

決策樹 與 隨機森林

Decision Tree and Random Forest

國立東華大學電機工程學系 楊哲旻

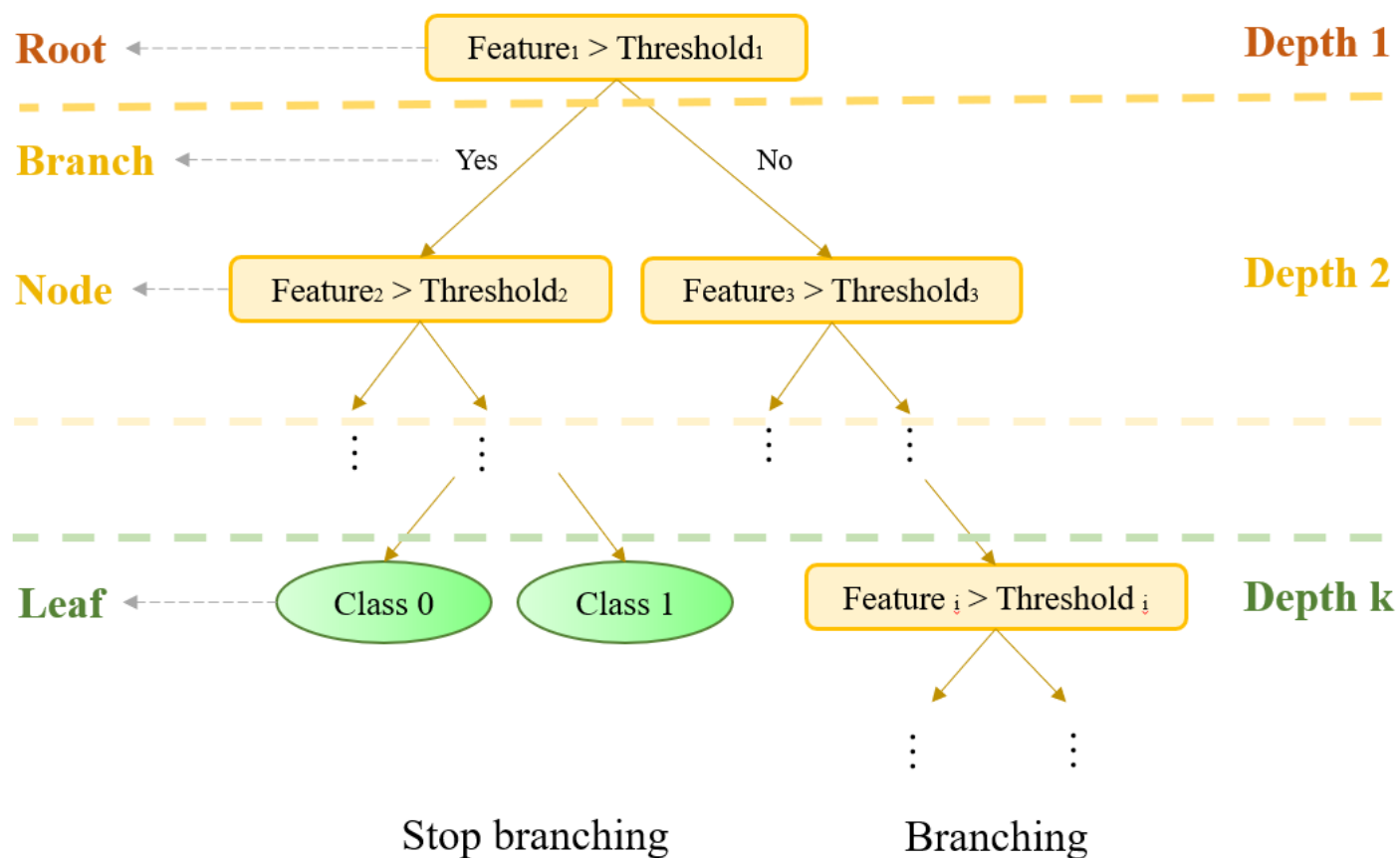
Outline



- 1 決策樹
- 2 交叉熵
- 3 基尼不純度
- 4 剪枝
- 5 隨機森林
- 6 集成學習

決策樹為監督式學習，類似於流程圖的樹結構 (根節點、節點、分支、樹葉)。

決策樹分支標準則有交叉熵與基尼不純度兩種方法。



交叉熵

交叉熵由R. Quinlan提出的分支架構，有處理離散與連續特徵的ID3與C4.5模型，皆根據交叉熵所計算的訊息增益量(Information Gain)來決定節點，透過A來做為節點分類獲取了多少訊息。訊息增益量越大，則這個特徵的分類性越好

$$\text{Information Gain}(A) = \text{Info}(D) - \text{Info}_A(D)$$

$$\text{Info}_A(D) = \sum_{i=1}^k \frac{N_i}{N} \text{Info}(D) \quad \text{Info}(D) = \sum_{i=1}^q p_i \log_r \frac{1}{p_i}$$

