



R 語言資料探勘實務

逢甲資工 助理教授
許懷中



R 語言資料工程及探勘實務

2016/7/14-2016/7/15



逢甲資工
助理教授
許懷中 博士

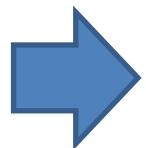
資料工程



DSP 智庫驅動
神一般的！奶爸
謝宗震 博士

資料探勘

我的背景



雲端計算、軟體工程

我的資料科學專案



線上遊戲玩家
黏著度分析



K-15 學生
快速程度評量



虛寶銷售預測與
成因分析



票房因素分析



企業、法人與政府
資料科學人才培訓

逢甲大學



李秉乾校長



高承恕董事長

蘋果即時



R 語言資料探勘實務

何為資料探勘？



2016/10/30



2016/10/30

9

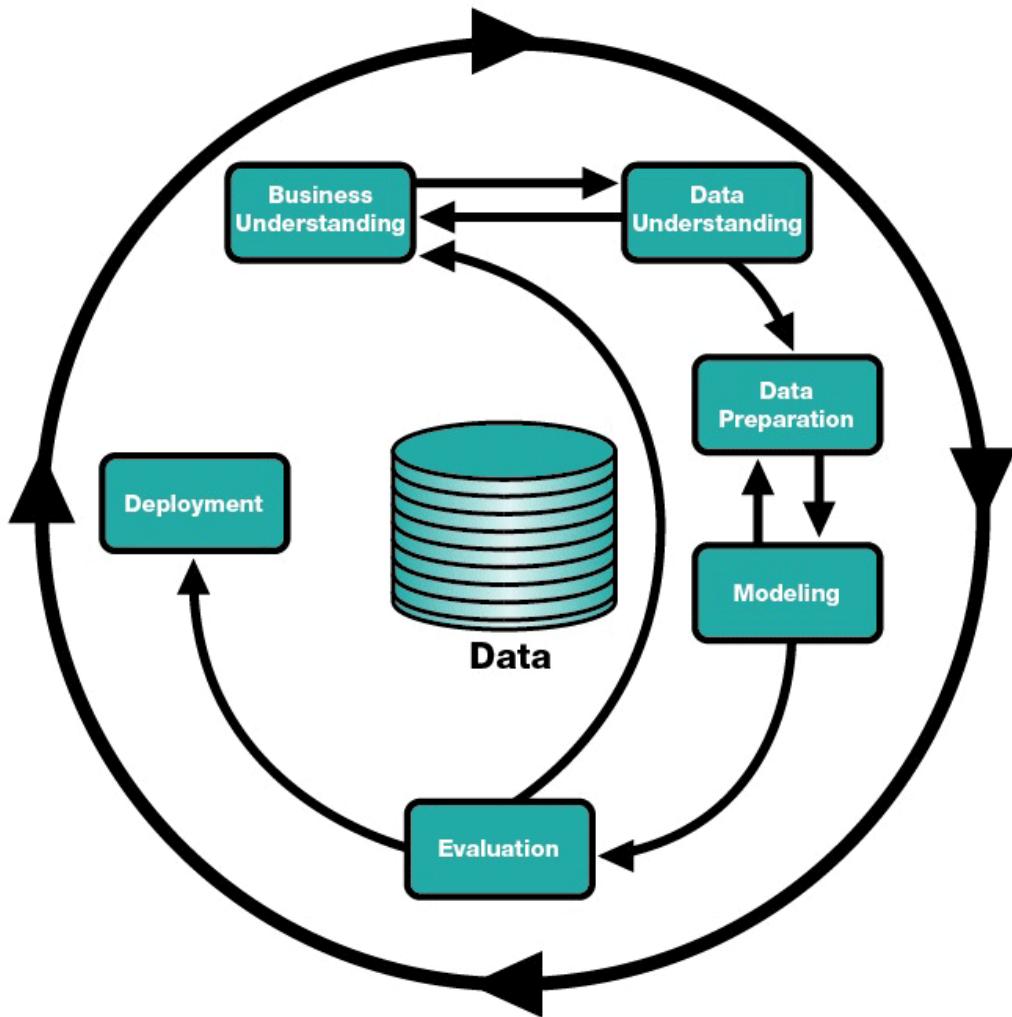


什麼是資料探勘？

- 從資料中萃取**有價值**的資訊
- 資料哪裡來？
 - 觀察而得 (observational data)
 - 刻意取得 (experimental data)
- 什麼資訊有價值？
 - 直接資訊 (應變數 y)、間接資訊 (自變數 x)
 - 趨勢 (Trend)、模式 (Pattern)、特徵 (Feature)
 - 關聯 (Relationship)

進行資料探勘的程序

- 商業理解
- 資料理解
- 資料預處理
- 建立模型
- 評價與解釋
- 部署



跨業別資料採礦標準程序 (CRISP-DM)

R 語言資料探勘實務

資料探勘基石

核心概念

- 機率 與 概似函數 (Probability & Likelihood)
- 距離 與 相似度 (Distance & Similarity)
- 成本 與 風險 (Cost & Risk)
- 準確度 與 精密度 (Accuracy & Precision)

機率與資料探勘

- 機率：事件 A 發生的機率記為 $P(A)$
 - 事件 A 發生的次數/全部事件的次數
- 條件機率：在事件 B 已發生的前提下，事件 A 發生的機率，記為 $P(A|B)$
 - 事件 A 與事件 B 皆發生的機率/事件 B 發生的機率
- 用被購買的機率排序熱門商品
 - $P(\text{奶綠}) > P(\text{檸檬汁}) > P(\text{冰淇淋})$
- 用條件機率篩選影響因子
 - $P(\text{會買冰淇淋}) < P(\text{會買冰淇淋}|\text{天氣熱})$
 - $P(\text{會買哈利波特}) < P(\text{會買哈利波特}|\text{會買魔戒})$

距離與資料探勘

- 距離：物件 A 與 B 間的距離，記為 $D(A, B)$
- 相似度：物件 A 與 B 之間的相似程度，記為 $C(A, B)$
 - 一般而言，相似度是一個介於 $[0, 1]$ 的值， 1 表示一模一樣， 0 表示毫無相似之處
- 用距離(相似度)作商品推薦
 - 推薦規格相仿的產品
 - 推薦訪客瀏覽行為相似的產品

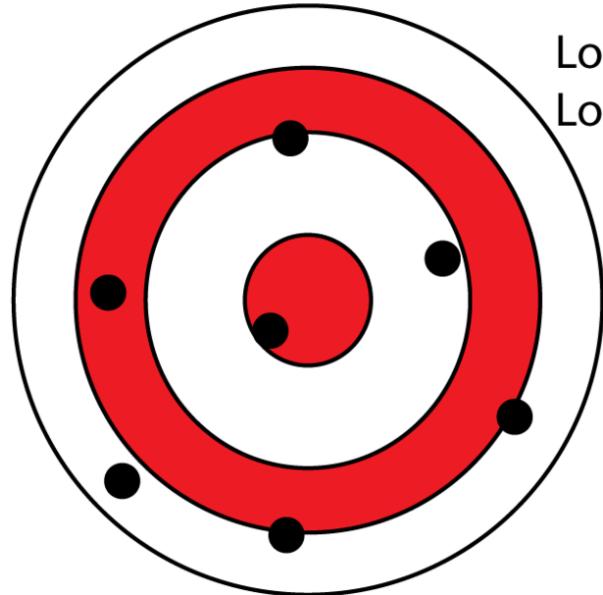


風險與資料探勘

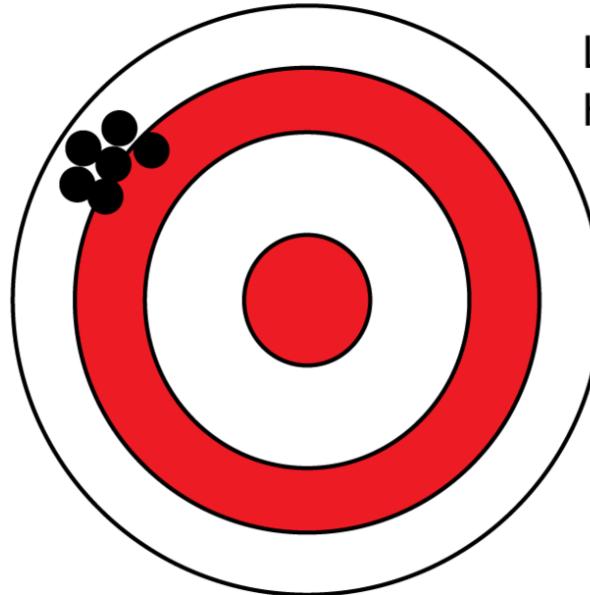
- 風險：事件發生時須付出的代價
 - 風險：不確定性(機率) \times 成本
- 建構企業信用風險模型
 - 違約機率模型 (PD) · e.g. 決策樹模型
 - 違約損失率模型 (LGD)
 - 風險 => 預期損失率 (EL) $EL = LGD \times PD$
 - $EL = LGD \times PD \times EAD$ (違約曝險額)

準確度與精密度

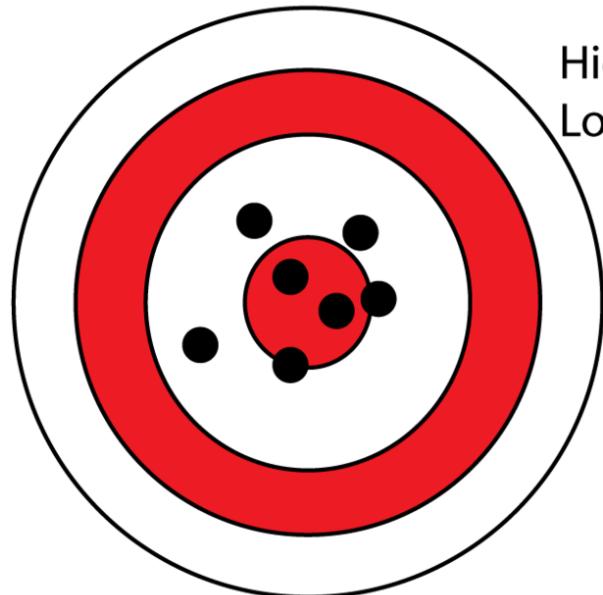
- 準確度 (Accuracy)
 - 測量平均值與真實值之間的差距
 - 又稱偏誤(Bias)
- 精密度 (Precision)
 - 獨立實驗數據分布的集中程度
 - 又稱偏差 (Deviation)
- 評量標準
 - 均方差： $MSE(T) = Var(T) + Bias(T)^2$



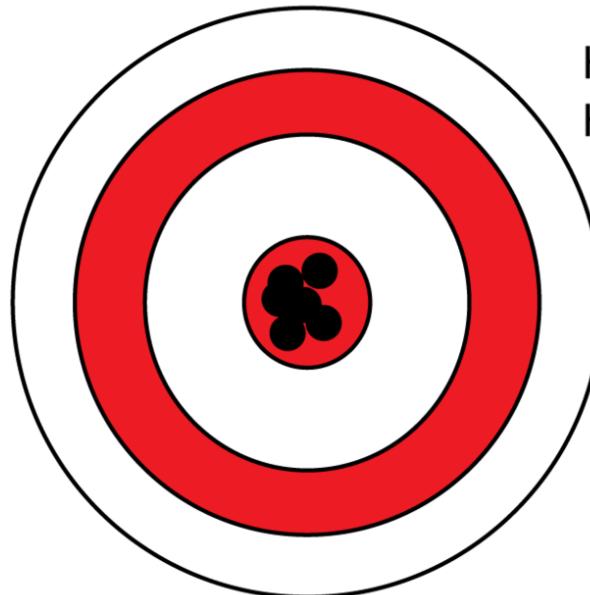
Low accuracy
Low precision



Low accuracy
High precision



High accuracy
Low precision



High accuracy
High precision

資料探勘應用舉例：行銷

- 經驗行銷
 - 依靠人的印象與直覺
 - 優秀店長的直覺有時比資料探勘還準
 - 但是優秀店長不多，一個人能接觸的客戶也有限
- 傳統行銷
 - 用客戶的背景資訊篩選與定義行銷對象
 - 年齡、性別、婚姻狀況、工作狀況、家庭結構
 - 向目標受眾推播商品資訊
 - 倚賴過去的經驗與對產品的想像
 - 相同背景的人就會有相同的需求？

資料探勘應用舉例: 行銷 (cont.)

- 資料探勘行銷
 - 計算人與人之間的距離 (User-based)
 - 計算商品與商品之間的距離 (Item-based)
 - 計算每個人購買種商品的機率
 - 計算每個人購買各種商品的最佳順序
 - 只要有資料的受眾就可以納入計算
 - 從資料出發，不受既定印象與偏見影響，因此可以發掘潛在客戶 (買LV包的大叔)
 - 成功要素：資料量、使用者涵蓋範圍、反應時間

因時制宜、因地制宜

- 小賣店靠優秀店長的經驗、印象與直覺
- 中型賣場靠傳統的行銷分析
- 大規模商業集團需要引進料探勘系統取得領先

R 語言資料探勘實務

關聯性分析





關聯性分析概念

- 分析不同品項之間的關係程度
- 常見問題
 - 如果消費者買了產品 A，那麼他有多大機率也會購買產品 B？
 - 如果消費者透買了產品 C 和 D，那麼他還會購買什麼產品？
 - 在店門口擺設什麼樣的廣告，可以促進產品 X 的促銷活動成效？

關聯性分析概念 (cont.)

- 品項 (Item)
 - 可購買的物品: 牛奶、麵包、啤酒、尿布等.....
- 交易記錄 (Transaction)
 - 一名顧客購買的物品清單: {啤酒, 尿布}
- 關聯規則 (Association Rule)
 - $X \Rightarrow Y$: 當X出現在一個交易記錄時 Y也會出現
 - 啤酒 \Rightarrow 尿布
 - X, Y 稱作品項集(Item Set) , 可以含有多個品項
 - X (LHS itemset) 和 Y (RHS itemset) 中不能包含相同的商品

人肉關聯性分析！

- 四筆交易記錄
 - {牛奶, 麵包}
 - {啤酒, 尿布}
 - {牛奶, 麵包, 奶油}
 - {麵包}
- 關聯規則？
 - 牛奶 \Rightarrow 麵包 (買牛奶的同時也會買麵包)

交易記錄 + 瀏覽記錄

You may also like...

What Other Items Do Customers Buy After Viewing This Item?



The Filter Bubble: What the Internet Is Hiding...

★★★★☆ (47)

\$16.53



The Filter Bubble: How Search Engines Decide What You See

\$9.91

> Explore similar items



The Shallows: What the Internet Is Doing to Our Brains

➤ Eli Pariser

★★★★☆ (47)

Hardcover

\$16.53



The Shallows: What the Internet Is Doing to Our Brains

➤ Nicholas Carr

★★★★☆ (146)

Paperback

\$9.45



The Power of Habit: Why We Do What We Do in Life and Work

➤ Charles Duhigg

★★★★☆ (201)

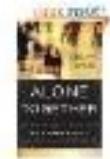
Hardcover

\$16.00



The Shallows: What the Internet Is Doing to Our Brains

by Nicholas Carr



Alone Together

by Sherry Turkle



Public Parts

by Jeff Jarvis

Continue Shopping: Customers Who Bought Items in Your Recent History Also Bought



The Shallows: What the Internet...

Nicholas Carr

★★★★☆ (146)

Kindle Edition

\$10.26

Fix this recommendation



Alone Together: Why We Expect...

Sherry Turkle

★★★★☆ (40)

Kindle Edition

\$11.55

Fix this recommendation



Public Parts

Jeff Jarvis

★★★★☆ (15)

Kindle Edition

\$14.99

Fix this recommendation



The Information Diet: A Case for...

Clay Johnson

★★★★☆ (32)

Kindle Edition

\$11.99

Fix this recommendation



In The Plex

Steven Levy

★★★★☆ (71)

Kindle Edition

\$16.99

Fix this recommendation

> See more recommendations

計算有用的關聯性規則

- 支持度 (Support)
 - 紿定品項集 X ，有多少比率的交易記錄 包含了 X ，寫作 $Supp(X)$
- 範例
 - $\{\text{牛奶}, \text{麵包}\}, \{\text{啤酒}, \text{尿布}\}, \{\text{牛奶}, \text{麵包}, \text{奶油}\}, \{\text{麵包}\}$
 - $Supp(\{\text{牛奶}\}) = 2/4$
 - $Supp(\{\text{牛奶}, \text{麵包}\}) = 2/4$

計算有用的關聯性規則 (cont.)

- 置信度

- 紿定關聯規則 $X \Rightarrow Y$

$$\text{置信度 } \text{Conf}(X \Rightarrow Y) = \frac{\text{Supp}(X \cup Y)}{\text{Supp}(X)}$$

- 也就是在交易記錄中含有X的前提下，該交易記錄也同時含有Y的條件機率 $P(Y|X)$

- 範例

- $\{\text{牛奶}, \text{麵包}\}, \{\text{啤酒}, \text{尿布}\}, \{\text{牛奶}, \text{麵包}, \text{奶油}\}, \{\text{麵包}\}$
 - $\text{Supp}(\{\text{牛奶}\} \Rightarrow \{\text{麵包}\}) = 2/2 = 1$
 - $\text{Supp}(\{\text{麵包}\} \Rightarrow \{\text{牛奶}\}) = 2/3$

計算有用的關聯性規則 (cont.)

