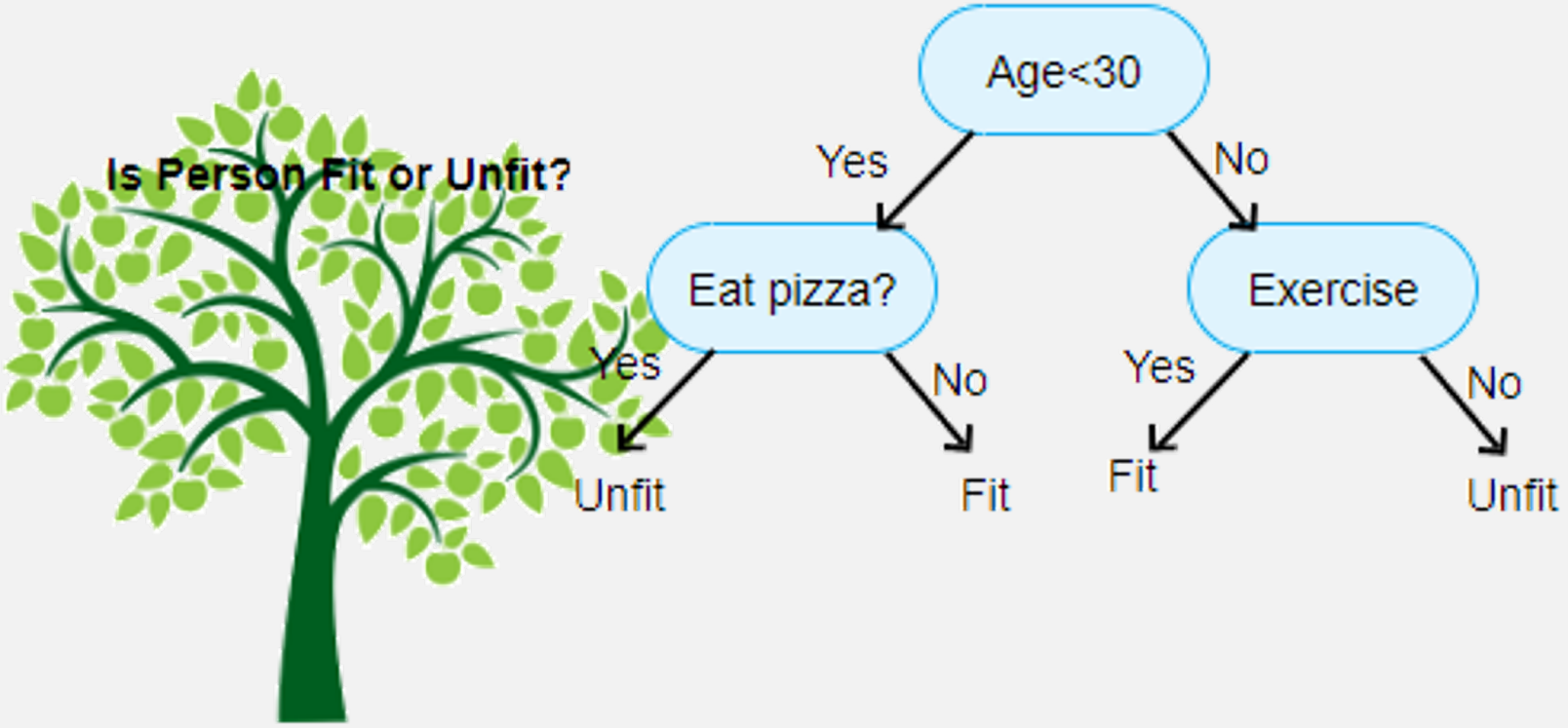
決策樹系列演算法

輔大智慧金融實驗室 黃弘開
Email: kai830227@gmail.com

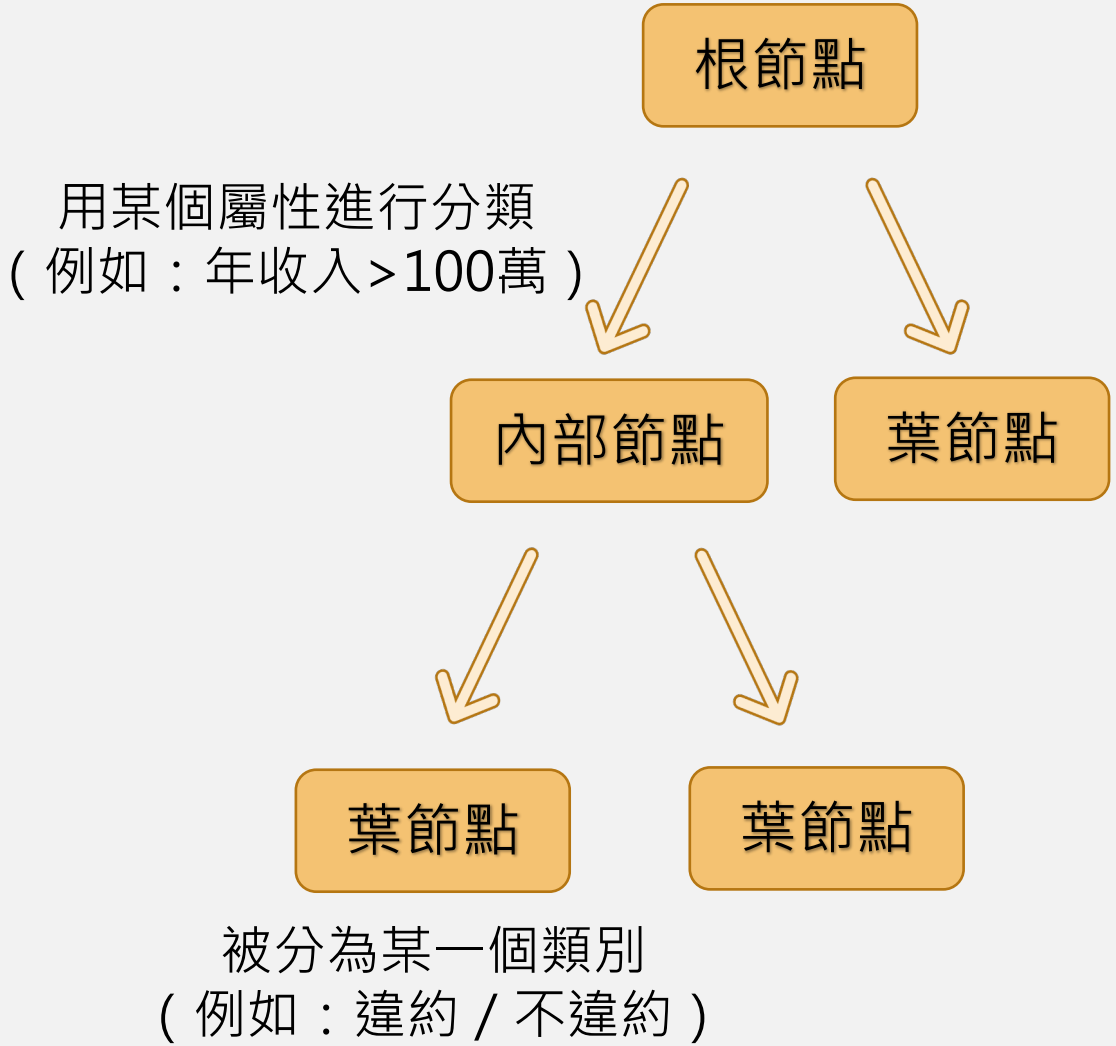
決策樹

當我們想要推論一個人胖不胖?



