人工智慧概論

CH04: 監督式與非監督式學習

National Taiwan Ocean University
Dept. Computer Science and Engineering

Prof. Chien-Fu Cheng

案例：用機器學習讀懂嬰兒哭聲

- 開發育兒APP的日本First-Ascent公司推出了一項新技術，可以根據嬰兒的哭聲分析哭泣的原因。父母可以用智慧型手機的麥克錄入嬰兒的哭聲，通過雲共享，First-Ascent公司人工智慧系統會對哭聲進行分析，爲面對夜晚大哭不止的嬰兒而束手無策的父母們提供幫助。
- 和人工智慧AlphaGo一樣，這款APP也導入了同樣的分析方式，可以「深度學習」。通過模仿人腦處理信息的方法，軟體可以在大量數據中尋找規律進行學習。哭聲的分析結果將按可能性從高到低排列，例如：「困70%，無聊25%，肚子餓5%」等。讓摸不清嬰兒夜晚哭泣原因的父母可以相對輕鬆應對。隨著樣本數量的增加，APP的分析準確度將越來越高。

4-1 什麼是機器學習

- 機器學習則是機器透過數據的學習並建立決策模型。
- 人工智慧也有快速決策的方式，那就是用「邏輯判斷」(Ruled-based) 的方式。

圖4-2 人工智慧與機器學習的關係。(資料來源：Intel 官網。)

4-1 什麼是機器學習

- Rule-based 的方式十分常被應用在市場上，但判斷常常發生錯誤。
  - 貓學狗叫容易造成誤判。
  - 用頭髮長度判斷男生、女生。

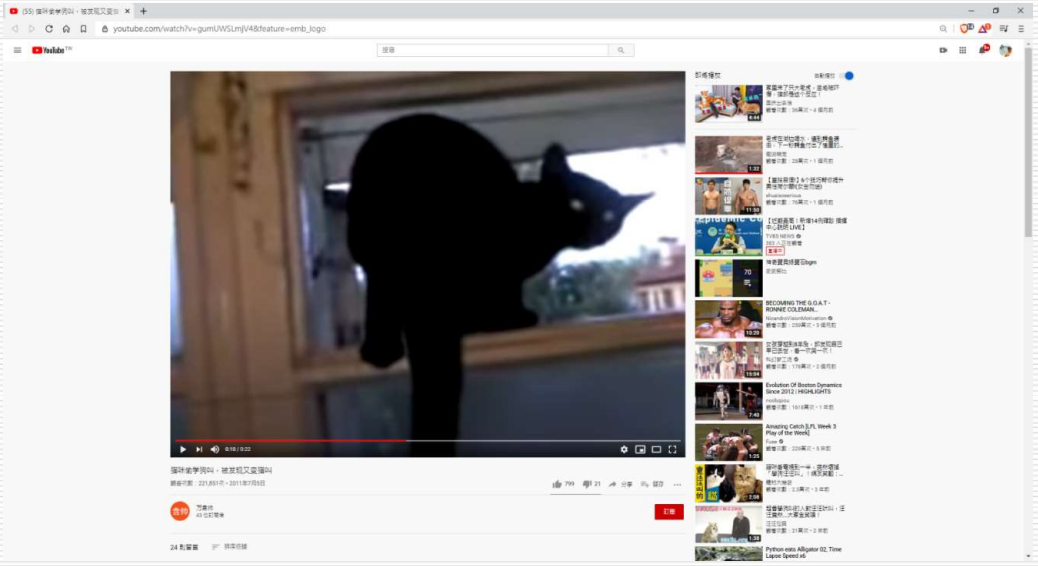


圖4-3 貓學狗叫容易造成誤判。(圖片來源：Youtube。)