- 交易記錄
 - {牛奶, 麵包}, {啤酒, 尿布}, {牛奶, 麵包, 奶油}, {麵包}
- 衡量關聯性規則
 - $X \Rightarrow Y, (\text{Conf}(X \Rightarrow Y), \text{Supp}(X \Rightarrow Y))$
 - {牛奶} \Rightarrow {麵包}, (2/2, 2/4)
 - {啤酒} \Rightarrow {尿布}, (1/1, 1/1)
 - {牛奶, 麵包} \Rightarrow {奶油}, (1/2, 1/4)

強關聯規則

- 規則的支持度 (Support) 越高，代表越有影響力
- 規則的置信度 (Confidence) 越高，代表越有可能是有效的規則
- 強關聯規則，通常支持度與置信度都很高
- 然而反過來卻未必成立！

支持度與置信度的不足

- 假設 100 筆交易記錄，60 筆包含牛奶，75 筆包含麵包，40 筆同時包含麵包與牛奶
 - $\{\text{牛奶}\} \Rightarrow \{\text{麵包}\}$: ($\text{Supp} = 40\%$, $\text{Conf} = 67\%$)
- 看起來很好？有 $2/3$ 的消費者，買了牛奶就會一起買麵包
- 但是，其實麵包本身的**支持度**有 75%，比這條規則的**置信度** 67% 還高
- $P(\text{麵包}|\text{牛奶}) < P(\text{麵包})$ ，也就是說購買牛奶，反而會降低購買麵包的機會

增益值 (Lift)

- 為了彌補支持度與置信度的不足，額外檢查增益值(Lift)
 - $Lift(X \Rightarrow Y) = \frac{Conf(X \cup Y)}{Supp(Y)} = \frac{P(X, Y)}{P(X)P(Y)}$
- Lift 其實就是 (X, Y) 的相關性指標
 - **增益值 > 1** : 表示 X 的銷售與 Y 的銷售呈現正相關，規則有用！
 - **增益值 $= 1$** : 表示 X 的銷售與 Y 的銷售無關，該規則接近亂數取得。
 - **增益值 < 1** : 表示 X 的銷售與 Y 的銷售呈現負相關，此規則不僅無用，反而可能有害。

關聯規則建立流程

- 找出大於設定之**最低支持度**(minimum support)的**高頻項目集合** (large itemset)
- 利用前述高頻項目集合，產生**關聯規則**，計算各規則置信度，若高於所設定之**最低置信度** (minimum confidence)，則為候選規則
- 計算各候選規則之**增益值**，確認其確實為有用之關聯性規則

實際操作關聯性規則

- 請各位完成 **RDM-01-Association-Rule**

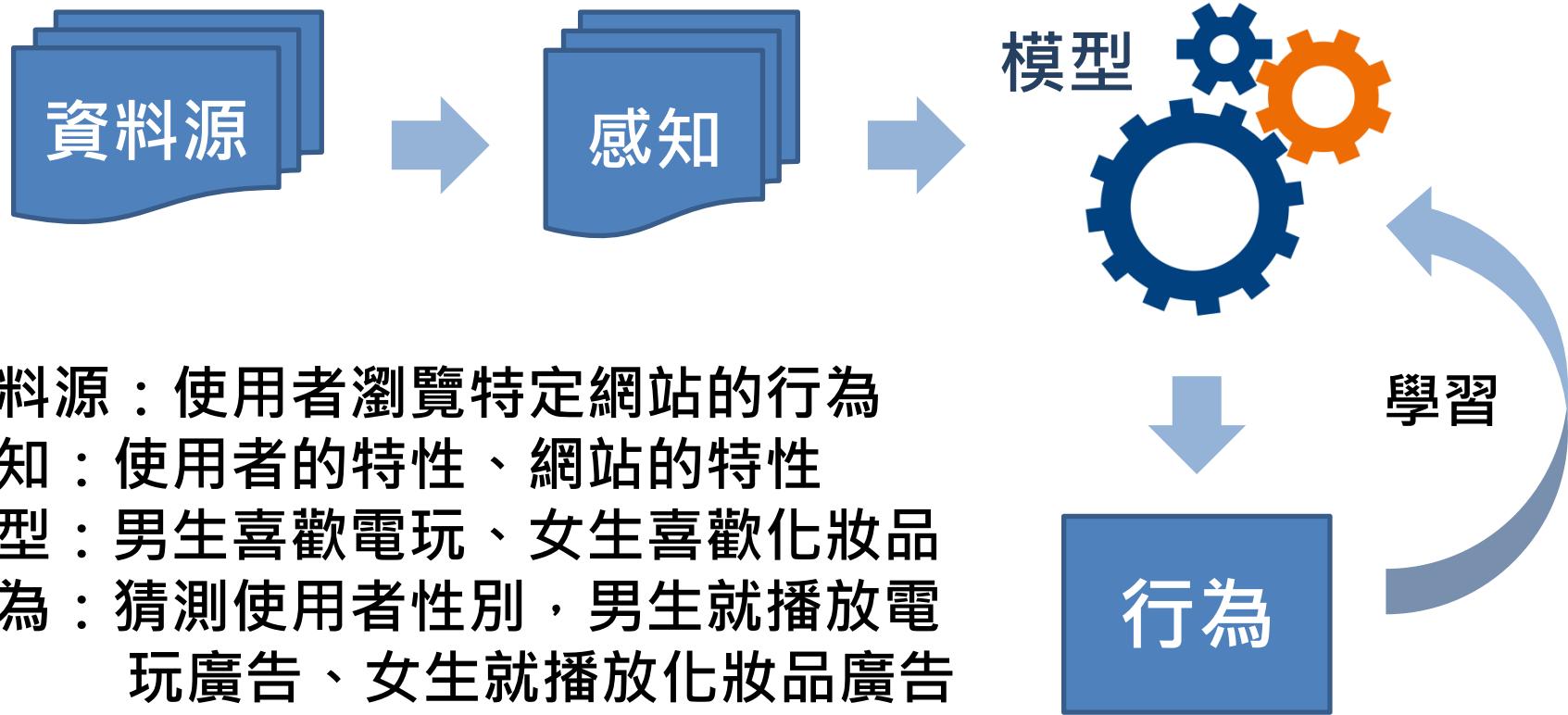
R 語言資料探勘實務

利用機器學習進行資料探勘

何謂機器學習？

- 可以從資料中學習既存之趨勢、模式與特徵，並進而做出預測的演算法
- 我們今天只探討怎樣利用機器學習進行資料探勘
- 如果想要深究機器學習原理，推薦從林軒田老師的**機器學習基石**入手
- 或者投書陳昇瑋會長，請他說服林軒田老師再開一場「**一天搞懂機器學習**」課程

機器學習應用範例



利用機器學習進行資料探勘？！

- 機器學習能解決的問題有限
 - 預測數值
 - 分類
- 商業問題需要進行轉換
 - 想要增加廣告營收
 - 廣告營收與點擊率有關
 - 推播廣告前，預測使用者點擊各種廣告的機率
 - 推播點擊機率高的廣告
- **預測廣告點擊率 → 增加廣告營收**

資料探勘與機器學習的關聯

- 藉由資料探勘進行特徵工程 (Feature Engineering)，讓機器學習可以預測得更準
- 進行解釋型分析，藉由機器所學到的趨勢、模式與特徵瞭解商業問題的核心
- 藉由機器學習研究人所無法消化與理解的大量資料以及複雜模型
- 藉由機器的自我學習與演變，適應環境的改變

機器學習的種類

- 常用名詞
 - Y: 目標變數、應變數、Response Variable
 - X: 解釋變數、自變數、獨立變數、Covariates
- 監督式學習：給定 X，預測 Y
 - 分類：Y是類別型變數
 - 迴歸：Y是連續型變數
- 非監督式學習：給定 X，對 X 分群

分類問題

- 二元分類
 - 點擊預測： $Y = \text{使用者點/不點廣告}$
 - 股票漲跌： $Y = \text{下一個單位時間，目標股票股價上漲/下跌}$
- 多元分類
 - 手寫辨識： $Y = \text{使用者所撰寫的字}$
 - 語音辨識： $Y = \text{使用者所說的話}$

迴歸問題

- 點擊率預測： $Y = \text{使用者點擊廣告的機率}$
- 股價預測： $Y = \text{下一個單位時間，特定股票的股價}$
- 銷售預測： $Y = \text{特定商品，在下一個單位時間可以賣出多少件}$

叢集問題

- 顧客分群：

$X = \text{顧客的特徵}$

目標是將特徵相近的顧客分群

- 廣告分群

$X = \text{廣告的特徵}$

目標是將特徵相近的廣告分群

應用機器學習於資料探勘

- 預測得準的模型未必有用
 - 太複雜的模型無法解釋
- 解釋型分析
 - 挖掘現象的因果
 - 預測公司的營收，預測到很準有沒有差？
 - 預測營收為 9-10 億 vs. 9.9 億 – 10.1 億
 - 找出影響營收的因素才真正有價值

解釋型分析 (cont.)

- 分析師透過模型尋找價值
- 例如，利用機器學習進行資安分析
 - 企業機器是否遭受攻擊
 - 企業機器的特徵
 - 利用模型，從特徵預測是否遭受攻擊
 - 分析模型，尋找其中的重要特徵，藉由該重要特徵，推敲系統弱點
- 模型的合理性重於準確度
 - **但不是說準確度不重要**

R 語言資料探勘實務

利用監督式學習幫助資料探勘

統計模型

- Y 我們感興趣的變數， X 是可能與 Y 有關的資料
- $Y = f(X) + \varepsilon$
 - $f(X)$ 描述我們能理解的變化，也就是可能可以用 X 解釋的變化
 - ε 描述我們無法理解的變化，也就是無法用 X 解釋的變化，通常是一個隨機變數

實際範例：cars 資料集

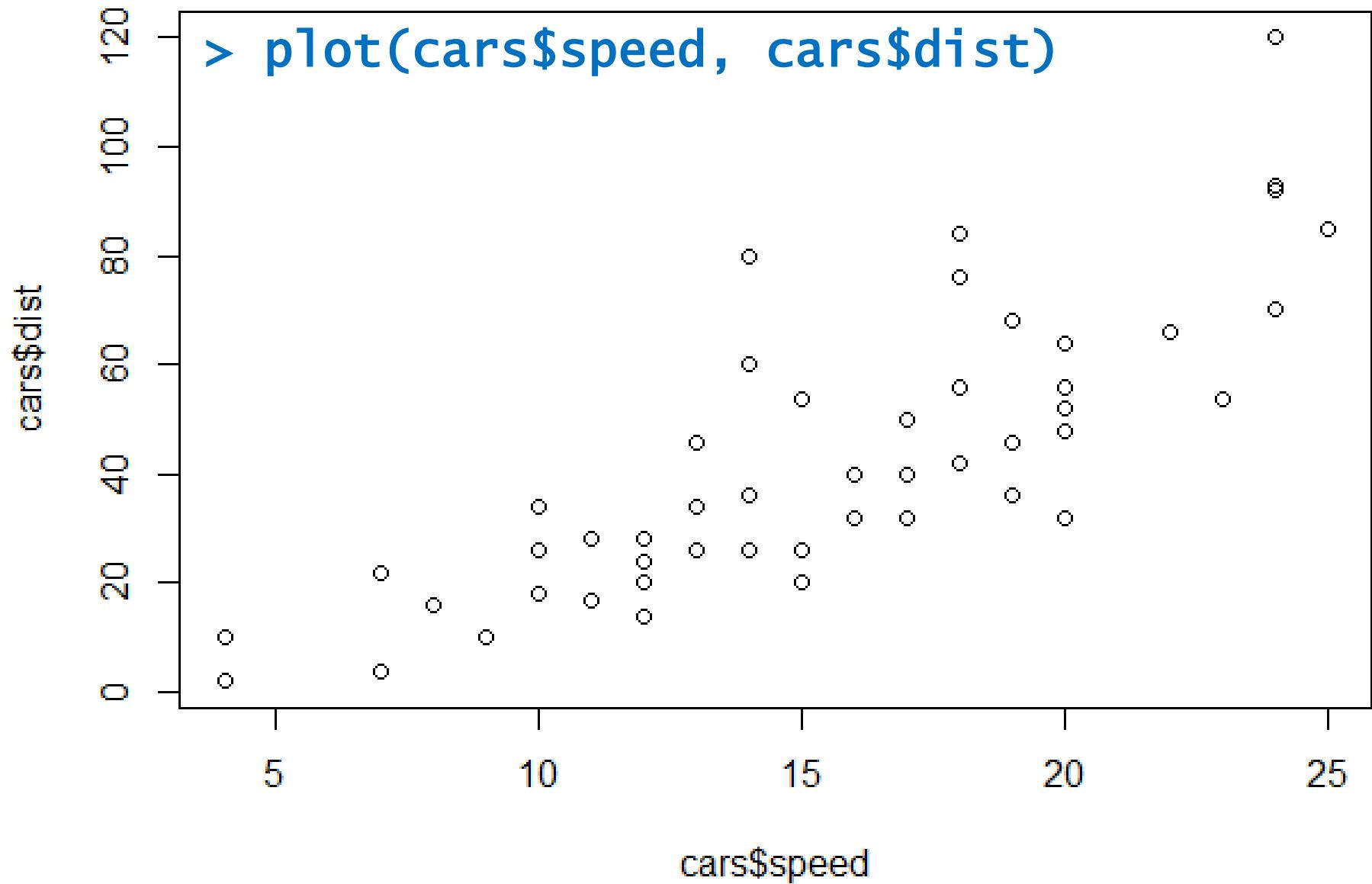
- > `data(cars)`
- > `view(cars)`
- 會看到 50 筆煞車距離 (dist) 與車速 (speed) 的資料

線性模型 (Linear Model)

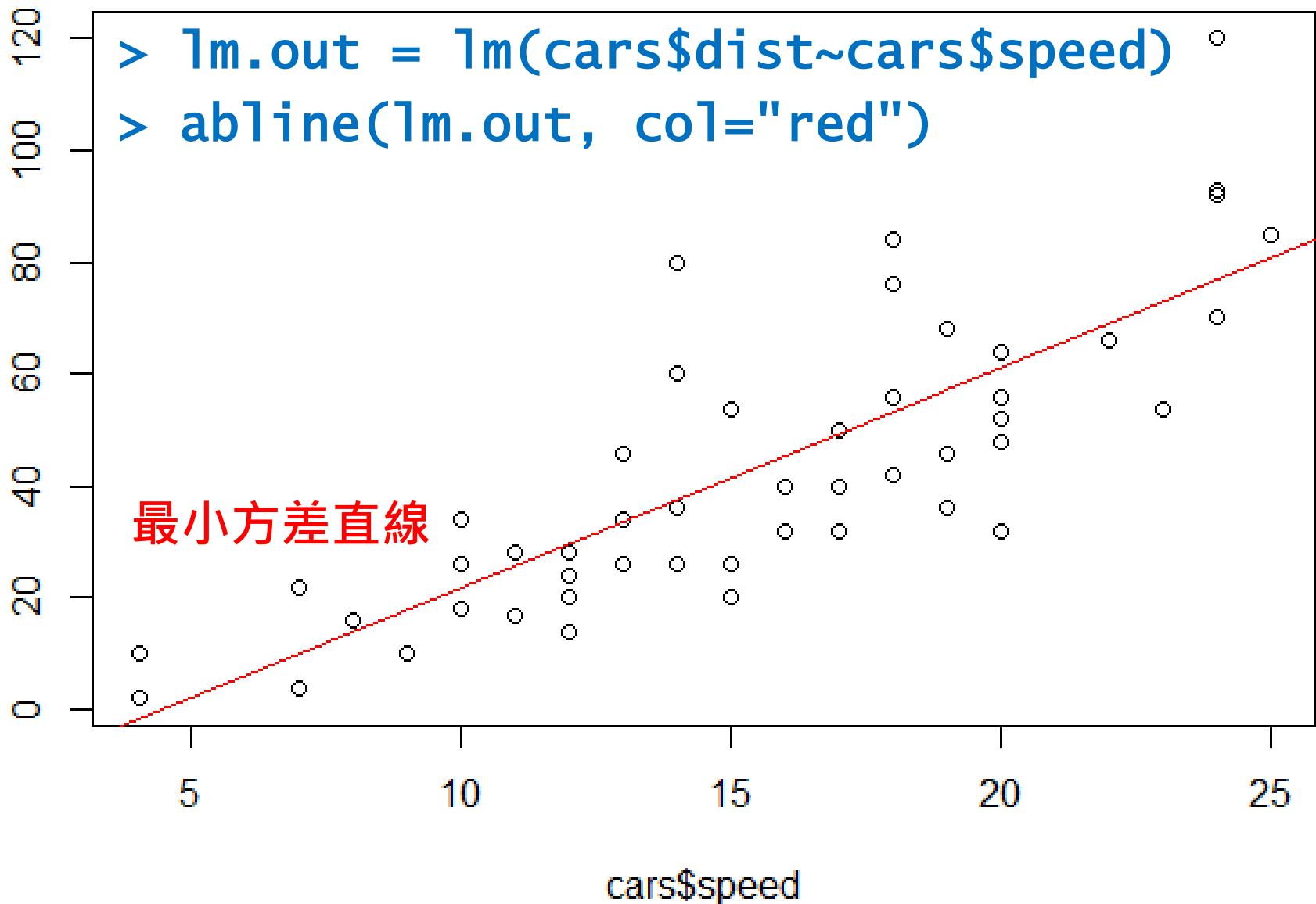
- 令 $Y = \text{cars\$dist}$, $X = \text{cars\$speed}$
- $Y = f(X) + \varepsilon$
- $f(X) = \beta_0 + \beta_1 X$

線性迴歸

```
> plot(cars$speed, cars$dist)
```



線性迴歸 (cont.)



線性迴歸 (cont.)

```
> lm.out = lm(cars$dist~cars$speed)
```

```
> print(lm.out)
```

```
Call: lm(formula = cars$dist ~ cars$speed)
```

```
Coefficients: (Intercept) cars$speed
```

-17.579	3.932
---------	-------

線性迴歸 (cont.)

