資料探勘:
概念與技術

— 第二章 —

第二章: 資料前處理

- 為何要資料前處理?
- 敘述資料彙總
- 資料清除
- 資料整合與轉換
- 資料縮減
- 離散化與產生概念階層
- 總結

為何要資料前處理?

- 真實世界的資料是不乾淨
  - 不完整:缺少屬性值或某些有興趣的屬性或僅包含聚合資料
    - 例, 職業=" "
  - 有雜訊:包含錯誤或離異值
    - 例, 薪資="-10"
  - 不一致:進行商品分類時部門代碼的差異
    - 例., 年齡="42" 生日="03/07/1997"
    - 例., 過去分類 "1,2,3", 現在分類 "A, B, C"
    - 例., 重複紀錄的不一致

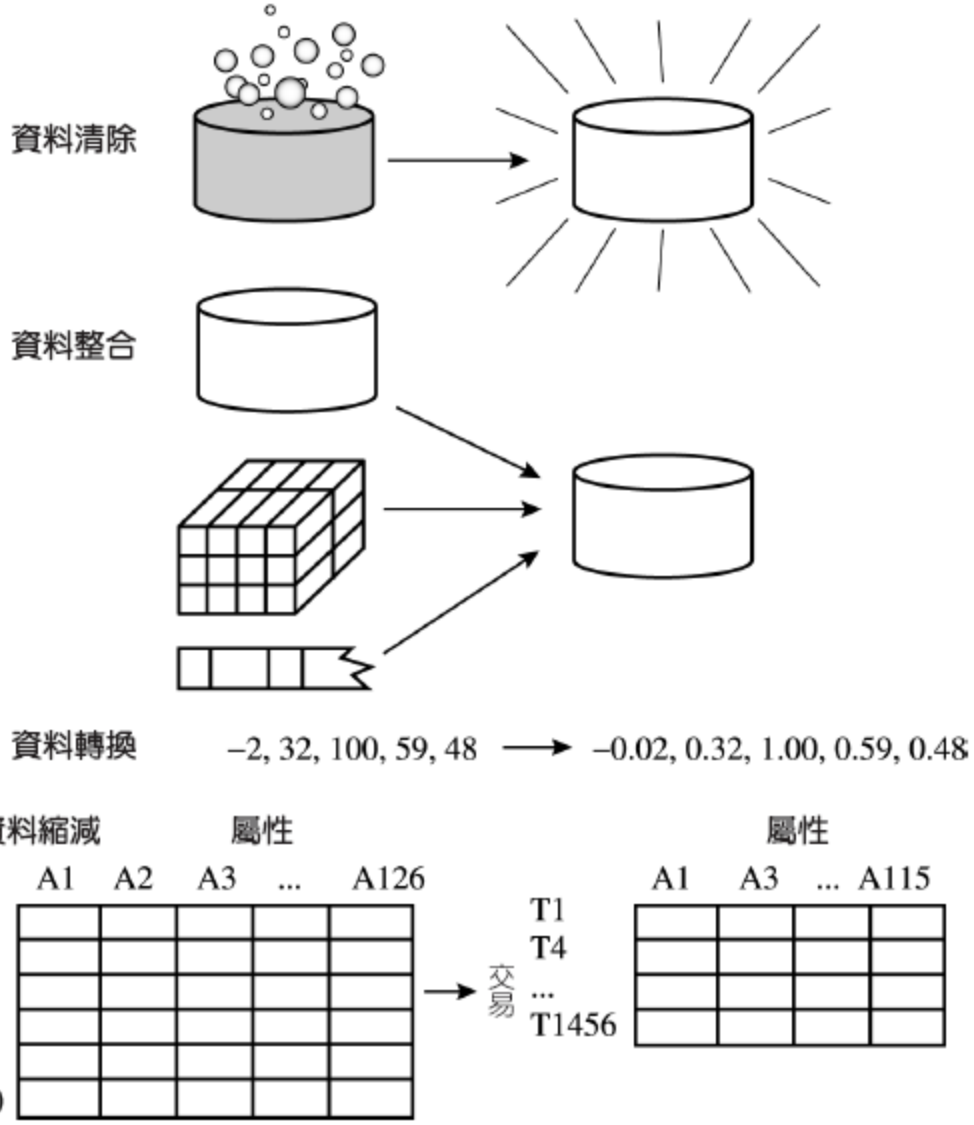
為什麼資料是不乾淨?

- 不完整來自
  - 輸入的時候認為它是不重要的項目
  - 相關資料由於誤解或設備故障沒有被記錄下來
  - 歷史資料的紀錄或修改並未仔細檢查
- 雜訊來自 (不正確值)來自
  - 資料收集工具的錯誤
  - 資料輸入時人為或電腦錯誤
  - 資料傳輸錯誤
- 不一致來自
  - 不同資料來源
  - 違反功能相依 (例, 修改一些連結資料)
- 重複紀錄需要資料清除

資料前處理主要工作

- 資料清除
  - 代入遺失值、將雜訊平滑化、找出且移除離異值、並解決不一致
- 資料整合
  - 整合許多資料庫、資料方塊或檔案
- 資料轉換
  - 正規化與聚合
- 資料縮減
  - 降低資料集的大小並能獲得相同的分析結果
- 資料離散化
  - 資料縮減一部分,但是對數值資料特別重要

資料前處理型式


資料清除
資料整合
資料轉換
–2, 32, 100, 59, 48 → –0.02, 0.32, 1.00, 0.59, 0.48
資料縮減
屬性
A1 A2 A3 ... A126
T1
T2
T3
T4
...
T2000
交易
屬性
A1 A3 ... A115
T1
T4
...
T1456
交易

第二章: 資料前處理

- 為何要資料前處理?
- 敘述資料彙總
- 資料清除
- 資料整合與轉換
- 資料縮減
- 離散化與產生概念階層
- 總結

衡量主要傾向

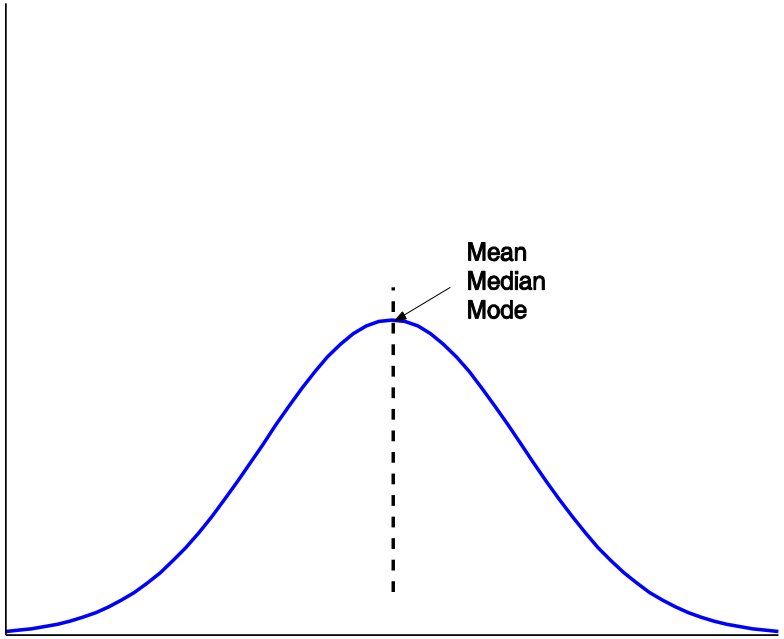
- 均值(代數式度量) (樣本或族群): $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \mu = \frac{\sum x}{N}$
  - 權重均值:
  - 刪除均值: 刪除極值

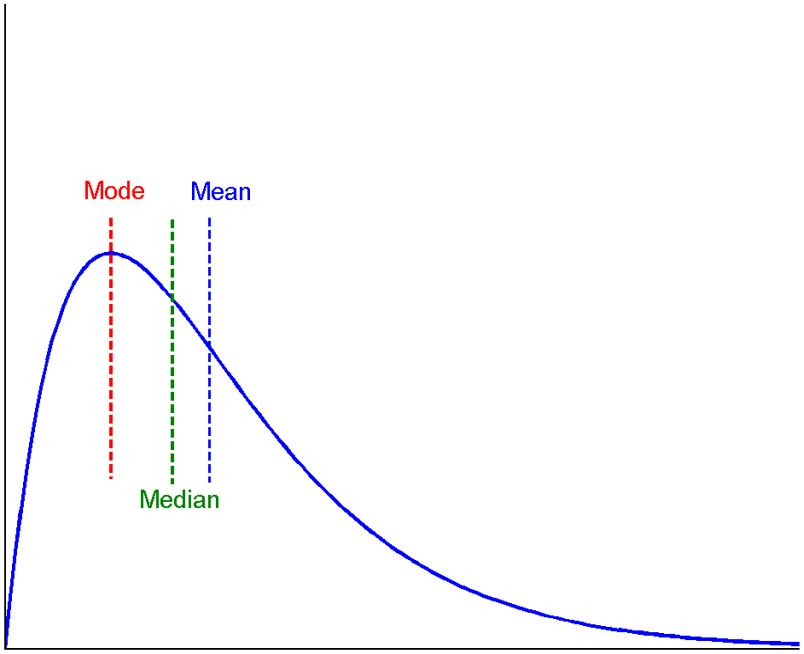
$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

