



EEC 巨量資料處理與分析 專業認證教師研習會

EEC-
企業電子化
資料分析師

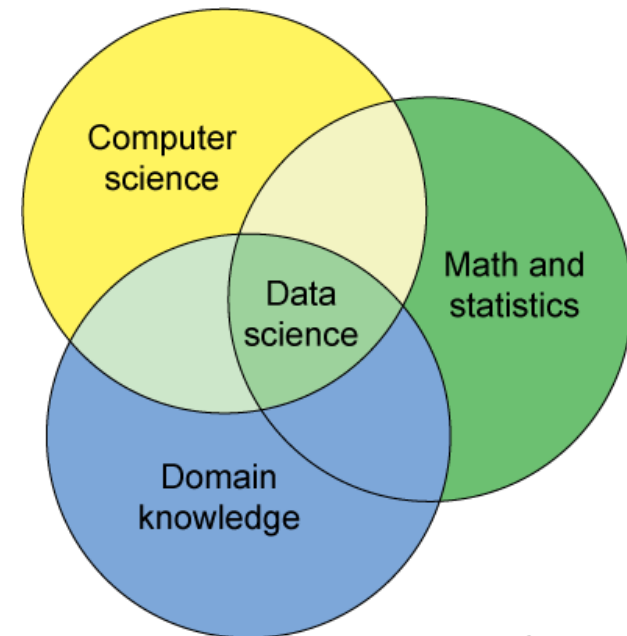


目錄

- 第一週：資料科學與巨量資料 (3hr)
- 第二週：資料處理的基礎 (3hr)
- 第三週：統計分析的基礎 (3hr)
- 第四週：統計分析的基礎 (3hr)
- 第五週：統計分析的基礎 (3hr)
- 第六週：資料探勘與視覺化基礎 (3hr)
- 第七週：資料探勘與視覺化基礎 (3hr)
- 第八週：資料探勘與視覺化基礎 (3hr)

何謂資料科學？

- 定義：研究如何從資料中萃取知識，融合數學、機率模型、統計學、機器學習、電腦程式設計、資料工程、型態辨識、非確定建模、資料倉儲與高效能計算等理論與技術，以從資料中萃取意義並創制資料產品為目標。(取自維基百科)



資料科學任務集

- 如同許多領域的問題一樣，各行各業的資料分析背後也有共同的任務集。

1. 資料縮減 (橫向精簡與縱向降維，後者又稱為屬性挑選與萃取) 將大的資料集替換為包含資料中重要訊息的小型資料集。

- 小型資料集較容易處理，且訊息揭露性較佳。
- 例如：顧客影片觀賞偏好資料集(顧客vs.影片)，可縮減為變數較少的潛在(latent)品味偏好資料集(eg. 顧客vs.影片風格 genre)。
- 資料縮減牽涉或多或少的訊息流失，重要的是如何權衡取捨以提高訊息的洞察力。

資料科學任務集

2. 相似度匹配 (問題解決的基本邏輯：異中求同、同中求異)

透過資料去衡量顧客間或品項間的相似程度，通常用相似度(similarity)或距離(是一種不相似度的衡量)來找尋類似實體

- 案例：
IBM 想找到與公司過去的最佳商業夥伴相似的公司，以讓銷售團隊專注於最佳的銷售機會。於是公司透過企業統計變項(firmographic)資料進行相似度分析。
- 相似度匹配是產品推薦系統的基礎（以相似的人喜歡或購買的商品進行推薦，或是推薦與過去購買商品相似的品項）。
- 相似度衡量亦是分類、迴歸、群集(clustering)等資料探勘任務的解決方法內涵。

資料科學任務集

3. 剖繪profiling(又稱行為描述)

描繪個體、群組或母體的典型行為

- 典型的問題是：這個顧客區隔 (segment)的典型手機用量為何？

描繪手機使用行為可能需要對夜間和週末通話時間的平均值、國際話費、漫遊費、簡訊用量等做分析，剖繪可對母體、或下至群組及個人的行為做描述。

- 剖繪經常用來建立屬於異常行為的偵測標準，例如異常偵測(anomaly detection)中的詐欺偵測和電腦系統的入侵偵測。
- 假若知道一個人通常使用信用卡進行哪些消費，我們便可以判斷新的消費是否符合使用者的消費行為剖繪，計算其不符合程度的分數，作為詐欺嫌疑分數，當分數太高時即發出警報

