資料探勘
DATA MINING

張家瑋 助理教授

國立臺中科技大學資訊工程系

國立成功大學工程科學系

jwchang@nutc.edu.tw

張家瑋.大平台.tw

分群
CLUSTERING

概念

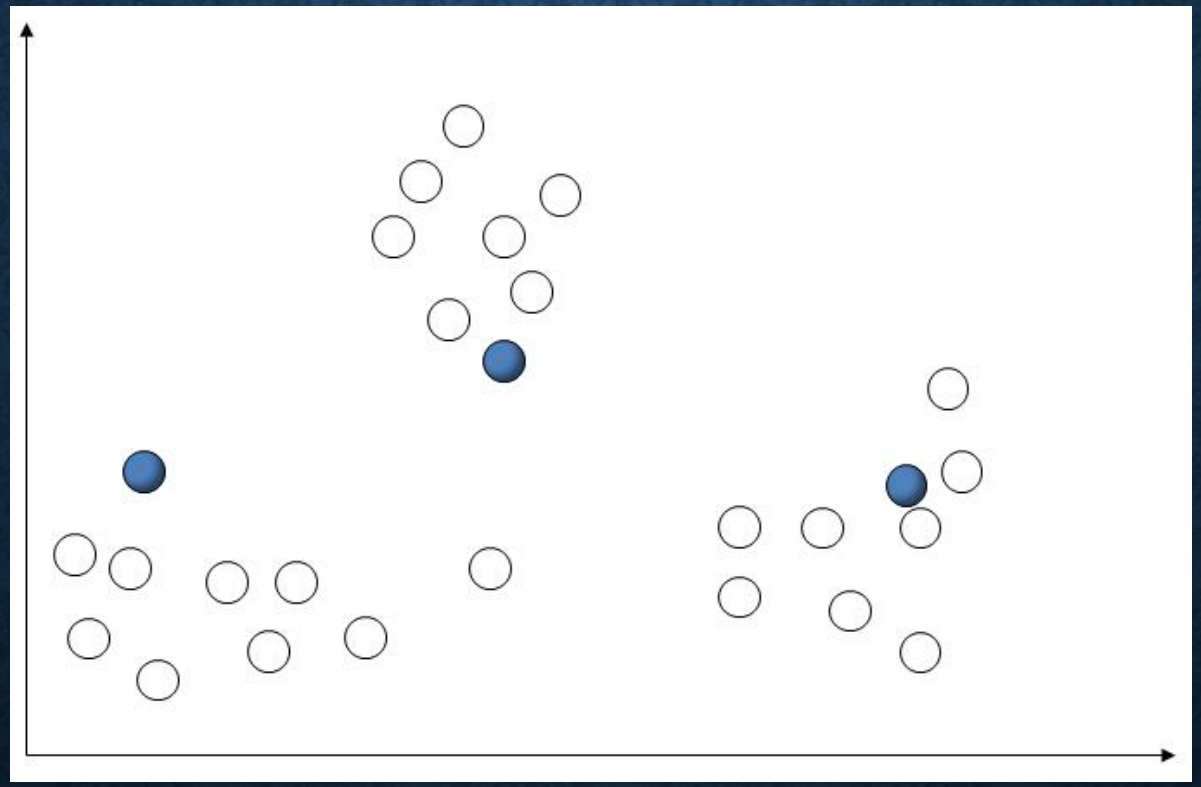
- 把許多事物按照某種標準歸為數個類別，其中較為相近/類似的聚為一類，反之較不相近的則聚為不同類。
- 目的是企圖從一大堆雜亂無章的原始資料中，找出少數幾個較小的群體，使得群體內的分子在某些變項的測量值均很類似，而群體與群體間的分子在該測量值上差異較大。
- 同一組樣本會因不同目的、資料輸入方式、所選擇分群特徵或資料屬性，形成不同的分群結果。