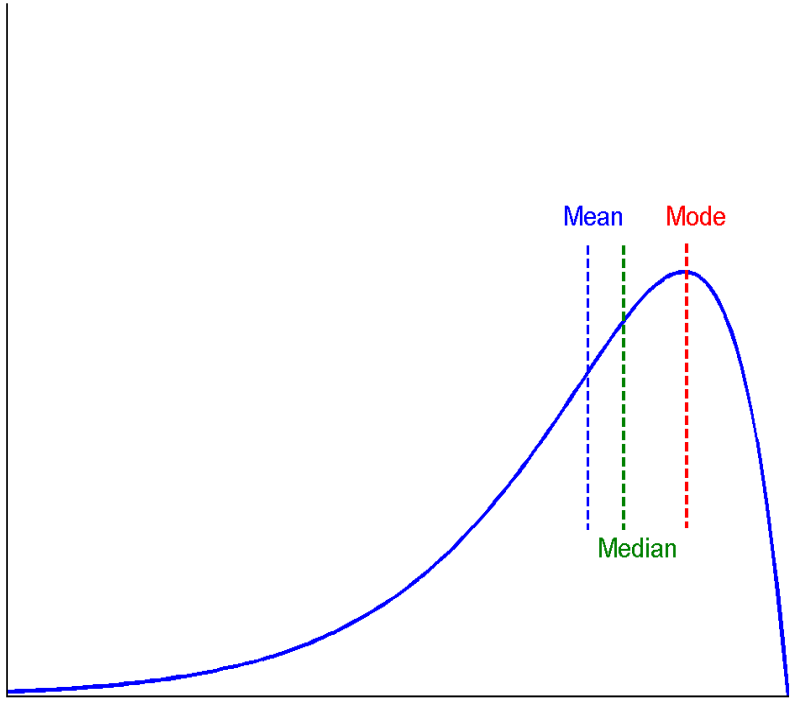
- 中值:整體度量
  - 如果數值個數為奇數則中值為所有數字的中值, 反之則為最中間的兩個數字平均
  - 由內插法來預估(群組資料): $median = L_1 + (\frac{n/2 - (\sum f)l}{f_{median}})c$
- 模式
  - 出現最頻繁的值
  - 單一模式,雙模式,三模式
  - 實證公式: $mean - mode = 3 \times (mean - median)$

對稱 與 偏斜資料

- 對稱中值, 均值與模式, 正向與負向偏斜資料







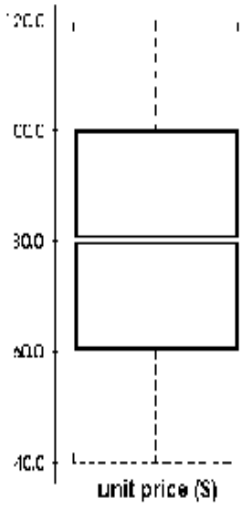
衡量資料分佈程度

- 四分位數,離異值與盒狀圖
  - 四分位數: Q_1 (第25個百分位數), Q_3 (第75個百分位數)
  - 四分位距: IQR = $Q_3 - Q_1$
  - 五個數字彙總: min, Q_1, M, Q_3, max
  - 盒狀圖:盒子的兩邊為四分位數,盒子中的一條線代表中值,盒子外的兩條線 (whiskers, 鬚晶) 延伸至觀察最小與最大值,離異觀察值單獨被畫出
  - 離異值: 高/低超過 1.5 x IQR
- 變異數與標準差 (樣本*: s, 族群: σ)*
  - 變異數: (代數, 可度量計算)

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}[\sum_{i=1}^{n}x_i^2 - \frac{1}{n}(\sum_{i=1}^{n}x_i)^2] \qquad \sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{n}x_i^2 - \mu^2$$

  - 標準差 *s (or σ)* 變異數 s^2 (or σ^2) 平方根

盒狀圖分析

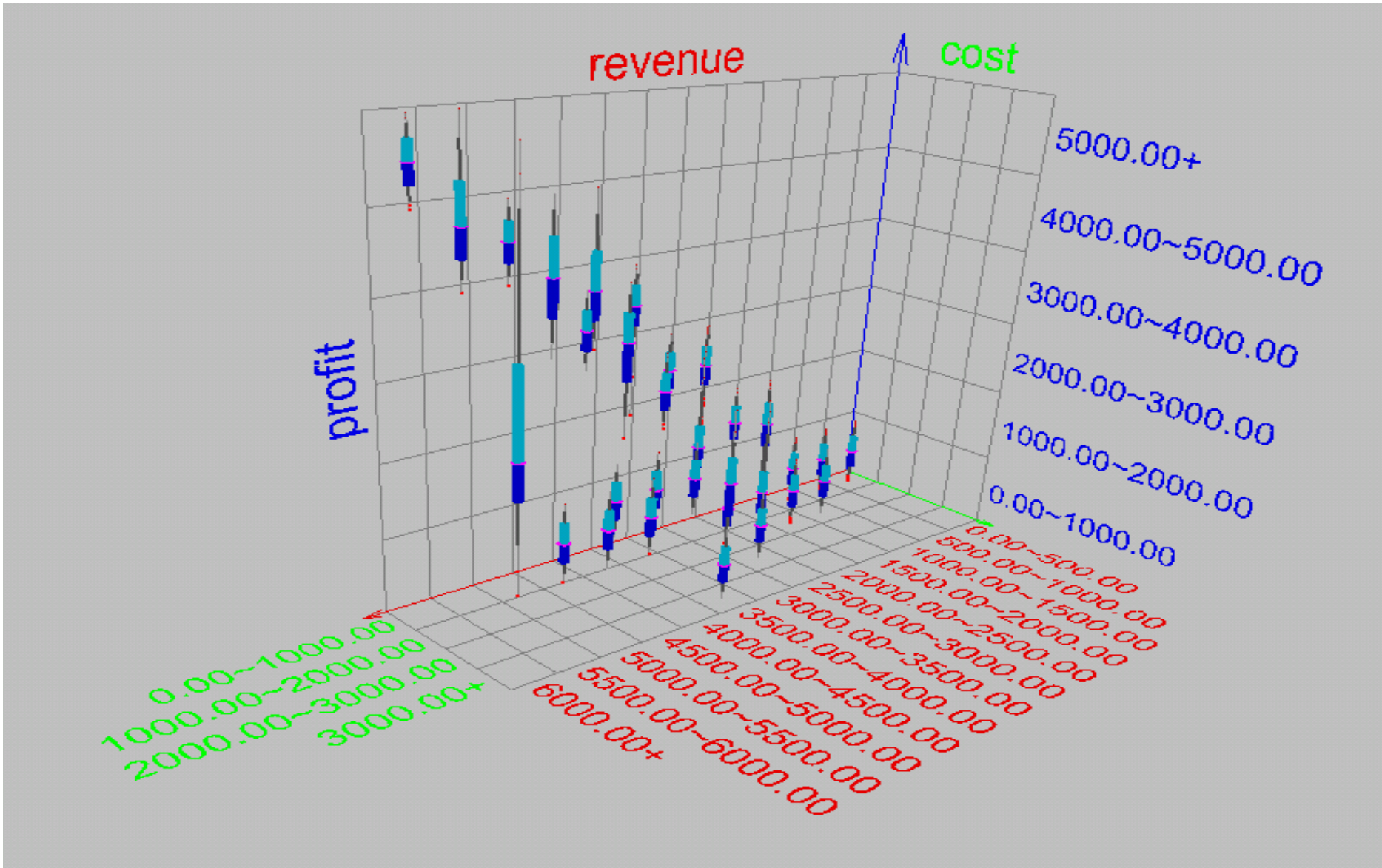


- 五個數字彙總:

  Minimum, Q1, M, Q3, Maximum

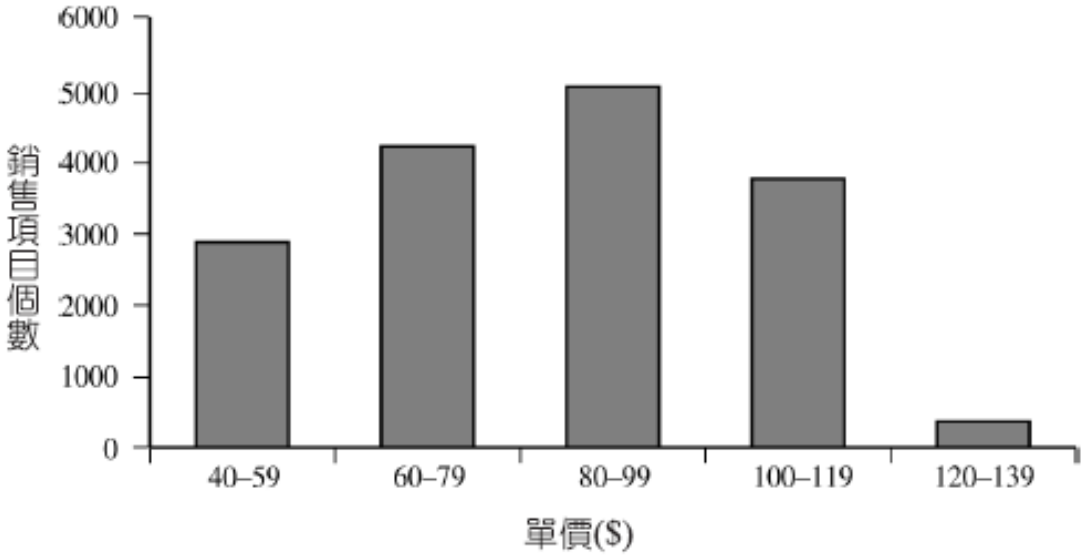
- 盒狀圖
  - 資料用盒狀表示
  - 盒子的兩邊為第一與第三個四分位數，盒子的長度為四分位距
  - 盒子中的一條線代表中值
  - 盒子外的兩條線 (whiskers, 鬚晶) 延伸至觀察最小與最大值

