

AI基礎訓練初級班

三、資料探勘簡介

制定部門：總管理處技訓中心
編定日期：2020年12月30日編印
版次：R1

本著作非經著作權人同意，不得轉載、翻印或轉售。

著作權人：台灣塑膠工業股份有限公司
南亞塑膠工業股份有限公司
台灣化學纖維股份有限公司
台塑石化股份有限公司

AI初級班課程項目：

- 一、人工智慧概論
- 二、指標衡量方法
- 三、資料探勘簡介
- 四、資料視覺化原理

課程目的

在本課程中，將簡介資料探勘技術，讓學員初步瞭解資料探勘技術與可解決的問題，並初步掌握資料探勘技術相關的專業詞彙。

課程大綱

- (一) 資料探勘基本介紹
- (二) 資料探勘與其他系統的比較
- (三) 知識發現的過程
- (四) 資料探勘常用的技術

目錄

(一) 資料探勘基本介紹

1. 資料探勘的目的與誘因·····	11
2. 資料探勘技術的成功經驗·····	12
3. 資料探勘技術的資料與應用·····	13
4. 資料探勘的興起·····	14
5. 資料探勘的功能·····	15
6. 何謂資料探勘·····	16~17

目錄

(二) 資料探勘與其他系統的比較

1. 資料探勘與人工智慧的不同點.....19
2. 資料探勘與智慧型決策支援系統的不同點.....20
3. 資料探勘與線上分析的不同點.....21
4. 資料探勘與統計分析的不同點.....22~23

(三)知識發現的過程

1. 知識發現的過程.....	25
2. 專業領域知識的重要性.....	26
3. 原始資料收集.....	27
4. 工廠數位資料的收集.....	28
5. 資料前置處理	29
6. 型樣評估	30
7. 結果的呈現.....	31

(四) 資料探勘的技術與功能

1. 資料探勘常用的技術	33~34
2. 資料探勘的功能	35
3. 資料分類	36
4. 資料分類技術的兩階段過程	37
5. 資料分群	38
6. 資料關聯分析(關聯規則探勘)	39
7. 循序性樣式探勘	40

(一) 資料探勘基本介紹

1. 資料探勘的目的與誘因

(1)目的：從大量資料中採取有價值的知識，供管理人員做為決策的參考，以開創企業新機或處理危機

(2)誘因：巨量資料的累積

- A.零售業：威名超市(Walmart)每小時約新增100萬筆交易資料
- B.臉書(Facebook)有超過 400億張的照片
- C.線上購物交易資料、網頁資料
- D.Line資料、雲端資料、健保資料
- E.將來來自穿戴式裝置、物聯網、自駕車的資料

2. 資料探勘技術的成功經驗

(1) 美國超級市場(Wal-Mart)

- A. 分析銷售資料，發現尿布和啤酒常會被顧客一起購買
- B. 於是將商品放在一起促銷，得到意想不到的業績成長

(2) 美國銀行(Bank of America)

- A. 從客戶資料中，找出既有客戶申請貸款的契機
- B. 規劃全新行銷方案，方案推出後，申貸率成長兩倍以上

3. 資料探勘的資料與應用

- 以PChome商店街之線上購物(零售業)為例

- (1)可獲取之資料

- A.交易紀錄(會員、購買的商品、購買的數量、購買時間、寄送地點等)

- B.會員顧客資料(居住地區、年齡、性別、工作性質)

- (2)可行之資料探勘應用：分析顧客型譜，了解顧客族群特性

- A.找出特定商品的潛在客戶名單

- B.建立個人化行銷模式

- C.預測目前哪些顧客可能流失



4. 資料探勘的興起

- 歸功於三類技術的普遍與成熟
 - (1)大量資料存取與管理的相關技術
 - A.關聯式資料庫(relational database)應用廣泛
 - B.資料庫系統與整合平台技術的成熟與穩定
 - C.網路發達
 - (2)高效平價的多處理器電腦架構
 - A.高效平價的CPU、記憶體、硬碟
 - B.這樣的電腦架構，使一般人能在可容忍的時間內完成大量資料的處理
 - (3)資料探勘相關的演算法
 - A.統計學(statistics)
 - B.人工智慧(artificial intelligence)
 - C.機器學習(machine learning)

