Day 15

決策樹 (分類器)

[全民瘋AI系列]

第12屆 iT邦幫忙 鐵人賽

Day 15 學習目標

01 **決策樹演算法介紹**

決策樹如何生成？如何處理分類問題？

02 **實作決策樹分類器**

觀察決策樹是如何生成的

Part 1

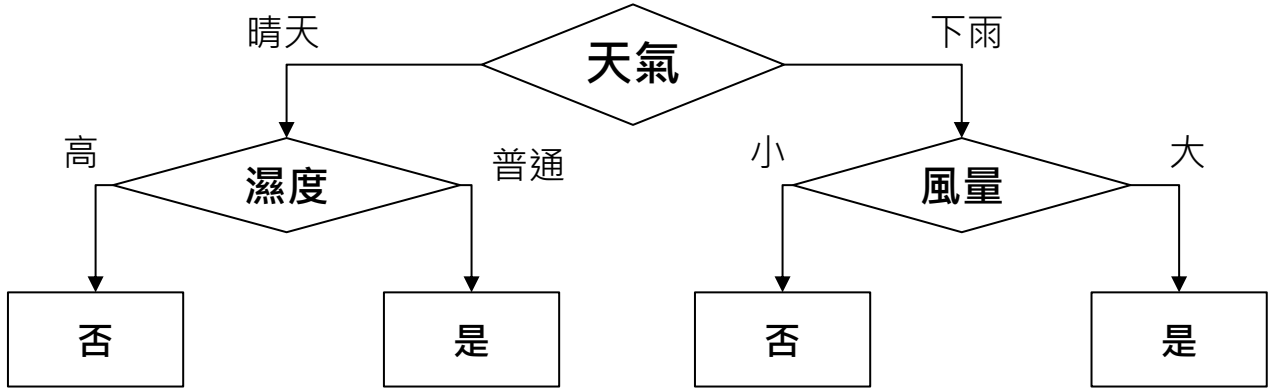
決策樹 (分類器)
觀念講解

第12屆 iT邦幫忙 鐵人賽

決策樹 Decision trees

- 將特徵以條件判斷的方式決定答案

天氣	濕度	風量	是否舉行
晴天	高	大	否
陰天	低	小	是

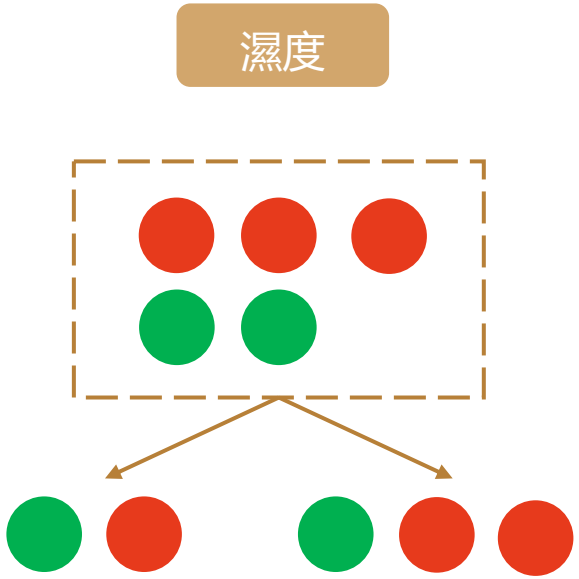


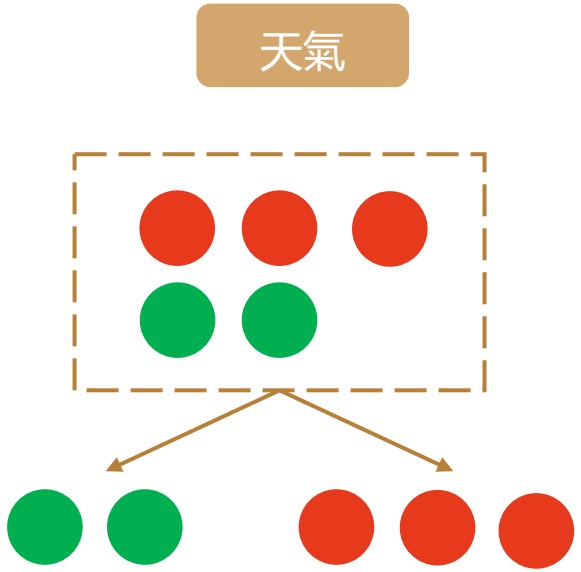
決策樹如何生成？

- 決策樹以貪婪法則來決定每一層要問什麼問題
- 尋找最有利的特徵可以快速的做分類

正常舉行

取消舉行





決策樹的混亂評估指標

我們需要客觀的標準來決定決策樹的分支

- Information gain (資訊獲利)
- Gain ratio (吉尼獲利)
- Gini index (吉尼係數) = Gini Impurity (吉尼不純度)