資料分布視覺化: 盒狀圖分析



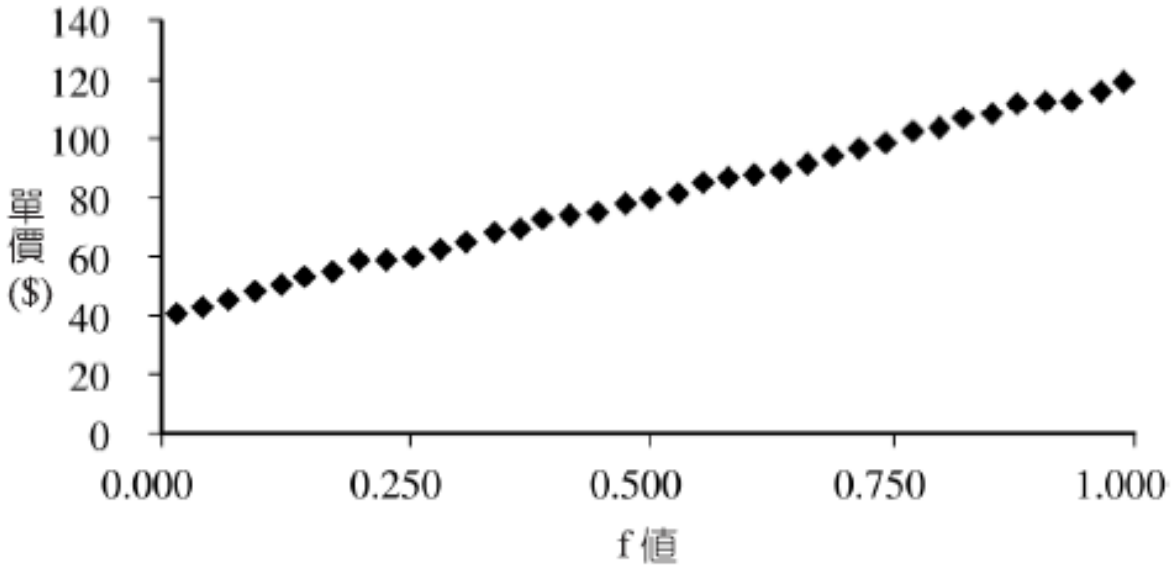
直方圖分析

- 基本統計類別敘述的圖示
  - 頻率直方圖
    - 單變量圖示方法
    - 儲存區以長方形表示，長方形的高度代表包含在儲存區的資料值的個數或相對頻率



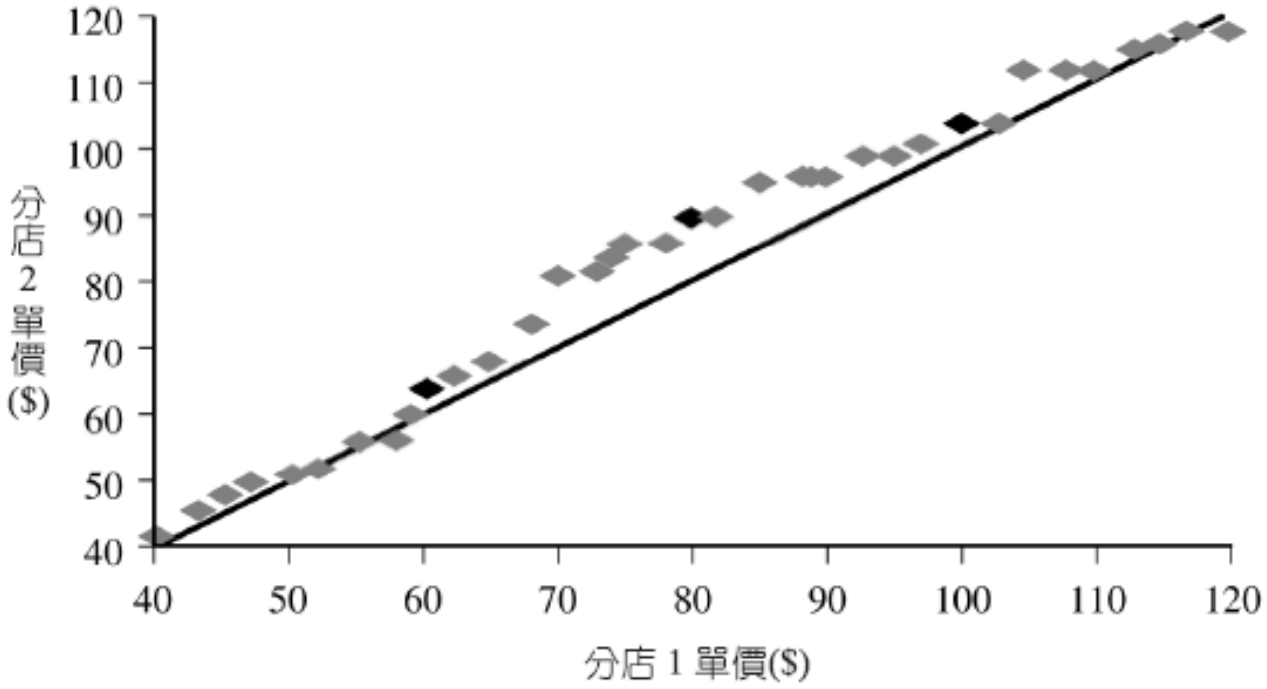
量分位圖

- 顯示特定屬性的所有資料 (允許使用者評估全部行為與不尋常出現)
  - 繪製量分位訊息
  - 將資料 x_i 遞增排序, f_i 表示有100% f_i的資料是低於x_i或與x_i相等



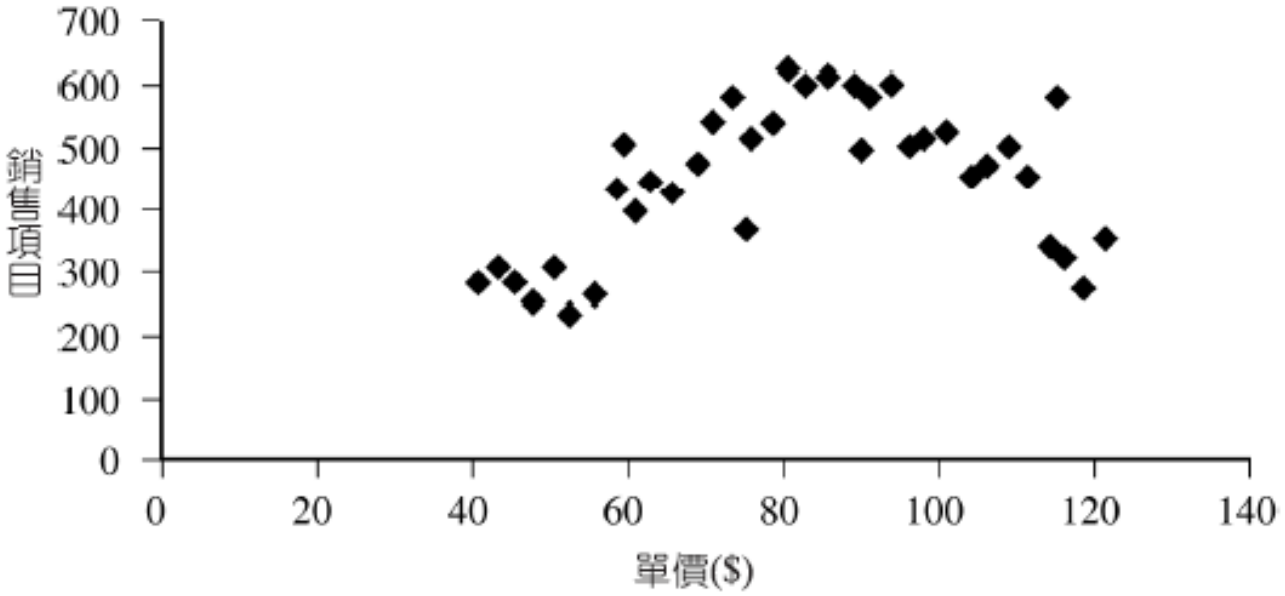
量分位-量分位圖(Q-Q)

- 將一個單變量的量分位與另一個相對應的量分位透過繪圖來進行比對
- 檢視資料從一個分佈到另一個分佈是否有位移



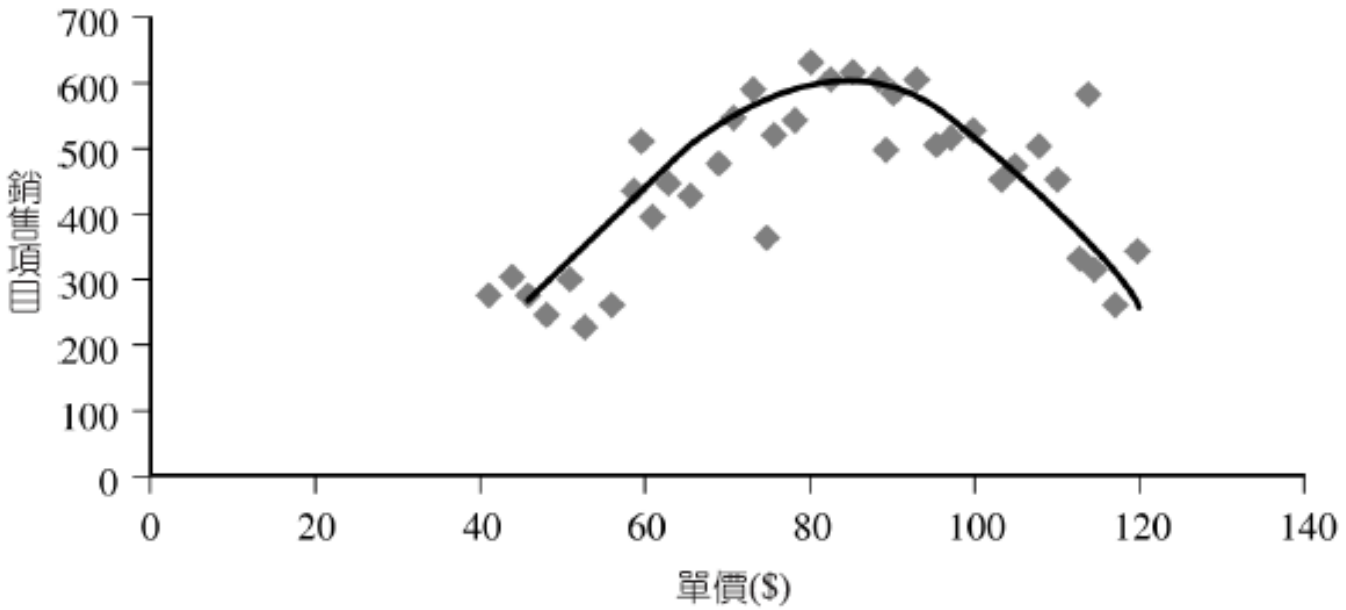
分佈圖

- 決定兩個數值屬性是否顯示一個關係、樣式或趨勢
- 每一對的屬性值視為代數座標，並將這些座標點畫在平面上



區域迴歸曲線

- 在分佈圖中加入一條平滑曲線，以提供樣式相依更好的知覺
- 為了代入區域迴歸曲線，我們需要設定兩個參數：為平滑參數，為代入迴歸式多項式的階次。



第二章: 資料前處理

- 為何要資料前處理?
- 敘述資料彙總
- 資料清除
- 資料整合與轉換
- 資料縮減
- 離散化與產生概念階層
- 總結

資料清除

- 重要性
  - "資料清除為資料倉儲中三個最大問題之一"—Ralph Kimball
  - "資料清除為資料倉儲首要問題"—DCI survey
- 資料清除工作
  - 填補遺失值
  - 找出利益值並淡化 (平滑) 雜訊
  - 修正資料的不一致
  - 解決資料整合所造成重複

如何處理遺失值?