5. 資料探勘的功能

(1) 採取知識、預測趨勢

- A. 有價證券(股匯)行情預測
- B. 天氣預測
- C. 地震預測
- D. 消費行為預測
- E. 商品價格/庫存/出貨量預測…等等

(2) 找出未知的樣式

- A. 找出會購買筆記型電腦的顧客特徵
- B. 依消費習性相近的顧客進行群組推薦
- C. 鑑別消費者可能會同時購買的商品組合…等等

6. 何謂資料探勘(1)

學者曾對資料探勘做過的定義

(1) Frawley

- 從資料庫中挖掘潛在、明確、或有用的知識的過程

(2) Grupe & Owrang

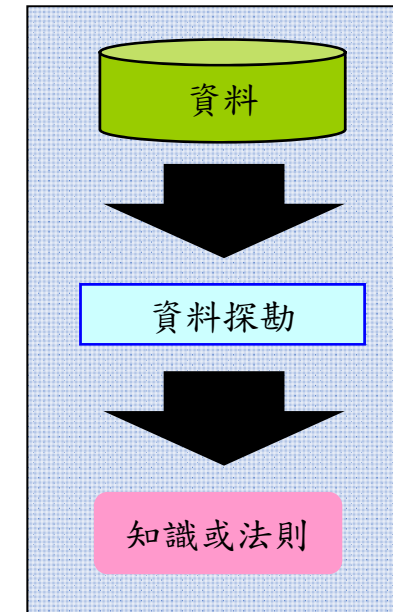
- 從已存在的資料庫當中挖掘出專家仍未知的新事實

(3) Fayyad

- 知識發現 (Knowledge Discovery) 為從大量資料中選取合適的資料，
進行資料處理、轉換等工作，再進行資料探勘與結果評估的一系列過程

(4) Berry & Linoff

- 使用自動或半自動的方法，對大量資料分析，找出有意義的關係或法則。



6. 何謂資料探勘(2)

資料探勘：資料庫之知識發現（Knowledge Discovery in Databases，簡稱KDD）

(1)從資料庫所儲存的資料中去萃取出有用、有趣的知識

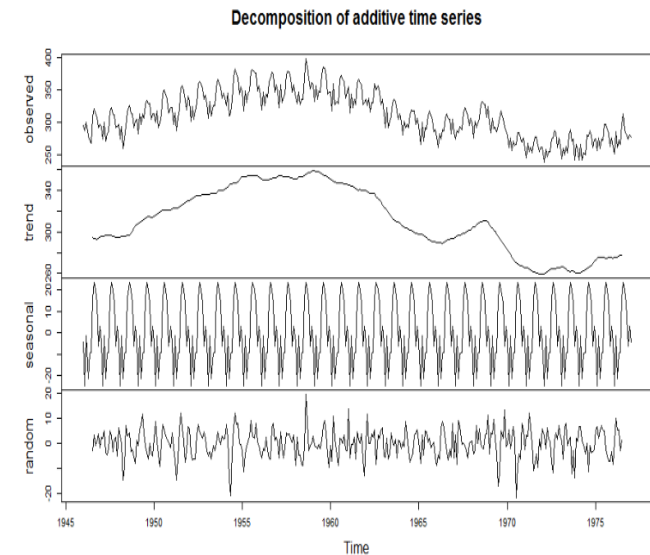
- 大型的資料庫如線上即時的資料庫（On-line Database）及資料倉儲（Data Warehouse）…等等

(2)知識的形式

- 規則、規律性、型樣、趨勢、傾向…等等

(3)知識的一個範例

- 「如果 顧客的年齡是在三十到四十歲之間，
而且 年收入是在四十萬到六十萬之間，
那麼 此顧客很有可能會購買筆記型電腦。」



(二) 資料探勘其他系統 的比較

1. 資料探勘與人工智慧的不同點

(1)人工智慧(artificial intelligence) 也被叫做機器智能(machine intelligence)

- 也就是電腦/機器(computers/machines)所展現的智能

(2)人工智慧的主要目標

- 讓電腦(機器)具有人類智慧形式的功能

(3)傳統的人工智慧研究問題(目標)

- 推理、知識表示、規劃、學習(機器學習/資料探勘)、自然語言處理、感知、影像識別、語音識別、運動和操控、認知、情緒智慧和社群智慧

(4)主要工具(方法)

- 搜尋、數學最佳化(optimization)、邏輯、人工神經網路和統計方法、機率與經濟學

2. 資料探勘與智慧型決策支援系統的不同點

(1) 智慧型決策支援系統

(Intelligent Decision Support System=DB Apl. + AI)