- $f(speed) = -17.5791 + 3.9324 * speed$
- $dist = f(speed) + \varepsilon$
- $f(speed)$ 是一條與 dist 形成最小方差的直線
($\sum (dist_i - f(speed_i))^2$ 殘差平方和最小)
- ε 是一個常態分布 (normal distribution)

線性迴歸 (cont.)

```
> summary(lm(cars$dist~cars$speed))
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(intercept)	-17.5791	6.7584	-2.601	0.0123*
Cars\$speed	3.9324	0.4155	9.464	1.48e-12***

該變數對於模型具有顯著的影響

Multiple R-squared: **0.6511**, Adjusted R-squared: **0.6438**

F-statistic: 89.57 on 1 and 48 DF, p-value: 1.49e-12

模型對於應變數具有預測力，但並不十分準確

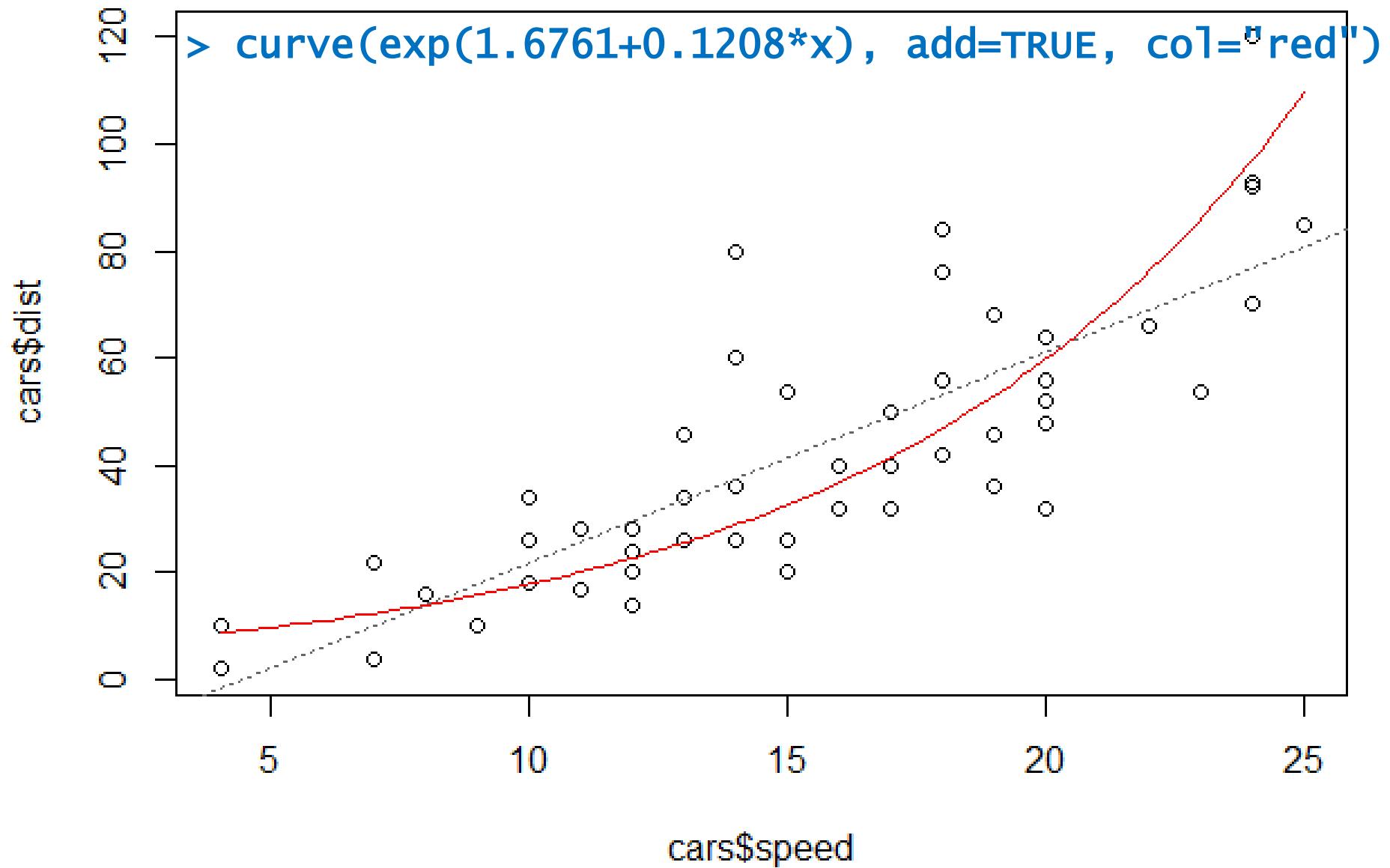
線性模型，不僅僅是直線

- $speed = 0$ 時，煞車距離 **-17.579** 與現實不符
- 既然 $dist$ 都是正的，我們可以取 \log 處理

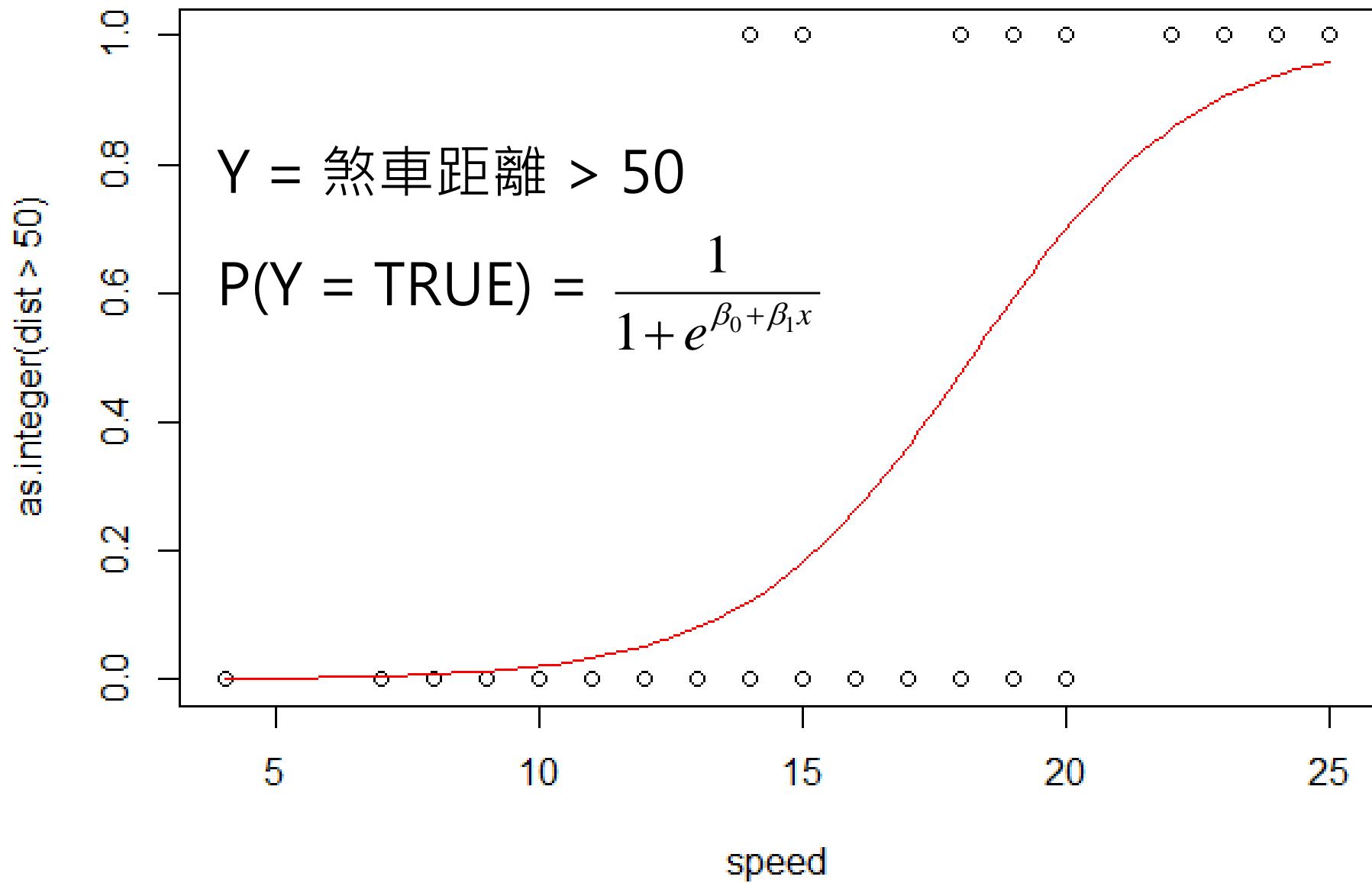
Log Linear

```
> lm.log = lm(log(cars$dist)~cars$speed)  
  