- 忽略這些值組:通常用於進行判別時值組的類別是遺失的狀況，這個方法不是很有效，除非許多值組的屬性包含遺失值.
- 利用人工方式填入遺失值:非常費時並不實際?
- 自動填入
  - 利用全域常數 (global constant) 填入遺失值：例., "未知", 新類別!
  - 使用屬性均值來填入遺失值
  - 使用相同類別值組的屬性均值: 較聰明方法
  - 使用最有可能的值來填入遺失值:可以透過迴歸、利用貝氏理論的推論式工具或決策樹推論來決定

如何處理雜訊資料?

- 箱狀法
  - 首先資料先經過排序，然後分成頻率相同箱子
  - 然後進行箱子均值平滑化法,箱子中值平滑化法,箱子邊界值平滑化法等
- 迴歸
  - 透過將資料對應至函數來進行平滑化
- 分群
  - 偵測並移除離異值
- 整合電腦與人檢驗
  - 利用人檢測有疑問的值 (例., 處理可能離異值)

簡單離散化方法: 箱狀法

- 等寬 (距離) 分割
  - 切割成N個等寬範圍: 均勻格線
  - 如果 *A* 與 *B* 為屬性的最低與最高值, 範圍寬度為: $W = (B - A)/N$.
  - 最直接但是離異值會主導表示
  - 對偏斜資料不能處理很好
- 等深 (頻率) 分割
  - 切割成N個範圍, 每個範圍有大約相同樣本數
  - 具好資料量度性
  - 對類別屬性會較難處理

利用箱狀法進行資料平滑化

儲存價格資料($) : 4, 8, 15, 21, 21, 24, 25, 28, 34

分割成(等頻率)箱子

箱子 1 : 4, 8, 15
箱子 2 : 21, 21, 24
箱子 3 : 25, 28, 34

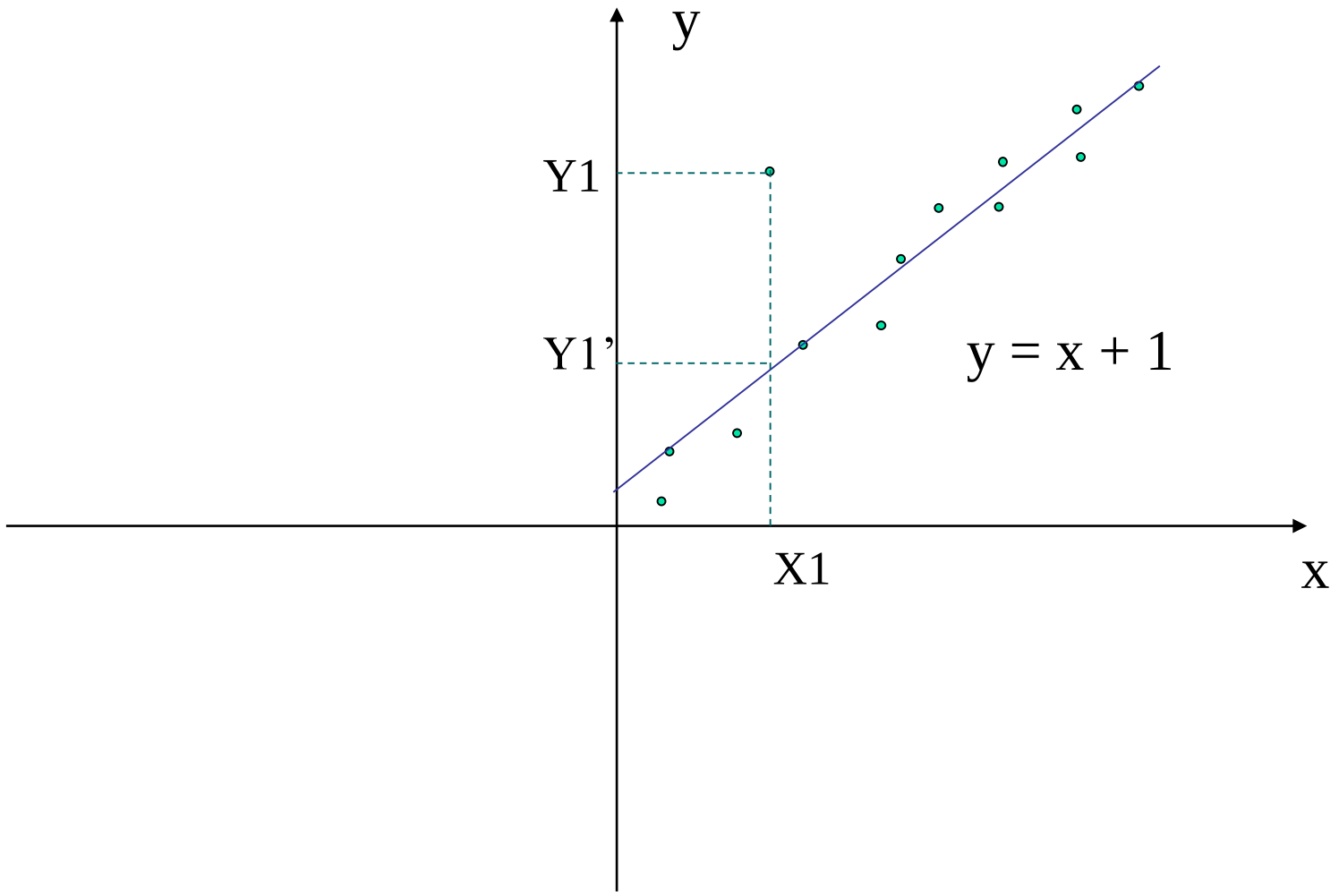
用箱子均值平滑化

箱子 1 : 9, 9, 9
箱子 2 : 22, 22, 22
箱子 3 : 29, 29, 29

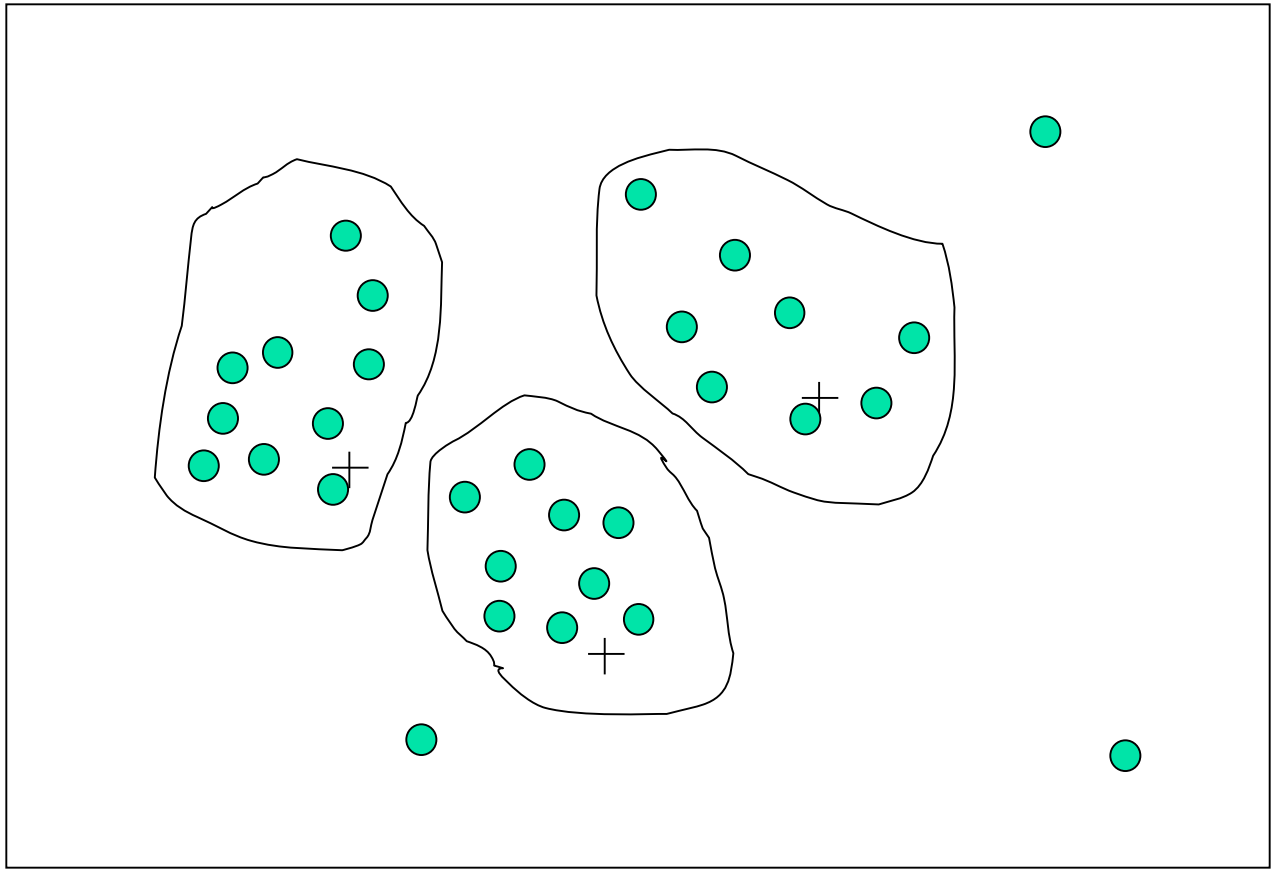
用箱子邊界值平滑化

箱子 1 : 4, 4, 15
箱子 2 : 21, 21, 24
箱子 3 : 25, 25, 34

迴歸



分群分析



資料清除過程

- 差異檢測
  - 使用詮釋資料 (例., 範疇, 範圍, 相依, 分佈)
  - 檢查欄位超載(field overloading)
  - 對獨特規則、連續規則與空規則來進行檢視
  - 使用商用工具
    - 資料擦洗:用簡單的範疇知識 (地址的知識或拼字檢查) 來發現錯誤並進行更正
    - 資料稽核工具:透過分析資料來找出規則與關係，並檢查違反條件的資料(例.,用統計分析來尋找相互關係，並利用分群來尋找離異值)
- 資料遷移與整合
  - 資料遷移工具:允許設定簡單轉換
  - ETL (Extraction/Transformation/Loading)工具:允許使用者透過圖形介面設定轉換
- 兩種過程整合
  - 重複與互動 (例., Potter's Wheels)