- 依決策模型或推論規則提供建議
 - A. 可以來自於領域專家(Domain Expert)的經驗法則
 - B. 可運用知識工程(Knowledge Engineering)的技術自專家腦中擷取
 - C. 以專家為基礎的(Expert-Based)

(2) 資料探勘

- 自動化的資料分析與預測
 - A. 資料驅動的(Data-Driven)

3. 資料探勘與線上分析的不同點



- 線上分析（On-Line Analytical Processing，簡稱OLAP）
 - 對資料庫制式化的資料作分析，統計數據、提供趨勢給決策人員參考
 - 若以零售業為例，可用以瞭解不同產品、銷售區域對於成本及營業毛利之影響，但無法解析顧客的購買行為模式

線上分析處理： <u>過去的事實</u>	資料探勘： <u>未來的預測</u>
多少人曾購買筆記型電腦？	哪些顧客可能會購買筆記型電腦？
上個月有多少顧客沒有進入網站瀏覽商品？	哪些顧客較有可能在未來三個月內不上站瀏覽商品？
顧客的平均單月消費總金額是多少？	哪些顧客下個月的消費有可能會超過一萬元？
哪些顧客訂單超過三天未付款？	哪些顧客較有可能延遲付款？
電子報的點閱率多少？	電子報行銷方式對那些會員較有效？
去年的銷售業績統計報表	明年預期之銷售業績額度。

4. 資料探勘與統計分析的不同點(1)

(1) 統計分析 (Statistical Analysis)

- 以假設 (Hypothesis) 及 驗證 (Verification) 為基礎

A. 可針對較少資料，進行統計分析或關連性分析

B. 由具專業背景的專家對統計結果加以檢測

(2) 資料探勘

- 以發現 (Discovery) 為基礎，著重「型樣辨識」 (Pattern Recognition)

A. 針對較大量資料

B. 也可供不具專業背景的使用者 (如高層決策人員) 使用

4. 資料探勘與統計分析的不同點(2)