> print(lm.log)  
Call: lm(formula = cars$dist ~ cars$speed)  
Coefficients: (Intercept) cars$speed  
                1.6761      0.1208
```

Log Linear



Logistic Regression



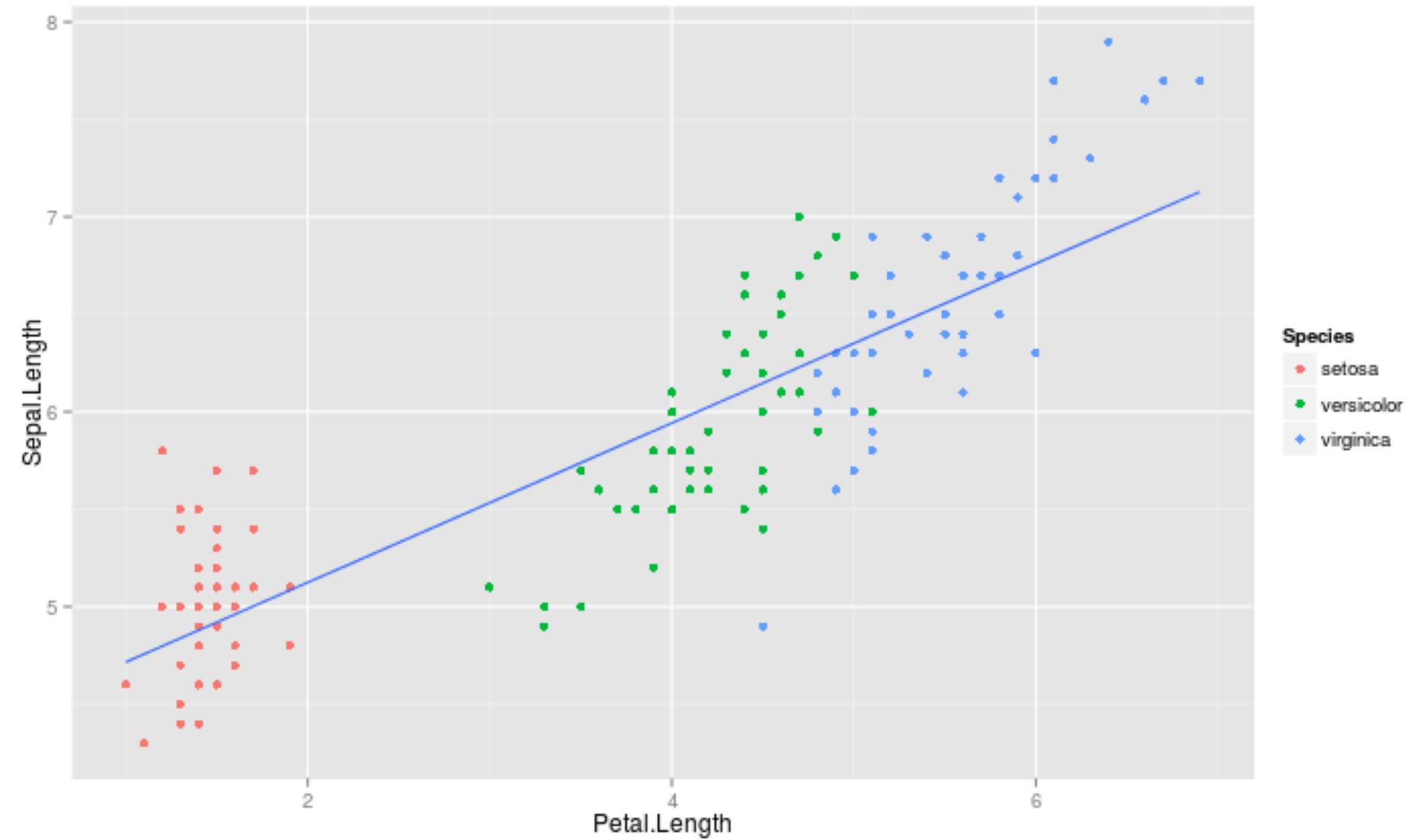
Logistic Regression

- $\beta_0 + \beta_1 X$: 線性
- $\frac{1}{1+e^{\beta_0+\beta_1 x}}$ 將實數線從 $(-\infty, +\infty)$ 轉換至 $(0, 1)$
- ε 則由銅板機率 (Bernoulli 分布) 取代
- 常用於疾病分析、廣告點擊率分析等

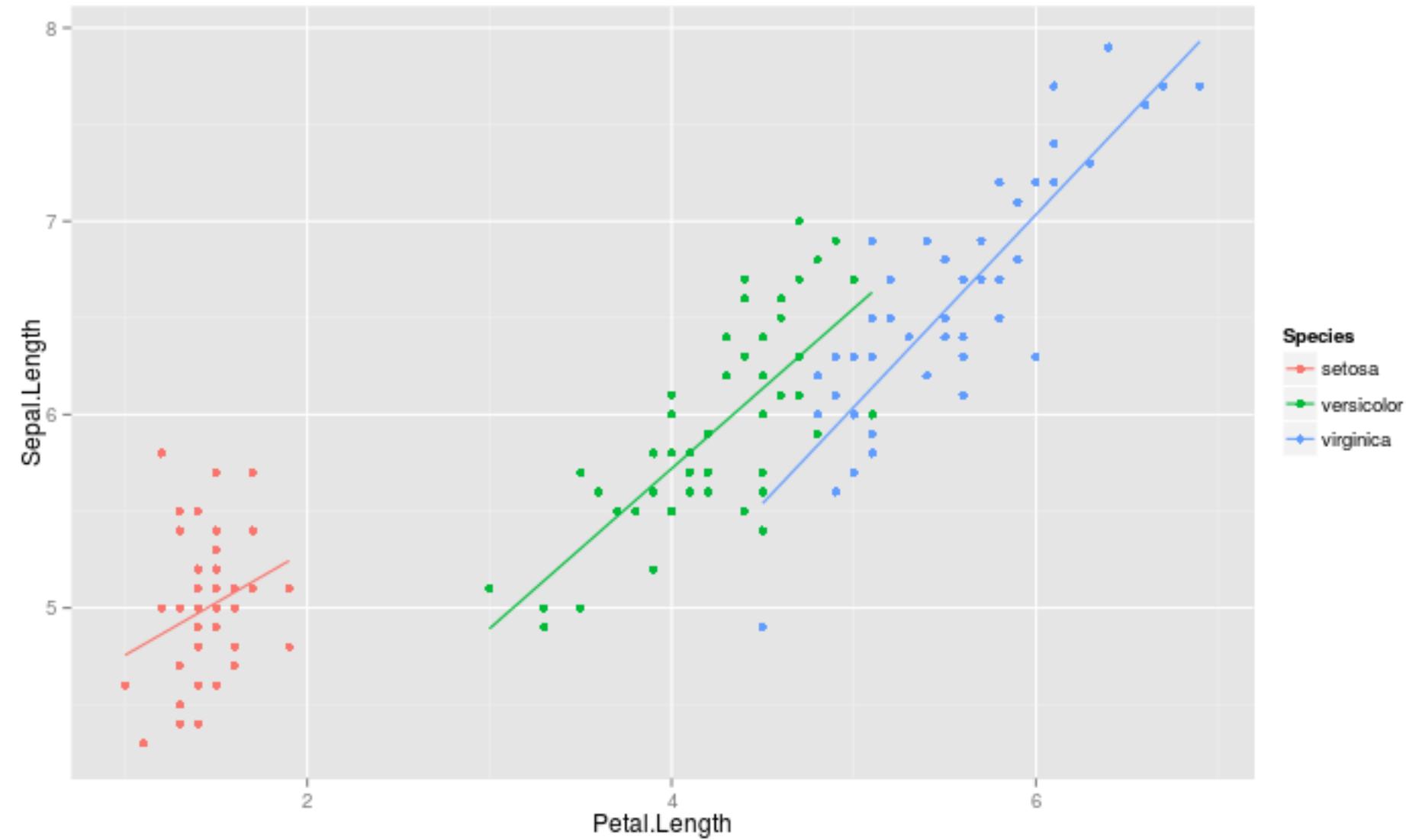
線性模型

- 線性模型還有很多種
 - Poisson Regression: 當資料為非負整數
 - Isotonic Regression: 當資料嚴格遞增
- X 的挑選對於線性模型好壞影響很大
- X 間的交互作用，或者形式轉換(例如：數值轉類別)是常用的技巧
- 經驗與領域知識十分重要，需要靠不斷的嘗試培養

不考慮交互作用



考慮交互作用



考慮交互作用的應用

- 推薦點擊率最高的廣告
- 廣告點擊率 = 廣告 + 網站
 - 所有網站上的廣告其播出邏輯都相同
- 廣告點擊率 = 廣告 + 網站 + 廣告*網站
 - 同時考慮廣告、網站分別的屬性，以及廣告與網站的交互作用
 - 考慮網站本身對於廣告影響力的影響
 - 不同網站會有不同的廣告播出邏輯

多種類別的分類

- 有 K 種類別
- 分別為類別 $1, 2, 3, \dots, K-1$ 各建立一個 logistic regression model
- 類別 $k \neq K$ 的機率為
$$\frac{e_K^{\beta x}}{1 + \sum_{i=1}^{K-1} e_i^{\beta x}}$$
- 類別 $k = K$ 的機率為
$$\frac{1}{1 + \sum_{i=1}^{K-1} e_i^{\beta x}}$$

實地操作學習線性模型

- 請各位完成
- RDM-o2-Supervised-Learning-o1-Linear-Model
- RDM-o3-Supervised-Learning-o2-Generalized-Linear-Model

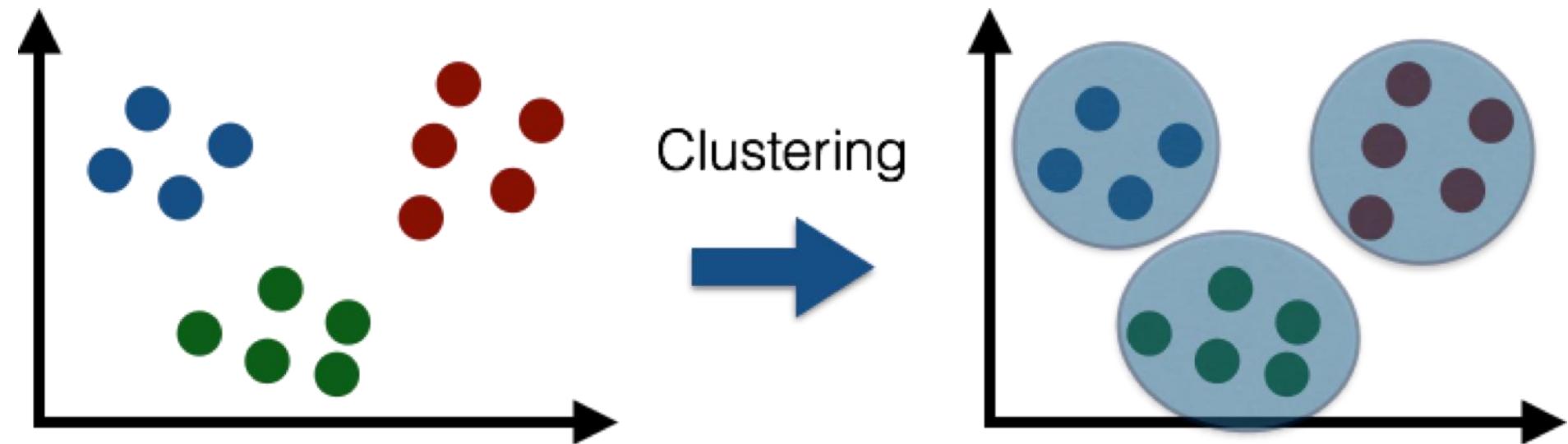
R 語言資料探勘實務

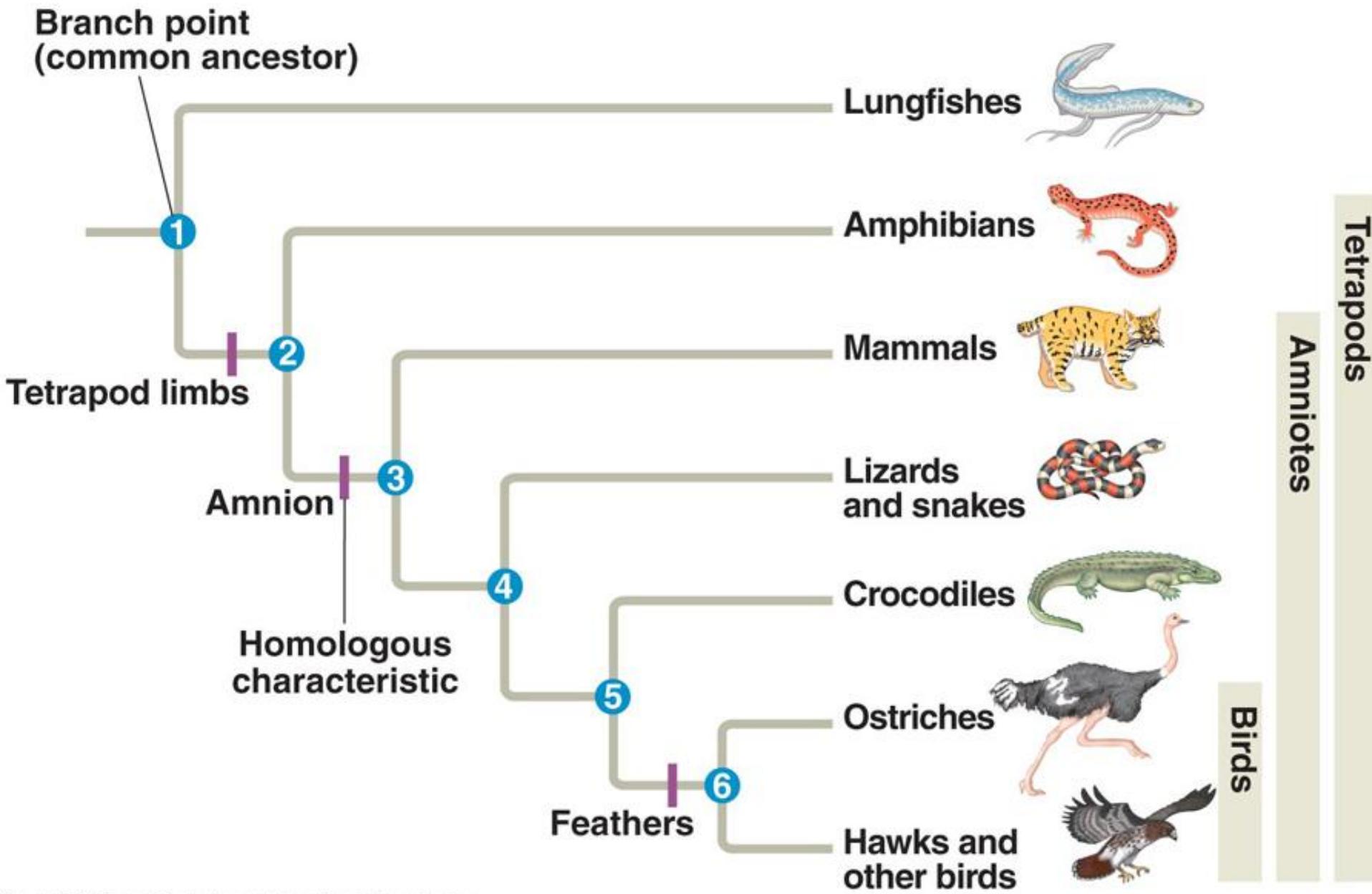
利用非監督式學習幫助資料探勘

叢集分析 (Clustering Analysis)

- 物以類聚，將類似的物件聚集在一起
- 常見應用
 - 電商：歸納消費行為類似的消費者
 - FinTech: 歸納客戶特徵類似的分行
- 有時會以群集之中心，代表各個群集
- **相似度、距離！**

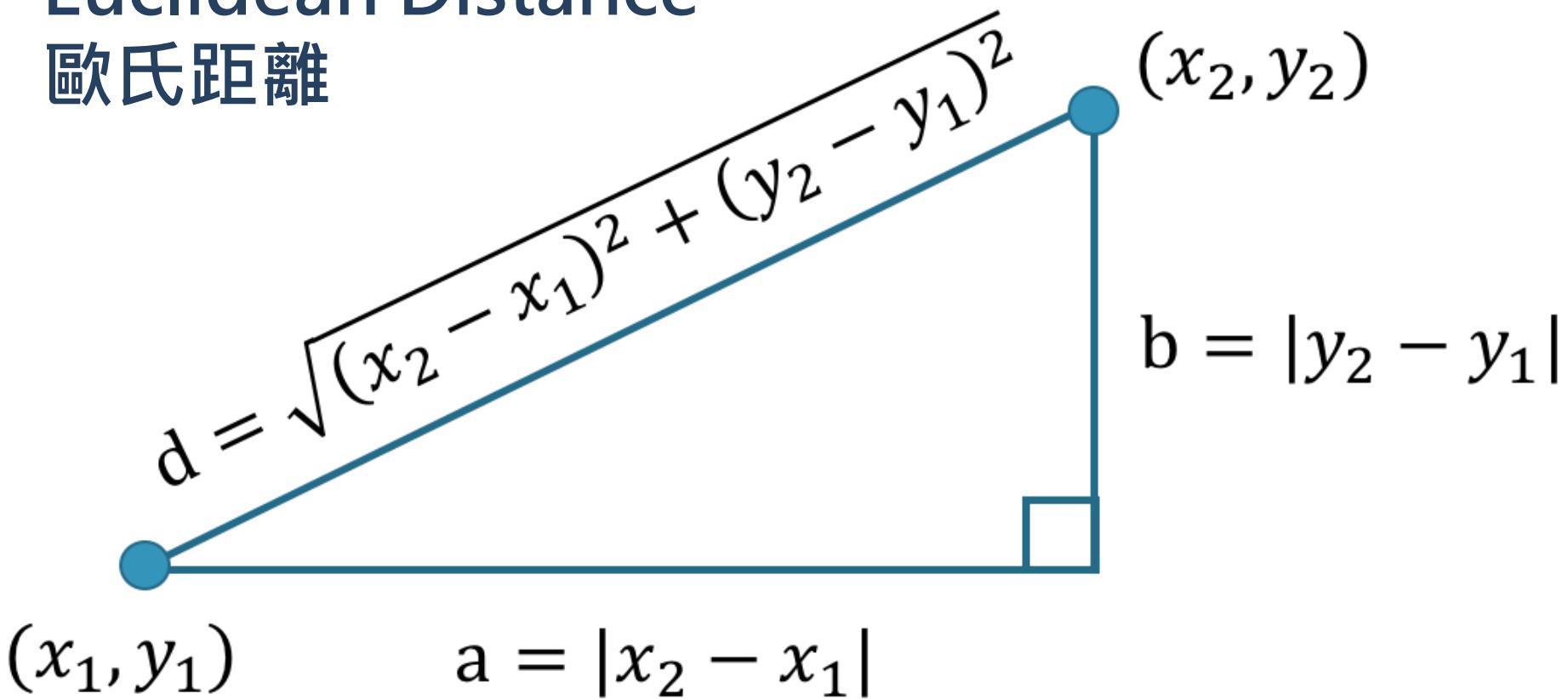
人肉分群





定義距離

Euclidean Distance
歐氏距離

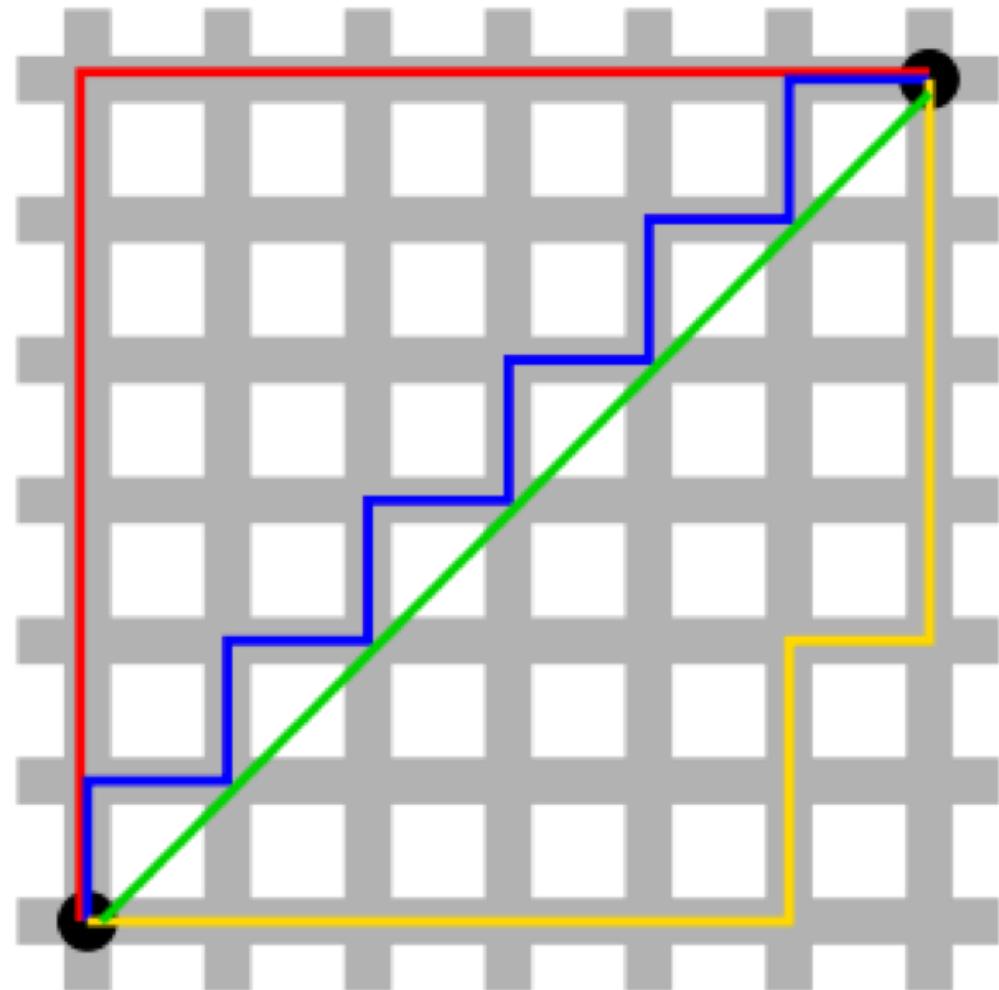


定義距離 (cont.)

Manhattan Distance

曼哈頓距離

- 投影到座標軸的長度和
- 適合度量網格間距



$$d = |x_2 - x_1| + |y_2 - y_1|$$

定義距離 (cont.)

用飛行時間當作距離
與歐氏距離並不成比例



定義距離 (cont.)

工具邦 / 交通地圖 / 台北捷運路線圖



用價錢計算距離

心的距離

MISS
YOU

E-MAIL
ME

定義相似度

- 相似度 (Similarity)：越高越像
 - 相似度一般介於 $[0, 1]$
- 相異度 (Dissimilarity)：越低越像
 - 相異度 = $1 - \text{相似度}$
- 相似度/相異度可以視為距離指標的特例
 - 經過正規化的距離
 - 距離僅為大於零的數值 $[0, \infty)$

二元類別型變數

公司是否參與標案

	標案1	標案2	標案3	標案4	標案5
公司A	1	0	0	0	0
公司B	1	1	0	1	0

$M_{11} = 1$, 公司A與B均參與的標案數

