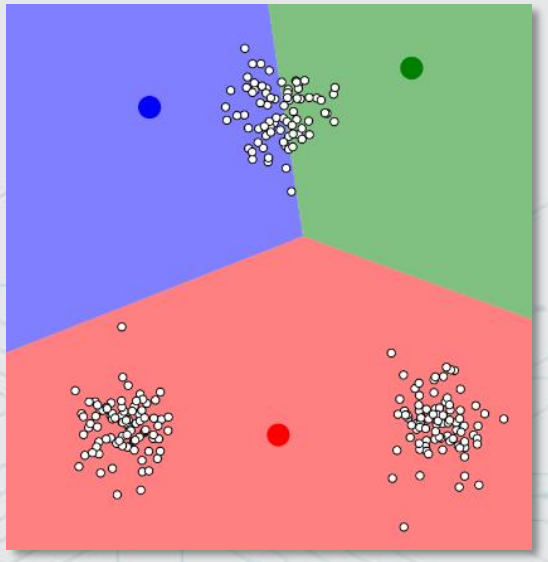
K-means 步驟說明

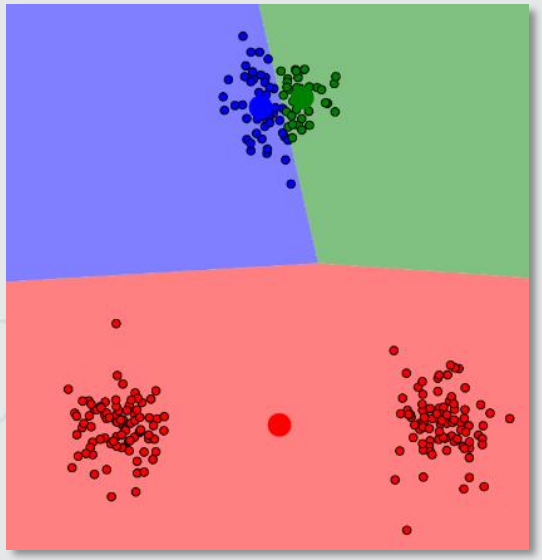
- 重複直到群組中心移動量小於設定的閾值



K-means 初始方法

- 初始群中心會影響到最終分群的**結果**和**收斂速度**，隨機選擇有可能會造成結果不好





K-means 初始方法

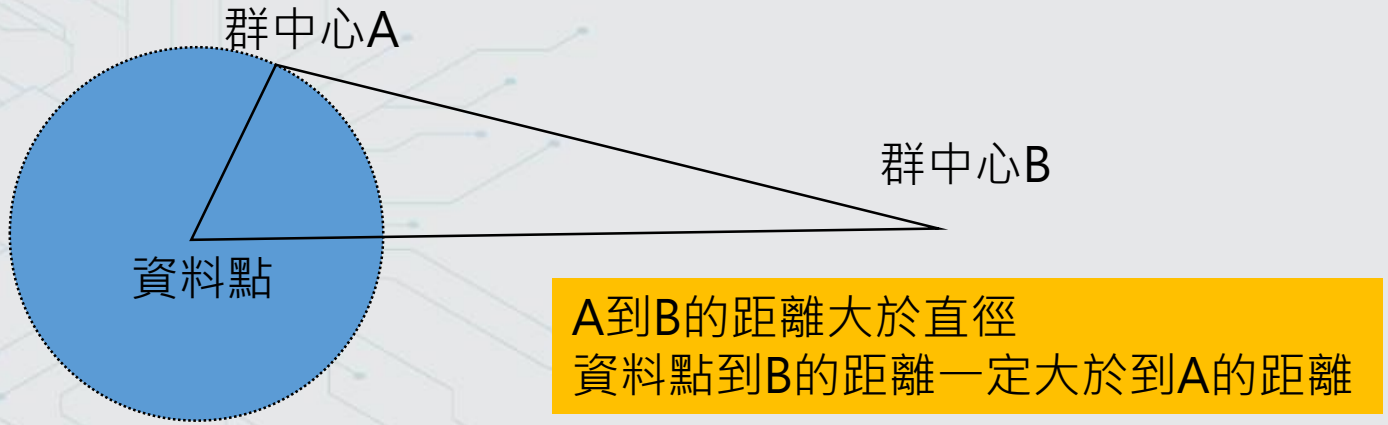
- K-means++想法為讓每個初始群中心盡可能相距越遠越好

步驟

- 從資料集中隨機選擇第一點
- 計算已選擇的群中心中最遠的資料位置為新的群中心
- 重複直到K個群中心產生

加速運算

- 在歸類資料所屬群中心時需要大量計算
  - 傳統作法：每筆資料和每個群中心計算距離。
  - elkan
    1. 預先計算各群中心的距離。
    2. 2倍資料點到群中心A的距離小於群中心A到群中心B的距離，則資料點到群中心B的距離一定比大於到A的距離。



優缺點

優

- 速度快
- 參數簡單

缺

- 無法區分噪點或離群點
- 不適用非凸資料分佈
- 資料不平均、各類變異不同