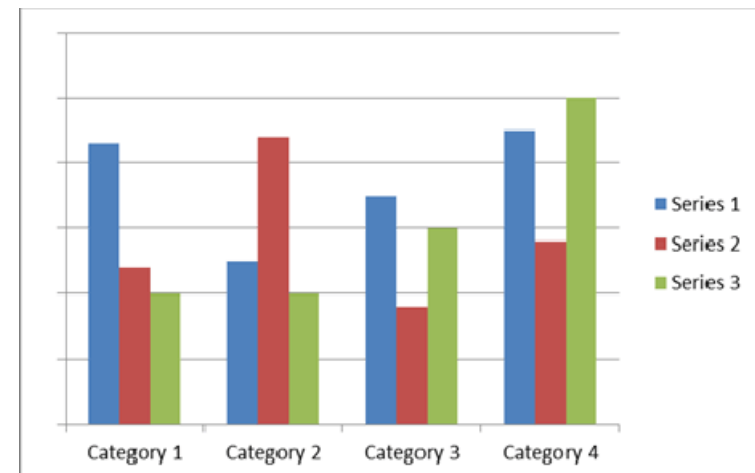
分析的例子

(1) 統計分析

- 不同特性的消費者在本月的消費總額、差異是否顯著
 - A. 不同年齡層的消費者在本月的消費總額
 - B. 不同性別消費者個別在本月的消費總額

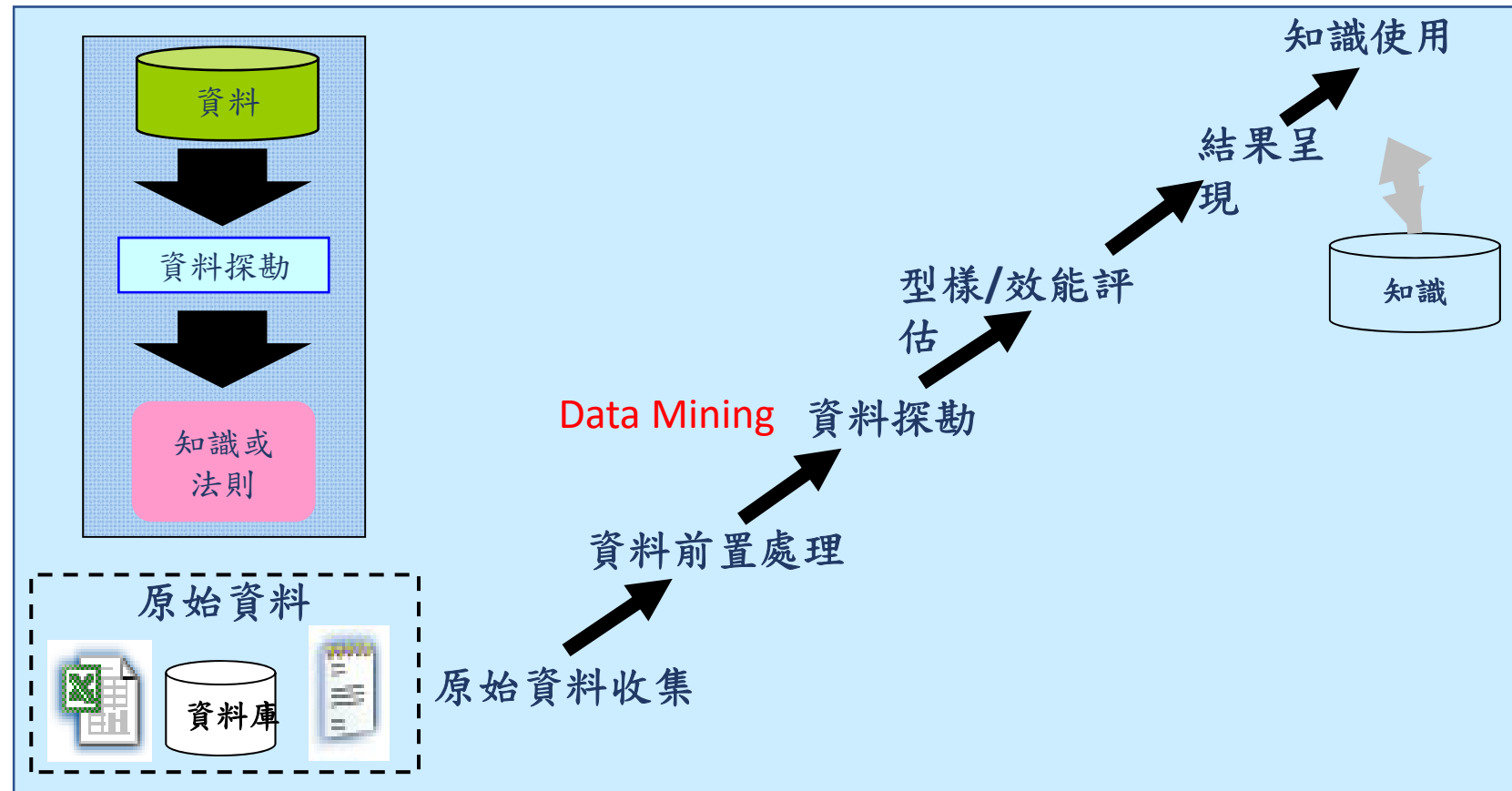
(2) 資料探勘

- 找出如下的規則：
 - IF 居住在台北
 - AND 性別是男性
 - AND 年齡介於 37到42歲之間
 - THEN 購買筆記型電腦的可能性是 85%



(三)知識發現的過程

1. 知識發現的過程



專業領域知識(Domain Knowledge)的取得

2. 專業領域知識的重要性

- (1) 以能夠辨識特定產業的重要知識
- (2) 以能夠運用特定產業的知識



3. 原始資料收集

- (1) 知識發現的第一個步驟
- (2) 原始的資料來源或格式
 - A. 資料庫
 - B. Excel表格
 - C. 文字檔
 - D. 網際網路網頁資料
 - E. 問卷調查…等等