其中A為一輸入特徵， q 為該標籤的類別數量， p_i 為該節點分支後樣本根據標籤計算的機率值， N 為該節點的總樣本數， N_i 為該節點分支後的樣本數， k 為該特徵的分支數量

範例：預測顧客是否買電腦？

RID	Age	Income	Student	Credit rating	Class: buy computer
1	Youth	High	No	Fair	No
2	Youth	High	No	Excellent	No
3	Middle aged	High	No	Fair	Yes
4	Senior	medium	No	Fair	Yes
5	Senior	Low	Yes	Fair	Yes
6	Senior	Low	Yes	Excellent	No
7	Middle aged	Low	Yes	Excellent	Yes
8	Youth	Medium	No	Fair	No
9	Youth	Low	Yes	Fair	Yes
10	Senior	Medium	Yes	Fair	Yes
11	Youth	Medium	Yes	Excellent	Yes
12	Middle aged	Medium	No	Excellent	Yes
13	Middle aged	High	Yes	Fair	Yes
14	Senior	Medium	No	Excellent	No

機器學習 – 決策樹

02. 交叉熵

- 如果還沒使用節點來分類時，訊息熵(Entropy)為：

$$Info(D) = -\left[\frac{9}{14} \log_2 \left(\frac{9}{14}\right) + \frac{5}{14} \log_2 \left(\frac{5}{14}\right)\right] = 0.940 \text{ bits}$$

- 如果根據**年齡**來分類，訊息熵(Entropy)為：

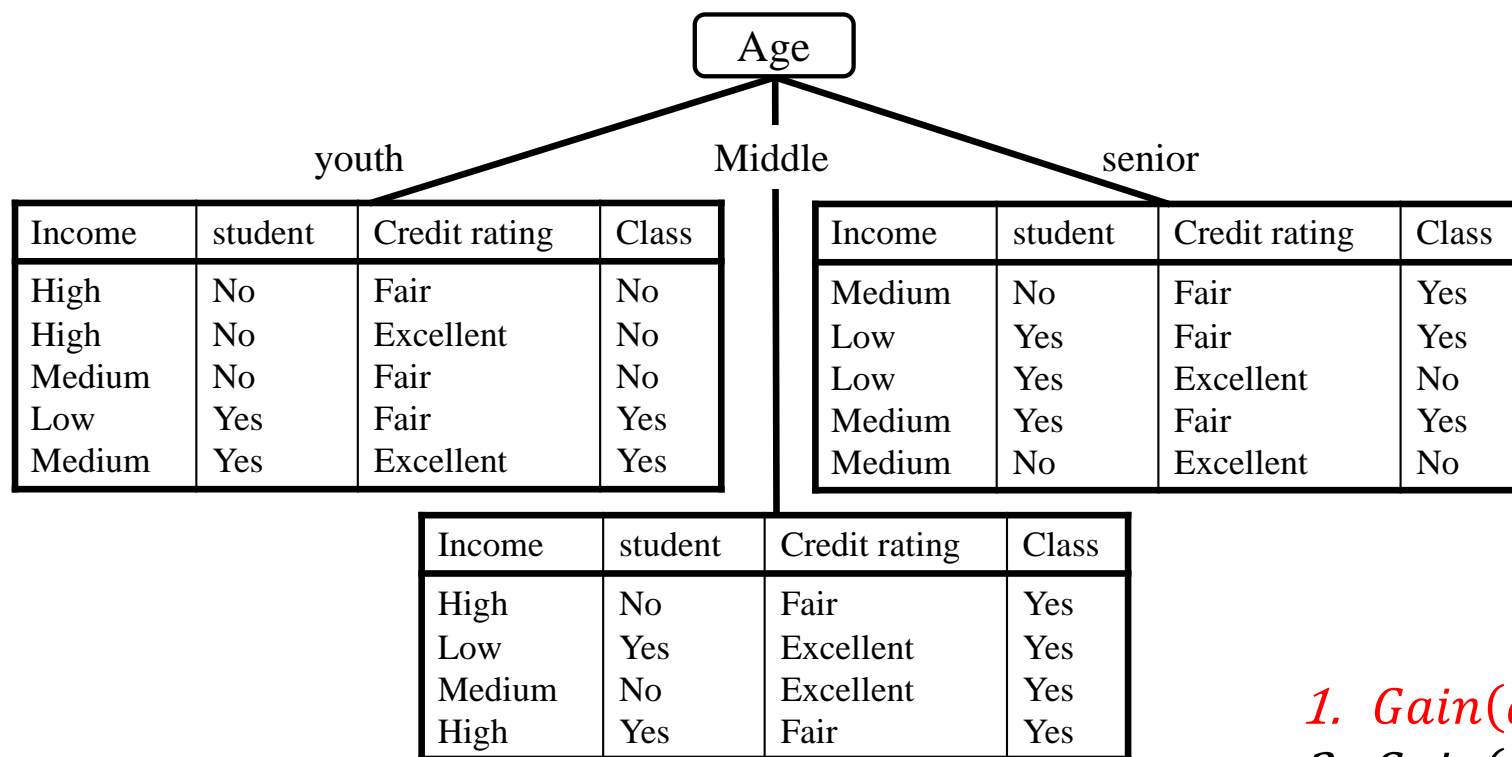
$$\begin{aligned} Info_{age}(D) &= \frac{5}{14} \times \left[-\left(\frac{2}{5} \log_2 \left(\frac{2}{5}\right) + \frac{3}{5} \log_2 \left(\frac{3}{5}\right)\right) \right] \\ &\quad + \frac{4}{14} \times \left[-\left(\frac{4}{4} \log_2 \left(\frac{4}{4}\right) + \frac{0}{4} \log_2 \left(\frac{0}{4}\right)\right) \right] \\ &\quad + \frac{5}{14} \times \left[-\left(\frac{3}{5} \log_2 \left(\frac{3}{5}\right) + \frac{2}{5} \log_2 \left(\frac{2}{5}\right)\right) \right] \\ &= 0.694 \text{ bits} \end{aligned}$$