4-1 什麼是機器學習

- 應該用更「智慧」的判斷方式，也就是叫機器去學習怎麼判斷。

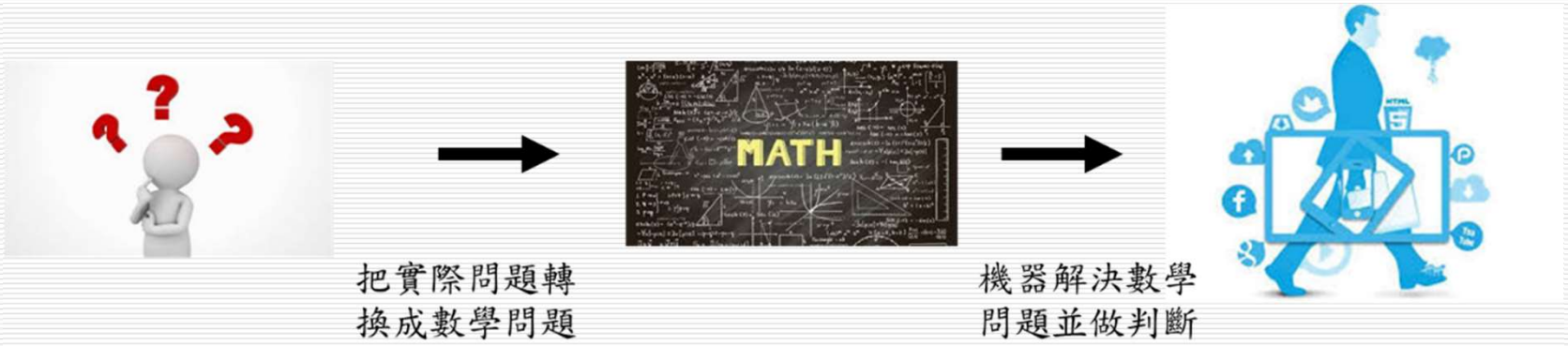


圖4-4 機器學習的基本思維。

4-1 什麼是機器學習

- 機器學習基本資料的兩個部分
  - 特徵值 (feature)
  - 標籤值 (label)
- 特徵值就是用來描述資料的特徵。
- 標籤值就是希望機器最後判斷出來的結果。
  - 特徵值就會有：「喵喵叫」、「汪汪叫」的叫聲數據
  - 標籤值就會有：「貓」和「狗」兩個
- 標籤值就是輸出資料，就是辨識結果，也就是貓或是狗。

圖4-5 機器學習的流程。

4-1 什麼是機器學習

- 機器學習本質上是資料分析，目的有：
  - 預測 (Predictive Modeling)
  - 關聯 (Association Rules)
  - 分群 (Clustering)
  - 異常偵測 (Anomaly Detection)

4-1 什麼是機器學習

- **預測**：就是先建立行為的模型，未來如果有資料時，我們就可以根據預測模型去知道這個資料屬於哪一個行為模型，便可以知道結果。例如：將天空中的溫度、溼度的資料進行分析後，就知道空氣中的溫溼度如何變化後，什麼地區的下雨量會是大雨、豪雨、還是豪大雨。。
- **關聯**：啤酒尿布法則就是一例。關聯分析的應用範例。
- **分群**：分群是一個比較偏向是數據特性探討的機器學習分析技術。分群技術是處理事先並不知道被歸在哪一群的資料，然後分群後可以進行探討。
- **異常偵測**：異常學習是透過模式建立，但如果在樣式識別後發現有資料與原本建立的模型差異太大，那可能就是異常的資料。
- 監督式學習 (Supervised Learning) 跟非監督式學習(Unsupervised Learning) 是機器學習的兩個主要學習的演算法類型。

4-2 監督式學習

- 監督式學習就是分類的學習，所以在學習時就必須知道這些資料是屬於哪一個類別的。
- 進行監督式學習必須知道每一筆資料的特徵以及標籤。
- 監督式學習有很多種演算法，比較常見的如表4-1：

表 4-1　比較常見的監督式學習。

	監督式學習的演算法	英文
1	K- 近鄰演算法	K-Nearest Neighbor Classification
2	決策樹	Decision Tree
3	支持向量機	Support Vector Machine
4	最小平方法	Least Squares Method
5	貝氏分類	Bayes Classifier
6	回歸法	Regression
7	類神經網路	Artificial Neural Network

4-2 監督式學習

- K- 近鄰演算法
  - K- 近鄰演算法通常直接寫作KNN。
  - KNN 的核心思想就是，當有一個新的數據需要進行分類時，參考最近的數據來決定新數據的歸屬。
  - 在KNN 的實際分類上，可以參考圖4-6 的圖。
  - KNN 的K = 3 時，表示新的數據要分到哪一類要看最鄰近的3個數據點。