4. 工廠數位資料的收集

- 智慧製造的重要核心概念：讓設備的數據可以被儲存、運算、分析、預測。
常見的應用是設備監診系統，數位資料可藉由感測器來收集，如

(1) ABB 轉動設備 數位訊號截取板

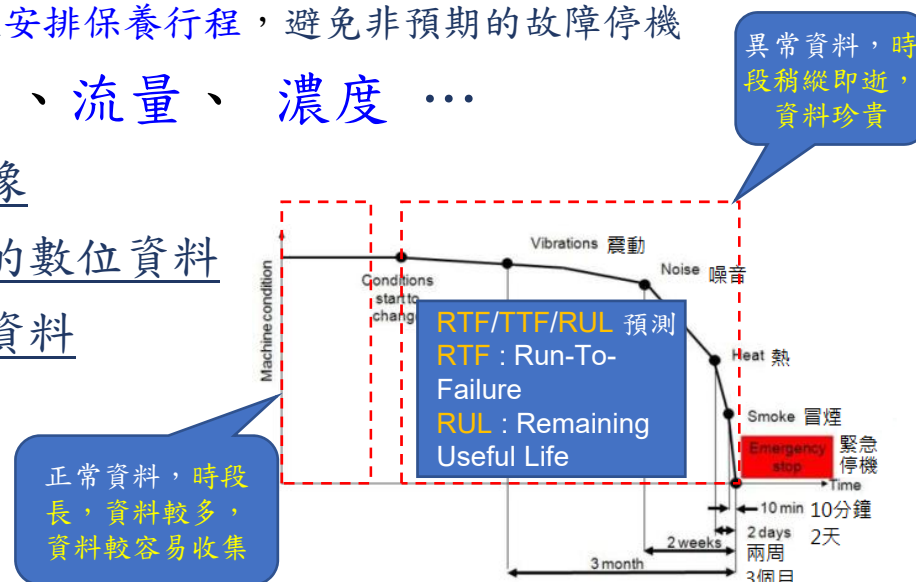
- A. 可取得轉動設備的相關資訊，包括轉速、溫度、轉子中心偏移值、異常震動、電力相位是否平衡等
- B. 在塑化廠區，小到抽水馬達、大至發電機的數位資料可被收集成資料庫
- C. 藉數據的分析，可預測設備壽命及安排保養行程，避免非預期的故障停機

(2) 製程運轉數據：溫度、壓力、流量、濃度 …

(3) 工廠無人機管路巡檢的數位影像

(4) 塑膠產品產製模穴壓力感測器的數位資料

(5) 儀器或載具的振動感測器數位資料



5. 資料前置處理

- (1)可能包含資料的整合、清理、格式轉換等前置作業
- (2)資料探勘有60~80%時間花費在資料收集與前置處理

原因

- A.真實的資料可能雜亂，需要彙整
- B.資料值有遺漏
- C.資料值不一致（度量衡值、命名不一致）
- D.資料不見得全可用

6. 型樣評估

(1) 型樣評估 (pattern evaluation)

A. 評估所挖掘的知識是不是真的有用？

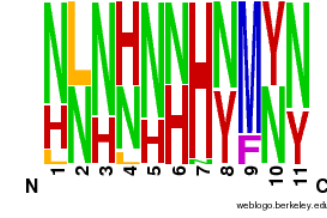
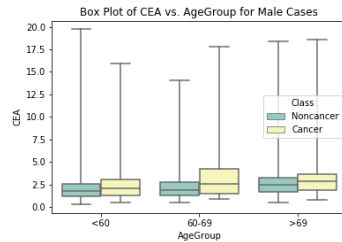
B. 過濾沒用的資訊，只挑出有價值的知識

(2) 型樣評估的範例

A. 「天氣好，旅遊人數就多；天氣差，旅遊人數就少」，這樣的探勘結果我們可能認為它“有趣”的程度並不高，因為它是屬於一般常識

7. 結果的呈現

Profile of healthy people

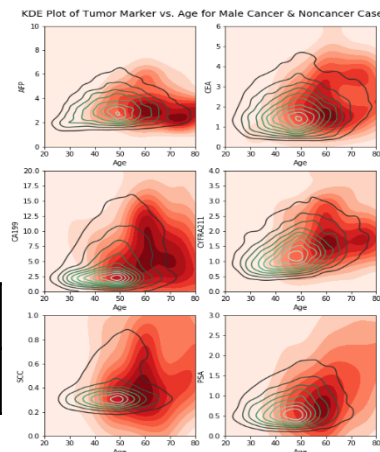
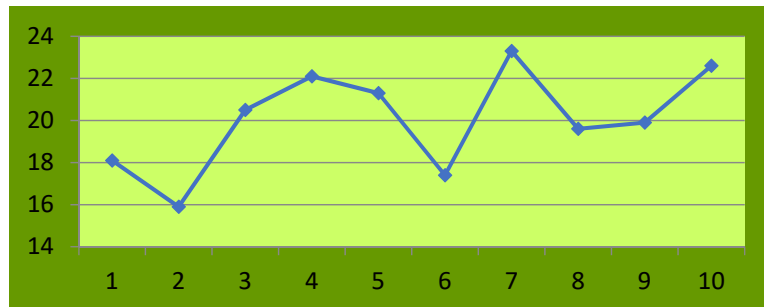