K-MEANS

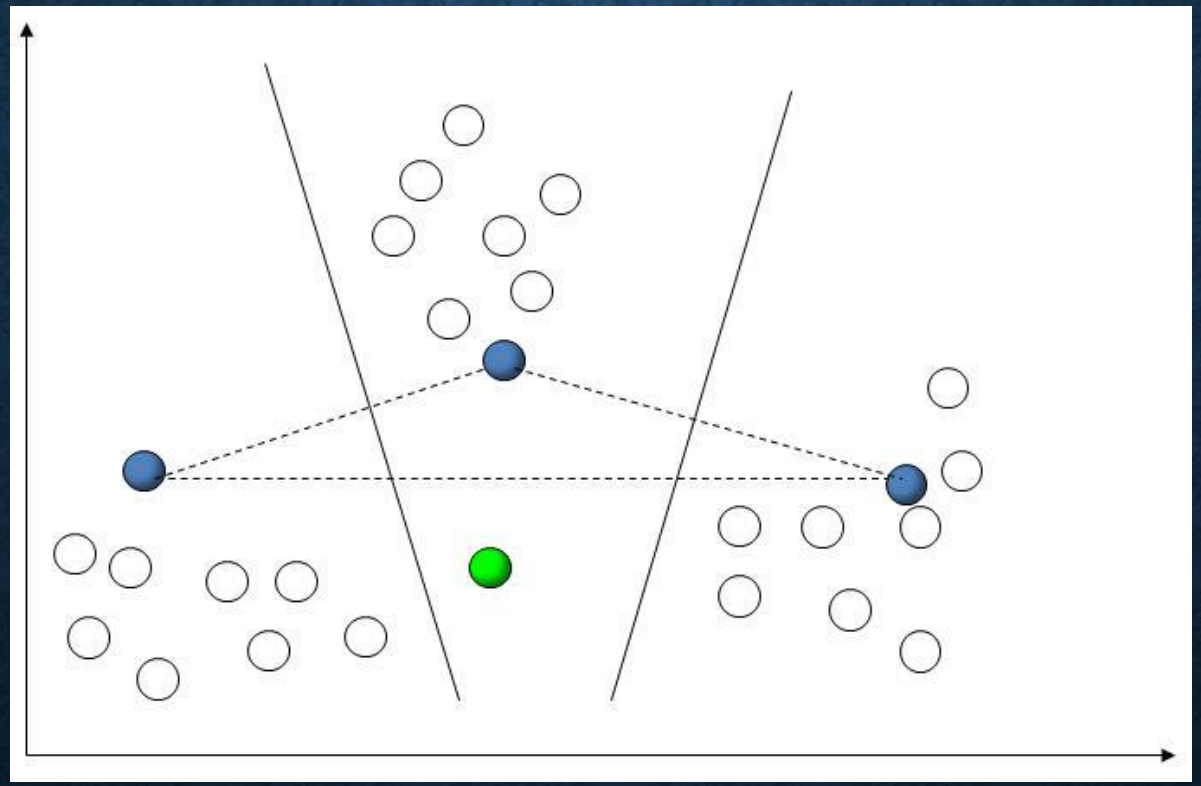
概念

- 隨機選取 k 個樣本作為起始中心點，將其餘樣本歸入相似度最高中心點所在的群；再計算目前群內樣本座標的平均值為新的中心點，依次循環反覆運算，直到所有樣本所屬的群不再變動。