$Gain(age) = Info(D) - Info_{age}(D) = 0.246 \text{ bits}$ ，其餘特徵類推。

機器學習 – 決策樹

02. 交叉熵

5



取最大訊息增益量作為
第一個根節點，即age

1. $Gain(age) = 0.246 \text{ bits}$
2. $Gain(income) = 0.029 \text{ bits}$
3. $Gain(student) = 0.151 \text{ bits}$
4. $Gain(credit \text{ rating}) = 0.048 \text{ bits}$

基尼不純度

基尼不純度是由L. Breiman所提出處理連續特徵的CART(Classification And Regression Tree)模型，計算基尼指數來決定節點

$$Gini(A) = \sum_{i=1}^2 \frac{N_i}{N} gini(A) \quad gini(A) = 1 - \sum_{j=1}^2 p_j^2$$

其中 A 為一輸入特徵， N 為該節點的總樣本數， N_i 為該節點分支後的樣本數， p_i 為該節點分支後樣本根據標籤計算的機率值


分支停止

1. 給定節點的所有樣本屬於同一類
2. 沒有剩餘特徵可以用來進一步劃分樣本或是已經剪枝，在此情況下使用多數表決

剪枝

由於決策樹的複雜度會隨著分支深度而增大，越深層的樹雖然在訓練集的預測能力高，但對於測試集或是新資料中的擬合程度可能偏低，因此需要做剪枝(Pruning)處理：

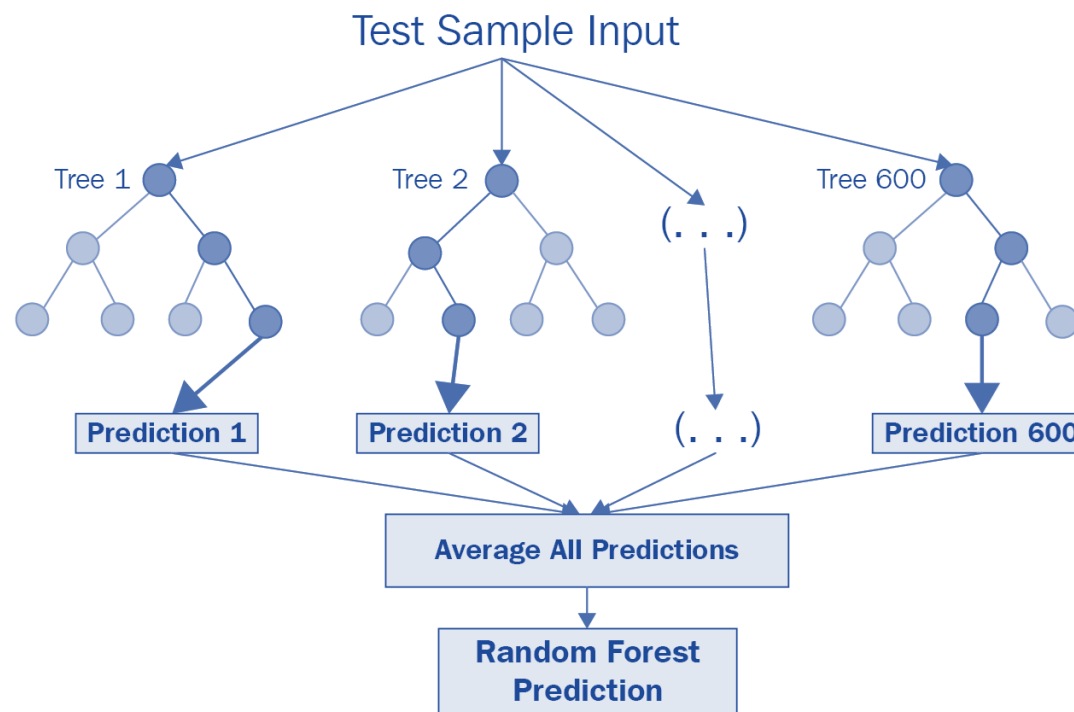
- 樹分支前就指定深度的**預剪枝(Pre-pruning)**
- 生長完的樹針對細節剪枝處理的**後剪枝(Post-pruning)**

