



機器學習基礎與演算法

Chapter 6 決策樹系列 (Tree Based Model)

[講師投影片Chapter6](#)

[課程投影片](#)

[資料與程式碼](#)

[播放清單](#)



「版權聲明頁」

本投影片已經獲得作者授權台灣人工智慧學校得以使用於教學用途，如需取得重製權以及公開傳輸權需要透過台灣人工智慧學校取得著作人同意；如果需要修改本投影片著作，則需要取得改作權；另外，如果有需要以光碟或紙本等實體的方式傳播，則需要取得人工智慧學校散佈權。

課程內容

6. 決策樹系列 (Decision tree)

- Decision tree
- Information gain
- Gain ratio
- Gini index
- Tree pruning
- Miscellaneous topics

[實作] 決策樹 (Decision Tree)

Code 放在Hub中的course內

- 為維護課程資料, courses中的檔案皆為read-only, 如需修改請cp至自身環境中
- 打開terminal, 輸入

`cp -r courses-tpe/Machine_Learning <存放至本機的名稱>`



Chapter 6 決策樹

(Decision tree)

- 範例程式(example)的檔名會以藍色字體顯示且旁邊附上
- 練習(exercise)的檔案以紅色字體顯示且旁邊附上

06-1: Decision tree

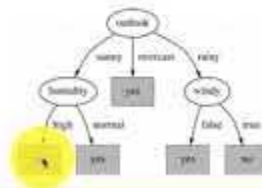


Decision tree introduction (1/2)

- Training set

Feature vector (x_i)				y_i : +1: Yes, -1: No
Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Med	Normal	False	Yes

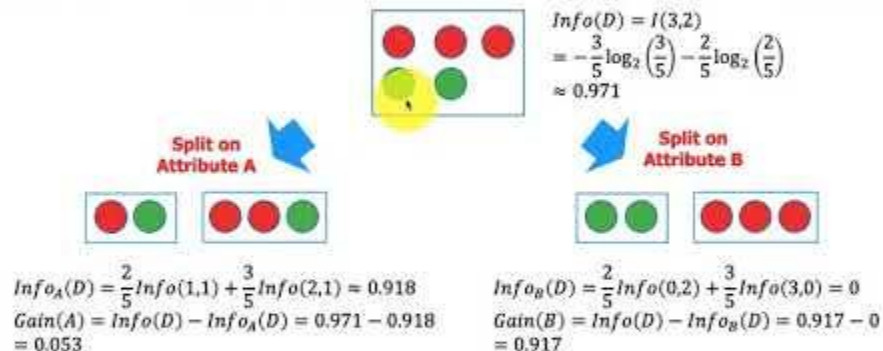
- Learned decision tree



06-2: Information gain



Information gain



06-3: Quiz



Quiz

- Which of the following has a higher entropy?
 - (O,O,X,X) vs (O,O,X,X,X) vs (O,O,O)
- Which attribute (A or B) will be selected by a decision tree classifier based on information gain?



- Can we apply Decision Tree Classifier on the datasets with only numerical attributes?



06-4: Answer



Quiz

- Which of the following has a higher entropy?
 - (O,O,X,X) vs (O,O,X,X,X) vs (O,O,O)
- Which attribute (A or B) will be selected by a decision tree classifier based on information gain?



- Can we apply Decision Tree Classifier on the datasets with only numerical attributes?



06-5: Gain ratio



Attribute Selection: Gain ratio

age	income	student	credit_rating	buys_computer
youth	high	no	fair	no
youth	high	no	excellent	no
middle_aged	high	no	fair	yes
senior	medium	no	fair	yes
senior	low	yes	fair	yes
senior	low	yes	excellent	no
middle_aged	low	yes	excellent	yes
youth	medium	no	fair	no
youth	low	yes	fair	yes
senior	medium	yes	fair	yes
youth	medium	yes	excellent	yes
middle_aged	medium	no	excellent	yes
middle_aged	high	yes	fair	yes
senior	medium	no	excellent	no

income=high: 4
income=medium: 6
income=low: 4

$$\bullet \text{ SplitInfo}_{\text{income}}(D) = -\frac{4}{14} \log_2 \frac{4}{14} - \frac{6}{14} \log_2 \frac{6}{14} - \frac{4}{14} \log_2 \frac{4}{14} \approx 0.926$$

$$\bullet \text{ GainRatio}(\text{income}) = \text{Gain}(\text{income}) / \text{SplitInfo}_A(D) = \frac{0.029}{0.926} \approx 0.031$$