(1)將數據圖形化：把有趣的數據呈現成容易觀察的知識例如：

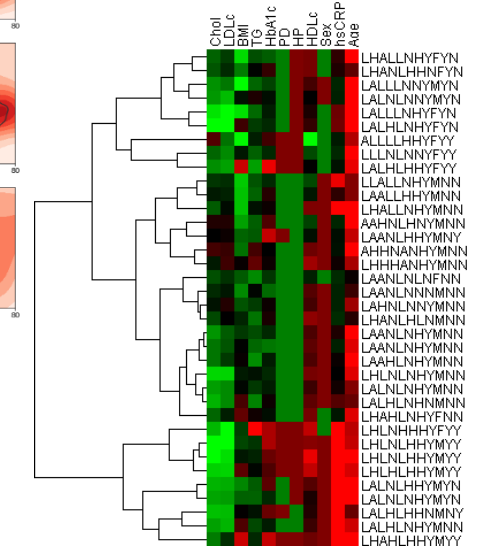
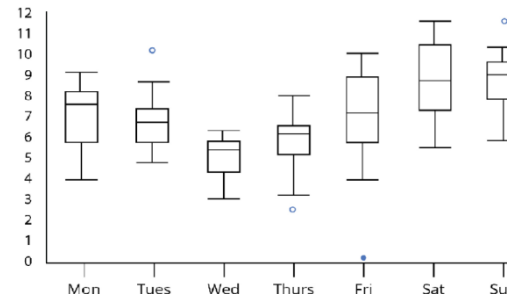
A.用折線圖呈現趨勢

B.用散佈圖呈現物件的分布等等

1	2	3	4	5	6	7	8	9	10
18.1	15.9	20.5	22.1	21.3	17.4	23.3	19.6	19.9	22.6



CAD Profile



Data matrix for visualization

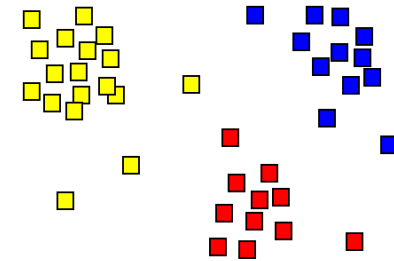
(四) 資料探勘的技術與功能

1. 資料探勘常用的技術

(1) 傳統技術

A. 以統計分析為代表，主要包括

- a. 敘述性統計(敘述性分析)、機率論、迴歸分析
- b. 類別資料分析 (Categorical Data Analysis)
- c. 因素分析 (Factor Analysis)：應用來精簡變數
- d. 群集分析 (Cluster Analysis)：探索性分析，應用來探索物件的
群聚特性
- e. 區別分析 (Discriminant Analysis)：預測性分析，應用來判別分類



1. 資料探勘常用的技術(續)



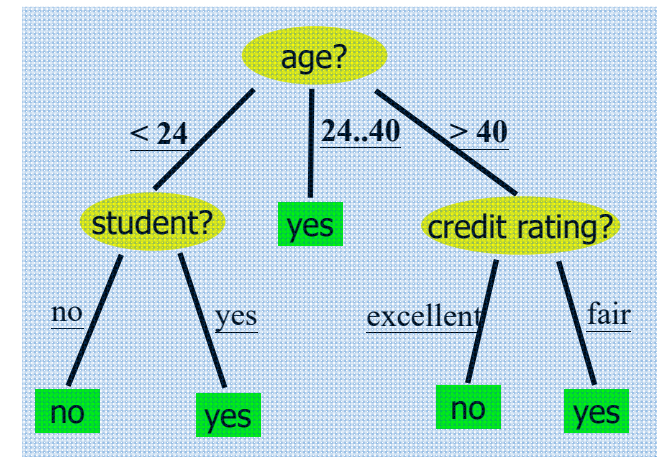
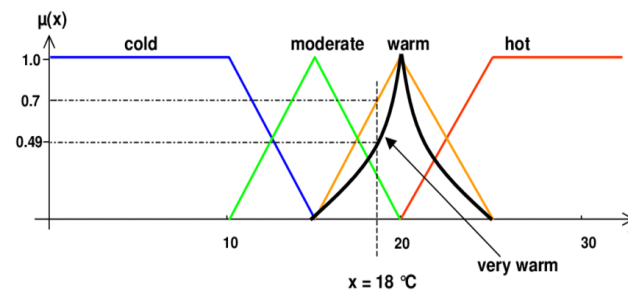
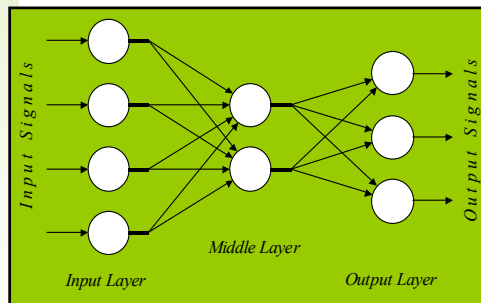
(2) 常見的新式技術