$M_{00} = 2$, A與B均未參與的標案數

$M_{10} = 0$, A參與、B未參與的標案數

$M_{01} = 2$, A未參與、B參與的標案數

- Jaccard Index:

$$\frac{M_{11}}{M_{10} + M_{01} + M_{11}}$$

- Jaccard Index 沒有考慮 M_{00} ?

第八屆立法委員投票行為分析

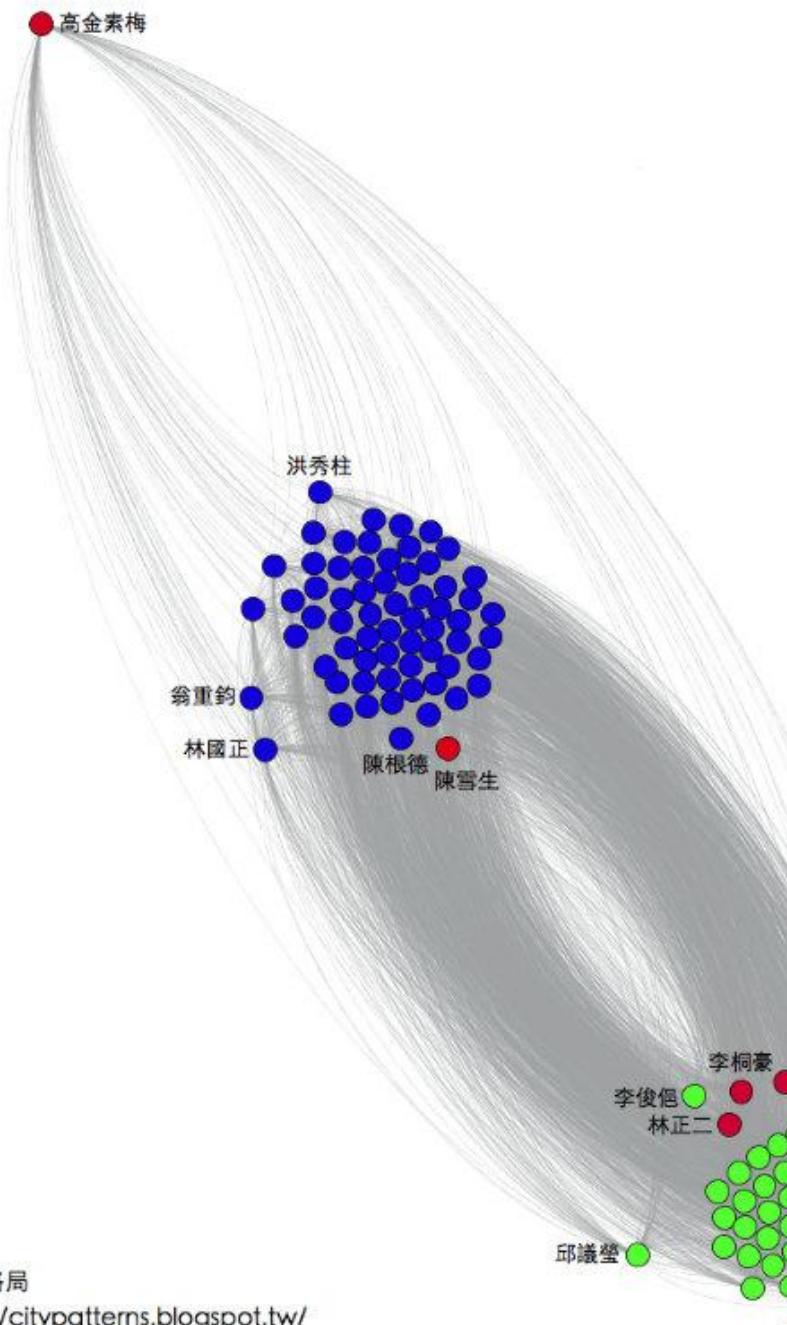
時間：2012.03.19 - 2012.11.28

法案數：47

立委人數：113

只投過1次票：高金素梅

完全沒投票：王金平（院長）、顏清標



資料來源：城市格局

<http://citypatterns.blogspot.tw/>

表決型態	倒數	00:10
員等43人提案	記名	計時
反對:紅色	棄權:黃色	
002 楊文玲	003 陳津鈞	004 張大千
006 王廷升	007 林德福	008 林海光
010 蔡正元	011 陳雪生	012 高金素梅
014 黃昭暉	015 楊瓊瓈	016 翁重鈞
018 劉秉達	019 黃偉哲	020 何欣純
022 朱學勤	023 林建鋐	024 高志鵬
026 陳其南	027	028 蘭波華
030 陳其南	031 直轄狀	032 李國樺
034 陳其南	035 江啟臣	036 林添駿
038 陳其南	039 陳其南	040 魏明谷
042 陳其南	043 陳孟安	044 邱志偉
046 江蕙貞	047 徐欣瑩	048
050 顏培陸	051 林國正	052 蘇清泉
054 陳其南	055 徐少萍	056 李貴敏
058 062 陳其南	063 陳子弘	064 陳其南

01-27-103 11:35:41

表決型態	倒數	00:00
員等提案	記名	計時
反對:紅色	棄權:黃色	
002 楊文玲	003 陳津鈞	004 張大千
006 王廷升	007 林德福	008 林海光
010 蔡正元	011 陳雪生	012 高金素梅
014 黃昭暉	015 楊瓊瓈	016 翁重鈞
018 劉秉達	019 黃偉哲	020 何欣純
022 朱學勤	023 林建鋐	024 高志鵬
026 陳其南	027	028 蘭波華
030 陳其南	031 直轄狀	032 李國樺
034 陳其南	035 江啟臣	036 林添駿
038 陳其南	039 陳其南	040 魏明谷
042 陳其南	043 陳孟安	044 邱志偉
046 江蕙貞	047 徐欣瑩	048
050 顏培陸	051 林國正	052 蘇清泉
054 陳其南	055 徐少萍	056 李貴敏
058 062 陳其南	063 陳子弘	064 陳其南

01-27-103 11:37:53

多元類別型變數

- Simple Matching Coefficient (SMC)
- SMC 範例
 - 使用者 A: 中年、公務員、大專畢業、未婚
 - 使用者 B: 高齡、自由業、大專畢業、已婚
- SMC: $\frac{\text{相同的屬性個數}}{\text{常數}} \propto \text{相同的屬性個數}$

多元標籤 轉換為 Jaccard Index

- 使用者A: {男性、單身、電玩、上班族}
- 使用者B: {男性、旅遊、電玩、學生}
- Jaccard Index:
$$\frac{|A \cap B|}{|A \cup B|}$$
- $J(\text{使用者A}, \text{使用者B}) = 2/6$
- $v := \{i_{\text{男性}}, i_{\text{單身}}, i_{\text{電玩}}, i_{\text{上班族}}, i_{\text{學生}}, i_{\text{旅遊}}, i_{\text{女性}}, \dots\}$
- $v_a = \{1, 1, 1, 1, 0, 0, 0, \dots\}$
- $v_b = \{1, 0, 1, 0, 1, 1, 0, \dots\}$

::: 首頁 > 標案查詢

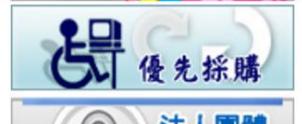
決標公告

[顯示簡要資料](#)[友善列印](#)[文字列印](#)

決標公告

公告日: 104/02/17

機關資料	機關代碼	A.7.31.1
	機關名稱	臺灣銀行股份有限公司
	單位名稱	臺灣銀行股份有限公司採購部
	機關地址	100臺北市中正區武昌街一段45號
	聯絡人	開標一科
	聯絡電話	(02)23497661
	傳真號碼	(02)23822010
	標案案號	GF4-103122
	招標方式	公開招標