 決策樹的超參數為剪枝深度

由於生長完的決策樹會有低偏差與高方差的特點導致過度擬合，L. Breiman提出裝袋(Bagging)的隨機森林之集成學習(Ensemble learning)架構，利用**取樣特徵與訓練集資料**來生成多個不同決策樹，最終預測以**多數決**方式來決定

🔗 隨機森林的超參數為

- 取樣比例
- 剪枝深度
- 決策樹棵樹





決策樹 與 隨機森林 — 實作

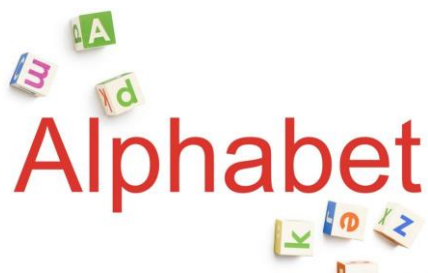
樹類模型 – 實作

A-Z Handwritten Alphabets Database

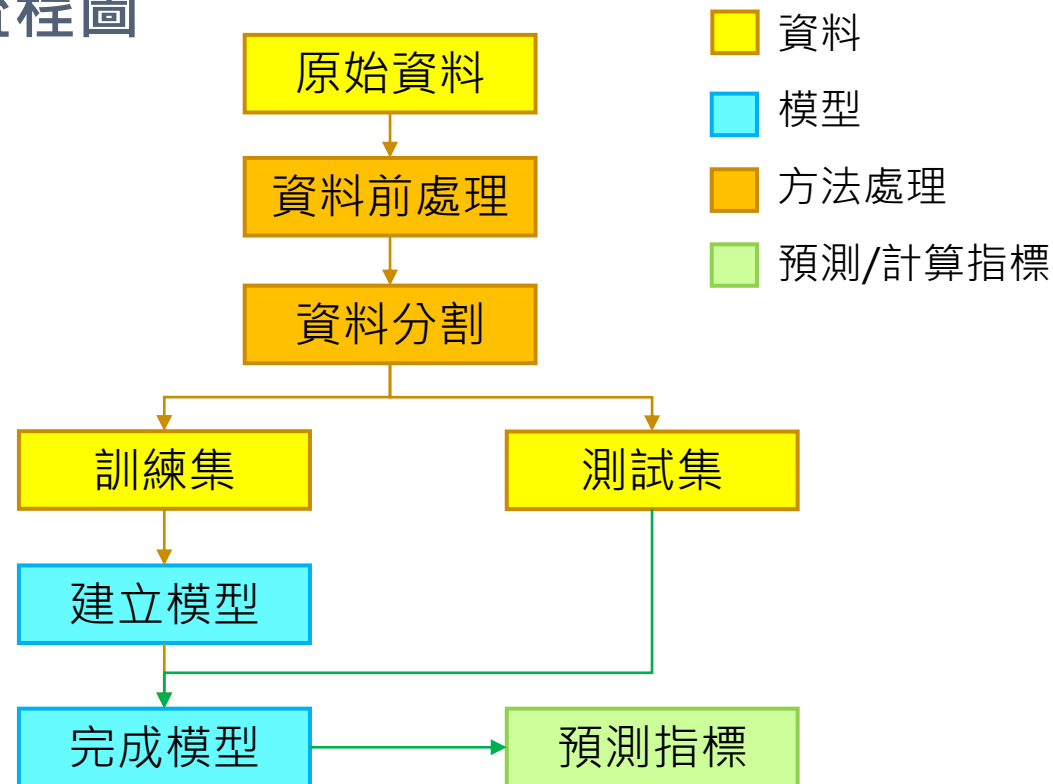
Kaggle Dataset

<https://www.kaggle.com/sachinpatel21/az-handwritten-alphabets-in-csv-format>

- 數據集包含26個手寫字母A-Z，大小為28×28的手寫灰階影像，共372451張，並以.csv檔儲存



流程圖



應用端