第二章: 資料前處理

- 為何要資料前處理?
- 敘述資料彙總
- 資料清除
- 資料整合與轉換
- 資料縮減
- 離散化與產生概念階層
- 總結

資料整合

- 資料整合:
  - 將許多來源的資料整合成一個連貫的資料
  - 綱目整合: 例., A.cust-id ≡ B.cust-#
  - 整合不同來源的詮釋資料
- 個體識別問題:
  - 樣對不同資料來源的真實個體進行比對, 例., Bill Clinton = William Clinton
- 發掘並解決資料值衝突問題
  - 真實世界的個體, 屬性值會來自不同來源
  - 可能原因: 不同表示, 不同量化, 例., 公制與英制

在資料整合中處理重複問題

- 當整合許多資料庫會導致資料重複
  - 個體識別: 在不同資料庫, 相同屬性或個體會有不同名稱
  - 推論資料: 某個屬性可以從另一個資料表的屬性推論得之, 例., 年盈餘
- 多餘屬性可以藉由相互關係分析 (correlation analysis) 找出
- 仔細從許多來源整合資料可以降低並避免重複與不一致, 有助於改善探勘的速度與品質

相互關係分析(數值資料)

- 相互關係係數(皮爾遜積差係數)

$$r_{A,B} = \frac{\sum (A-\overline{A})(B-\overline{B})}{(n-1)\sigma_A \sigma_B} = \frac{\sum (AB) - n\overline{A}\overline{B}}{(n-1)\sigma_A \sigma_B}$$

N 爲值組個數，a_i 與 b_i 爲屬性 A 與 B 在值組 i 的值，$\overline{A}$ 與 $\overline{B}$ 爲屬性 A 與 B 的均值，σ_A 與 σ_B 爲 A 與 B 的標準差，$\sum(a_i b_i)$ 爲 AB 交互乘積(cross-product，對每個值組，屬性 A 的值乘以屬性 B 的值)的總和。請注意 $-1 \le r_{A,B} \le +1$。當 $r_{A,B}$ 大於 0 表示 A 與 B 爲正相關，它表示當 A 的值增加時 B 的值也會跟著增加，$\gamma_{A,B}$ 值越高表示兩者關係越強，因此一個很高的值顯示 A (或 B) 是多餘可被移除的。如果 $r_{A,B}$ 值爲 0，表示兩者之間爲獨立，兩者之間並沒有任何關係。如果 $r_{A,B}$ 值小於 0，則兩者爲負關聯，也就是一個值增加則另一個值就減少，也就是說一個屬性會攔阻另一個屬性。分佈圖也可用於檢視屬性的相互關係

相互關係分析(類別資料)

- X^2 (chi-square) 檢定

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

- X^2 值愈大, 變數愈有關聯
- 對X^2值有貢獻的儲存格是它的個數與期望個數差異很大
- 相互關係並沒有暗示有因果關係
  - 醫院的個數與代表汽車失竊的個數是相關的
  - 這兩個屬性都與另一個屬性 (人口數) 有因果的連結

Chi-Square 計算: 範例

	Play chess	Not play chess	Sum (row)
Like science fiction	250(90)	200(360)	450
Not like science fiction	50(210)	1000(840)	1050
Sum(col.)	300	1200	1500

- X^2 計算 (括弧的數字代表期望頻率, 由兩個類別的資料分佈計算得之)

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 507.93$$

- 在這一群組中, 它顯示 like_science_fiction 與 play_chess 是相關

資料轉換

- 平滑化:移除雜訊
- 聚合:匯總或建立資料方塊
- 資料一般化: 攀緣概念階層
- 正規化:將屬性資料值轉換到較小的設定範圍
  - min-max 正規化
  - z-score正規化
  - 十進位正規化
- 屬性建立
  - 從既有屬性建立新的屬性

資料轉換: 正規化

- min-max 正規化: 轉換至 [new_min$_A$, new_max$_A$]

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

  - 例. 將收入範圍 $12,000 to $98,000 正規化至 [0.0, 1.0]. 則 $73,000 會對應至 $\frac{73{,}600 - 12{,}000}{98{,}000 - 12{,}000}(1.0 - 0) + 0 = 0.716$

- Z-score正規化 (μ: 均值, σ: 標準差):

$$v' = \frac{v - \mu_A}{\sigma_A}$$

  - 例. 當 μ = 54,000, σ = 16,000. 則 $\frac{73{,}600 - 54{,}000}{16{,}000} = 1.225$

- 十進位正規化

$$v' = \frac{v}{10^j}$$ j 為當 Max(|v'|) < 1時最小的整數

第二章: 資料前處理

- 為何要資料前處理?
- 敘述資料彙總
- 資料清除
- 資料整合與轉換
- 資料縮減
- 離散化與產生概念階層
- 總結

資料縮減策略

- 為何要資料縮減?
  - 資料倉儲中選擇進行分析的資料極為龐大
  - 複雜的資料分析與探勘龐大的資料會花費很長的時間
- 資料縮減
  - 希望獲得一個比原始資料小很多的縮減資料集，並且它幾乎能保持原始資料的完整性
- 資料縮減策略
  - 資料方塊聚合
  - 子屬性集選擇，例. 移除不重要屬性
  - 維度縮減
  - 數值縮減 — 例., 將資料帶入模型
  - 離散化或概念階層建立

資料方塊聚合

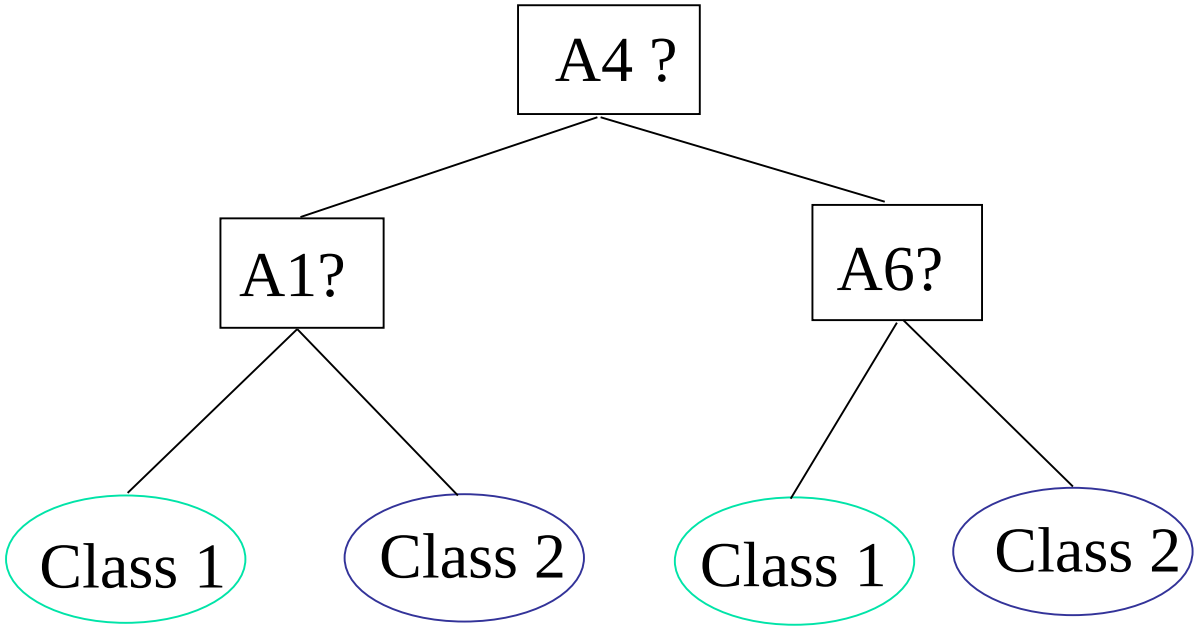
- 資料方塊最底層 (基礎長方體)
  - 有趣個體的聚合資料
- 資料方塊多層聚合
  - 進一步降低要處理資料的大小
- 參照適當層次
  - 使用足以處理問題的最小表示
- 相關於聚合的查詢, 儘可能使用資料方塊

子屬性集選擇

- 屬性選擇 (例., 屬性子集合選擇):
  - 希望找到最小的屬性集，並且這個最小的屬性集的結果要與使用全部屬性集的結果盡量接近
  - 降低探勘樣式屬性的數目，這樣會讓探勘的樣式更容易被了解
- 啟發式 (由於有冪次方數目的選擇):
  - 逐步向前選擇
  - 逐步向後刪除
  - 向前選擇與向後刪除的組合
  - 決策樹歸納

決策樹歸納範例

起始屬性集:
{A1, A2, A3, A4, A5, A6}

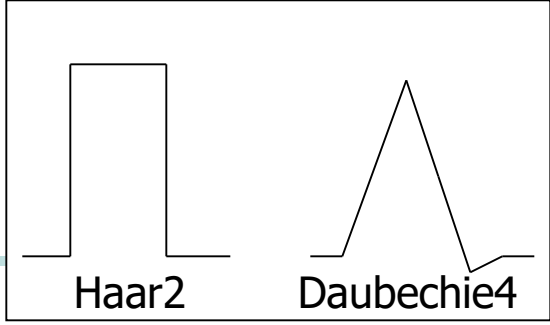


-------> 縮減屬性集: {A1, A4, A6}

啟發式屬性選擇方法

- 在d個屬性中有 2^d 種可能得子屬性集
- 許多啟發式屬性選擇方法:
  - 經由屬性獨立假設找出最好屬性: 透過顯著性檢定進行選取
  - 最佳逐步屬性選取:
    - 首先選擇最佳屬性
    - 在第一個屬性條件下選擇次佳屬性, …
  - 最佳逐步屬性刪除:
    - 重複刪除最差屬性
  - 最佳選擇與刪除的組合
  - 最佳分枝與界限:
    - 使用屬性移除與退卻

維度縮減:小波轉換



- 離散小波轉換(DWT): 線性訊號處理, 多分辨率分析
- 壓縮式近似: 僅儲存少量最強的小波係數
- 類似於離散傅立葉轉換(DFT), 但有較好的損耗壓縮與空間區域化
- 方法:
  - 長度, L, 必須是 2 的正冪次方(可以透過在資料向量加入0來完成)
  - 每個轉換牽涉到兩個函數:平滑化, 差異化
  - 套用至所有 成對資料，進而形成兩組長度為L/2的資料
  - 前一次獲得的資料會遞迴的代入這兩個函數，直到結果資料的長度為直到達到預設長度