06-6: Quiz



Quiz

- Given students' ID, height, weight, and gender as the training data, you are asked to build a decision tree classifier to predict a student's gender based on her/his ID, height, and weight
- Which attribute (ID, height, or weight) is likely to be selected first if you use information gain as the attribute selection method?



06-7: Answer



Quiz

- Given students' ID, height, weight, and gender as the training data, you are asked to build a decision tree classifier to predict a student's gender based on her/his ID, height, and weight
- Which attribute (ID, height, or weight) is likely to be selected first if you use information gain as the attribute selection method?



06-8: Gini index



Example of Gini index

age	income	student	credit rating	buys computer
youth	high	no	fair	no
youth	high	no	excellent	no
middle_aged	high	no	fair	yes
senior	medium	no	fair	yes
senior	low	yes	fair	yes
senior	low	yes	excellent	no
middle_aged	low	yes	excellent	yes
youth	medium	no	fair	no
youth	low	yes	fair	yes
senior	medium	yes	fair	yes
youth	medium	yes	excellent	yes
middle_aged	medium	no	excellent	yes
middle_aged	high	yes	fair	yes
senior	medium	no	excellent	no

When income = "high"
→ 2 "yes" and 2 "no"

When income = "medium"
→ 4 "yes" and 2 "no"

When income = "low"
→ 3 "yes" and 1 "no"

$$\bullet \text{Gini}_{\text{income}}(D) = \frac{4}{14} \left(1 - \left(\frac{2}{4} \right)^2 - \left(\frac{2}{4} \right)^2 \right) + \frac{6}{14} \left(1 - \left(\frac{4}{6} \right)^2 - \left(\frac{2}{6} \right)^2 \right) + \frac{4}{14} \left(1 - \left(\frac{3}{4} \right)^2 - \left(\frac{1}{4} \right)^2 \right) = 0.325$$

$$\bullet \text{Gain}(\text{income}) = \text{Gini}(D) - \text{Gini}_{\text{income}}(D) = 0.134$$



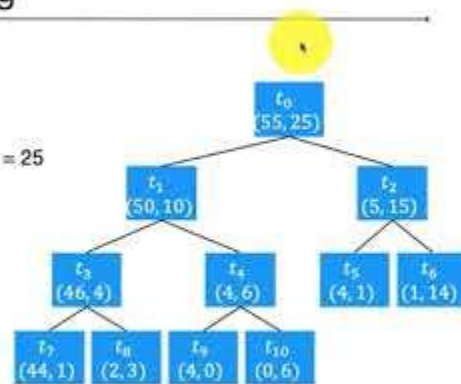
06-9: Tree pruning



Example of post-pruning

$$\alpha(t_i) = \frac{\# \text{ error after cut} - \# \text{ error before cut}}{\# \text{ leaves been cut} - 1}$$
$$\alpha(t_0) = ?$$

- If we cut the tree and leave only t_0 , # errors = 25
- If we don't cut the tree, # errors = $1+2+0+0+1+1=5$
- If we cut the tree, 6 leaves are cut
- $\alpha(t_0) = \frac{25-5}{6-1} = 4$



06-10: Miscellaneous topics



DT is constructed in a “greedy” manner

- Greedy: pick a feature to split the data best on the **current** information
- This may lead to a local optimal



06-11: Quiz



Quiz

- If all features are numerical, which of the following classifier requires more time in the prediction phase? KNN, logistic regression, or decision tree classifier?



06-12: Answer



Quiz

- If all features are numerical, which of the following classifier requires more time in the prediction phase? KNN, logistic regression, or decision tree classifier?