4-2 監督式學習

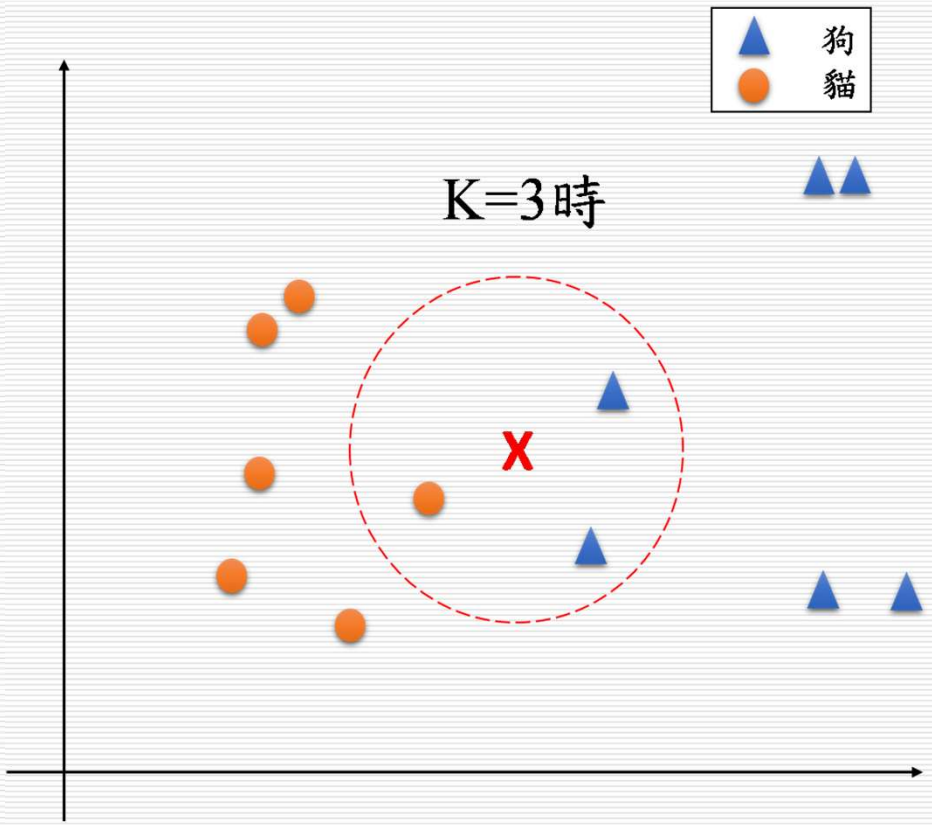


圖4-6 KNN分類，K＝3。

4-2 監督式學習

- K = 7 時呢？卻有不一樣的分類結果。K = 7 時被分類成貓了。

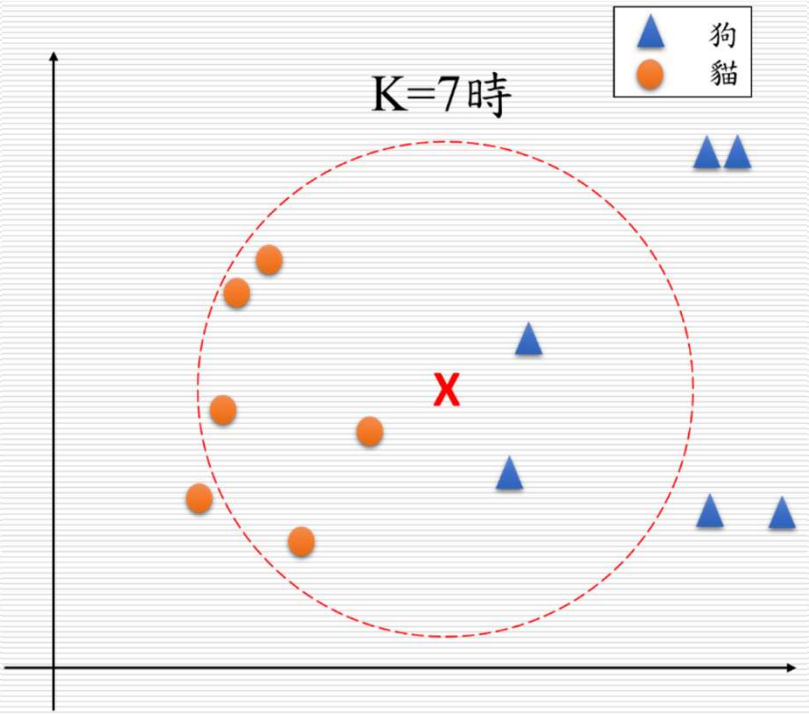


圖 4-7 KNN 分類，K = 7

4-2 監督式學習

- 所以使用KNN 學習有幾點需要注意：
  1. K 的選擇影響至關重要。
  2. KNN 演算法中的K，一般不會使用偶數。
     - ✓ 可能同票
  3. KNN 需要跟K 個鄰近資料作比較，要耗費比較大的計算能量。