資料科學任務集

4. 集群

將一群物體中相似的個體(eg. 顧客)聚集在一起，集群並非因為特定的目的而驅動，故屬於非監督式的學習

- 集群問題：我們的顧客是否形成自然群體或是自然區隔？
 - 集群常用在前期的探勘，可以觀察有哪些自然群體存在，分群結果可用來建議其他資料探勘的任務
- 分群也可以當作決策程序的輸入
 - 例如：公司應該提供或開發何種產品？
 - 公司的顧客服務團隊（或銷售團隊）應該如何建構？

資料科學任務集

5. 共同出現分組(co-occurrence or affinity grouping)

又稱為頻繁項目集探勘、關聯規則探勘和購物籃分析，透過與實體有關的交易找出實體間的關聯

- 例如：哪些品項經常會一起被購買。
- 分群基於資料物件屬性上的相似性，而共同出現分組以是否一起交易來衡量物件的相似性。
- 如何針對所發現的關聯來擬定行動方案需要創造力，例如：特惠活動、產品陳列或組合優惠等的規劃。
- 有些推薦系統也通過關聯分組來運作的，例如兩本書經常被同一個人購買，因此構成推薦的基礎。
- 共同出現情況的描述包含共同出現的頻率(即關聯規則的支持度)，以及共同出現的驚奇度(即關聯規則新奇的程度)。

資料科學任務集

6. 分類及類別的機率估計

用來預測群體中的每一位顧客屬於何種類別，類別集合中的類別通常為互斥的。

- 分類問題：電信公司的顧客對於顧客保留方案是否會回應？據此將顧客分為兩個類別—回應與不會回應
- 對於分類任務而言，資料探勘程序產生一分類模型(classifier)，用以判定新的顧客屬於哪一個類別？例如：流失或不會流失、違約或不會違約、購買或不會購買
- 另一相關的任務為評分(scoring)或分類機率估計(class probability estimation)
- 上述兩者密切相關！A model that can do one can usually be modified to do the other.

商業問題的資料探勘任務

7. 迴歸（值估計）

通過一些變量預測或估計另一變量

- 迴歸問題：特定顧客會使用多少服務？
變量為服務用度(service usage)，可透過母體中其他相似顧客的過去用度去建立模型
- 迴歸與分類相關，但兩者是不同的。
 - 分類：預測事件是否會發生，或屬於哪一類別
 - 迴歸：預測事件發生多少量

商業問題的資料探勘任務

8. 因果模式

了解哪些事件或行為會實際影響其他事件或行為

- 使用預測模式以鎖定目標消費者族群進行廣告促銷，並發現被鎖定的消費者的確購買機率提高
- 但這是因為廣告影響顧客的購買意願嗎？還是預測模型能精準的找出無論如何都會購買的顧客？
- 因果建模包括實驗法(網頁最佳化A/B測試)和觀察法，它們嘗試去了解某一處理(treatment)其各種情況下的差異為何。

資料科學任務集

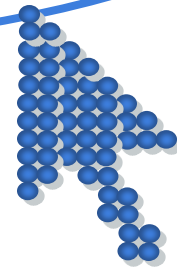
9. 鏈結預測 link prediction

(圖形資料探勘graph mining的用途之一)

預測網路節點間的連結可能性，並估計連結的強度

- 經常用於社群網路的好友推薦，例如：你和凱倫有10位相同的朋友 (shared nearest neighbor clustering based on graph)，極有可能你與凱倫會成為好友。
- 估計鏈結強度，例如：推薦給顧客的電影可想成一張顧客與其看過或評分過的電影所形成的網路圖。尋找圖中應該存在且關係強烈，但目前不存在的顧客電影連結，則這些連結就可以成為推薦的基礎。
- 許多模型背後通常有假設，資料科學家要了解假設可能使得分析結論是無效的 (因為你的資料與情境，很可能不符合模型的假設！)