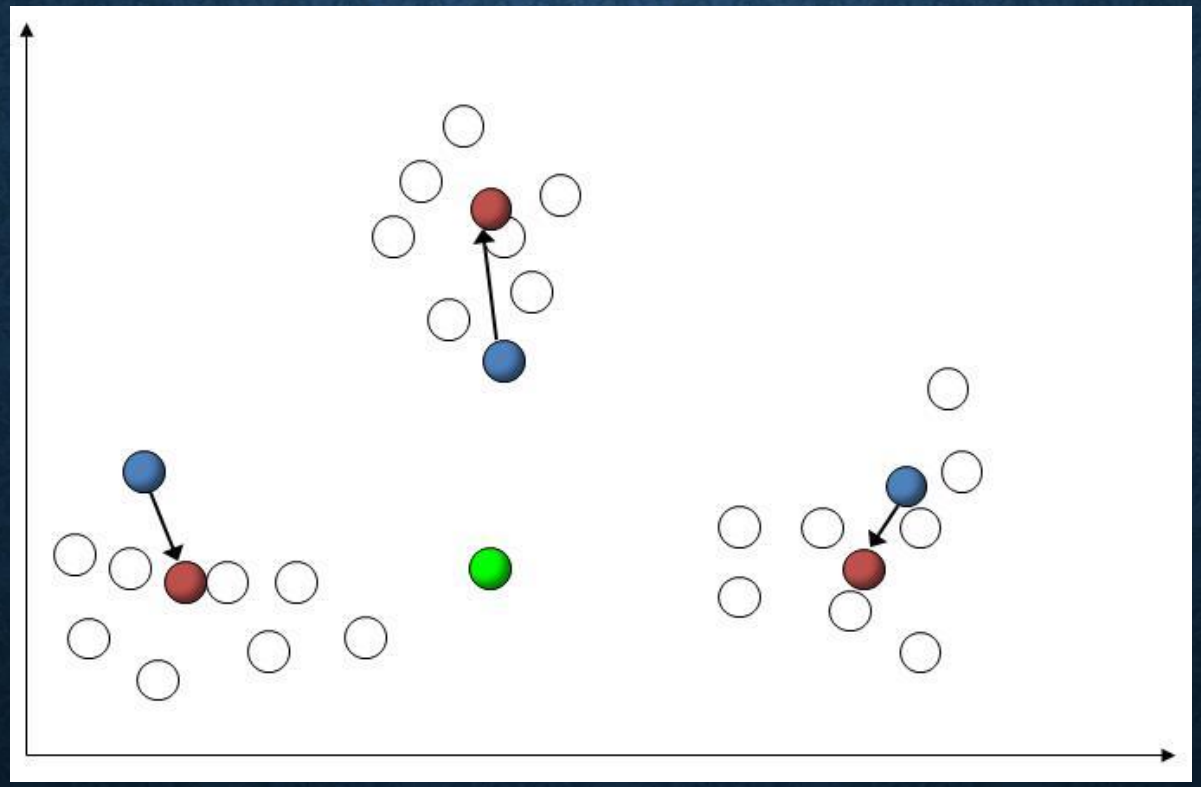
STEP 1. 隨機指派群集中心



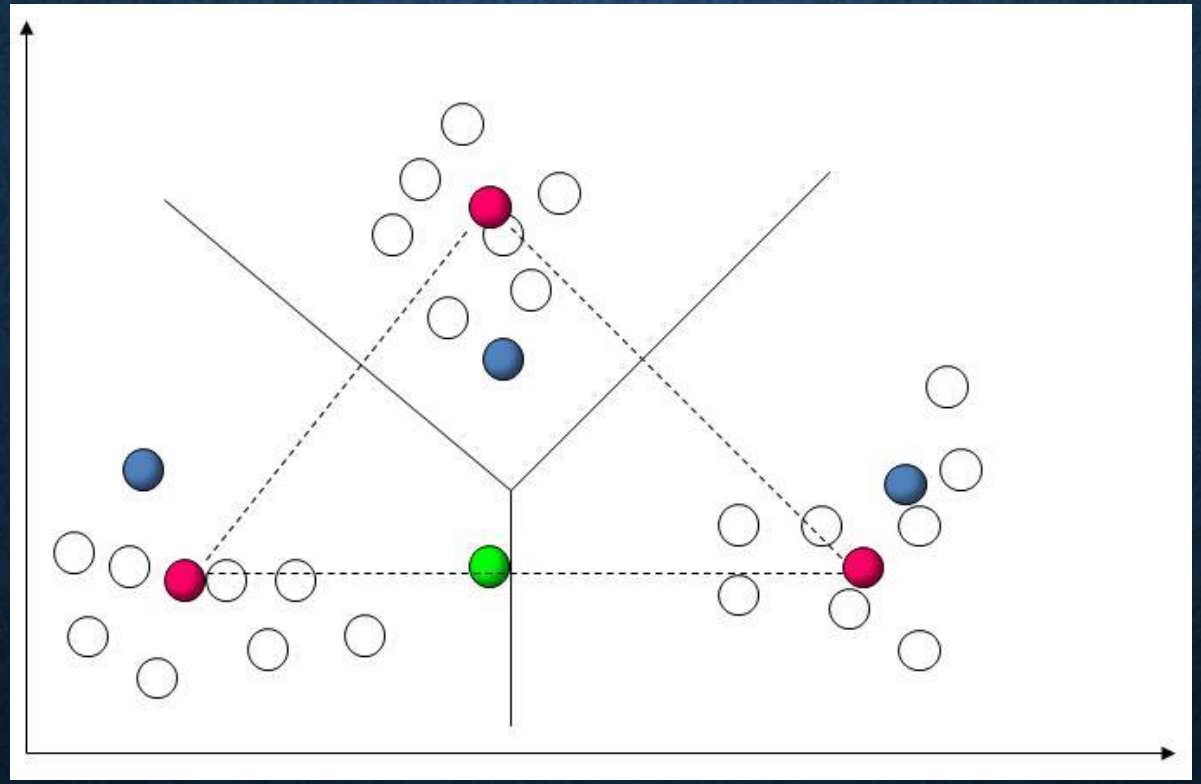
STEP 2. 產生初始群集



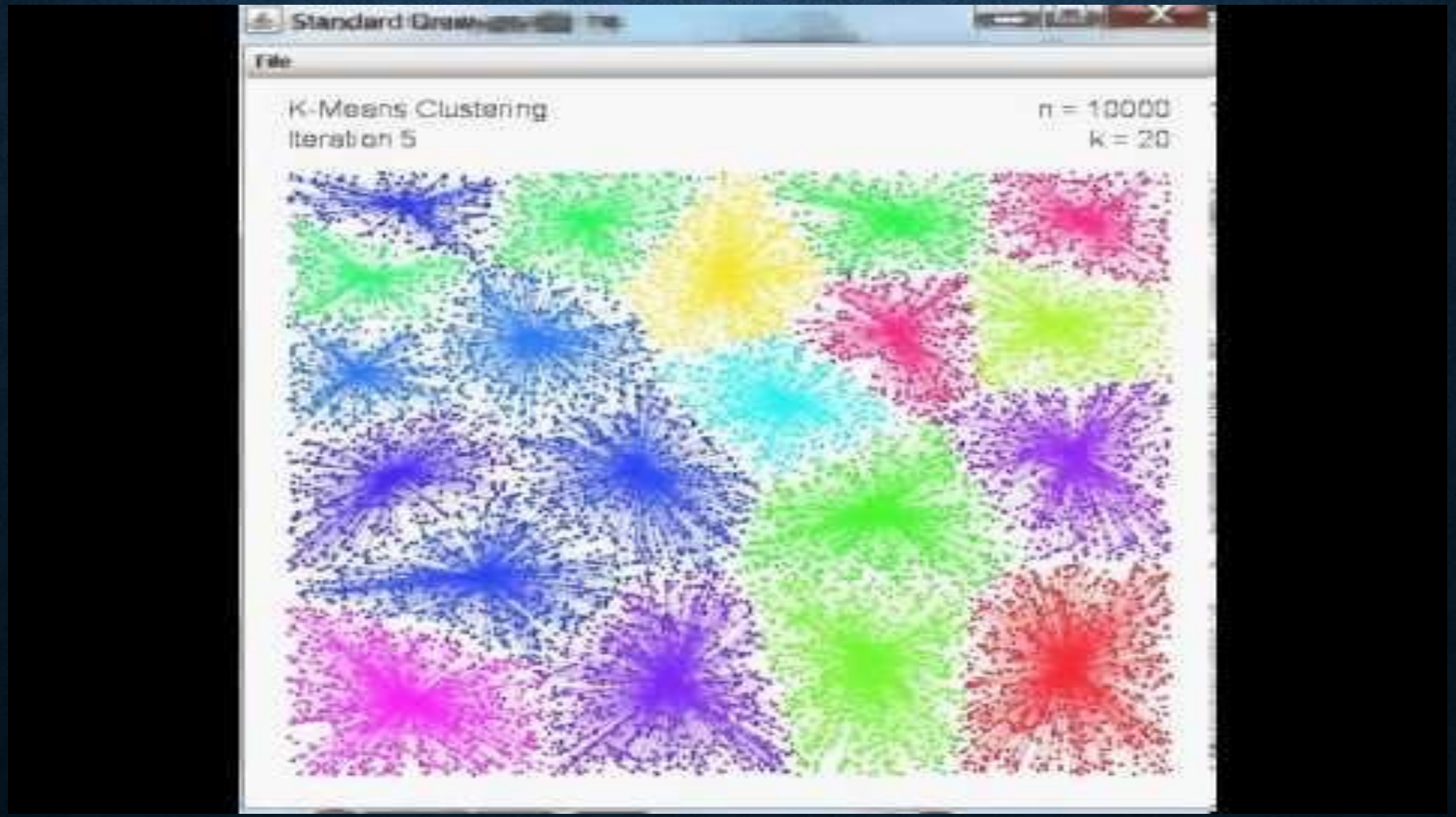
STEP 3. 產生新的質量中心



STEP 4. 變動群集邊界



EXAMPLE



參考來源

1. https://rpubs.com/skydome20/R-Note9-Clustering
2. https://jgpan.gitbooks.io/the-study-of-r/content/clustering.html
3. K-Means Clustering Example
4. http://ccckmit.wikidot.com/ai:kmeans