- KNN 是分類法上最簡單且有效率的演算法，除了在實作上較為容易外，同是分類精準度也較高，是一個相當常用的監督式學習方法。

4-2 監督式學習

➢ 決策樹

- 決策樹 (Decision Tree) 也是一個常用的分類方式。
- 每一個子節點按照某一個特徵屬性進行分類，典型的方式有三個。

表 4-2　典型決策樹特徵屬性演算法。

	名稱	特徵屬性演算法
1	ID3	資訊增益（Information Gain）
2	C4.5	增益比（Gain Ratio）
3	CART	基尼指數（Gini Index）

4-2 監督式學習

- 以最常使用到的 ID3 為例來建構決策樹。從根節點開始每次根據「最大資訊增益」選取當前最佳的特徵來分割資料，並按照該特徵的所有取值來切分成若干類，然後用同樣方法建構子樹。
- 以下面電商客戶消費紀錄資料為例進行決策樹的分類分析。

4-2 監督式學習

表 4-3　電商客戶消費紀錄彙整表。

	年齡層	平均消費	是否為會員	消費頻次	是否為潛在 VIP
1	青年	高	否	低	不是
2	青年	高	否	高	不是
3	中年	高	否	低	是
4	老年	中	否	低	是
5	老年	低	是	低	是
6	老年	低	是	高	不是
7	中年	低	是	高	是
8	青年	中	否	低	不是
9	青年	低	是	低	是
10	老年	中	是	低	是
11	青年	中	是	高	是
12	中年	中	否	高	是
13	中年	低	是	低	是
14	老年	中	否	高	不是