常見的 資料視覺化圖形

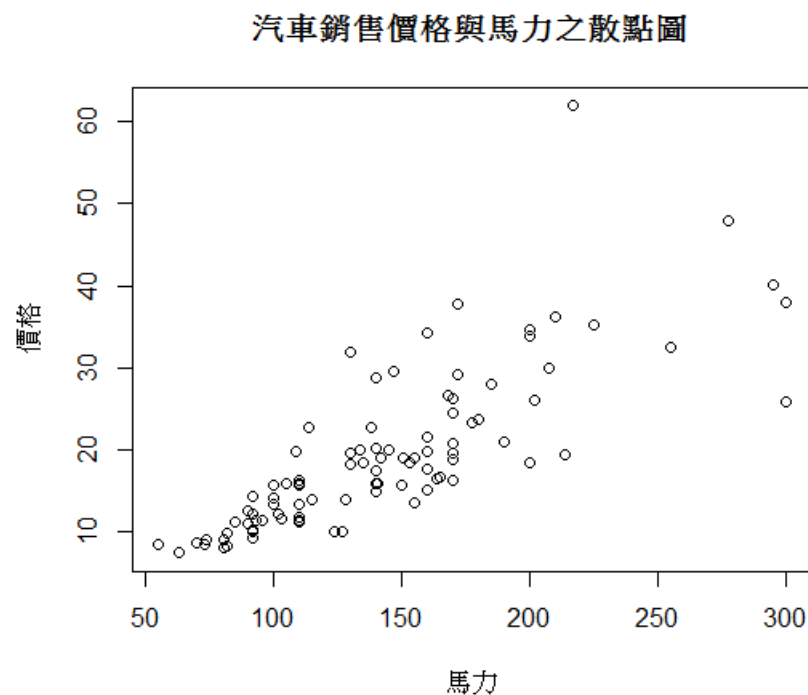


散佈圖

- 散佈圖：又稱為散點圖(scatter plot)，用來說明兩個變量的關係圖，散佈圖將資料繪製在2D或3D圖形。
 - 只畫資料點，點與點之間不連接。
 - 變數關係包括以下三種情形：
 - 無明顯關係
 - 呈現正相關
 - 呈現負相關

散佈圖(續)