決策樹結構

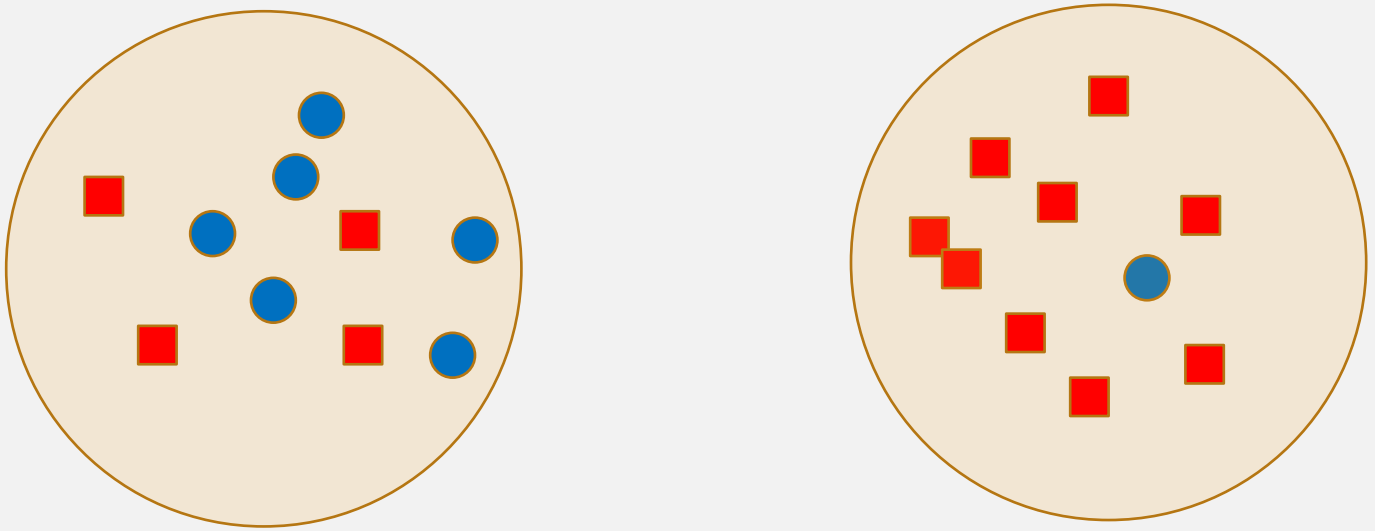
在每個節點中，找出分類能力最好的屬性作為分類依據，分類後產出分支，遞迴上述過程直到滿足終止條件。



如何衡量分類效果的好壞?

分割結果中，若具有較高同質性類別的節點，分割結果較佳，
以節點的不純度當作衡量屬性分類能力的標準（越低越好）

哪一個分類有較低的不純度？



要怎麼長出一顆決策樹?

生成一顆決策樹，需要考慮三個層面

如何生長（分割指標）

訓練模型時，依據什麼樣的分割指標去劃分

如何停止（剪枝規則）

若沒有任何停止劃分的規則，容易發生過擬合

預剪枝：決策樹停止生成的規則

後剪枝：樹生成後，以某種規則剪枝

如何處理缺失值

當選擇劃分屬性時，缺失值該怎麼處理

要對測試樣本進行劃分時，缺失值怎麼處理

ID3

如何生長
（分割指標）

如何停止
（剪枝規則）

如何處理缺失值

計算該節點的熵和所有分割屬性分割後的熵，取資訊增益最大的屬性作為分割屬性。

$$\textbf{Entropy(S)=} \sum_{i=1}^{c} - p_i \log_2 p_i$$

$$Gain(S, A) = Entropy(S) - \sum_{j=1}^{v} \frac{|S_j|}{|S|} Entropy(S_j)$$

缺點：資訊增益準則會偏好分類較多的特徵，例如「交易編號」

C4.5

如何生長
（分割指標）

如何停止
（剪枝規則）

如何處理缺失值

以資訊增益率作為屬性分類標準，在資訊增益的基礎上加入懲罰項，克服ID3演算法偏好分類較多的特徵的缺點

資訊增益率 = 資訊增益 ÷ 分類屬性的分割資訊值

$$GainRatio(S,A) = \frac{Gain(S,A)}{SplitInformation(S,A)}$$

$$SplitInformation(S,A) = -\sum_{i=1}^{c} \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|}$$

C4.5

如何生長 (分割指標)	資訊增益率 (Gain Ratio)
如何停止 (剪枝規則)	採用悲觀剪枝法 (Pessimistic Error Pruning) 由下往上檢視每一個內部節點，如果用葉節點替換後分類錯誤率下降，則進行剪枝。
如何處理缺失值	• 計算資訊增益時須乘上無缺失樣本的佔比 • 劃分樣本到子節點時，需要加入權重