4-2 監督式學習

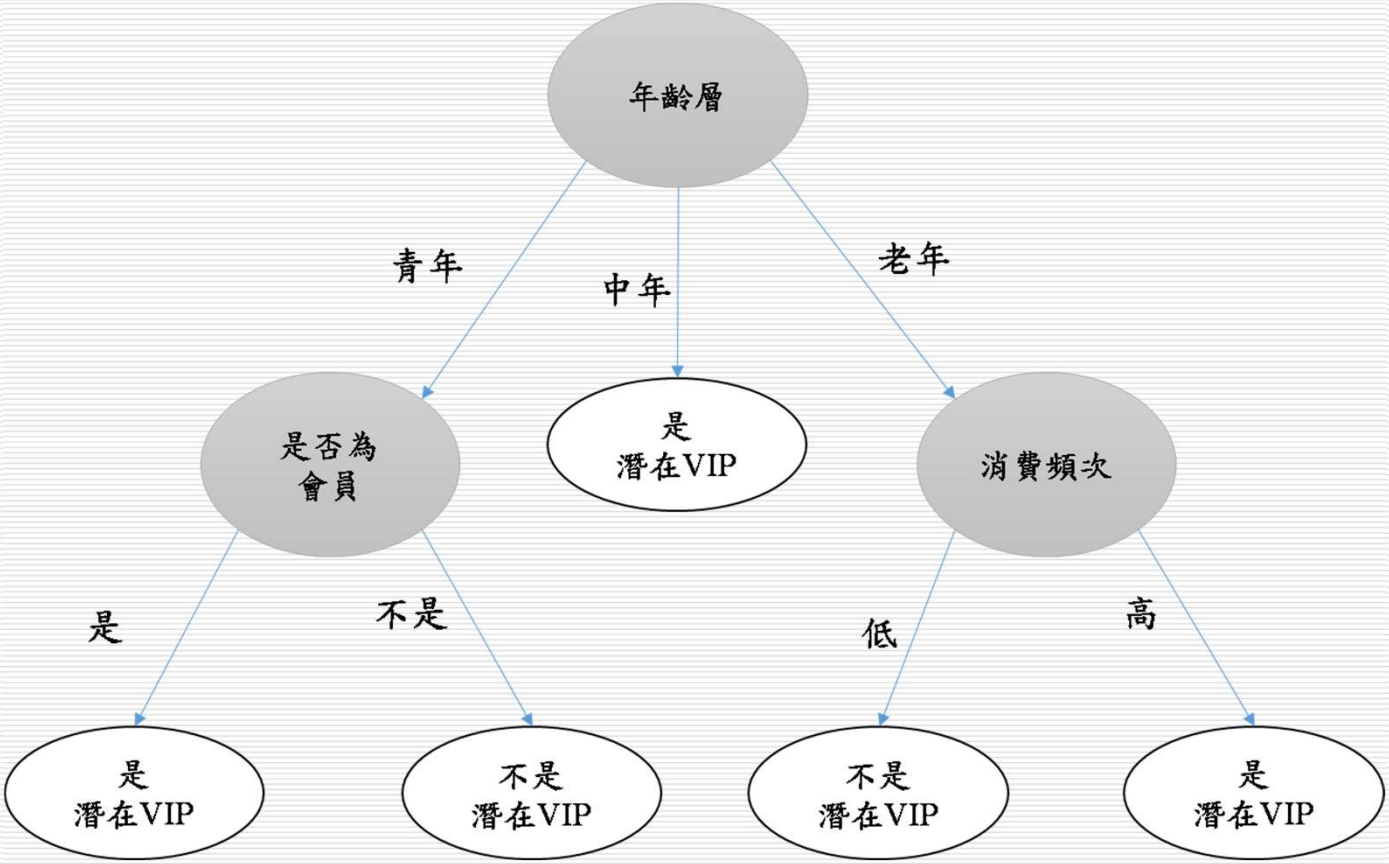


圖**4-8** 電商客戶消費紀錄根據ID3 產生的決策樹。

4-3 非監督式學習

- 非監督式學習只是靠資料特性，看看哪些資料表現是比較接近的，而把這些表現比較接近的資料視為同一群。
- 所以非監督式不用標籤 (Label)。

表 4-4　比較常見的非監督式學習。

	非監督式學習的演算法	英文
1	K- 平均分群法	K-Means Clustering
2	階層式分群法	Hierarchical Clustering
3	模糊 C- 平均分群法	Fuzzy C-MeansClustering
4	奇異值分解	Singular Value Decomposition

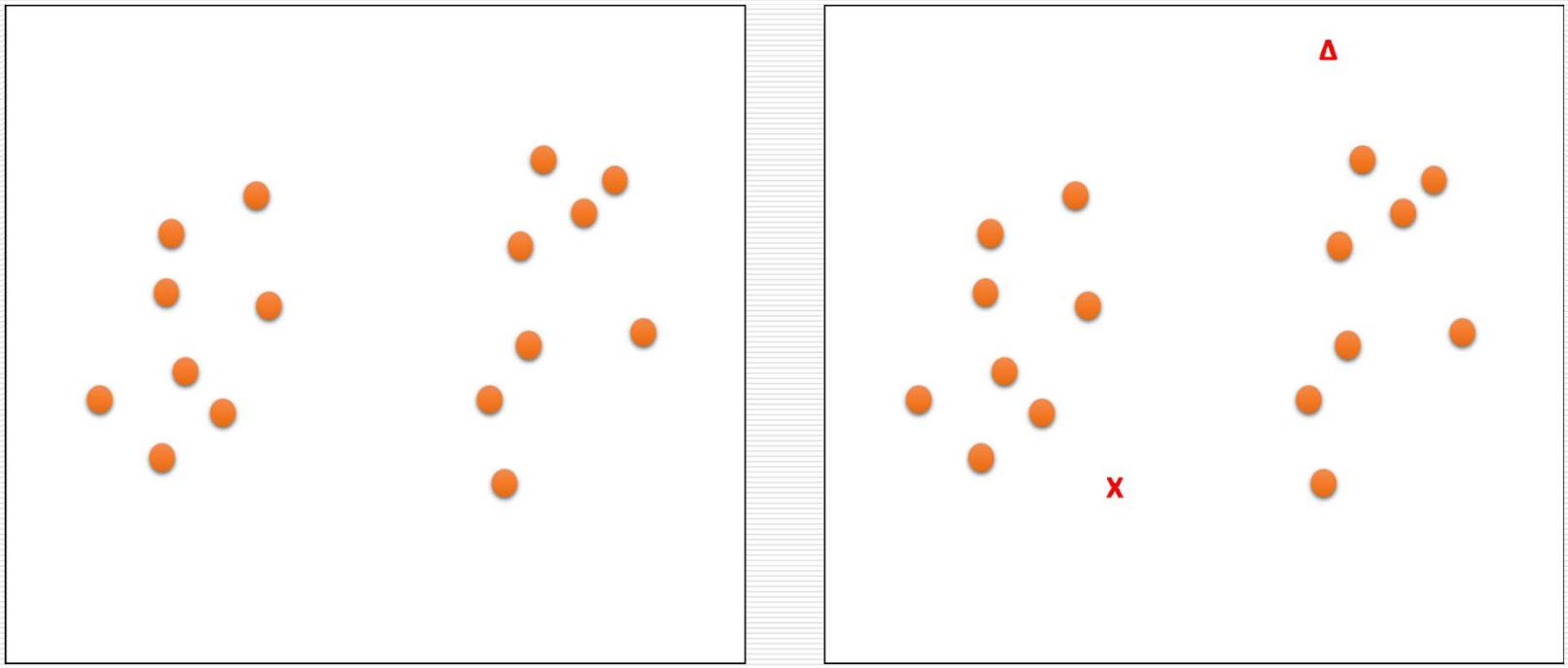
4-3 非監督式學習

K- 平均分群法

- K-Means 是將有類似特徵值的資料放在同一群 (Cluster)。
- 在K-Means 分群演算法中，K 代表要分成幾群，這個在演算法開始計算時就要決定了。
- 假設K-Means 分群要將資料分成 2 群。