A. 類神經網路 (Artificial Neural Network)

B. 決策樹推導演算法 (Decision Tree Induction Algorithms)

C. 基因演算法 (Genetic Algorithms)

D. 模糊邏輯理論 (Fuzzy Logic Theory)

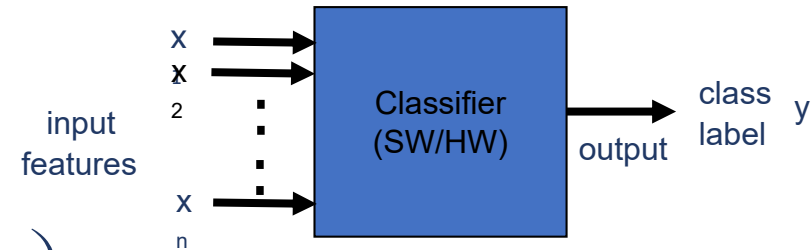


2. 資料探勘的功能

主要含以下四種常用資料分析功能

- (1) 資料分類 (Data Classification)
- (2) 資料分群 (Data Clustering)
- (3) 資料關聯分析 (Data Association)
- (4) 循序性樣式探勘 (Sequential Pattern Mining)

3. 資料分類



(1) 分類技術 (classification)

A. 屬於監督式學習 (supervised learning)，進行預測性分析

B. 分析資料的屬性，探索分門別類的規則，
以建立分類的預測模型

(2) 適用領域

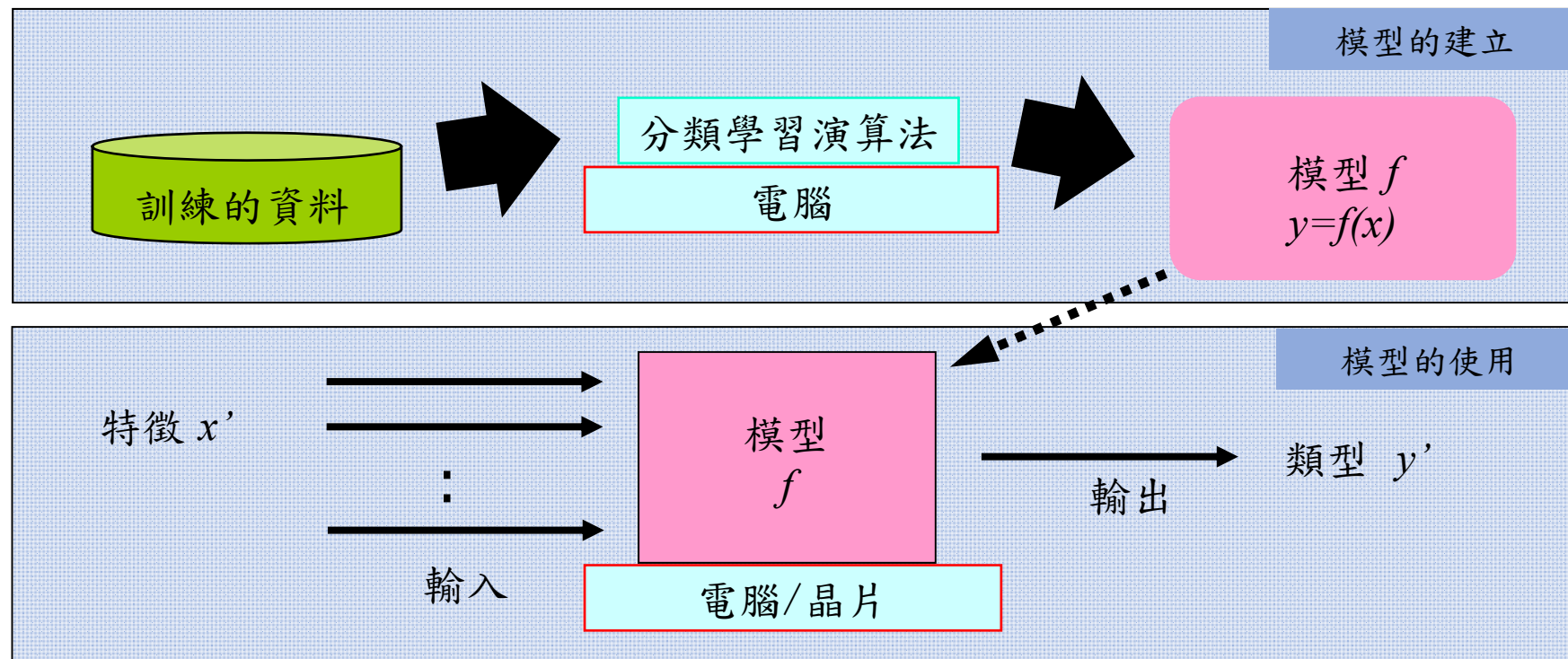
A. 顧客分類、疾病類別鑑別…等

(3) 範例

A. 將信用卡申請者的風險屬性，區分為高度風險、中度風險、
低度風險申請者



4. 資料分類技術的兩階段過程



110年10月18日 星期一

淺談機器學習技術

5. 資料分群

(1) 資料分群(Clustering)

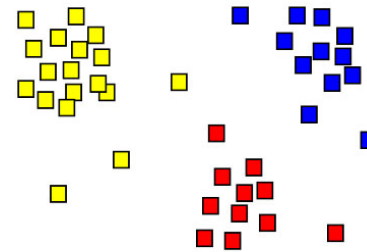
- A. 屬於非監督式學習(unsupervised learning)，進行探索性分析
- B. 透過觀察事物群集，探索了解事物間的相似與相異性關係，以及這些關係將會如何影響預測的結果

(2) 適用領域

- A. 顧客分群、群組商品推薦…等等

(3) 範例

- A. 一群特性相近的人，駕駛相近的汽車，使用相近家電，並且食用相近的食物。