	決標方式	最高標 本案屬收入性質採購



爬蟲取得董監事名單

ID	NAME	PARENT	BIRTHDAY	MAGNATE
00000000	復華廣告有限公司	NA	1976-05-24	
00000016	富台機械開發建設有限公司	NA	1979-04-30	王振林
00000022	泰煜建材股份有限公司	NA	NA	
00000037	茂盛工程有限公司（同名）	NA	1978-07-08	
00000043	啟猛股份有限公司（無統編）	NA	1984-05-22	鄭添發
00000058	詠詳鐵工廠股份有限公司（無統編）	NA	1984-03-07	吳秋進,吳戴麗珍,謝素梅,吳秋龍

利用 Jaccard Index 計算董監事相似度

	ID	NAME	PARENT	BIRTHDAY	MAGNATE
555426	27229231	尚達塩業股份有限公司	NA	2005-05-30	吳秀里, 周永紹, 周博元, 周碩良
1067348	70794974	上達糧業國際股份有限公司	NA	2002-01-08	吳秀里, 周永紹, 周博元, 周碩良

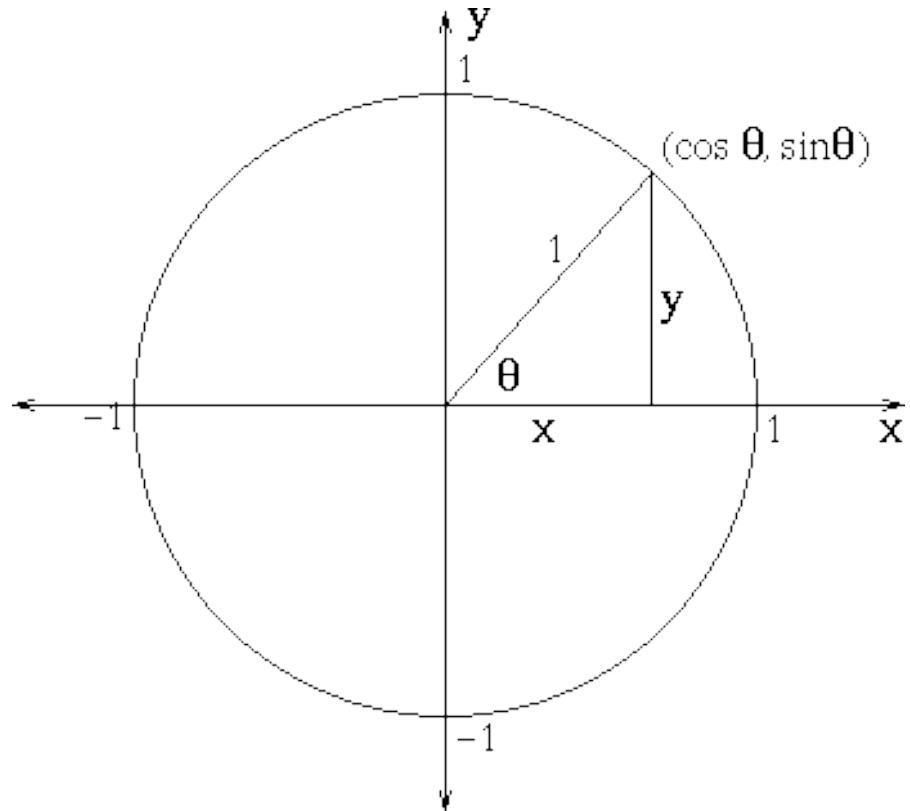
[http://web.pcc.gov.tw/tps/main/pms/tps/atm/atmAwardAction.do?
newEdit=false&searchMode=common&method=inquiryForPublic&pk
AtmMain=51493408&tenderCaseNo=GF4-103122](http://web.pcc.gov.tw/tps/main/pms/tps/atm/atmAwardAction.do?newEdit=false&searchMode=common&method=inquiryForPublic&pkAtmMain=51493408&tenderCaseNo=GF4-103122)

數值型變數

- Cosine Similarity

$$X_1, X_2 \in \mathbb{R}^d$$

$$\frac{X_1 \cdot X_2}{\|X_1\| \|X_2\|}$$

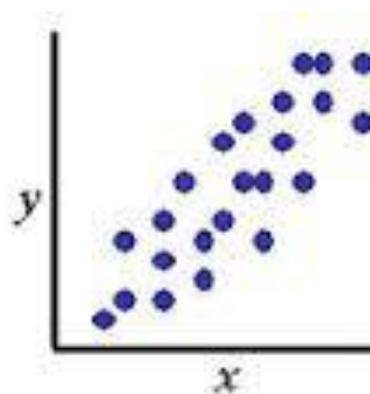


數值型變數

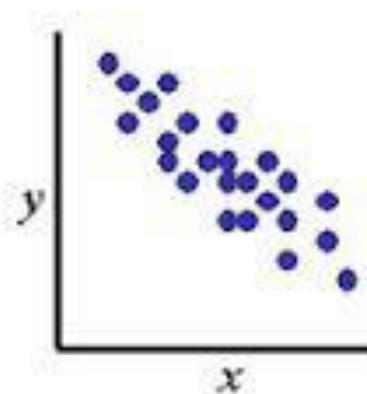
- 相關性 (Correlation)

$$X_1, X_2 \in \mathbb{R}^d$$

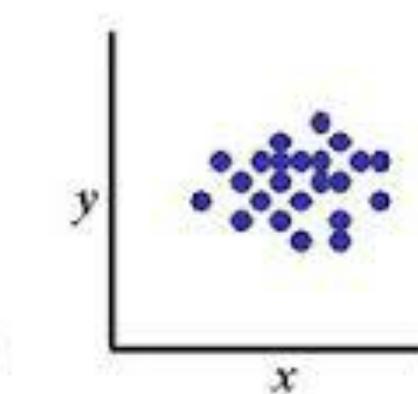
$$\text{Correlation} = \frac{(X_1 - \bar{X}_1) \cdot (X_2 - \bar{X}_2)}{\sigma(X_1)\sigma(X_2)}$$



Positive correlation



Negative correlation



No correlation

Gower 相異性係數

- 同時考量類別型與數執行資料的相異性

$$d(i, j) = \frac{1}{M} \sum_{k=1}^M d_{ijk}$$

i, j 是第 i, j 筆資料
 M 是變數總數
 K 是第 k 個變數

如果變數 k 是數值型 : $d_{ijk} = \frac{|x_{ik} - x_{jk}|}{|\max(x_k) - \min(x_k)|}$

如果變數 k 是類別型 : $d_{ijk} = \frac{|x_{ik} \cap x_{jk}|}{|x_{ik} \cup x_{jk}|}$

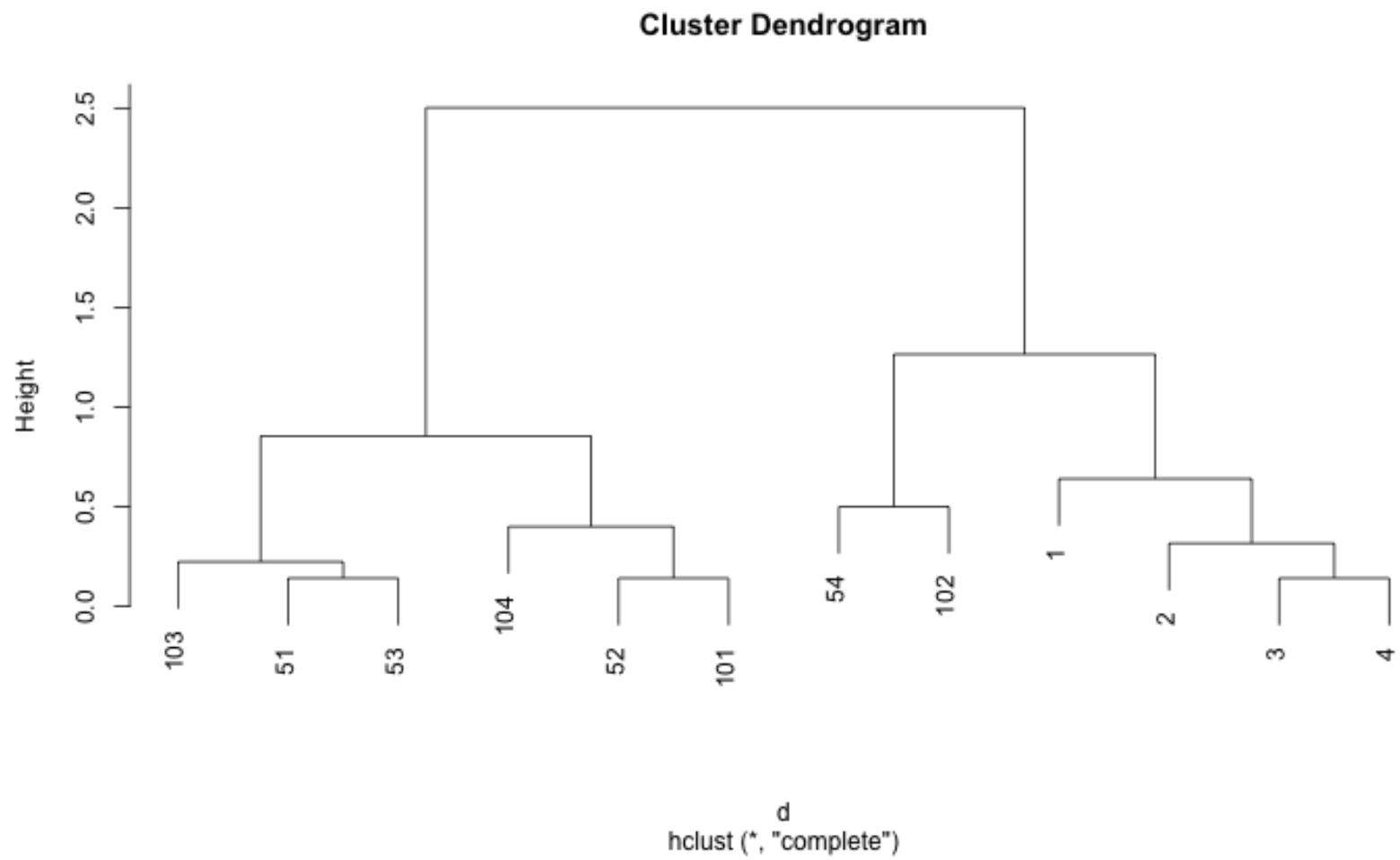
Gower 相異性係數 (cont.)

- R functions
 - `cluster::daisy(x, metric="gower")`
 - `vegan::vegdist(x, metric="gower")`

如何挑選相似度指標？

- 是目標與應用而定
 - 利用 $1NN$ 的分類結果來評估 (後面介紹)
- 試著利用領域知識，以相似度解釋資料
- 數值變數的 baseline
 - 先標準化
 - 以歐氏距離計算

階層式分群法 (Hierarchical)



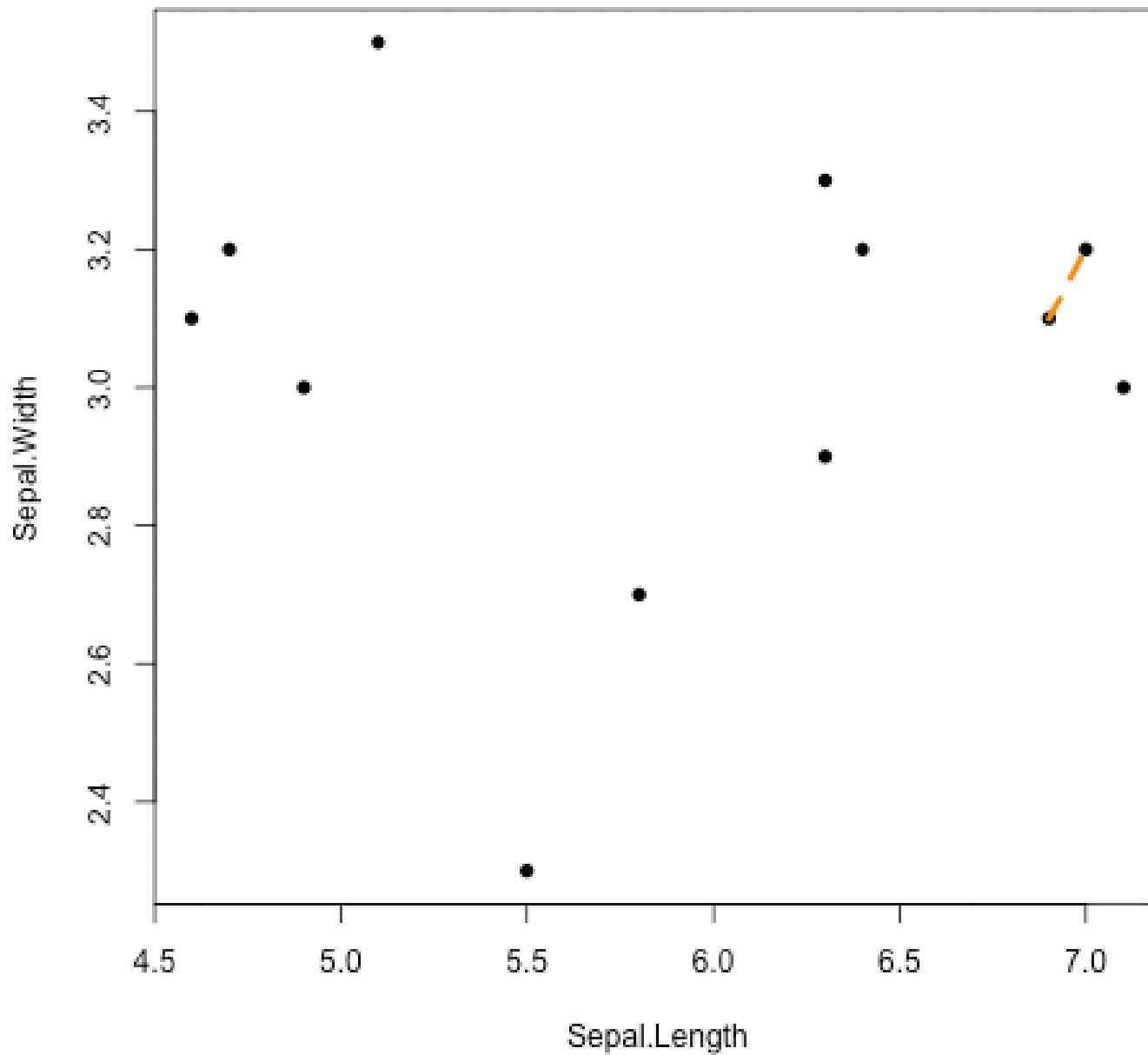
階層式分群法 (cont.)

- 概念簡單
- 不需要事先決定要分幾個 clusters
- 不需要資料點的實際座標，只需要資料點兩兩間的距離