1. 分類前的原始資料在特徵空間如圖4-9(a)。
2. 首先要在特徵空間中隨機產生 2 個點。

4-3 非監督式學習

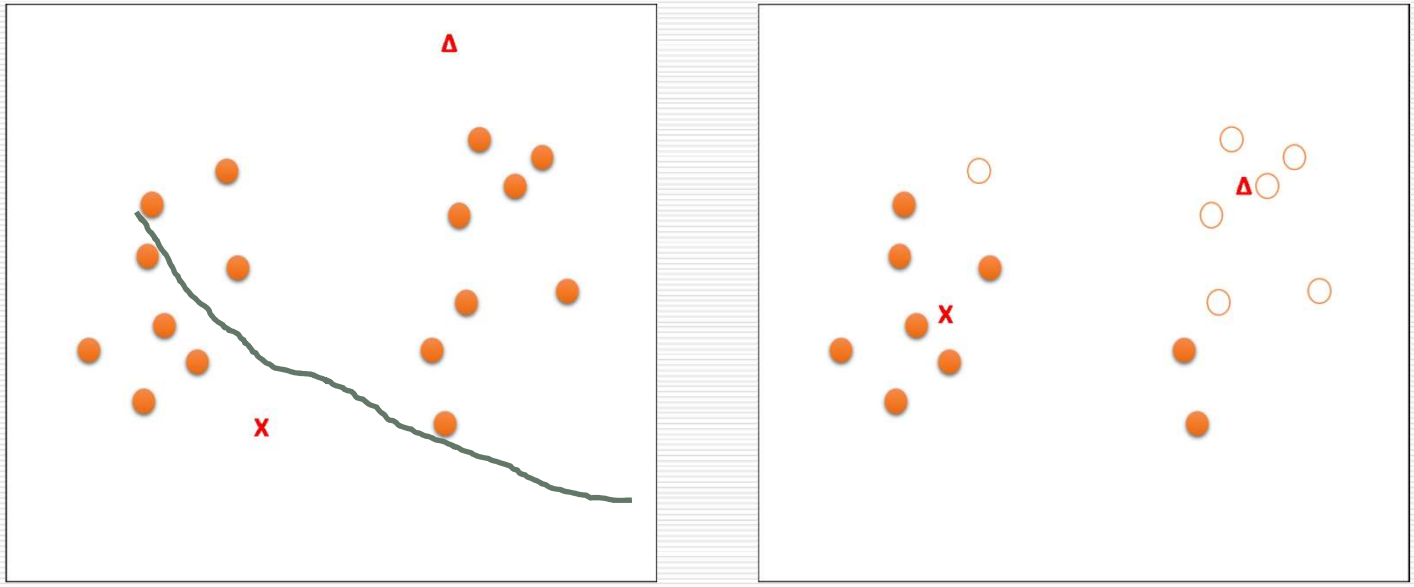


(a) (b)