THINKING

重點？

重點

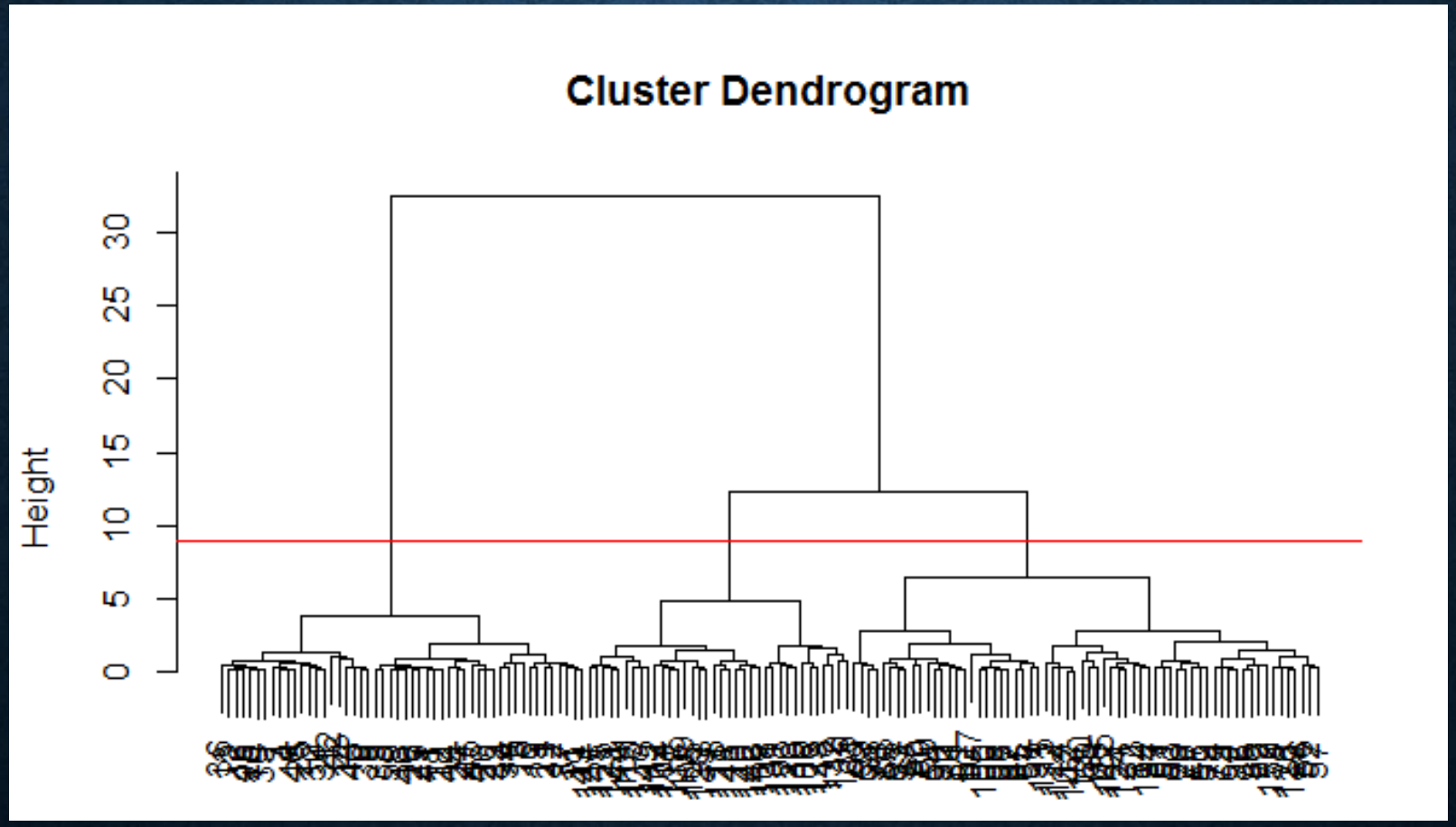
1. K 如何決定？
2. 相似度的方法

階層式分群法
HIERARCHICAL CLUSTERING

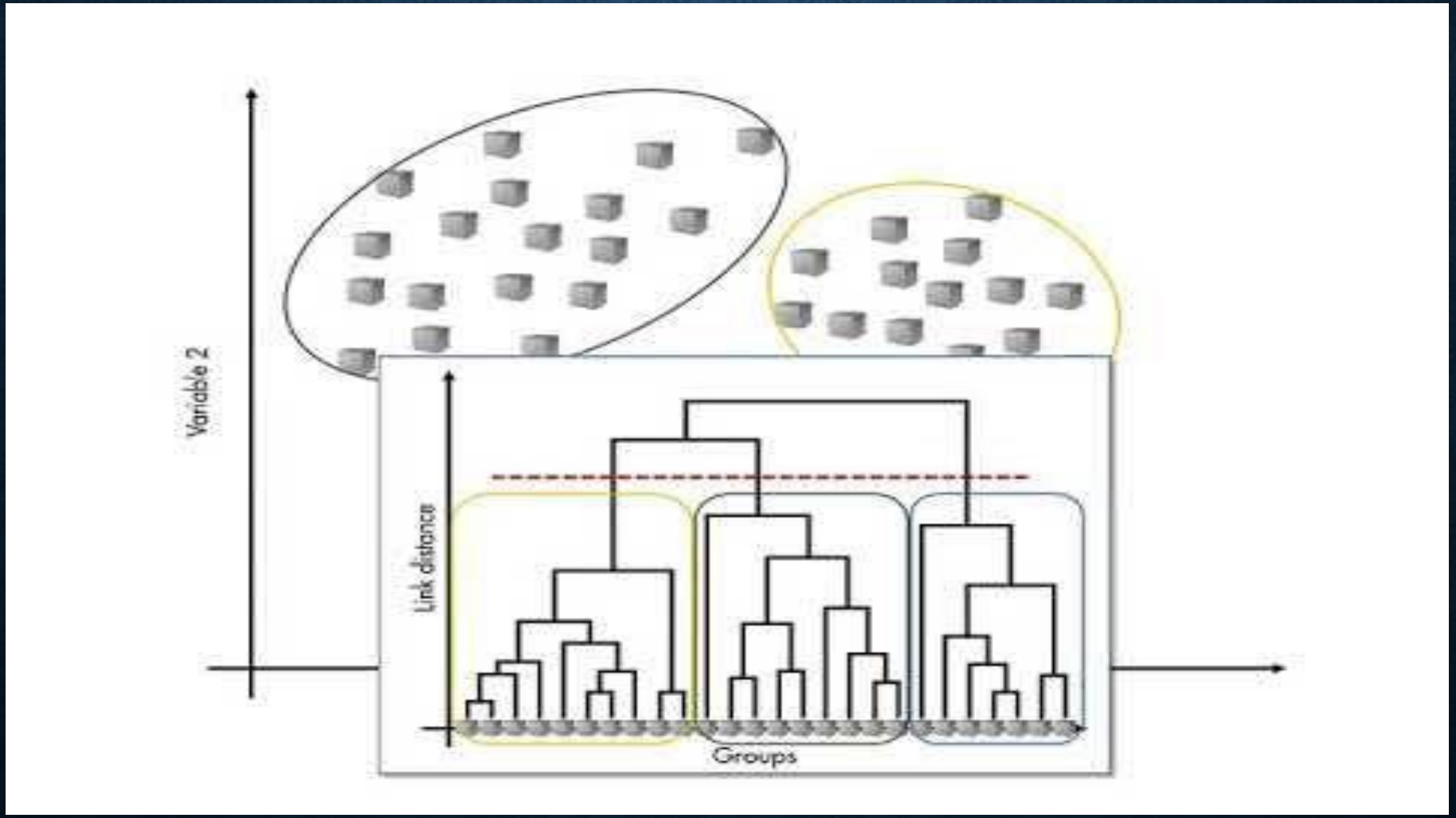
概念

- 不須事先設定群數 k，每次反覆運算過程僅將距離最近的兩個樣本/群聚為一類，直到符合設定的群集數條件
  - 由下往上聚合: 從樹狀結構底部開始，將資料或各分群逐次合併，一開始將每個資料都視為一個獨立的分群，然後依據分群間相似度計算公式，不斷合併兩個最相似的資料/分群，直到所有資料/分群都合併成一個大的群集或達到所訂定的停止條件（設定的數量）為止。

PROCESSES



EXAMPLE



參考來源

1. https://rpubs.com/skydome20/R-Note9-Clustering
2. https://jgpan.gitbooks.io/the-study-of-r/content/clustering.html
3. MATLAB skills, machine learning, sect 5: Hierarchical Clustering