- 資料量大時，效率不佳
- R packages
 - `stats::hclust`

階層式分群法 (cont.)

- 決定資料點間的距離
- 將相鄰的資料點合併成 cluster
- 決定資料點與 cluster 之間的距離
- 決定 cluster 與 cluster 之間的距離
- 距離由短到常依序合併資料點與 cluster
- 會製成樹狀圖 (Dendrogram)

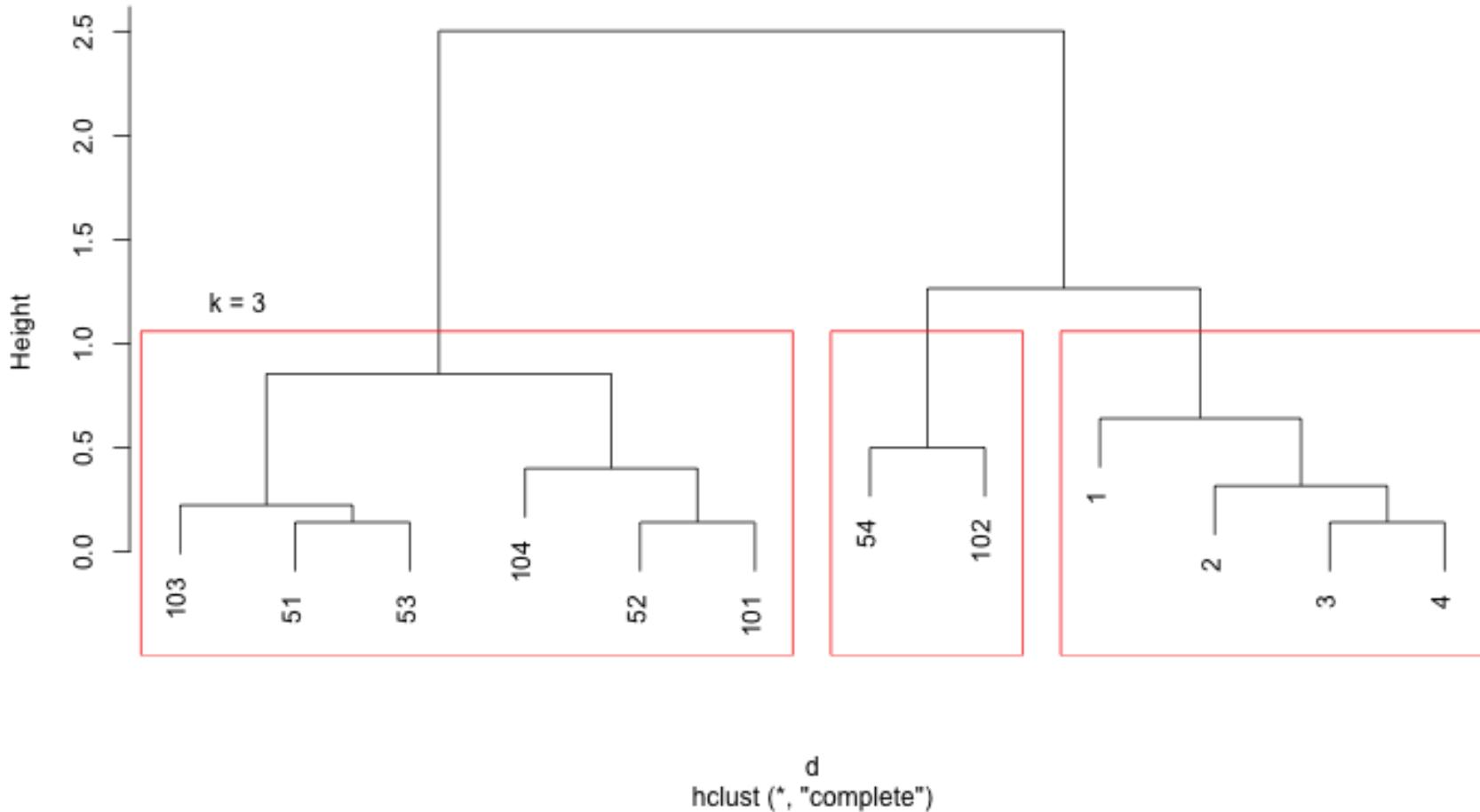
圖解 UPGMA 演算法



階層式分群法 (cont.)

想找 K 群？找到 K 群時停下來即可

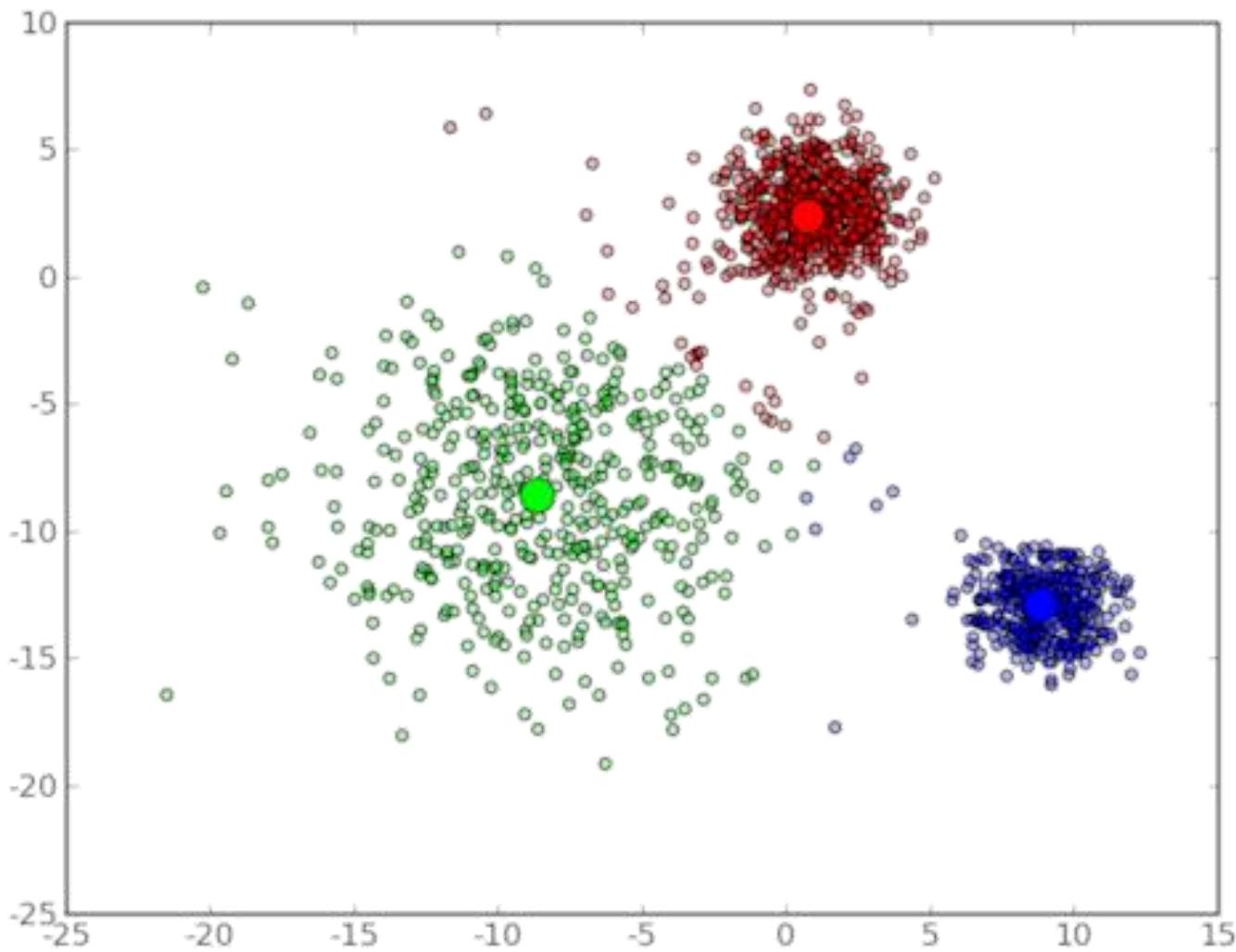
Cluster Dendrogram



階層式分群法 (cont.)

- 如何評斷分群結果好壞？
 - 群內的距離要短
 - 群間的距離要常
- 如何挑選分群的個數
 - 挑選分群結果好的群組個數
 - 透過樹狀圖的高度差距來比較

K-Means 分群法



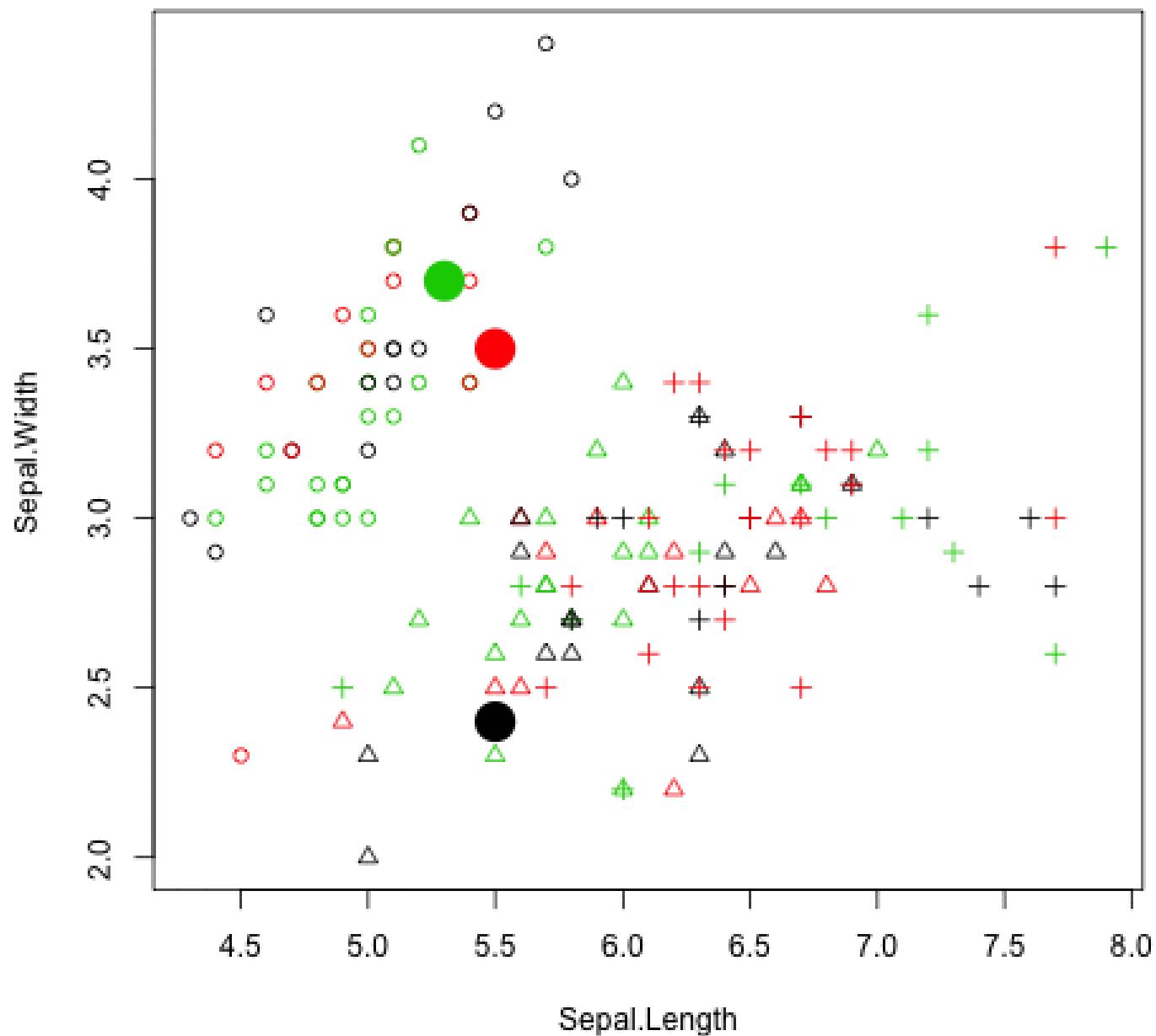
K-Means 分群法 (cont.)

- 以中心點為計算基礎的分群法
 - 基本、快速
 - 需要事先決定分幾群
 - 起始分群中心會影響分群結果
 - 離群值會有影響
 - 非球狀、群集大小不均會有問題
 - K-Means 一定會收斂，但只有局部最佳解
- R package
 - **stats::kmeans**

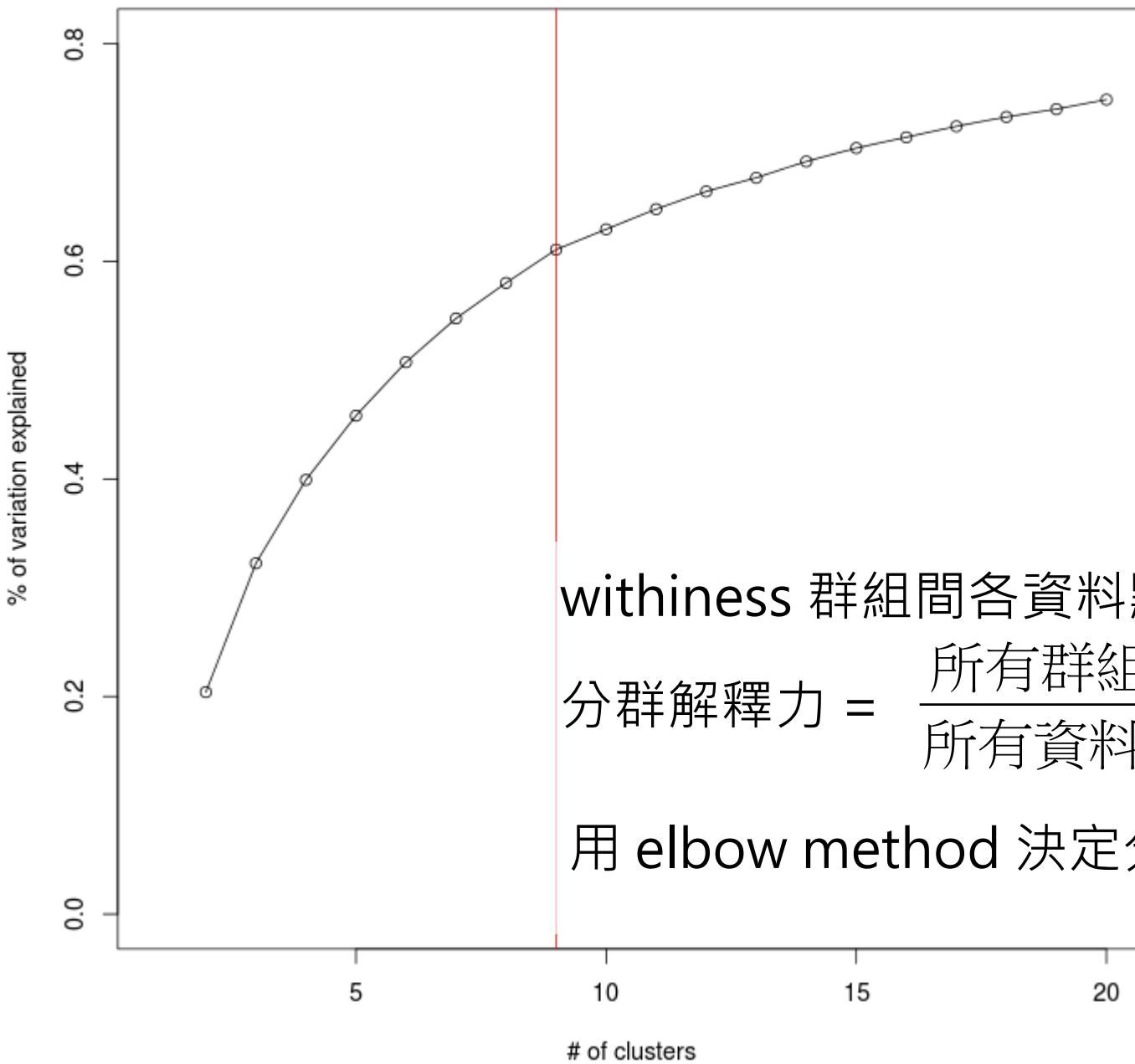
K-Means 分群法 (cont.)

- 指定群聚數目 k 與起始分群中心
- 利用距離指標定義**誤差函數**
- 藉由反覆迭代運算，逐次降低誤差值
- 直到目標函數不再變化時結束

k-means



Choosing K

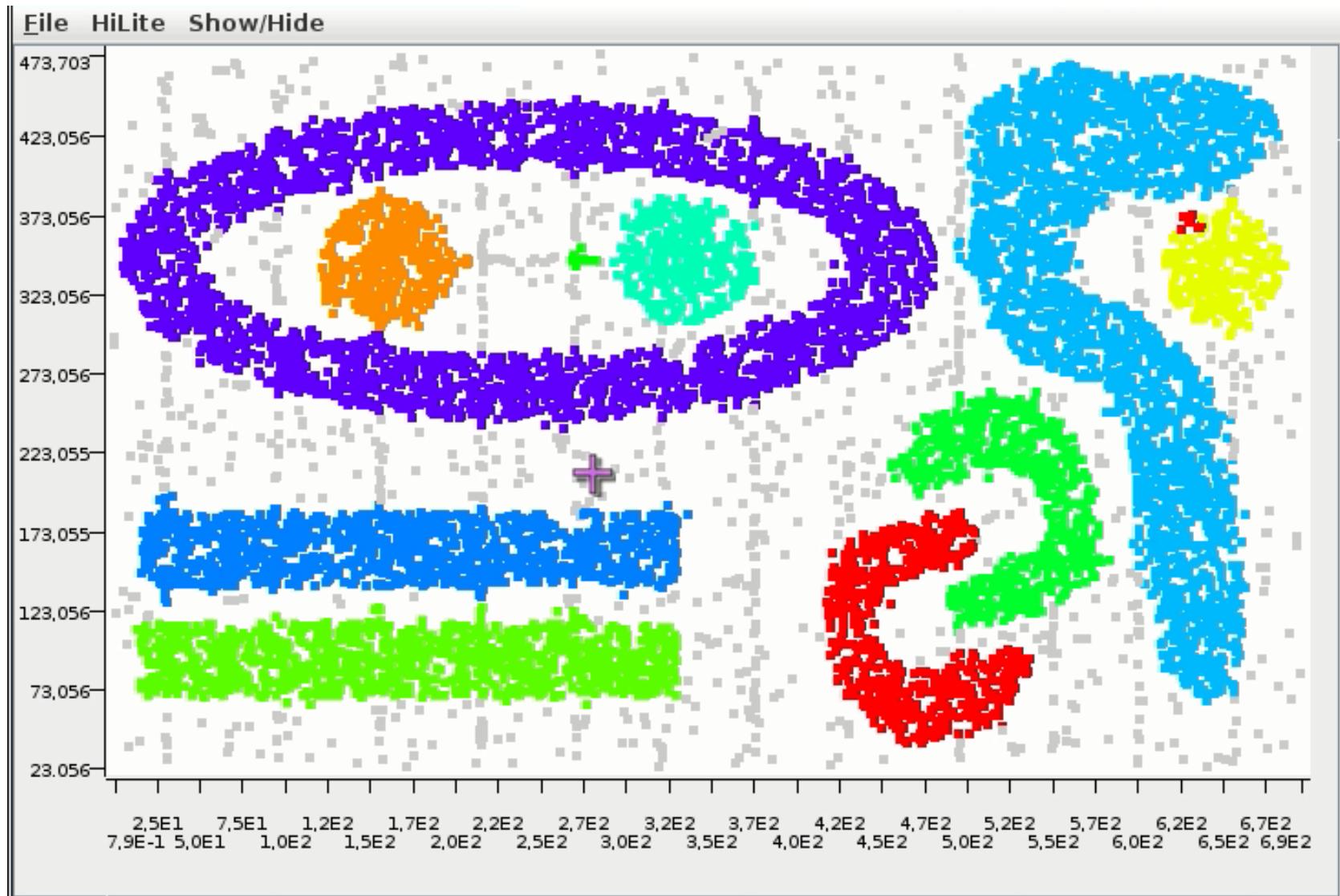


withiness 群組間各資料點距離的平方和

分群解釋力 = $\frac{\text{所有群組 withiness 總和}}{\text{所有資料點距離平方和}}$

用 elbow method 決定分群數

DBScan 分群法



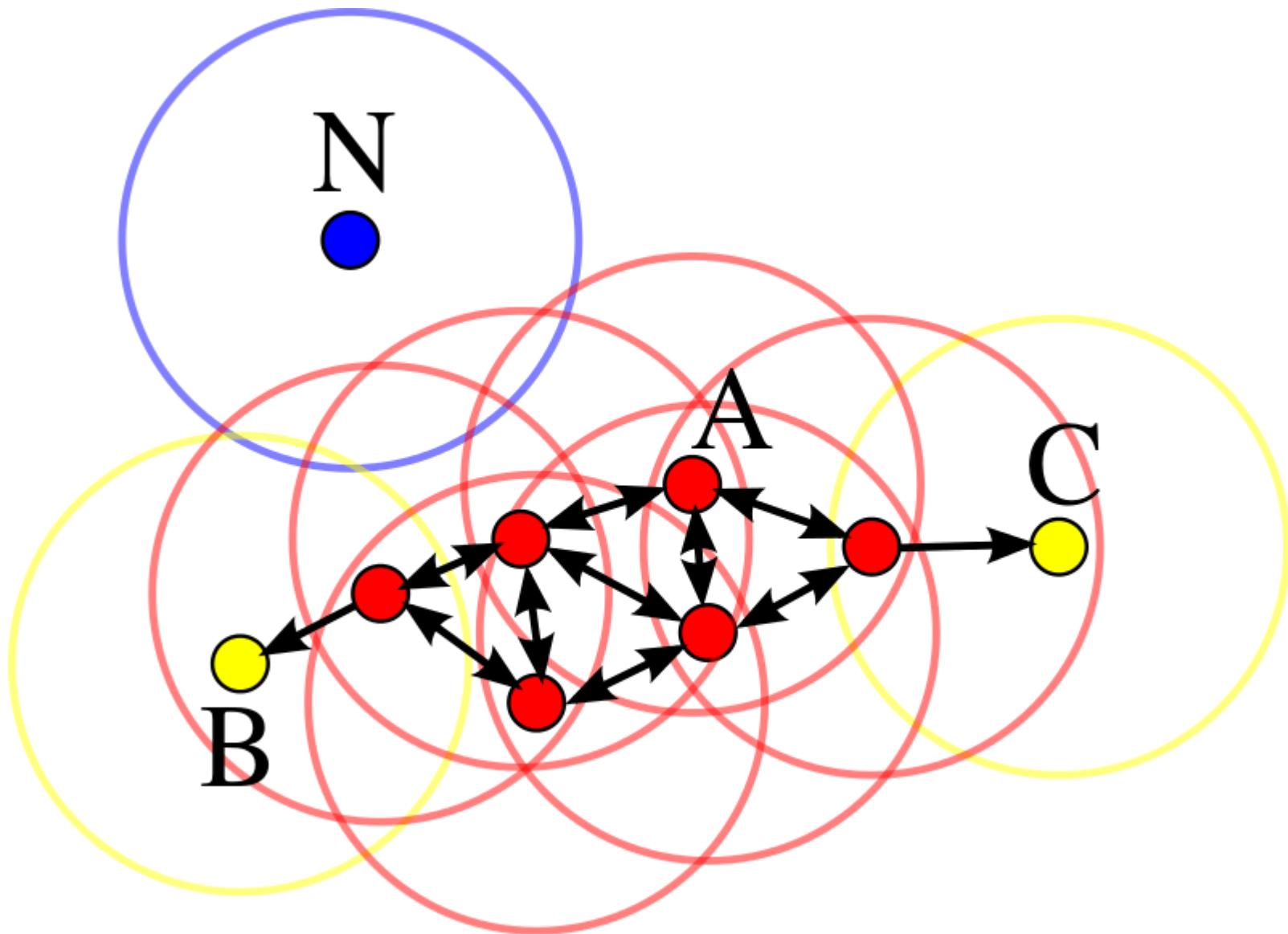
DBScan 分群法 (cont.)

- 以密度為基礎的分群法
 - 不需要事先決定分幾群
 - 可以分割出任意形狀的群集(環形、文字)
 - 會過濾雜訊 (noise)
 - 需要兩個參數：掃描半徑 (eps) 與半徑內最小包含點數 (MinPts)
 - 缺點：高維度資料、密度分布不均的資料以及資料量大的資料，效率不佳
- R packages
 - **fpc**
 - **dbSCAN**

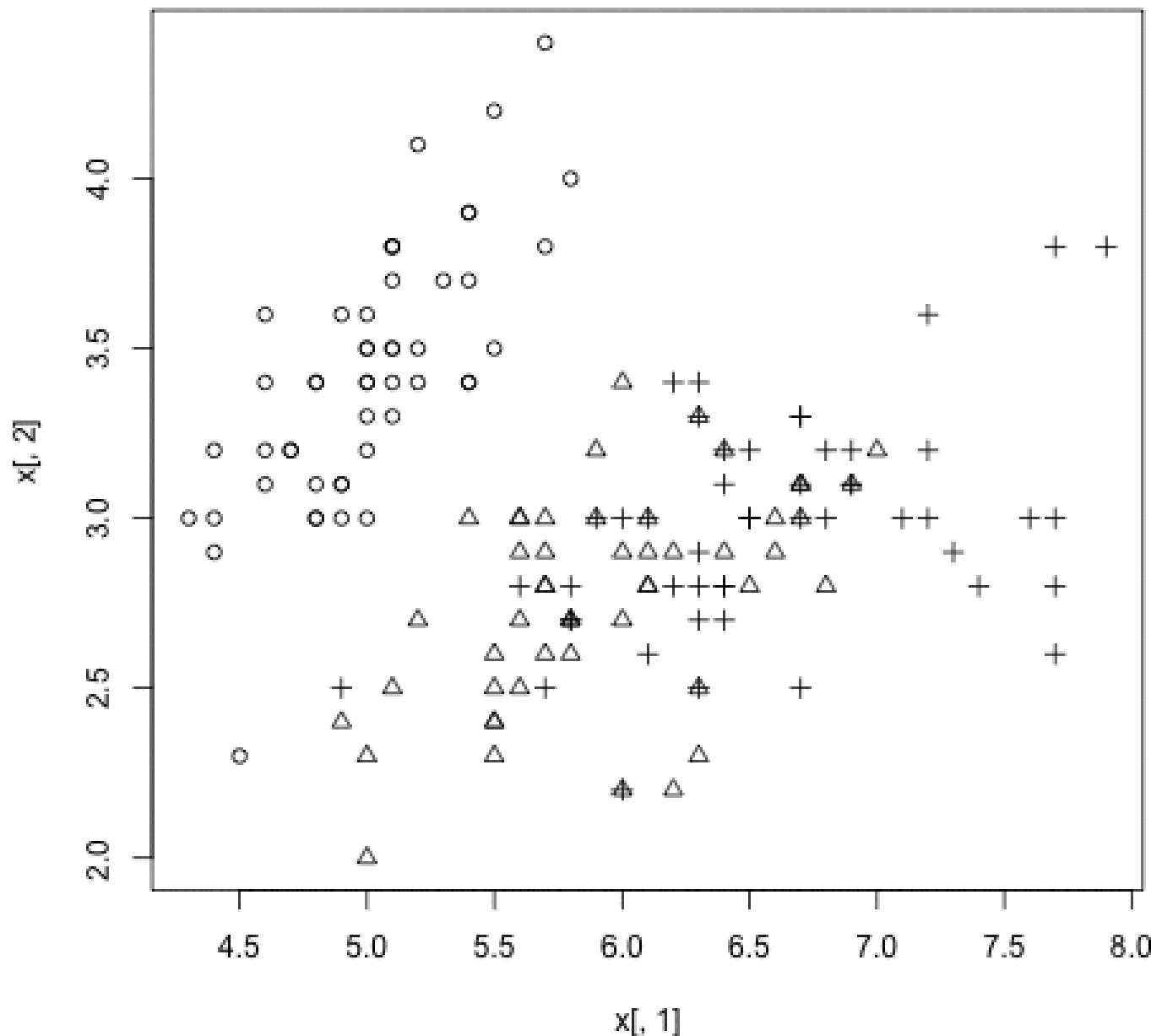
DBScan 分群法 (cont.)

- 紿定初始參數 eps (決定是否同群) 與 MinPts (濾除雜訊)
- 掃描所以觀察值，找出核心點(core)、邊界點(border)、雜訊點(noise)
- 移除雜訊點，並將核心點歸類到同一群

DBScan 分群法



dbSCAN



分群總結

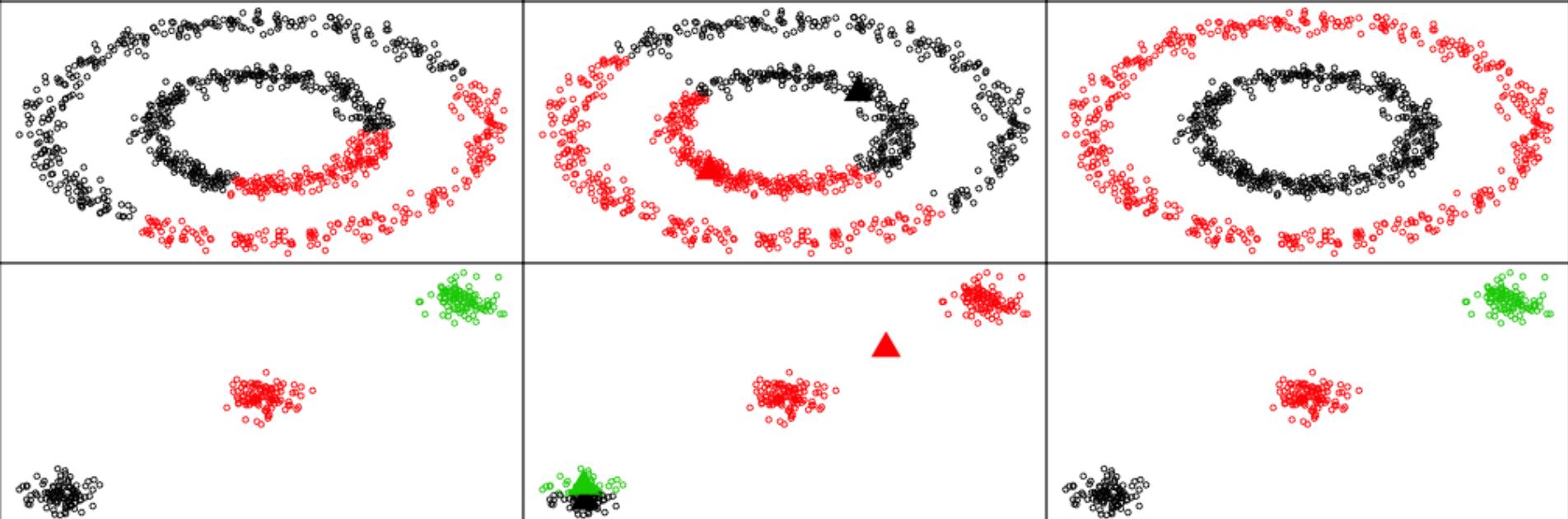
- 對資料表中的觀察值賦予群集的標籤
- 需要資料間距離的定義
(Distance/Similarity)
- 各種分群演算法
 - Hierarchical、K-Means、DBScan
- 驗證分群結果是否合理
 - 群組內差異小
 - 群組間差異大

分群演算法比較

Hierarchical Clustering

Center-based Clustering

Density-based Clustering



實際操作分群演算法

- 請各位完成 **RDM-o4-Unsupervised-Learning-Clustering**

R 語言資料探勘實務

歸類演算法

歸類演算法 (Classification)

- 歸類法則
 - 按照分析對象的屬性分門別類加以定義，建立類組 (Class)
- 常見問題
 - 判斷貸款申請者的風險程度
 - 透過基因表現量，預測癌症類別
 - 透過既有分類法則找除離群值(肥羊、異常客戶)
- 常用方法
 - 最近鄰居法 (K-NN)、Logistic Regression、SVM、Decision Tree、Gradient Boosted Decision Tree、Random Forest