圖4-9 分群初始狀態。

4-3 非監督式學習

3. 計算每一個特徵點X 與Δ的距離，離Δ比較近時則歸為Δ群。
4. 這時K-Means 分群演算法要重新計算各群的中心。

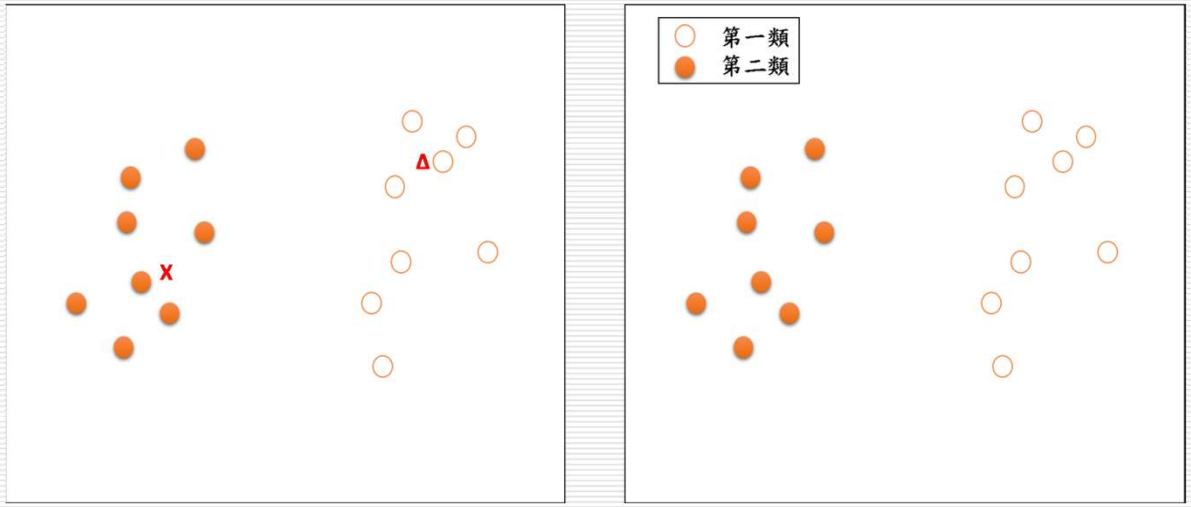


(a) (b)

圖**4-10** K-Means 分群演算法計算第一次分群。

4-3 非監督式學習

5. 再重新分群一次，這種重新分群、重複執行稱作迭代執行 (Iterative Process)。
6. 重新得到 2 群的分群。如圖4-11(a)。
7. 直到 2 群的群中心都不再有變化的時候，代表 2 群的分群已經穩定、明確。



(a) (b)

圖**4-11** K-Means 迭代執行收斂2 群結果。

4-3 非監督式學習

- K-Means有幾個重點需要注意：
  - K-Means 的K 值是由人給定的，這個在應用時會有一些限制。
  - K-Means 的各群中心，如果一開始設定的不好，分群效果可能也不盡理想。

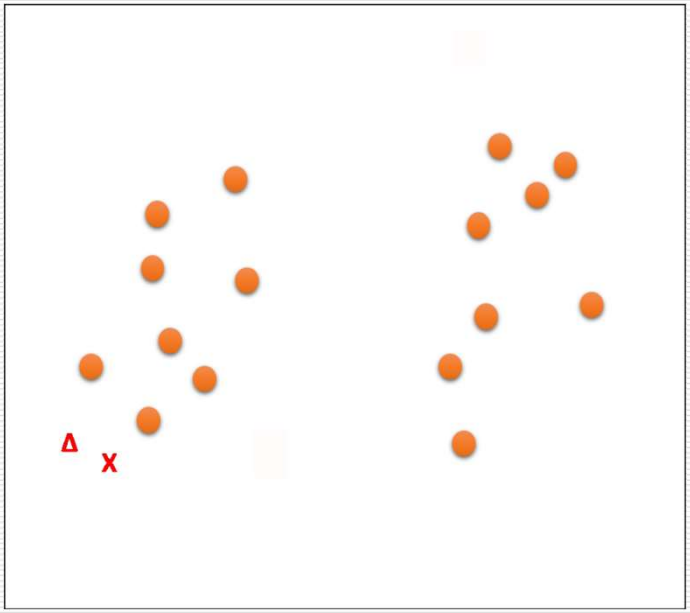


圖4-12 K-Means 不恰當的隨機初始2 個群中心。

4-3 非監督式學習

- 機器學習的方法有非常多種：
  - K-mean 平均演算法 (K-mean algorithm)
  - 決策樹學習問題 (Problems with decision tree learning)
  - 支持向量機 (Support Vector Machine)
  - 爬山演算法 (Hill climbing)
  - 模擬退火演算法 (Simulated annealing)
  - 螞蟻演算法 (Ant algorithm)
  - 分群演算法 (Swarm algorithm)
  - 感知器 (Perceptron)
  - 誤差反向傳播演算法 (Error backpropagation)
  - 梯度下降法 (Gradient decent)

4-4 機器學習的衡量方法

- 一般衡量方式有下面幾個：
  - Accuracy：正確率
  - Recall：召回率
  - Precision：準確率

表 4-5　預測準確與否的評估實例。

		懷孕真實情況	
		實際有懷孕	實際沒懷孕
驗孕棒測試結果	驗孕棒測試有懷孕	測試 / 判斷正確	測試 / 判斷錯誤
	驗孕棒測試沒懷孕	測試 / 判斷錯誤	測試 / 判斷正確