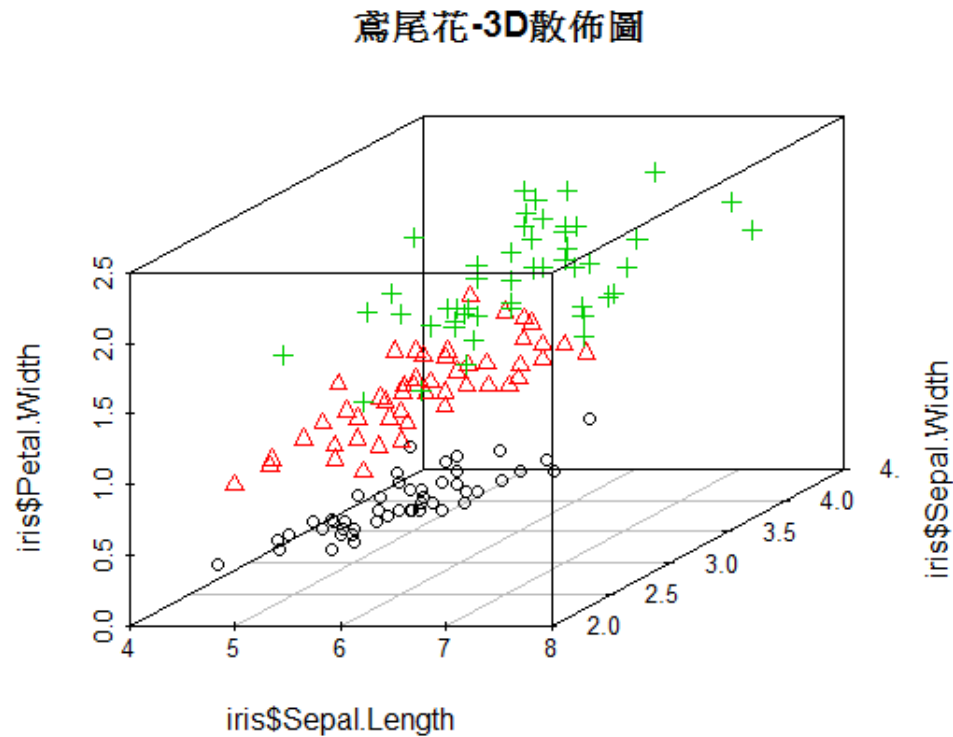
- 繪製汽車馬力與價格散佈圖。



呈現正相關

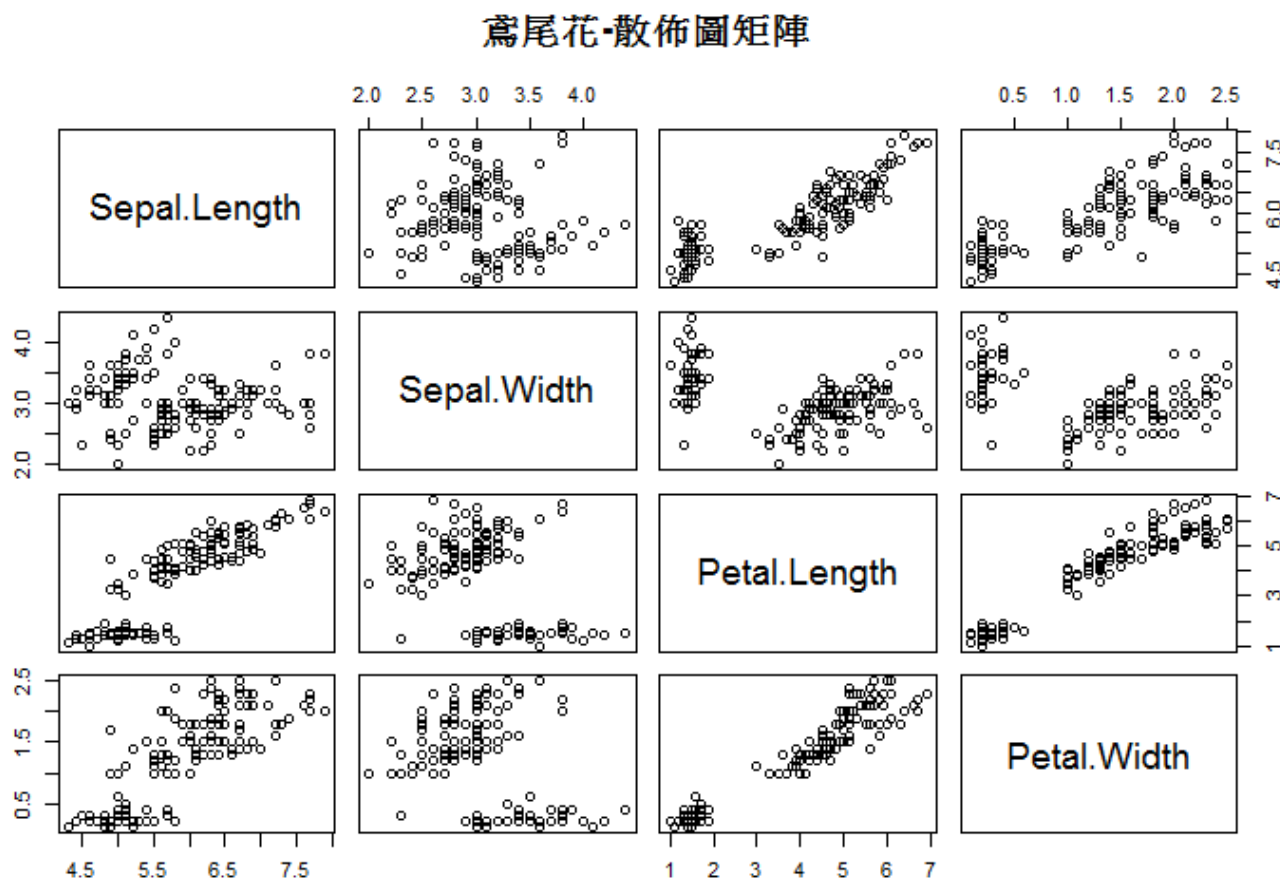
3D散佈圖

- 繪製鳶尾花-3D散佈圖。



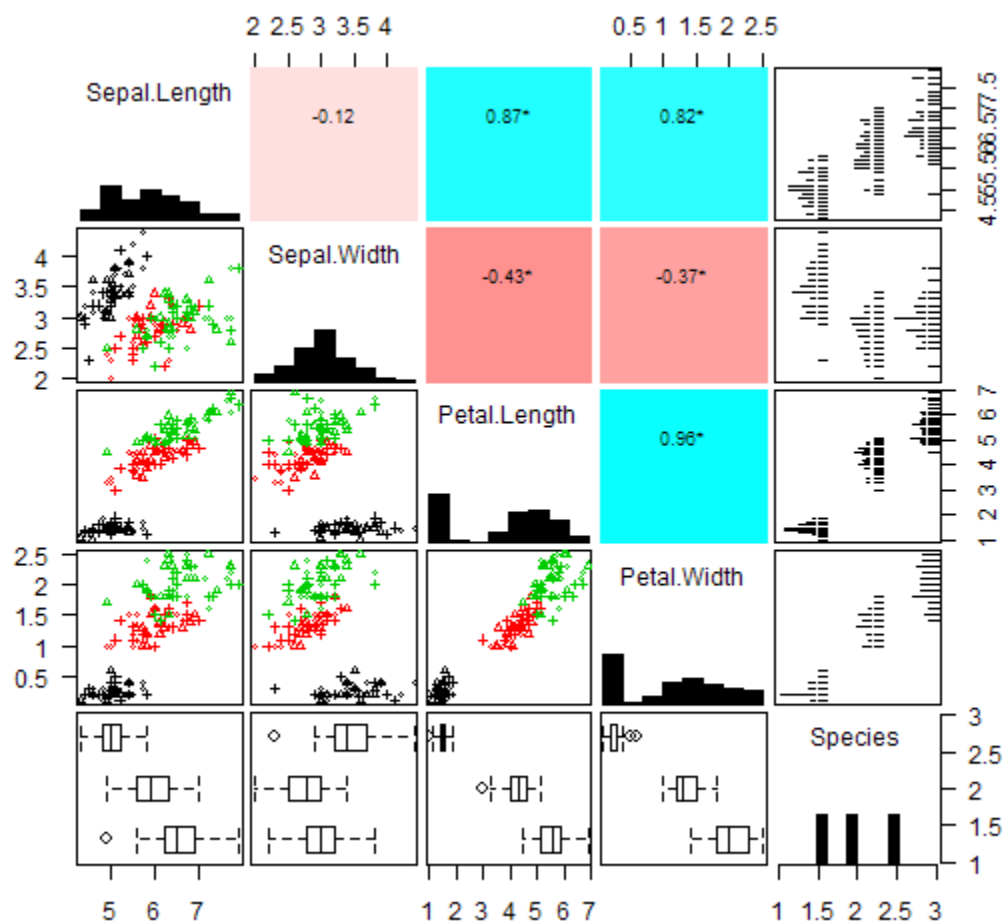
散佈圖矩陣

■ 繪製鳶尾花-散佈圖矩陣