- B. 另一群從事相同行業的人，家庭成員人數接近，年收入接近，出國次數也接近。



6. 資料關聯分析(關聯規則探勘)

(1) 資料關聯分析

- A. 分析資料項目(item)間的關聯性，找出資料項目同時出現的特性
- B. 也叫做關聯規則探勘(Association Rule Mining)，進行敘述性分析

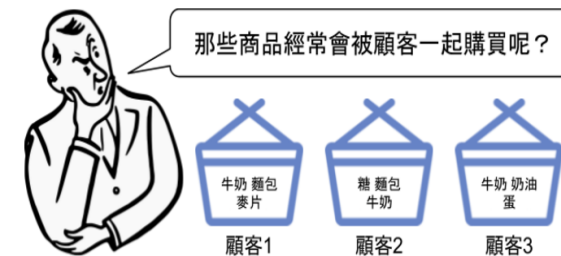

(2) 適用領域

- A. 購物籃分析(Market Basket Analysis)
 - 幫助零售業者瞭解客戶的消費行為

(3) 關聯規則範例

- A. 如果顧客買筆記型電腦，則同時購買隨身碟的機率是80%
- B. 如果顧客買全麥麵包及低脂優酪乳，則同時也買低脂牛奶的機率是85%

C. buy(T, "Beer") \Rightarrow buy(T, "Diaper") [support = 2%, confidence = 70%]

Rule	Support	Confidence	Lift
$A \Rightarrow D$	2/5	2/3	10/9
$C \Rightarrow A$	2/5	2/4	5/6
$A \Rightarrow C$	2/5	2/3	5/6
$B \& C \Rightarrow D$	1/5	1/3	5/9

Graph for 13 rules

size: support (0.333 - 0.333)
color: lift (1.333 - 3)



7. 循序性樣式探勘

(1) 循序性樣式探勘

A. 項目序列資料：假設每一筆項目序列是前後有序的項目串列

B. 找出常出現的循序性項目樣式

- 與關連法則不同的是，在循序性樣式探勘中，相關的項目是前後有序的

(2) 適用領域

A. 疾病發展歷程、行為預測…等等

(3) 範例

A. 代謝症候群 à 糖尿病 [support = 1.3%, confidence = 70%]

B. 糖尿病 à 視網膜病變 [support = 1.5%, confidence = 65%]