4-4 機器學習的衡量方法

- 以驗孕棒驗孕為例，共有四種狀況。
  - 第一種狀況是孕婦有懷孕，驗孕棒也測得有懷孕，這時候測試的結果是正確。
  - 第二種狀況是孕婦沒懷孕，驗孕棒卻測得有懷孕，這時候測試的結果是錯誤。
  - 第三種狀況是孕婦有懷孕，驗孕棒卻測得沒懷孕，這時測試的結果也同樣是錯誤。
  - 第四種狀況是孕婦沒懷孕，驗孕棒也測得沒懷孕，這時候測試的結果是正確。

4-4 機器學習的衡量方法

- 以驗孕棒驗孕為例，共有四種狀況。
  - 第一種狀況是孕婦有懷孕，驗孕棒也測得有懷孕，這時候測試的結果是正確。
  - 第二種狀況是孕婦沒懷孕，驗孕棒卻測得有懷孕，這時候測試的結果是錯誤。
  - 第三種狀況是孕婦有懷孕，驗孕棒卻測得沒懷孕，這時測試的結果也同樣是錯誤。
  - 第四種狀況是孕婦沒懷孕，驗孕棒也測得沒懷孕，這時候測試的結果是正確。

- 所以驗孕棒驗孕的正確率（ Accuracy ）是多少呢？
  - 驗孕棒的正確率：(第一種狀況 + 第四種狀況)／總驗孕人數
- 驗孕棒的召回率（ Recall ）：有懷孕且驗孕棒測得正確的人數，除以驗孕棒測得有懷孕的總數。
  - 驗孕棒的召回率：第一種狀況／(第一種狀況 + 第三種狀況)
- 準確率（ Precision ）是預測有懷孕的人，有多少比例是真正有懷孕的。
  - 驗孕棒的準確率：第一種狀況／(第一種狀況 + 第二種狀況)

4-4 機器學習的衡量方法

- 混合矩陣總共會有 4 個狀況。

表 4-6　機器學習模型的評估混合矩陣。

		真實情況	
		True	False
預測模型	True	True Positive （TP）	False Positive（FP）
	False	False Negative （FN）	True Negative （TN）

4-4 機器學習的衡量方法

- 正確的狀況有兩個：
  - True Positive (TP)，也稱作真陽性。
  - True Negative (TN)，也稱作真陰性。
- 預測失真：
  - False Positive (FP) 偽陽性。
  - False Negative (FN) 偽陰性。

4-4 機器學習的衡量方法

- 要衡量機器學習是否為好的預測模型，則用下面三個評估方式：
  - Accuracy (正確率)：(TP + TN)／Total Samples
  - Recall (召回率)：TP／(TP + FN)
  - Precision (準確率)：TP／(TP + FP)

4-4 機器學習的衡量方法

- 正確率是一個最直覺而且最簡單的評估方式。
- 這兩個評估模型在機器學習中則有交互影響的作用在。
- 要提高召回率就必須提高TP 且同時要降低FN，也就是要提高機器學習預測模型的敏感度。
- 提高機器學習預測模型的敏感度（此時，FP會變高），Recall (召回率) 會提高，自然Precision (準確率) 就會下降。
  - 以COVID-19的快篩來看，如果提高快篩的敏感度，勢必就會有很多原本快篩是陰性的民眾被誤判為陽性，也就是FP提高了，因此準確率自然就下降了。

- Accuracy (正確率)：(TP + TN)/Total Samples
- Recall (召回率)：TP/(TP + FN)
- Precision (準確率)：TP/(TP + FP)

4-5 機器學習最常使用的演算法

- 監督式學習是一種分類的學習方式。非監督式學習是分群的演算法。非監督式學習會依據資料的特徵屬性的相似度。
- 機器學習的模型訓練通常用交叉驗證的方式。提高正確率是機器學習模型必然的目標，但召回率及準確率則可以依據應用而有所不同。

4-5 機器學習最常使用的演算法

檢驗方法	檢驗時間	時　機	敏感性	特異性
Real-time RT-PCR	2~4 hrs	發病早期 防疫圍堵與阻斷傳播	> 95%	> 95%
抗原 (快篩)	15 min	發病早期 高盛行區	70%	> 95%
抗體 (快篩)	15 min	發病 7 天後 了解是否感染	75%	> 95%

圖**4-13** 新冠肺炎的RT-PCR 快篩檢測的敏感度（即召回率）與特異度（即準確率）。(資料來源：中央流行疫情指揮中心。)

Sources

- 投影片資料來源說明：
  - 本投影片之內容出自於書商所提供之投影片，並根據實際授課需求進行補充及修改。