評估分割資訊量

透過從訓練資料找出規則，讓每一個決策能夠使訊息增益最大化。

- 資訊獲利 (Information Gain)
- Gini不純度 (Gini Impurity)

$$Entropy = -\sum_j p_j \log_2 p_j$$

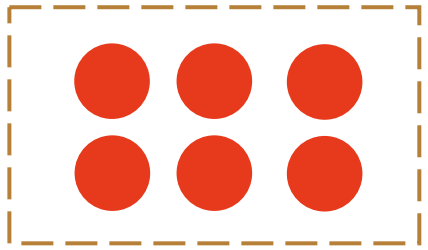
$$Gini = 1 - \sum_j p_j^2$$

熵 (Entropy)

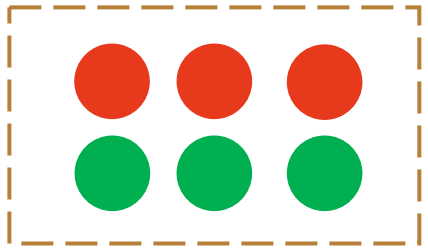
Entropy = $-\sum p_j \log_2 p_j$

Information Gain = $-p * \log_2 p - q * \log_2 q$

p：是的機率　q：否的機率



Info(6, 0) = $-\frac{6}{6}\log_2(\frac{6}{6}) - \frac{0}{6}\log_2(\frac{0}{6})$
= 0



Info(3, 3) = $-\frac{3}{6}\log_2(\frac{3}{6}) - \frac{3}{6}\log_2(\frac{3}{6})$
= 1

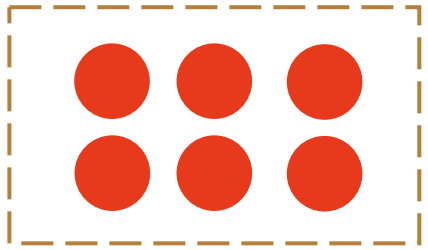
當所有的資料都是相同一致，它們的Entropy就是0
如果資料各有一半不同，那麼Entropy就是 1

Gini不純度 (Gini Impurity)

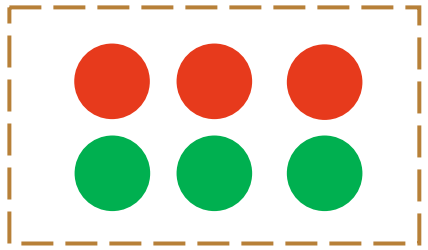
Gini = $1-\sum p_j{}^2$

Gini Impurity = $1-(p^2+q^2)$

p：是的機率　q：否的機率



Info(6, 0) = $1-(1^2+0^2)$
= 0



Info(3, 3) = = $1-(0.5^2+0.5^2)$
= 0.5

數字越大，代表序列中的資料越混亂

決策樹模型的優缺點

優點

- 簡單且高度可解釋性
- 低計算時間複雜度
- 每個決策階段都相當的明確清楚
- 幾乎沒有要調整的超參數

缺點

- 模型容易過度擬合
- 當標籤類別種類多時樹會很複雜

觀察決策樹是如何生成

Example : 鳶尾花朵

Part 2

決策樹 (分類器)

程式實作

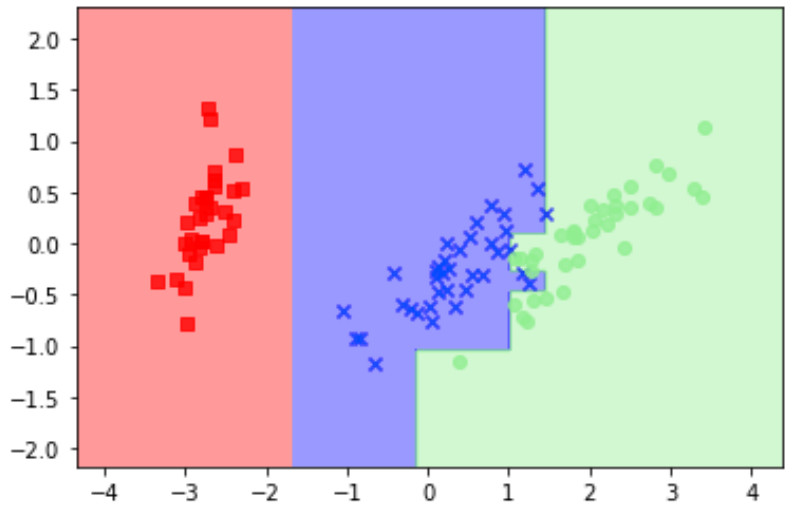
Decision Tree Classifier

```python
from sklearn.tree import DecisionTreeClassifier

# 建立Logistic模型
decisionTreeModel = DecisionTreeClassifier(criterion = 'entropy', max_depth=6, random_state=42)
# 使用訓練資料訓練模型
decisionTreeModel.fit(train_reduced, y_train)
# 使用訓練資料預測分類
predicted = decisionTreeModel.predict(train_reduced)
```

