





### 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

Application of the second		The second second		The Real Property lies	
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	3140	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	3140	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	3140	中等	否	良好	会购买
13	3140	高	是	一般	会购买
14	>40	中等	否	良好	不会购买



#### 資訊熵&資訊量增益

$$H(p_1, \cdots, p_n) = -K \sum_{i=1}^n p_i \log p_i$$

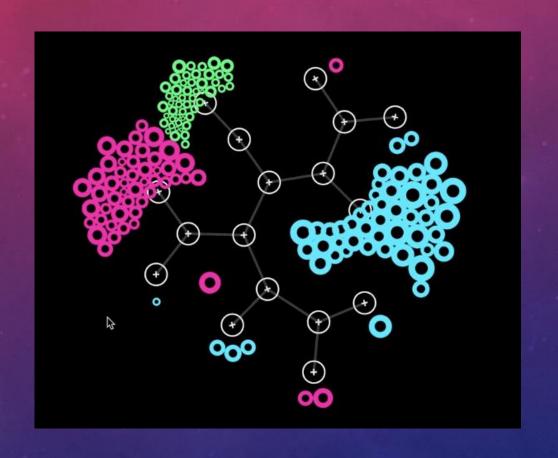
$$H(D) = -\frac{5}{14} \log_2 \frac{5}{14} - \frac{9}{14} \log_2 \frac{9}{14} = 0.94$$
 不買與買的資訊熵

$$H_{age}(D_{youth}) = -\frac{3}{5}\log\frac{3}{5} - \frac{2}{5}\log\frac{2}{5} = 0.971$$
 年輕人、不買與買的資訊熵

$$H_{age}(D) = \frac{5}{14} * 0.971 + \frac{4}{14} * 0 + \frac{5}{14} * 0.971 = 0.694$$
 年紀、不買與買的總資訊熵

Gain(age) = 0.94 - 0.694 = 0.246Gain(student) = 0.94 - 0.789 = 0.151Gain(credit\_rating) = 0.94 - 0.892 = 0.048Gain(income) = 0.94 - 0.911 = 0.029

## 視覺化決策樹



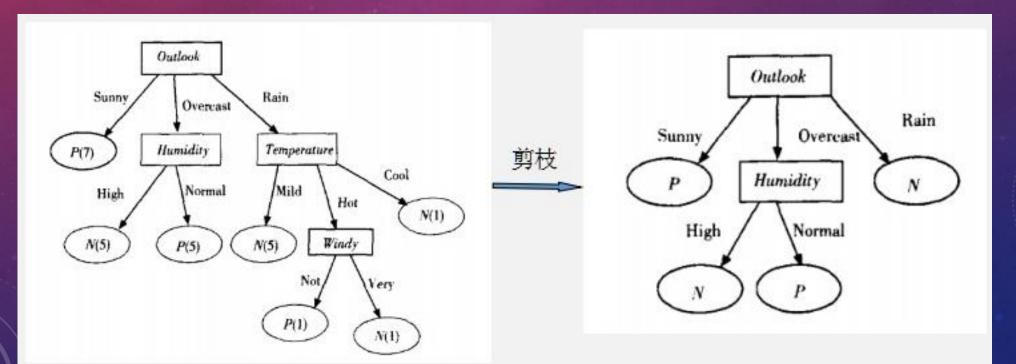
## 參考來源

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



#### 重點

- Overfitting 過度擬合
- Pruning 剪枝



#### 實作範例-Wine Dataset

[1.207e+01, 2.160e+00, 2.170e+00, 2.100e+01, 8.500e+01, 2.600e+00, 2.650e+00, 3.700e-01, 1.350e+00, 2.760e+00, 8.600e-01, 3.280e+00, 3.780e+02]

- (1) Alcohol  $\rightarrow$  1.207e+01
- $(3) Ash \rightarrow 2.170e+00$
- (5) Magnesium  $\rightarrow$  8.500e+01
- (7) Flavanoids  $\rightarrow$  2.650e+00
- (9) Proanthocyanins  $\rightarrow$  1.350e+00
- $(11) \text{Hue} \rightarrow 8.600 \text{e-} 01$
- $(13) Proline \rightarrow 3.780e + 02$

- (2) Malic acid  $\rightarrow$  2.160e+00
- (4) Alcalinity of ash  $\rightarrow$  2.100e+01
- (6) Total phenols  $\rightarrow$  2.600e+00
- (8) Nonflavanoid phenols  $\rightarrow$  3.700e-01
- (10)Color intensity  $\rightarrow$  2.760e+00
- (12)OD280/OD315 of diluted wines  $\rightarrow$  3.280e+00

# 延伸閱讀

- 1. C4.5 決策樹 (GAINRATIO)
- 2. 隨機森林 RANDOM FOREST
- 3. K NEAREST NEIGHBOR (KNN)



### 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  $y=eta_1+eta_2x$ ,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  $eta_1$  和  $eta_2$ :

$$\beta_1 + 1\beta_2 = 6$$

$$\beta_1 + 2\beta_2 = 5$$

$$\beta_1 + 3\beta_2 = 7$$

$$\beta_1 + 4\beta_2 = 10$$

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:

$$egin{aligned} S(eta_1,eta_2) = & [6-(eta_1+1eta_2)]^2 + [5-(eta_1+2eta_2)]^2 \ & + [7-(eta_1+3eta_2)]^2 + [10-(eta_1+4eta_2)]^2. \end{aligned}$$