THINKING

重點？

重點

1. 由上往下分裂？
2. 與 K means 的差異？

分類
CLASSIFICATION

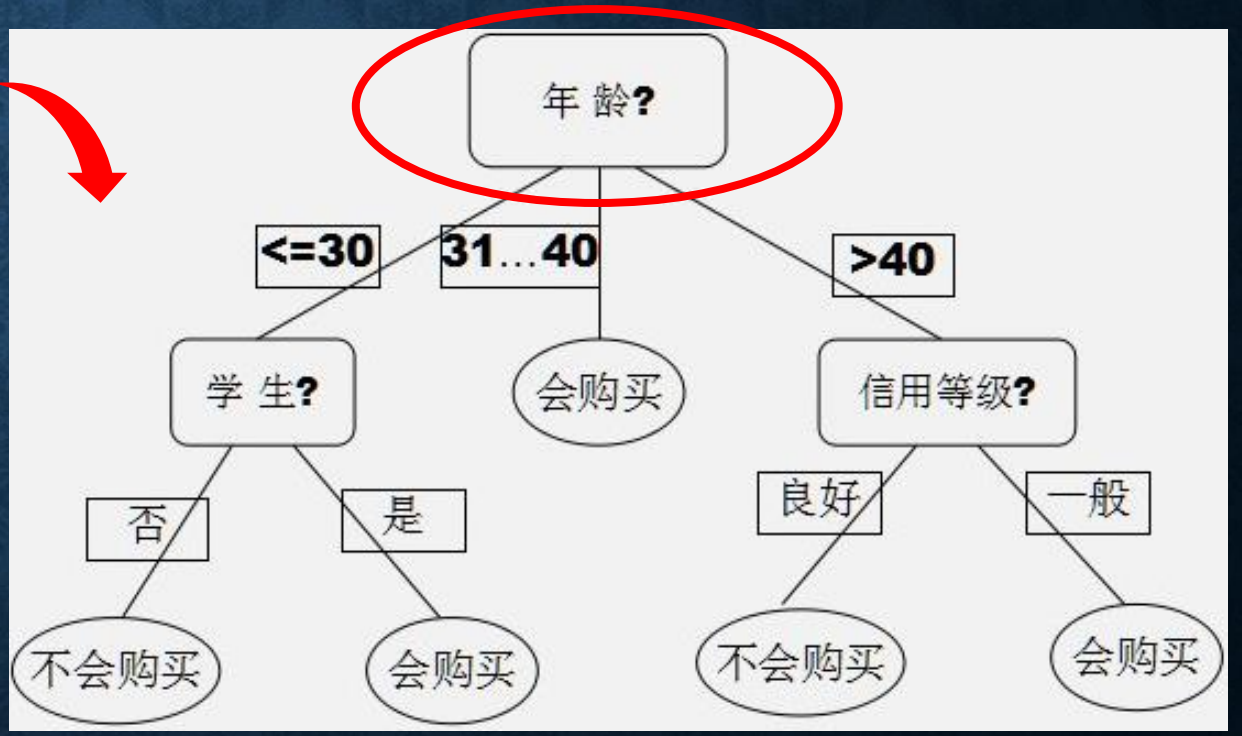
決策樹
DECISION TREE

概念

- 該決策樹方法先根據訓練集數據形成決策樹，如果該樹不能對所有對象給出正確的分類，那麼選擇一些例外加入到訓練集數據中，重複該過程一直到形成正確的決策集。

ID3 EXAMPLE

编号	年龄	收入	学生	信用等级	类别：购买电脑
1	<=30	高	否	一般	不会购买
2	<=30	高	否	良好	不会购买
3	31...40	高	否	一般	会购买
4	>40	中等	否	一般	会购买
5	>40	低	是	一般	会购买
6	>40	低	是	良好	不会购买
7	31...40	低	是	良好	会购买
8	<=30	中等	否	一般	不会购买
9	<=30	低	是	一般	会购买
10	>40	中等	是	一般	会购买
11	<=30	中等	是	良好	会购买
12	31...40	中等	否	良好	会购买
13	31...40	高	是	一般	会购买
14	>40	中等	否	良好	不会购买



三個年紀區間

9個買，5個不買

資訊熵 & 資訊量增益

$$H(D) = -\frac{5}{14}\log_2\frac{5}{14} - \frac{9}{14}\log_2\frac{9}{14} = 0.94$$

買與不買的資訊熵

$$H_{age}(D_{youth}) = -\frac{3}{5}\log\frac{3}{5} - \frac{2}{5}\log\frac{2}{5} = 0.971$$

年輕人、買與不買的資訊熵

$$H_{age}(D) = \frac{5}{14} * 0.971 + \frac{4}{14} * 0 + \frac{5}{14} * 0.971 = 0.694$$

年紀、買與不買的總資訊熵

Gain(age) = 0.94 – 0.694 = 0.246
Gain(student) = 0.94 – 0.789 = 0.151
Gain(credit_rating) = 0.94 – 0.892 = 0.048
Gain(income) = 0.94 – 0.911 = 0.029

參考來源

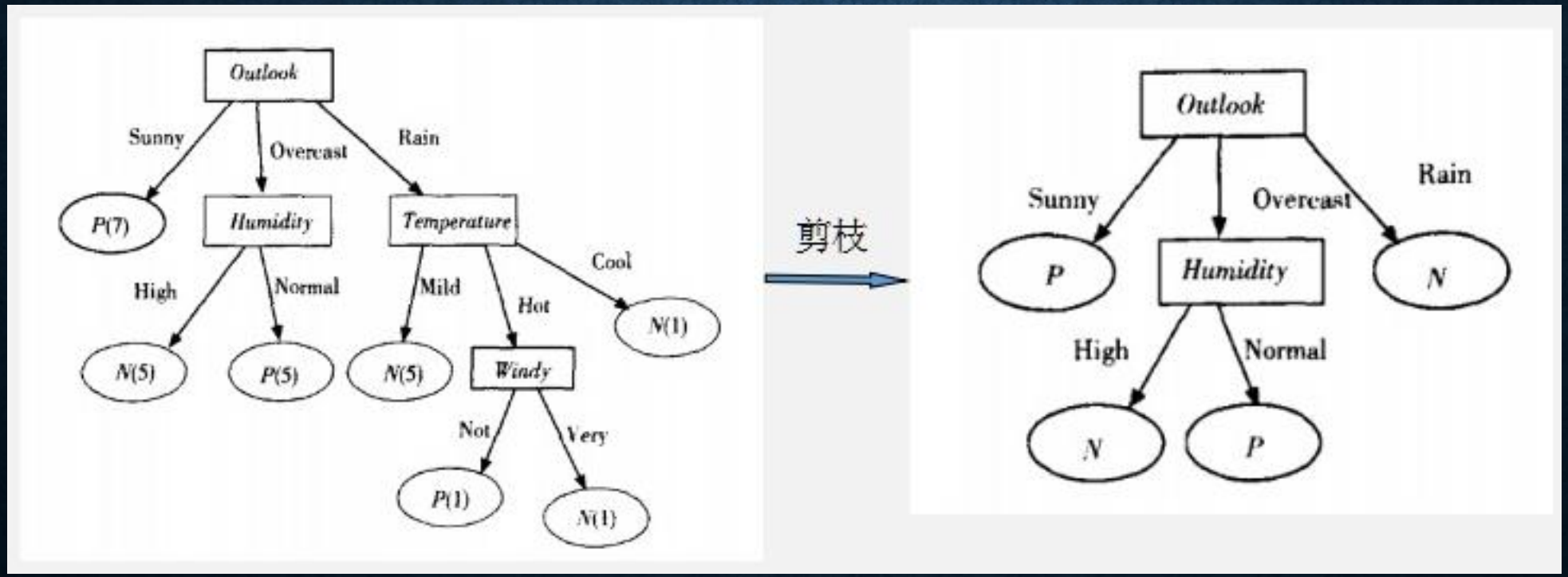
1. Visualizing a Decision Tree - Machine Learning Recipes #2
2. Decision Analysis 3: Decision Trees
3. C4.5決策樹算法