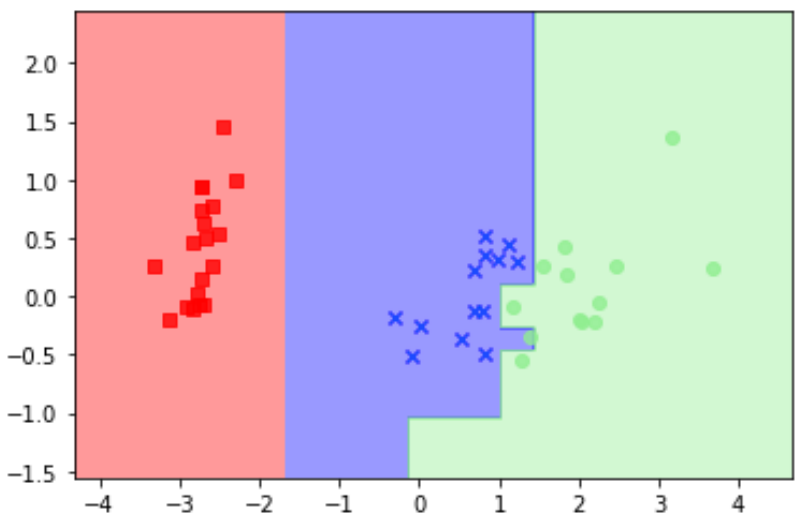
訓練集

```
train set accurancy:  1.0
```



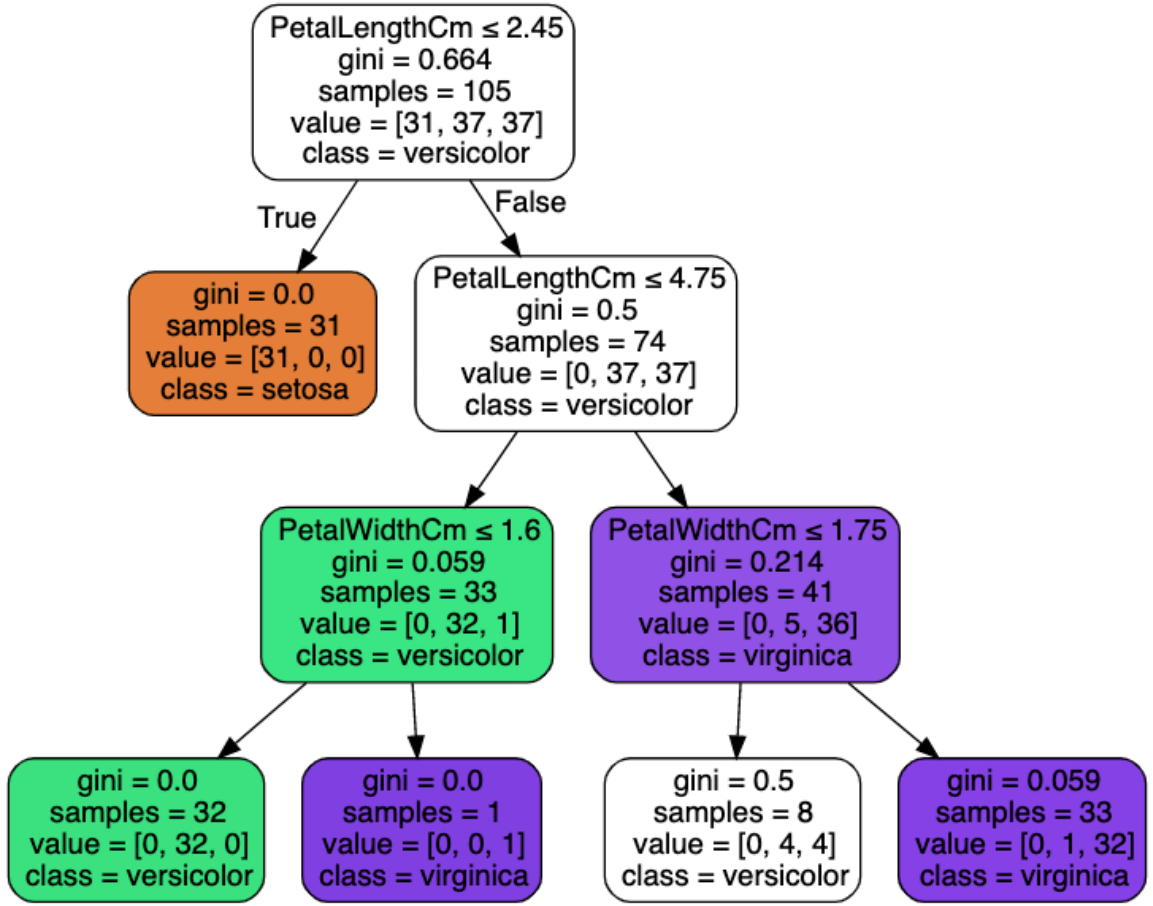
測試集

```
test set accurancy:  0.9777777777777777
```



視覺化決策樹

Graphviz 是視覺化決策樹的套件，可參考 Graphviz 官網的介紹。以下範例使用 iris 資料集，採用四個特徵下去做訓練，並繪製一棵樹。



Thanks

PRESENTED BY 10程式中

第12屆 iT邦幫忙 鐵人賽