最小值可以通過對  $S(eta_1,eta_2)$  分別求  $eta_1$  和  $eta_2$  的偏導數,然後使它們等於零得到。

$$rac{\partial S}{\partial eta_1} = 0 = 8eta_1 + 20eta_2 - 56$$

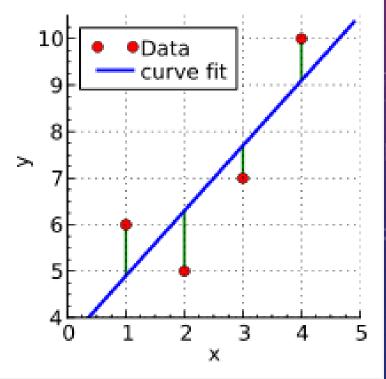
$$rac{\partial S}{\partial eta_2} = 0 = 20eta_1 + 60eta_2 - 154.$$

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:

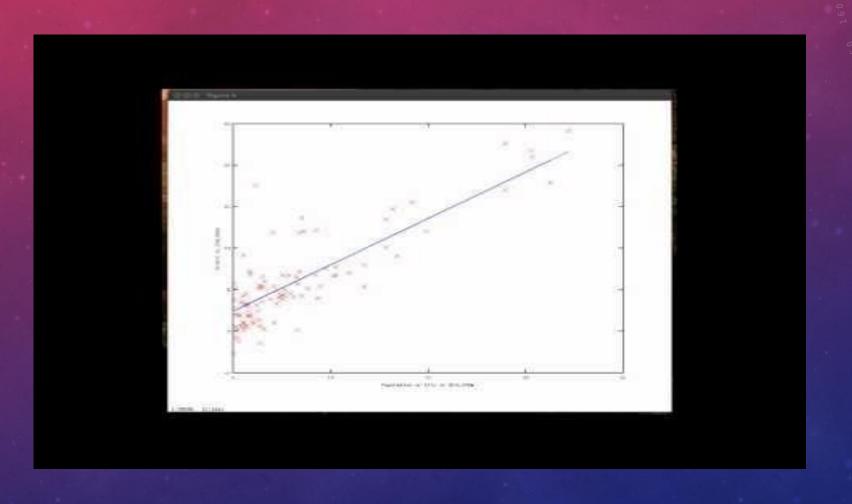
$$\beta_1 = 3.5$$

$$\beta_2 = 1.4$$

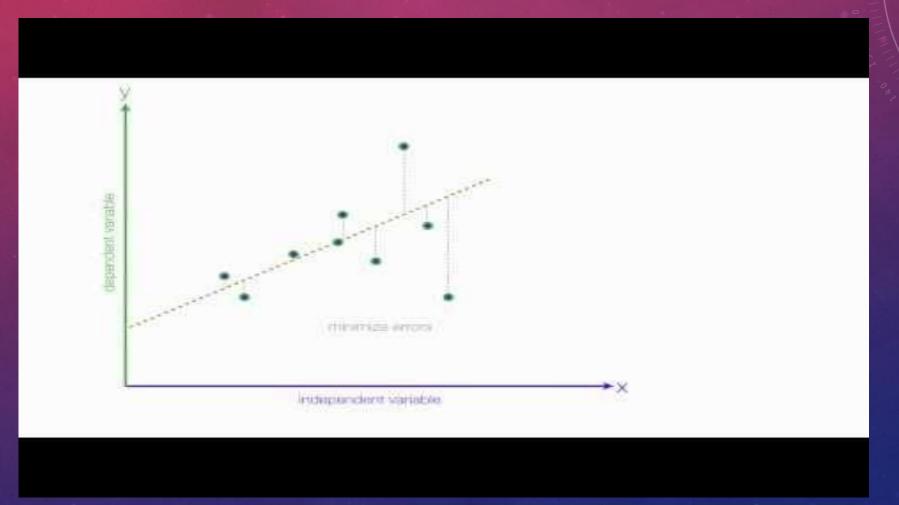
也就是說直線 y = 3.5 + 1.4x 是最佳的。



# 視覺化線性迴歸



## EXAMPLE



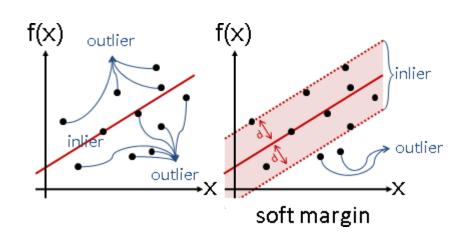
16

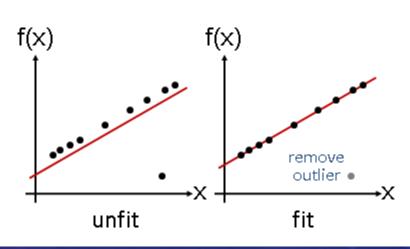
## 參考來源

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



#### 重點









#### **WEKA**

- 1. Tutorial on K Means Clustering using Weka
- 2. <u>algoritma c4 5 in weka</u>
- 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

## 實作範例

- 分類任務 kNN for Wine Data [link]
- 分類任務 Decision Tree for Wine Data [link]
- 數值預測 Linear Regression for Boston Price [link]