台灣人工智慧學校



[實作課程] 決策樹 (Decision Tree)

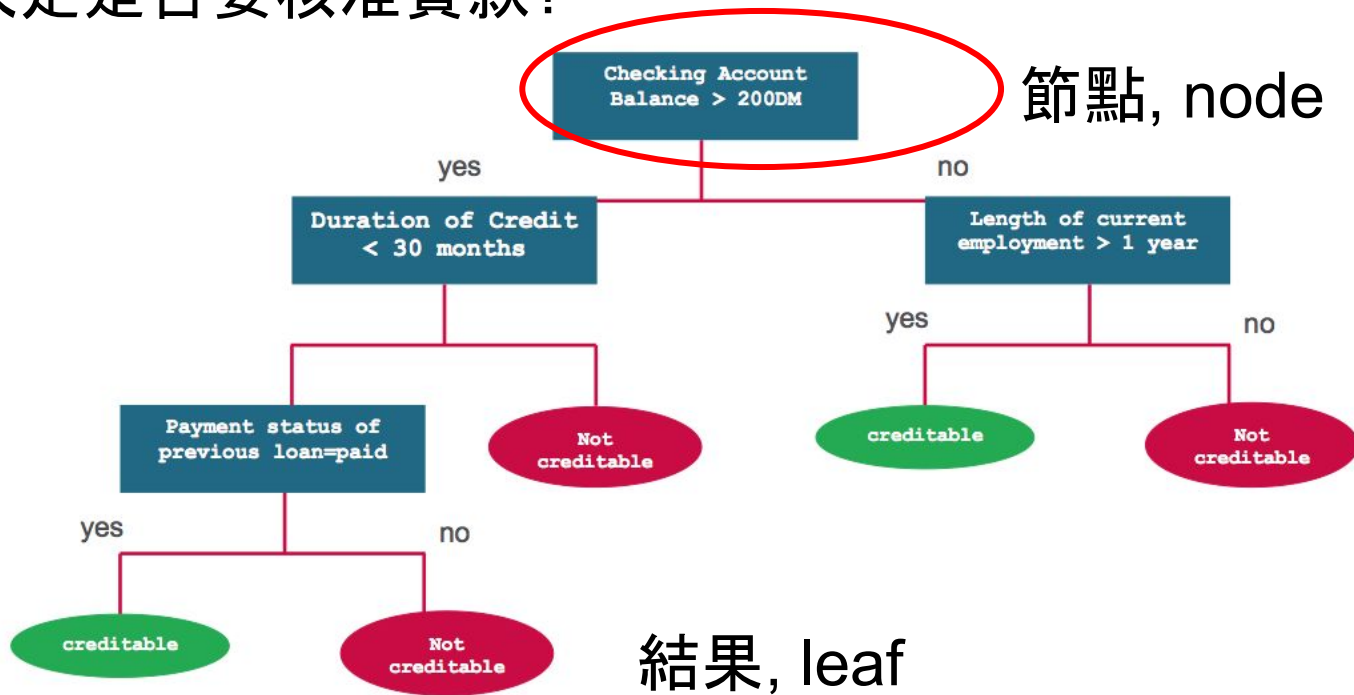
甚麼是決策樹？

- 決定是否要核准貸款？



甚麼是決策樹？

- 決定是否要核准貸款？



如何做決策？

- 該怎麼知道要用哪個 feature? 要用多少的值來做出我們的決策呢?
- 透過從訓練資料找出規則, 讓每一個決策能夠使**訊息增益** (Information gain) 最大化
- 如何衡量訊息增益?
 - 吉尼不純度, Gini impurity
 - 熵, Entropy



吉尼不純度 (Gini impurity)

- 數字越大，代表序列中的資料越混亂

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

	Parent
C0	6
C1	6
Gini = 0.5	

Gini :
 $1 - (6/12)^2 - (6/12)^2$
= 0.5



熵 (Entropy)