DWT 用於影像壓縮

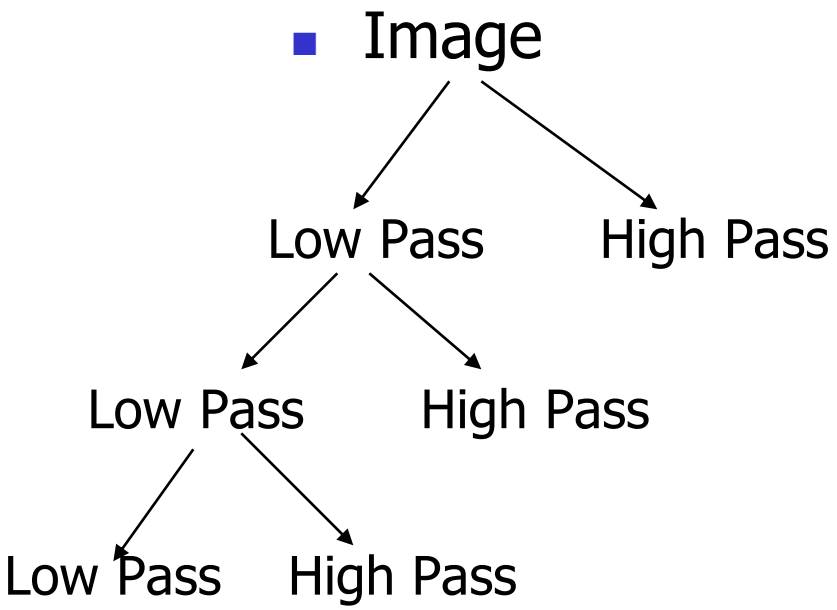
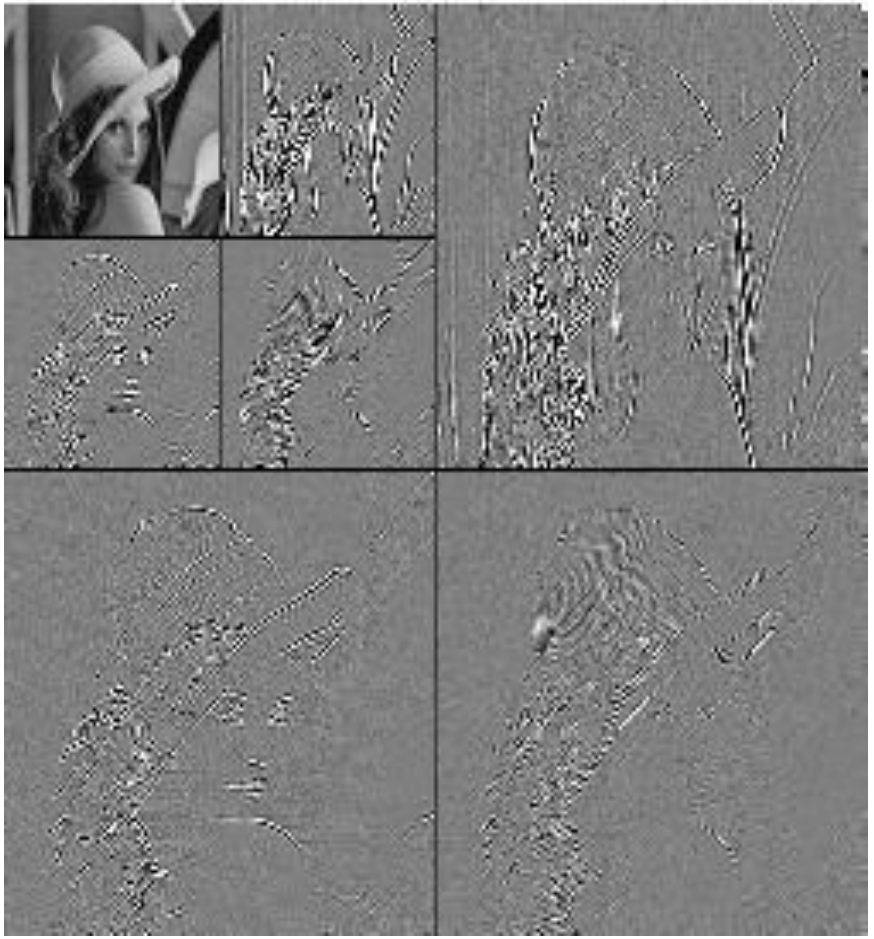


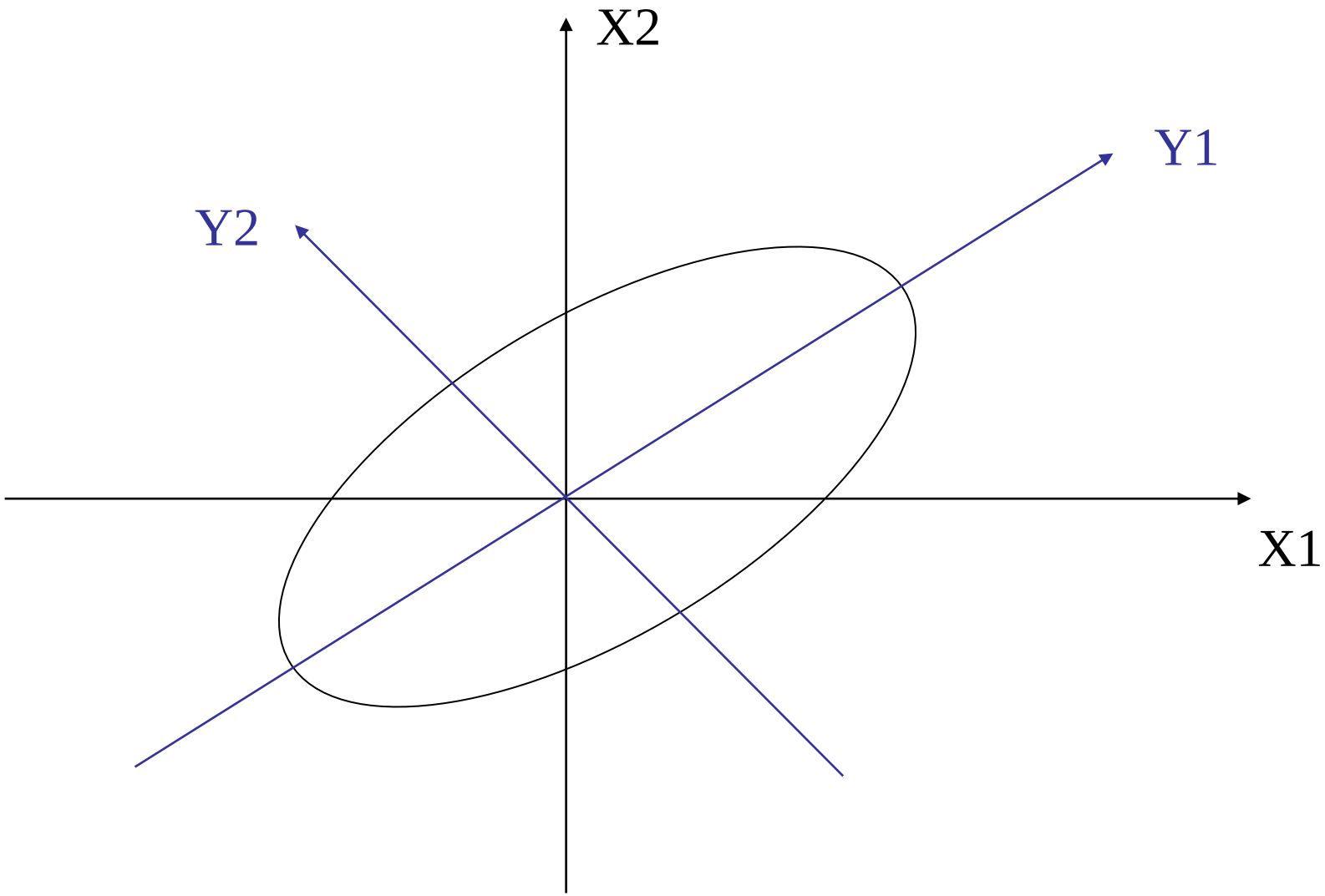
Image
Low Pass
High Pass
Low Pass
High Pass
Low Pass
High Pass



維度縮減: 主成分分析法 (PCA)

- 搜尋最能表示資料的k個維度正交向量, $k \leq n$
- 步驟
  - 輸入資料進行正規化: Each attribute falls within the same range
  - 計算k個正交向量, 也就是主成分
  - 輸入資料為主要成分的線性組合
  - 主要成分按照顯著性遞減排序
  - 因為成分按照顯著性遞減排序，所以透過移除較弱的成分就可以達到資料縮減的目的。使用最強的主成分，它應該足以近似於原始資料
- 僅適用於數值資料
- 當維度個數很大時使用

主成分分析



數值縮減

- 透過比較小的資料表示形式來表示縮減的資料
- 參數式方法
  - 一個模型會被用於估計原始資料，因此我們只需儲存模型的參數，而不需儲存原始資料 (除了可能離異值)
  - 範例: Log-linear 模型— 在m維空間資料點的值可由適當邊際子空間乘積獲得
- 非參數式方法
  - 不假設模型
  - 主要成員: 值方圖, 分群, 取樣