THINKING

重點？

重點

- Overfitting 過度擬合
- Pruning 剪枝



延伸閱讀

1. C4.5 決策樹 (GainRatio)
2. 隨機森林 Random Forest
3. K Nearest Neighbor (KNN)

LINEAR REGRESSION

數值型輸出

概念

某次實驗得到了四個數據點 (x, y)：$(1, 6)$、$(2, 5)$、$(3, 7)$、$(4, 10)$（右圖中紅色的點）。我們希望找出一條和這四個點最匹配的直線 $y = \beta_1 + \beta_2 x$，即找出在某種「最佳情況」下能夠大致符合如下超定線性方程組的 β_1 和 β_2：

$$\begin{aligned}
\beta_1 + 1\beta_2 &= 6 \\
\beta_1 + 2\beta_2 &= 5 \\
\beta_1 + 3\beta_2 &= 7 \\
\beta_1 + 4\beta_2 &= 10
\end{aligned}$$

最小平方法採用的手段是儘量使得等號兩邊的方差最小，也就是找出這個函數的最小值：

$$\begin{aligned}
S(\beta_1, \beta_2) = &[6 - (\beta_1 + 1\beta_2)]^2 + [5 - (\beta_1 + 2\beta_2)]^2 \\
&+ [7 - (\beta_1 + 3\beta_2)]^2 + [10 - (\beta_1 + 4\beta_2)]^2.
\end{aligned}$$

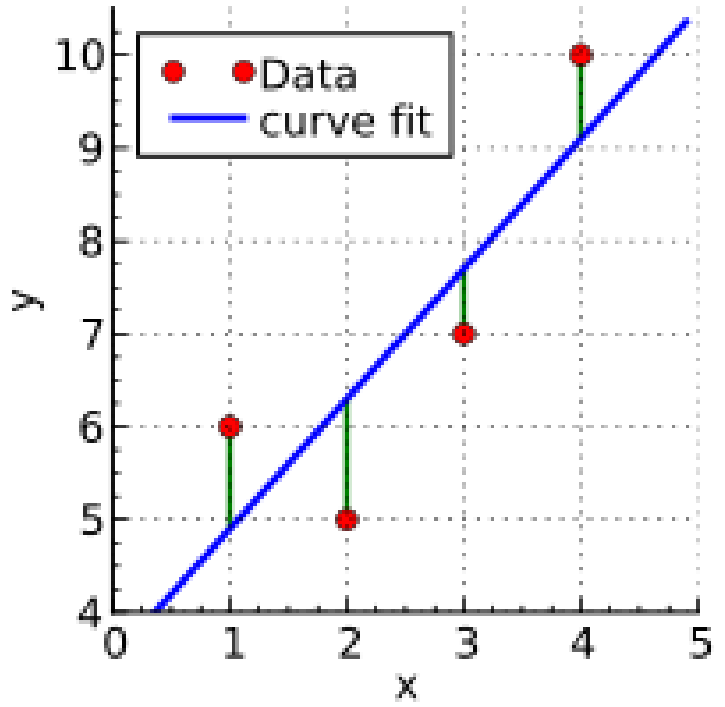
最小值可以通過對 $S(\beta_1, \beta_2)$ 分別求 β_1 和 β_2 的偏導數，然後使它們等於零得到。

$$\begin{aligned}
\frac{\partial S}{\partial \beta_1} &= 0 = 8\beta_1 + 20\beta_2 - 56 \\
\frac{\partial S}{\partial \beta_2} &= 0 = 20\beta_1 + 60\beta_2 - 154.
\end{aligned}$$