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

- 如果序列中所有 sample 都是同一個類別

$$entropy = -1 \log_2 1 = 0$$

- 若序列中各有一半的 sample 分屬不同的類別

$$entropy = -0.5 \log_2 0.5 - 0.5 \log_2 0.5 = 1$$



Gini vs. Entropy

- 都是在衡量一個序列中的混亂程度，越高越混亂
- 數值皆為 0 ~ 1 之間。0 代表序列都是同樣的值
- Scikit-learn 預設使用 Gini

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

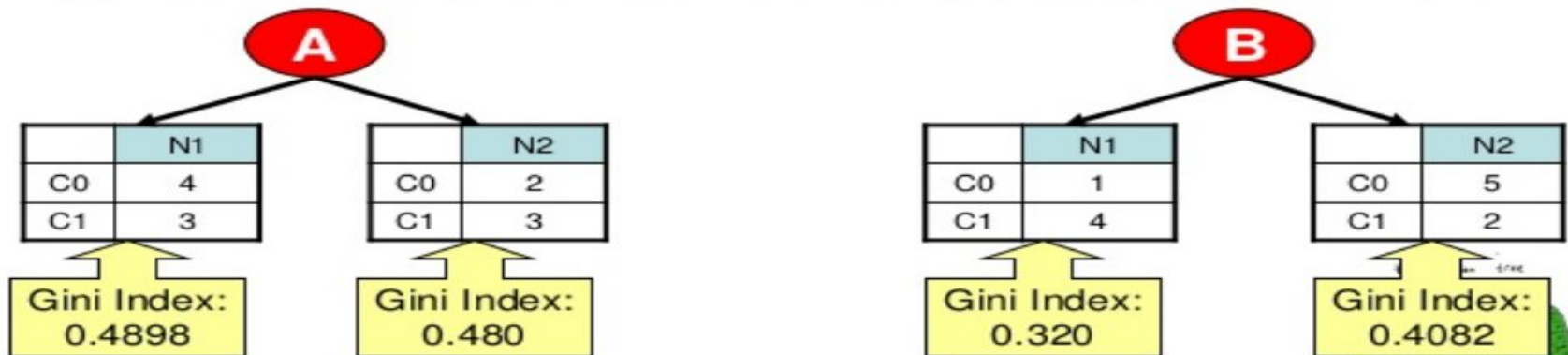
$$Entropy = - \sum_j p_j \log_2 p_j$$



Information Gain 訊息增益

- 決策樹中，試著用 feature 將資料做切分，選取的 feature 必須能最大化訊息增益。而訊息增益則是由 Gini 或 Entropy 衡量，我們希望切分後的資料越純越好 ($Gini=0$)

Suppose there are two ways (A and B) to split the data into smaller subset.



Which one is a better split??