資料縮減方法 (1): 迴歸與對數線性模型

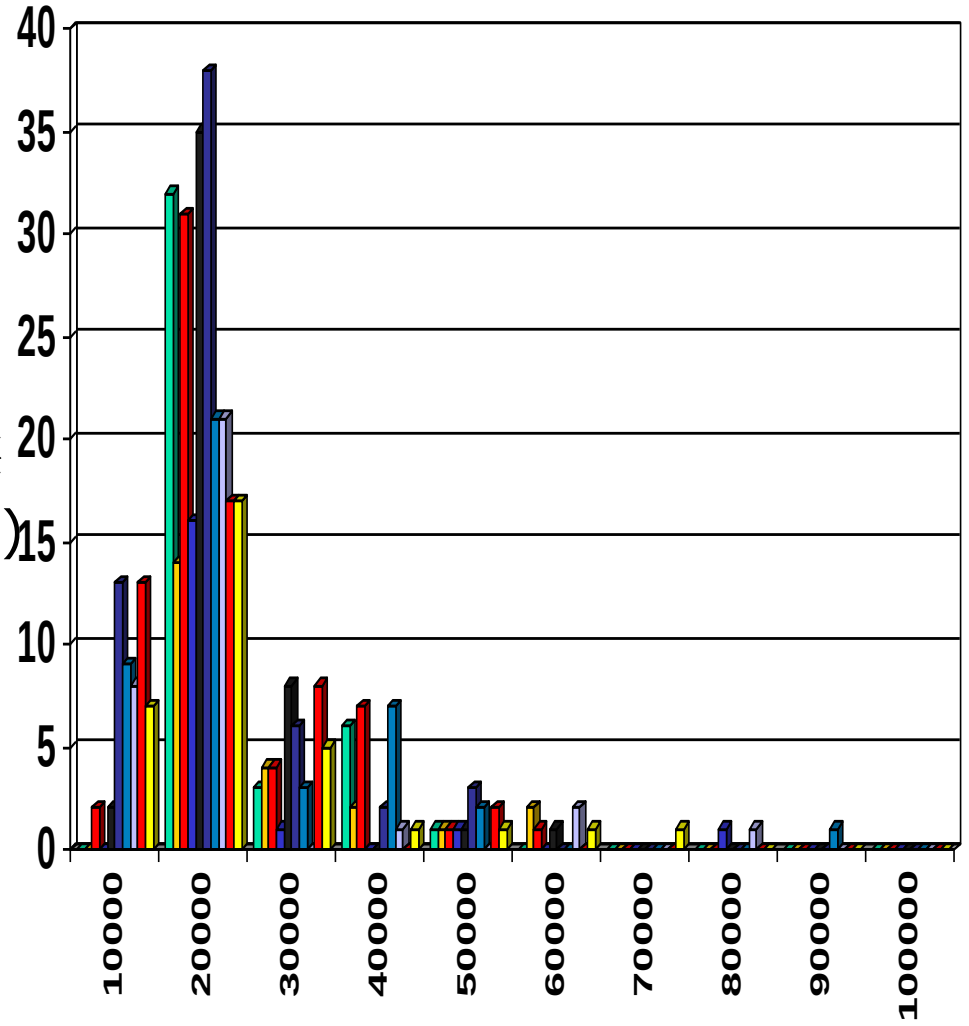
- 線性迴歸:資料被對應至一條直線
  - 通常使用最小平方差的方法
- 多元線性迴歸:顯示利用兩個或兩個以上的預測變數，來對回應變數設定線性模型
- 對數線性模型:預估離散多維度機率分佈

迴歸分析與對數線性模型

- 線性迴歸: $Y = w X + b$
  - 兩個迴歸係數w與b用於設定迴歸線, 而這兩個係數是用現有資料估計得來
  - 對已知的 $Y_1, Y_2, \ldots, X_1, X_2, \ldots$ 使用最小平方差條件
- 多元線性迴歸: $Y = b0 + b1\ X1 + b2\ X2.$
  - 可轉換許多非線性函數
- 對數線性模型:
  - 對一組離散屬性根據最小維度組合的子集合，來估計每一點在多維度空間的機率
  - 機率: $p(a, b, c, d) = \alpha ab\ \beta ac \chi ad\ \delta bcd$

資料縮減 (2): 直方圖

- 將資料分佈分割成箱子，在每個箱子儲存平均或總合
- 分割規則:
  - 等寬: 每個箱子範圍
  - 等頻率(等深)
  - V-極值:直方圖變異是最小的 (每個箱子所代表原始值的權重總和)
  - MaxDiff:箱子邊界會建立在能產生β-1的最大箱子差距



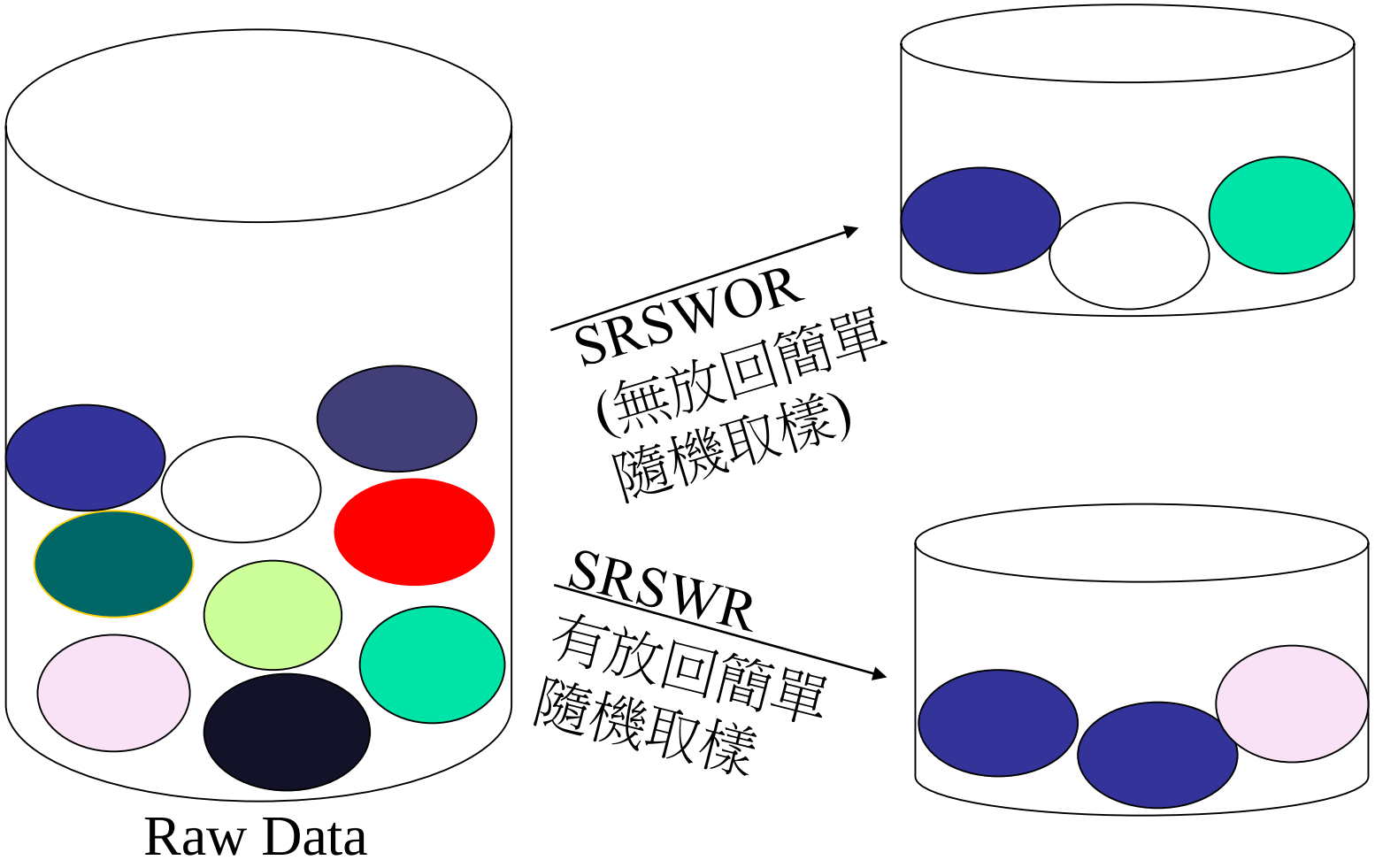
資料縮減 (3): 分群

- 根據相似度進行分群並儲存群組表示 (例.,中心與直徑)
- 如果資料是可群組會非常有效, 但是資料為不乾淨則否
- 可執行階層式分群並將結果存成多維度索引樹架構
- T有許多分群定義與分群運算法則的選擇
- 分群分析會在第七章詳述

資料縮減方法 (4): 取樣

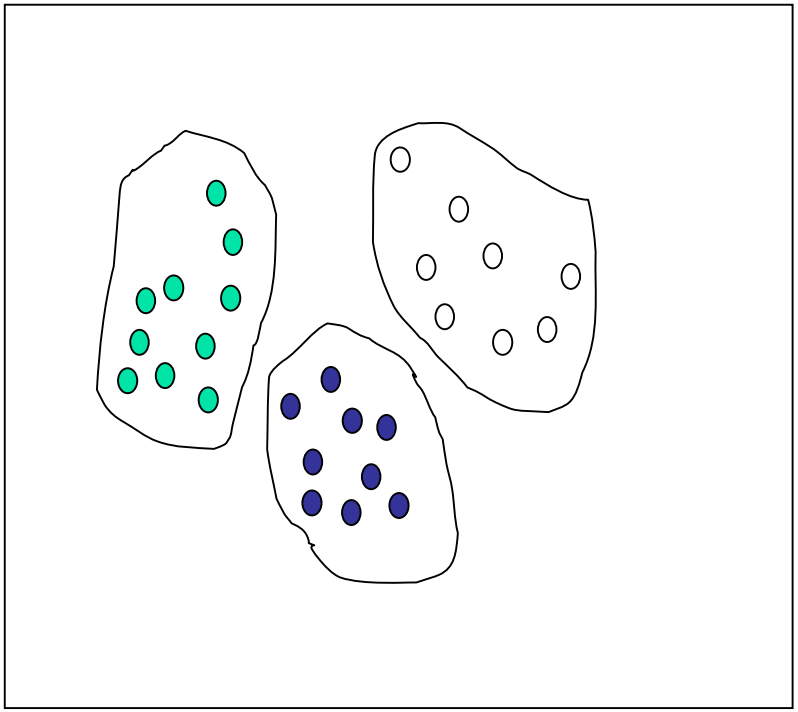
- 取樣:用比較小的原始資料的隨機樣本*s*來代表原始資料 *N*
- 讓探勘方法的計算的複雜度與資料維度個數成次線性成長
- 選取代表資料的子集合
  - 當資料包含偏斜時,簡單隨機取樣的效能會非常差
- 發展適合取樣方法
  - 分層抽樣:
    - 在所有的資料庫中對每個類別(有興趣的子族群)保留大約相同的百分比
    - 與偏斜資料一起使用
- 注意: 取樣或許不會降低資料庫 I/Os