資料集

- 基本上來說，分類演算法算是一種監督式學習
- 所以用來訓練模型的資料，會包含應變數(Y)的標籤
- 一般而言資料集會分成三類
 - 訓練集 (Training Dataset): 建模用
 - 驗證集 (Validation Dataset): 模型篩選用
 - 測試集 (Testing Dataset): 用來驗證最終模型

資料競賽

- 訓練集
 - 每個資料點均有：屬性X、標籤Y (類別型變數)
 - 需要自行從訓練集中割出一部分驗證集
- 測試集
 - 只有X沒有Y
 - 用訓練集所訓練的模型，輸入X以預測Y
 - 以測試集的**準確度**決勝

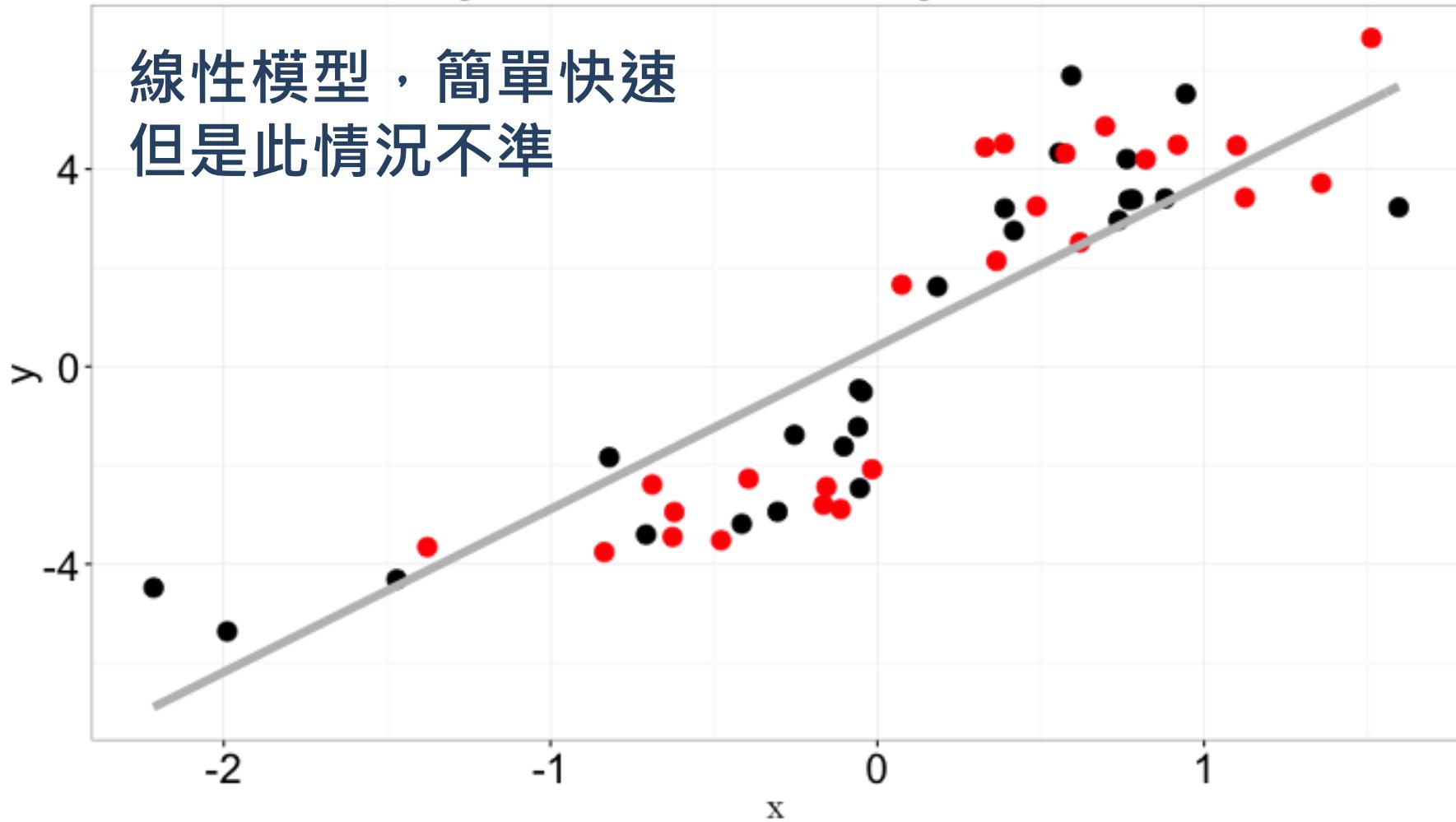
如何製作 驗證集

- N-fold Validation
 - 亂數將資料分割成 N 份
 - 以其中的 $N-1$ 份做訓練集，剩下的做驗證集
 - 可以重複 N 次
- Bootstrapping
 - 亂數取出一定比例的資料作為訓練集，剩下的做驗證集
 - 可以多次重複進行驗證

為什麼需要驗證？

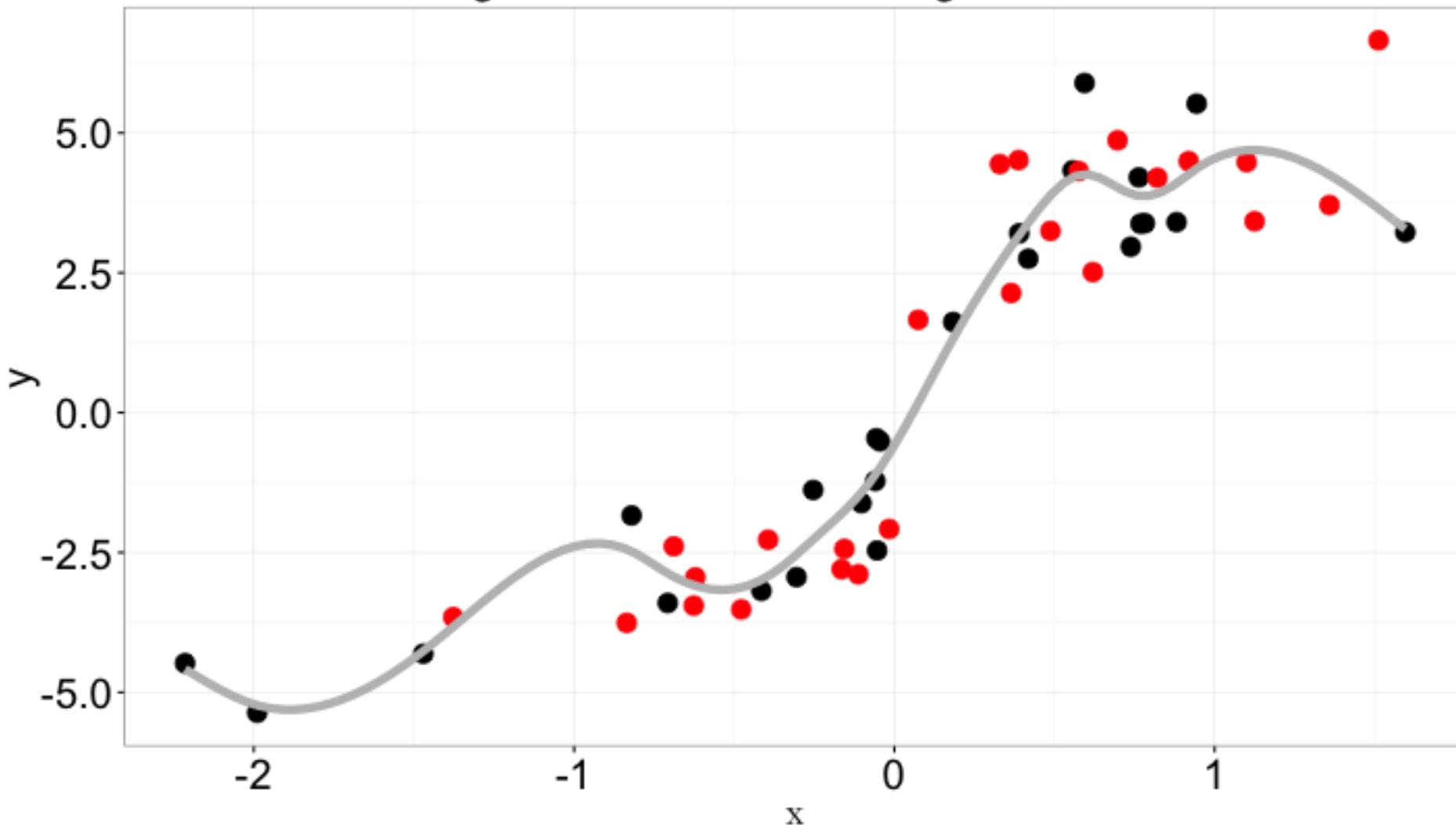
Training Loss: 65.8119 Testing Loss: 75.9753

線性模型，簡單快速
但是此情況不準



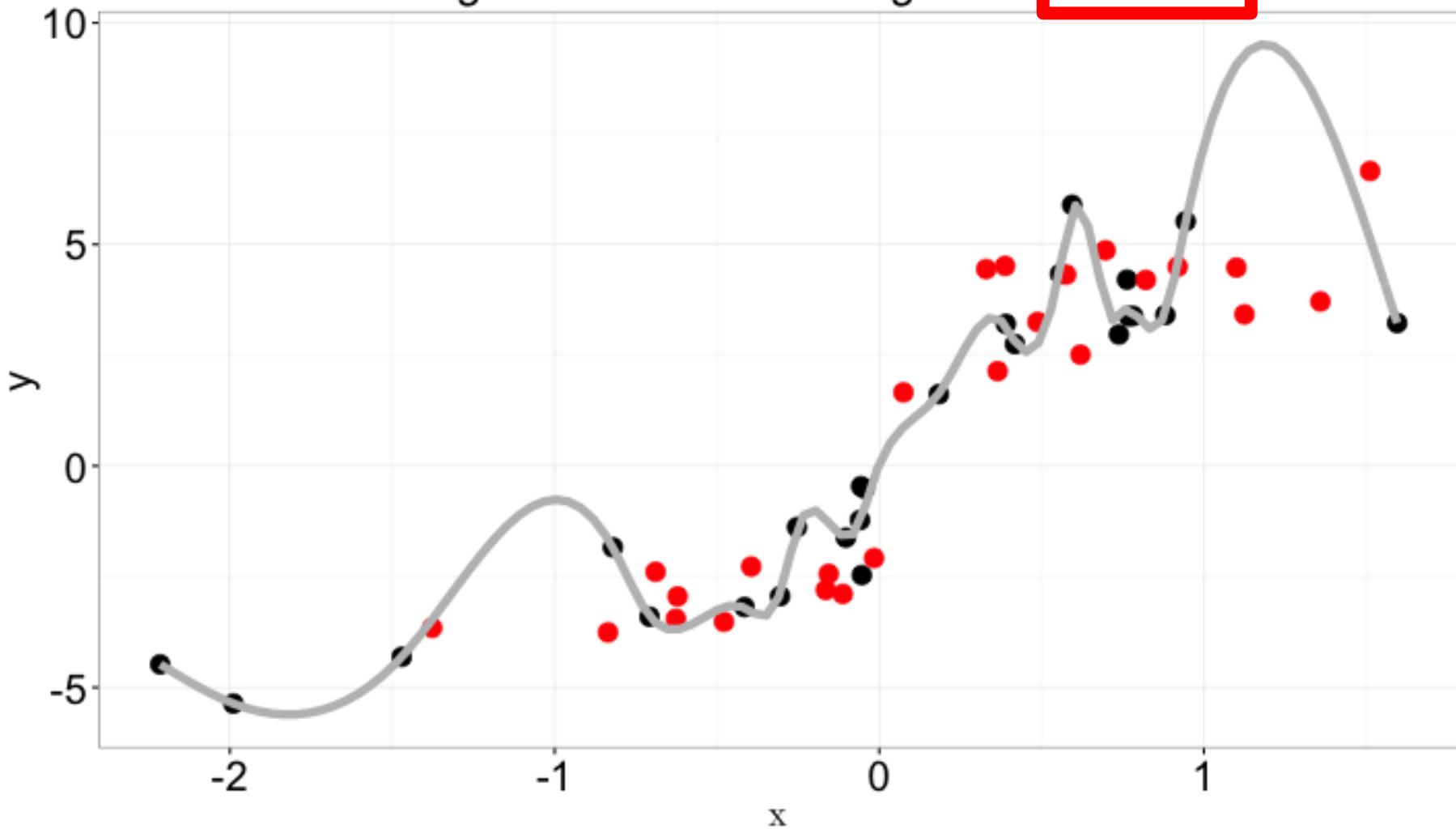
在訓練集與測試集都不錯的模型

Training Loss: 10.8080 Testing Loss: 31.0918



過度訓練 (Overfitting) 的模型

Training Loss: 3.0447 Testing Loss: 112.8659

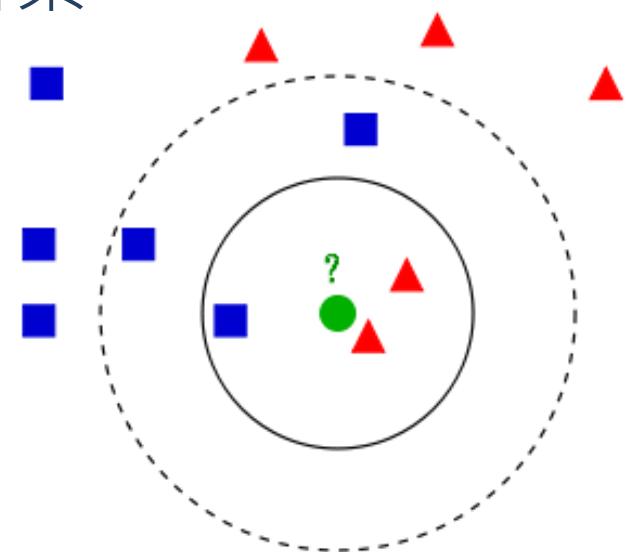


最近鄰居法 (Nearest-Neighborhood)

- 標的類別是由最近的一個鄰居賦予
 - 給定測試集，從訓練集中找出與測試集各資料點距離最近的鄰居
 - 用鄰居的類別猜測測試集資料點的類別
- R package
 - `class`

K-NN

- 標的的類別是尤其最近的 K 個鄰居多數決
 - 給定測試集，從訓練集中找出與測試集各資料點距離最近的前 K 個鄰居
 - 以該 K 個鄰居中，出現次數最多的類別，作為猜測測試集各資料點類別的答案

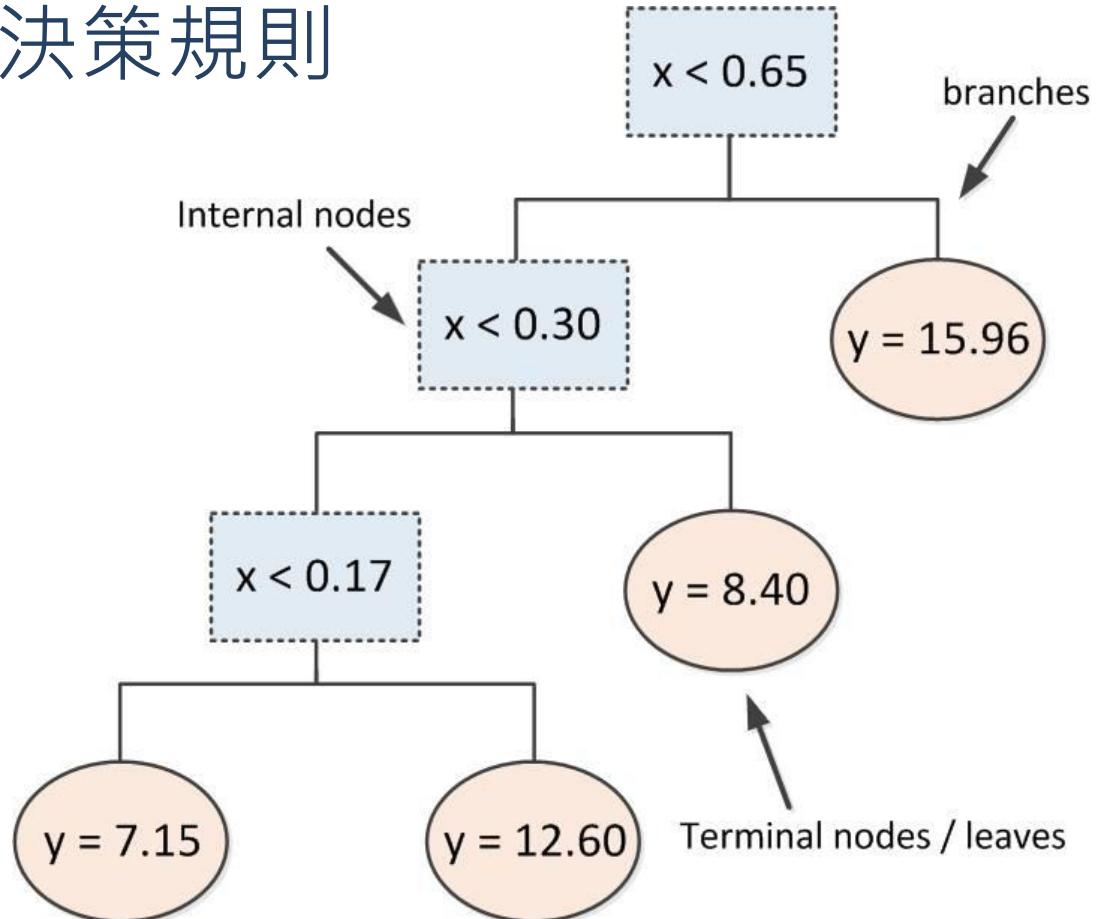


K-NN 常用於評量相似度指標

- 核心概念：物以類聚，簡單沒有冗餘
- 用**距離/相似度**決定鄰居
- 相似度指標的挑選，影響 K-NN 效果極大，因此可以用來挑選相似度指標

決策樹 (Decision Tree)

- 樹狀結構建立決策規則
- R package
 - c50
 - rpart



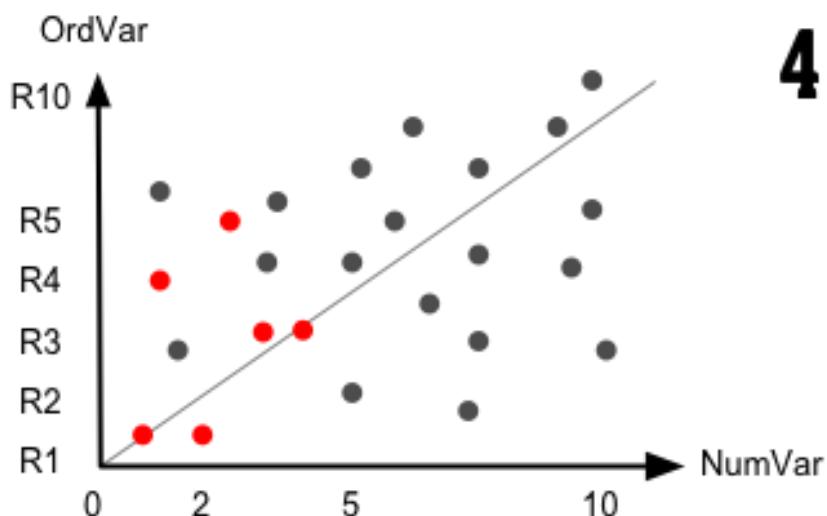
如何建構一顆樹

- 定義亂度指標
 - 亂度越大、隨機性越高
 - 常用系統的熵(Entropy)表示
- 一次做一種切割 (一次長一個分支)
 - 每次分割，應該要能有效減少系統的總亂度
 - 分割前系統亂度 - 分割後系統亂度 = 增益值
 - 每次分割要能獲得最大的增益
- 當無法再切割時停止

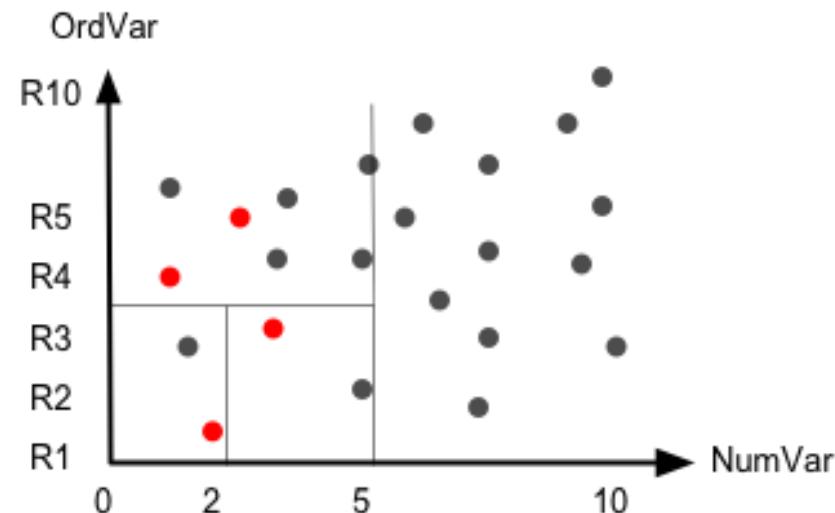
1

NumVar	OrdVar	CatVar
1	R2	Good
1.5	R4	Bad
2	R1	Bad
2.5	R5	Good
3	R3	Good
3.5	R3	Bad
⋮	⋮	⋮

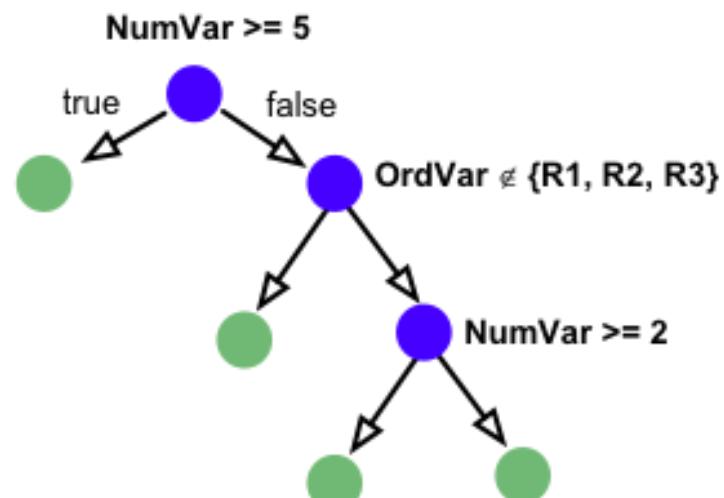
2



3



4



實際操作歸類演算法

- 請各位完成 **RDM-05-Classification**

R 語言資料探勘實務

最後的資料小語

總結

- 資料探勘的核心精神：機率、距離、風險、誤差
- R 語言對於資料探勘是非常優秀的工具
- 資料探勘並非資料科學的全部

小小叮嚀

- 資料探勘不僅僅是**技術**或**軟體工具**
- 資料探勘並非無所不能
- 資料探勘是從資料中挖掘有價值的**假設**，
但未必能**驗證**假設
- 不是只有資料分析師可以做資料探勘



資料是一種信仰

謝 謝 各 位

Q & A



許懷中 Hwai-Jung Hsu

hjhsu@mail.fcu.edu.tw

<https://tw.linkedin.com/in/hjhsu>