CART

如何生長
（分割指標）

如何停止
（剪枝規則）

如何處理缺失值

以吉尼係數（吉尼不純度）作為屬性分類標準。

吉尼係數表示隨機抽取兩樣本，兩者類別不同的機率，吉尼係數越低表示樣本不純度越低。

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

$$Gini_A(S) = \frac{|S_1|}{|S|} Gini(S_1) + \frac{|S_2|}{|S|} Gini(S_2)$$

CART

如何生長
（分割指標）

如何停止
（剪枝規則）

如何處理缺失值

採用代價複雜度剪枝

1. 選擇代價複雜度最小的內部節點進行剪枝，每次剪枝後會生成一顆子樹，不斷進行剪枝直到剩下根節點
2. 使用測試資料評估所有樹，選出分類效果最佳的樹

$$\alpha = \frac{R(T) - R(T_t)}{L(T) - 1}$$

$$R(T) = e(T) \times p(T)$$

$$R(T_t) = \sum_{t \in T} e(t) \times p(t)$$

CART

如何生長 (分割指標)	吉尼係數 (Gini Index)
如何停止 (剪枝規則)	代價複雜度剪枝 (Cost-Complexity Pruning)
如何處理缺失值	• 計算吉尼係數時須乘上無缺失樣本的佔比 • 使用替代劃分屬性劃分樣本到子節點 • sklearn的CART沒有考慮缺失值

CART

CART 演算法也可用於迴歸問題

1. 於任意一點作為劃分點，計算子節點目標變數的平均值作為該節點的輸出值
2. 計算均方差 (MSE)
3. 選擇 MSE 最小的劃分點作為劃分屬性

決策樹演算法比較

	ID3	C4.5	CART
應用場景	分類	分類	分類、迴歸
樣本類型	離散型	離散型、連續型	離散型、連續型
分割標準	資訊增益	資訊增益率	吉尼係數
剪枝策略	無	後剪枝	後剪枝
缺失值處理	無，對缺失值敏感	有	有
運算速度	慢	慢	快

決策樹的優缺點

優點

- 可以處理非線性可分的分類問題
- 適用於類別型變數
- 易於解釋（可視覺化呈現樹的結構）

缺點

- 樣本發生一點變動，就會導致樹結構的改變
- 容易 overfitting

隨機森林

隨機森林

隨機森林是由多顆 CART樹所組成，且每個樹是互相獨立的，透過每棵樹投票作為最終的分類結果。

1. 建立N個決策樹
2. 隨機抽樣樣本
3. 隨機選擇features
4. 以所有數的平均值當預測結果

隨機森林

Id	Survive	Pclass	Age	Fare
1	0	3	22	7.25
2	1	1	38	71.28
3	1	3	26	7.93
4	1	1	35	53.10
5	1	3	35	8.05
6	0	3	49	8.46
7	1	1	54	51.86
8	1	3	2	21.08
9	1	3	27	11.13
10	1	2	14	30.07

Id	Survive	Pclass	Age	Fare
1	0	3	22	7.25
2	1	1	38	71.28
3	1	3	26	7.93
4	1	1	35	53.10
5	1	3	35	8.05
6	0	3	49	8.46
7	1	1	54	51.86
8	1	3	2	21.08
9	1	3	27	11.13

Id	Survive	Pclass	Age	Fare
1	0	3	22	7.25
2	1	1	38	71.28
3	1	3	26	7.93
4	1	1	35	53.10
5	1	3	35	8.05
6	0	3	49	8.46
7	1	1	54	51.86
8	1	3	2	21.08
9	1	3	27	11.13

Id	Survive	Pclass	Age	Fare
1	0	3	22	7.25
2	1	1	38	71.28
3	1	3	26	7.93
4	1	1	35	53.10
5	1	3	35	8.05
6	0	3	49	8.46
7	1	1	54	51.86
8	1	3	2	21.08
9	1	3	27	11.13

隨機森林

survived

deceased

survived

survived

Predict: Survived

隨機森林

生存機率
30%

生存機率
87%

生存機率
63%

生存機率
28%

Predict: 52%