取樣: 是否放回

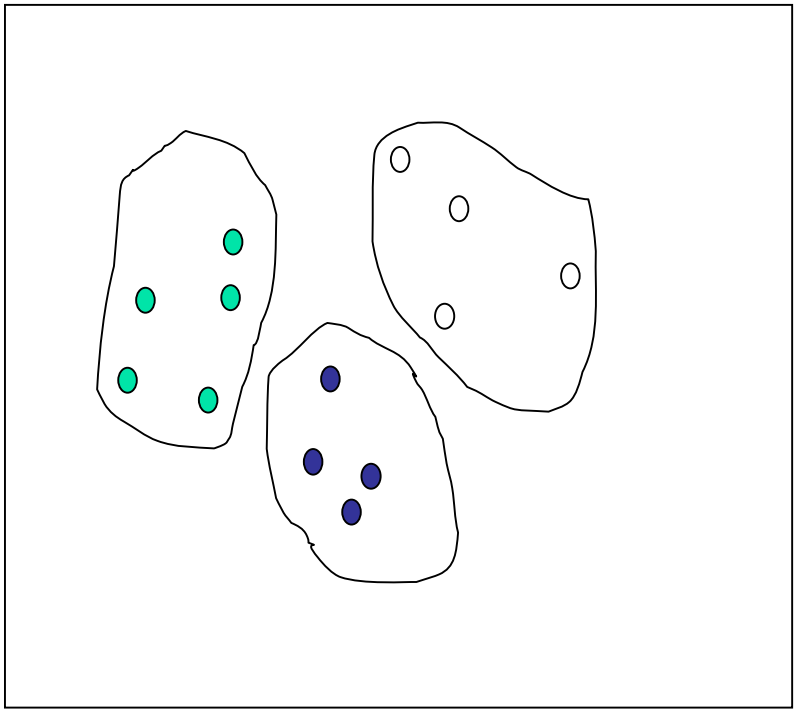


取樣: 分群或分層抽樣

原始資料



分群/分層抽樣



第二章: 資料前處理

- 為何要資料前處理?
- 敘述資料彙總
- 資料清除
- 資料整合與轉換
- 資料縮減
- 離散化與產生概念階層
- 總結

資料離散化與概念階層產生

- 離散化
  - 將一個連續值屬性的範圍分割成許多區間 (intervals) 來縮減屬性的值
  - 區間的標籤可以用來取代原始資料值
  - 監督式或無監督式
  - 分割 (由上而下) 或合併 (由下而上)
  - 離散化可以重複地套用在一個屬性
- 概念階層形成
  - 概念階層可以透過用較高層次的概念 (如青年、中年、老年) 取代較低層次的概念 (年齡屬性的值) 來進行資料縮減

數值資料的離散化與概念階層的產生

- 典型方法: 所有方法可重複套用
  - 箱狀法
    - 由上而下的分割,無監督式
  - 直方圖分析
    - 由上而下的分割,無監督式
  - 分群分析
    - 可以由上而下的分割或由下而上合併,無監督式
  - 熵式離散化:監督式,由上而下的分割
  - χ^2分析的區間合併:無監督式 ,由下而上合併
  - 直覺分割離散法:由上而下的分割,無監督式

熵式離散化

- 給定樣本 S, 如果 S 使用邊界 T 分割為兩個區間 S_1 與 S_2 , 分割後資訊增益為

$$I(S,T) = \frac{|S_1|}{|S|} Entropy(S_1) + \frac{|S_2|}{|S|} Entropy(S_2)$$

- 熵是根據集合中樣本類別分佈計算得之. 假設有 *m* 類別, S_1 的熵為

$$Entropy(S_1) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

p_i 為類別 *i* 在 S_1 中機率

- 在所有邊界值中,使用能最小化熵函數的邊界值當作二元離散化
- 重複套用分割直到停止條件滿足為止
- 這樣邊界值可以縮減資料大小並改善判別正確性

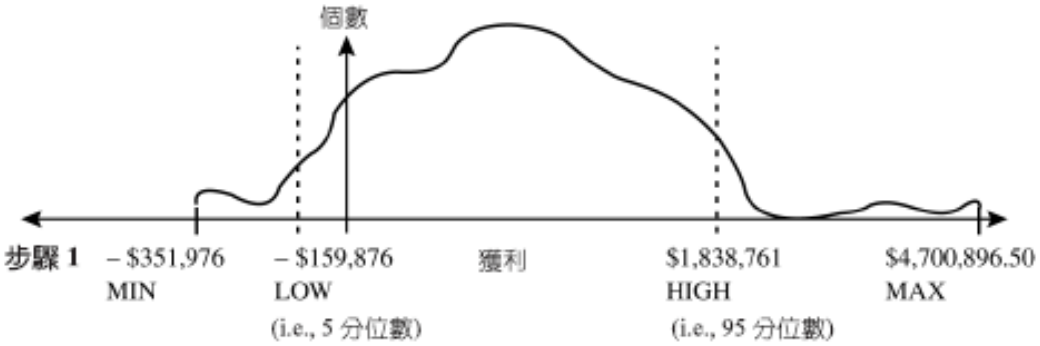
χ^2分析的區間合併

- 合併 (由下而上)
- 合併:重複地尋找最好的鄰居區間，然後將它們合併成較大的區間
- ChiMerge [Kerber AAAI 1992, See also Liu et al. DMKD 2002]
  - 一開始每個屬性的不同值A都視為一個區間
  - 對每個相鄰的值進行χ^2 檢測
  - 擁有最小值χ^2的相鄰值會進行合併
  - 合併的過程會重複進行一直到停止條件滿足為止(如顯著程度, 最大區間, 最大不一致)

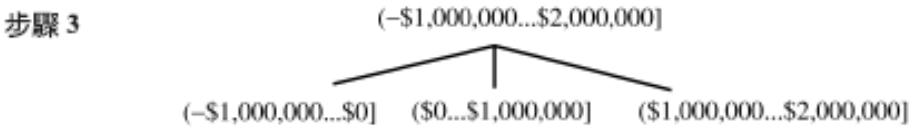
直覺分割離散法

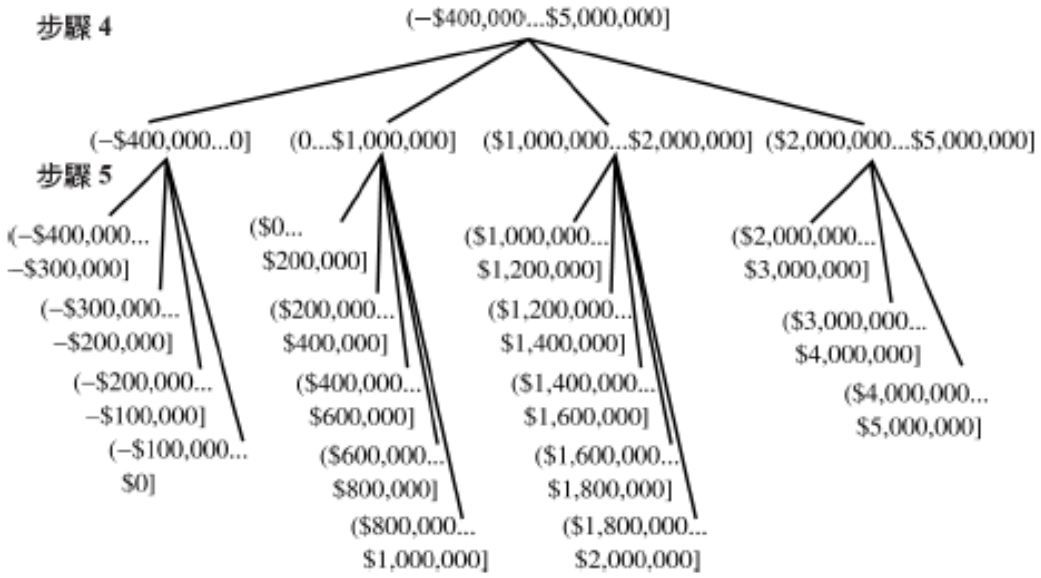
- 可用於將數值資料切割成相當均勻、相當自然的區間.
  - 區間包含3, 6, 7或9個不同值，則將範圍分割成3區間
  - 區間包含2, 4或8個不同值，則將範圍分割成4個等寬區間
  - 區間包含1, 5或10個不同值，則將範圍分割成5等寬區間

3-4-5 規則範例



步驟 2 msd = 1,000,000 LOW' = –$1,000,000 HIGH´ = $2,000,000



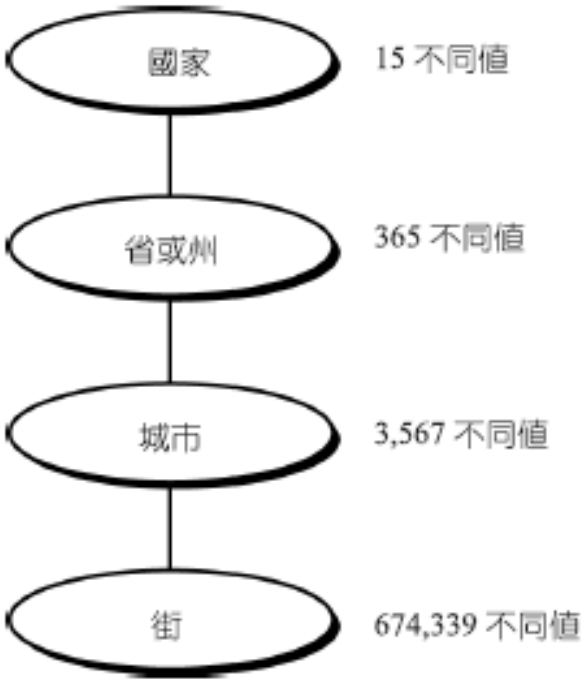


類別資料概念階層產生

- 由使用者或專家明確地在綱目層次中設定屬性部分順序
  - 街 < 城市 < 州 < 國家
- 透過明確的資料群組來設定一部分階層
  - {Urbana, Champaign, Chicago} < Illinois
- 僅設定部分屬性
  - 例., 僅有 街 < 城市, 其他沒有
- 概念階層可以根據屬性集中屬性包含的不同值來自動建立
  - 例., 一組屬性: {街 , 城市, 州, 國家}

自動產生概念階層

- 概念階層可以根據屬性集中屬性包含的不同值來自動建立
  - 擁有最多不同值的屬性會在階層的最下層
  - 例外, 例., 星期, 月, 季, 年



第二章: 資料前處理

- 為何要資料前處理?
- 敘述資料彙總
- 資料清除
- 資料整合與轉換
- 資料縮減
- 離散化與產生概念階層
- 總結

總結

- 資料前處理對資料倉儲與資料探勘是一項重要的議題
- 敘述資料彙總提供資料前處理分析的基礎
- 資料前處理包含
  - 資料清除與整合
  - 資料縮減與屬性選擇
  - 離散化
- 雖然已經發展許多資料前處理的方法，由於龐大數量不一致或不乾淨的資料與問題的複雜度，資料前處理仍然是一個活躍的研究領域