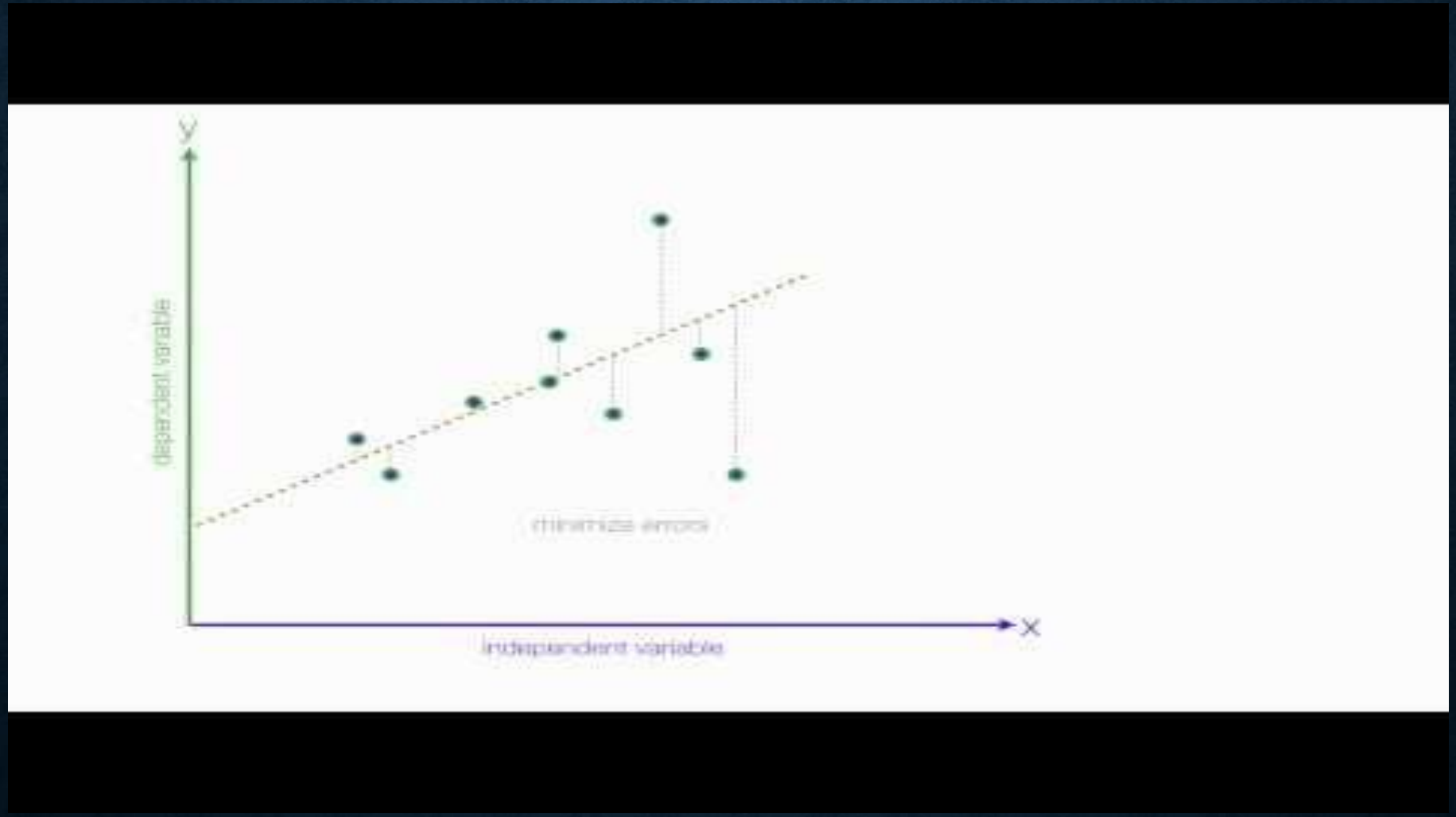
如此就得到了一個只有兩個未知數的方程組，很容易就可以解出：

$$\begin{aligned}
\beta_1 &= 3.5 \\
\beta_2 &= 1.4
\end{aligned}$$

也就是說直線 $y = 3.5 + 1.4x$ 是最佳的。



EXAMPLE


Y
dependent variable
minimize errors
X
independent variable

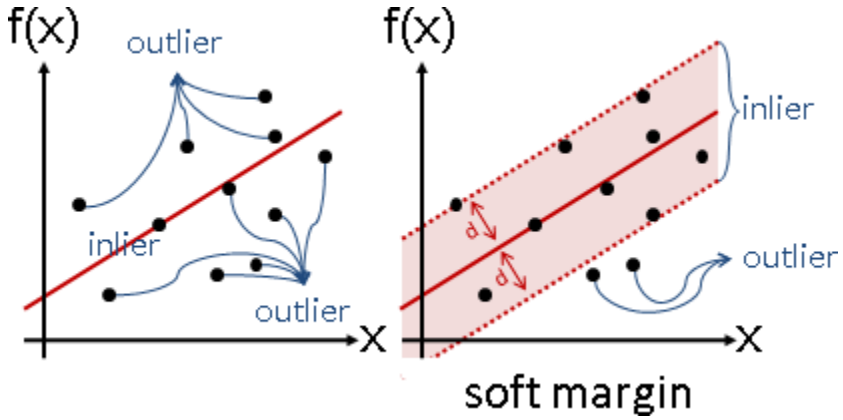
參考來源

1. An Introduction to Linear Regression Analysis
2. 最小平方法
3. http://www.csie.ntnu.edu.tw/~u91029/Regression.html

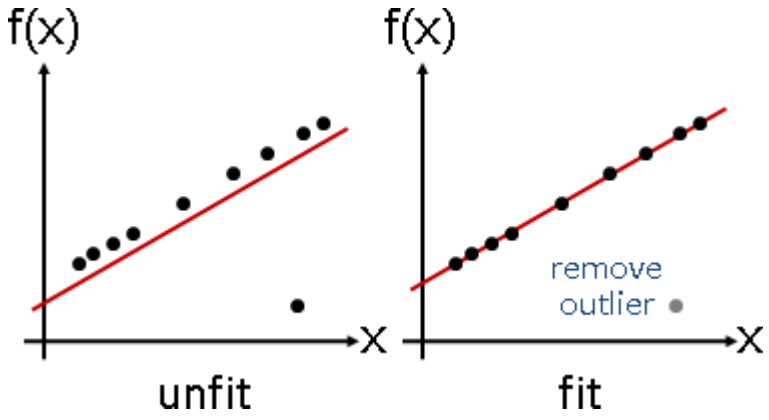
THINKING

重點？

重點



soft margin



延伸閱讀

1. Logistic Regression
2. Support Vector Regression

實作參考

WEKA

1. Tutorial on K Means Clustering using Weka
2. algoritma c4 5 in weka
3. Linear Regression Example in Weka : Weka Tutorias # 4

PYTHON

1. 莫煩- sklearn常用屬性與功能 (Linear Regression 範例)
2. Scikit-Learn 教學：Python 與機器學習

THANK YOU