Compute the **weighted average of the Gini index** of both attribute



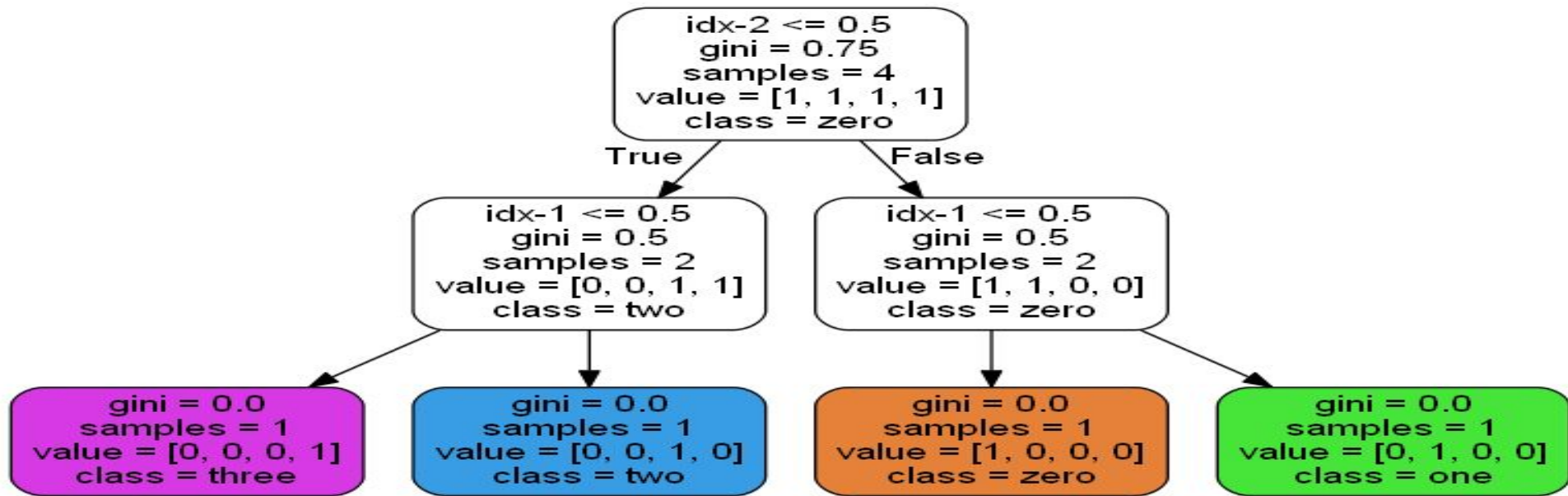
決策樹建立 (1/2)

- 不斷尋找 feature 進行決策, 試著將資料切分為同一個類別 (minimize Gini)
 - 這樣會造成甚麼後果?



決策樹建立 (2/2)

- 當我們拿一批訓練資料給決策樹進行分類時，若沒有給定任何條件，決策樹會不斷進行分枝，直到所有 leaf 的資料都屬於同一個類別為止



決策樹 in Scikit-learn

- 兩行 code 建立決策樹

```
from sklearn.tree import DecisionTreeRegressor
```

```
from sklearn.tree import DecisionTreeClassifier
```

```
clf = DecisionTreeClassifier()
```



決策樹模型中的參數

```
from sklearn.tree import DecisionTreeClassifier

clf = DecisionTreeClassifier(
    criterion = 'gini',
    max_depth = None,
    min_samples_split = 2,
    min_samples_leaf = 1,
)
```



Feature Importance

- 決策樹的另一優點是，我們可以從構建樹的過程中，透過 feature 被用來切分的次數，來得知哪些 features 是相對有用的
- 所有 feature importance 的總和會是 1
- 實務上，我們經常會用 feature importance 來排序 feature 的重要性以及選取要使用的 feature

```
# feature importance  
clf.feature_importances_
```



決策樹實戰

```
from sklearn.datasets import load_iris
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score

iris = load_iris()
print(iris.data.shape, iris.target.shape)
(150, 4) (150,)

x_train, x_test, y_train, y_test = train_test_split(iris.data, iris.target)

print("shape of x_train:", x_train.shape)
shape of x_train: (120, 4)

print("shape of x_test:", x_test.shape)
shape of x_test: (30, 4)

clf = DecisionTreeClassifier()
clf.fit(x_train, y_train)

DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=None,
                        max_features=None, max_leaf_nodes=None,
                        min_impurity_decrease=0.0, min_impurity_split=None,
                        min_samples_leaf=1, min_samples_split=2,
                        min_weight_fraction_leaf=0.0, presort=False, random_state=None,
                        splitter='best')

y_pred = clf.predict(x_test)
```

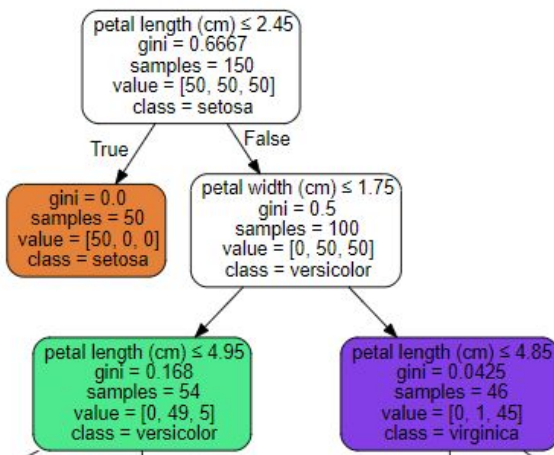
*影片中 code 有誤: **accuracy_score(y_test, y_pred)**