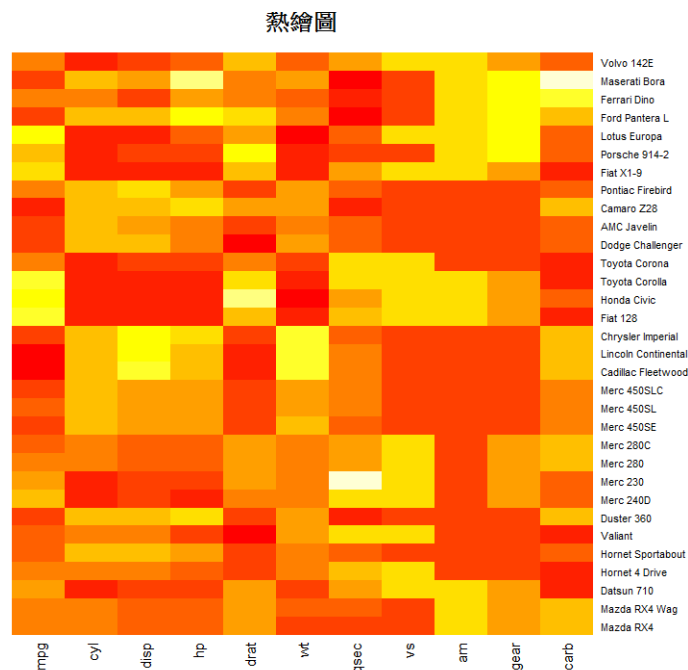
散佈圖矩陣(續)

■ 繪製鳶尾花-散佈圖矩陣(續)



熱繪圖

- 熱繪圖(Heat map)是將二維數值矩陣以顏色表示。
 - 較大的值由小暗灰色或黑色方塊（像素），較小的值由更淺的方塊表示。



文字雲

- 文字雲(Word cloud)是分析文章內容，然後組合成一幅圖，圖中會將文章的高頻字用比較大的字體顯示，低頻字就用比較小的字體。