決策樹視覺化

- 生成好的樹，可以用額外的套件 graphviz，自動從 code 繪製成圖形，讓我們了解決策樹究竟學到了甚麼決策

```
>>> dot_data = tree.export_graphviz(clf, out_file=None,  
                                     feature_names=iris.feature_names,  
                                     class_names=iris.target_names,  
                                     filled=True, rounded=True,  
                                     special_characters=True)  
  
>>> graph = graphviz.Source(dot_data)  
>>> graph
```



決策樹小結

決策樹 summary

- 掃過所有 feature 與對應的值將資料做切分
- 希望資料盡可能分開, 透過切分後的資料純度 (Gini or Entropy) 來衡量
- 如果不對決策樹進行任何限制 (樹的深度、葉子至少要有多少樣本), 容易造成 Overfitting
- 透過 feature importance 來排序重要性



決策樹 Summary

- 掃過所有 feature 與對應的值將資料做切分
- 希望資料盡可能分開, 透過切分後的資料純度 (Gini or Entropy) 來衡量
- 如果不對決策樹進行任何限制 (樹的深度、葉子至少要有多少樣本), 容易造成 Overfitting
- 透過 feature importance 來排序重要性



決策樹進化! Ensemble

- 決策樹有著非常容易被理解的優點, 但是通常預測結果不會那麼準確
- 之後的學者想出方法, 把樹結合起來 (ensemble) 做改進
 - **Bagging (Bootstrap aggregating)**: Fit many large trees to bootstrap-resampled versions of the training data, and classify by majority vote.
 - **Boosting**: Fit many large or small trees to reweighted versions of the training data. Classify by weighted majority vote.



練習 `decision_tree_example.ipynb`



- 請使用 Iris Dataset, 建立決策樹模型, 試著更改 Decision Tree 中的 **criterion**, **max_depth**, **min_samples_split** 等參數, 並評估不同的參數是否會影響以下結果
 - training error / loss
 - testing error / loss
 - training speed (可用 `%%timeit` 計算 cell 執行的速度)



Write a Decision Tree from Scratch (optional, but 推薦)



The image shows a man smiling on the left. To his right is a decision tree diagram. At the top is a table of fruit data. Below the table is a decision node: "Is diameter ≥ 3 ". The "False" branch leads to a leaf node with "R 1 Grape" and "R 1 Grape". The "True" branch leads to a leaf node with "G 3 Apple", "Y 3 Apple", and "Y 3 Lemon". A red banner across the bottom of the diagram area contains the text "{ML} Let's Write a Decision Tree from Scratch". Below the banner, there are two boxes: "Apple 100%" with a green leaf icon, and "Predict Apple 50% Lemon 50%" with a green leaf icon.

Color	Diam	Label
Green	3	Apple
Yellow	3	Apple
Red	1	Grape
Red	1	Grape
Yellow	3	Lemon

Is diameter ≥ 3 ?

R 1 Grape
R 1 Grape

G 3 Apple
Y 3 Apple
Y 3 Lemon

{ML} Let's Write a Decision Tree from Scratch

Apple 100%

Predict
Apple 50%
Lemon 50%



補充閱讀

- 如果前面助教講的影片你都聽不懂，肯定是因為助教講的不夠清楚，只好幫各位找一些寫的不錯的文章，給大家參考
 - [決策樹 \(Decision Tree\)](#) - 中文
 - [how decision tree works](#) - 英文



思考問題

- 在分類問題中，若沒有任何限制，決策樹有辦法把訓練資料的 loss 完全降成 0 嗎？
- 決策樹做分類問題時，資料的不純度比較容易計算（是否屬於同一個類別）。那如果變成回歸問題，這時切分後的資料不純度該如何計算？樹建置完成後，又該如何進行預測呢？

