

機器學習

決策樹



出題教練

楊証琨

知識地圖 機器學習- 模型選擇 - 樹狀模型 - 決策樹(Decision Tree)



機器學習概論 Introduction of Machine Learning

監督式學習 Supervised Learning



非監督式學習 Unsupervised Learning



模型選擇 Model selection

概論

驗證基礎

預測類型

評估指標

基礎模型 Basic Model

線性回歸 Linear Regression

邏輯斯回歸 Logistic Regression

套索算法 LASSO

嶺回歸 Ridge Regression

樹狀模型 Tree based Model

決策樹 Decision Tree

隨機森林 Logistic Regression

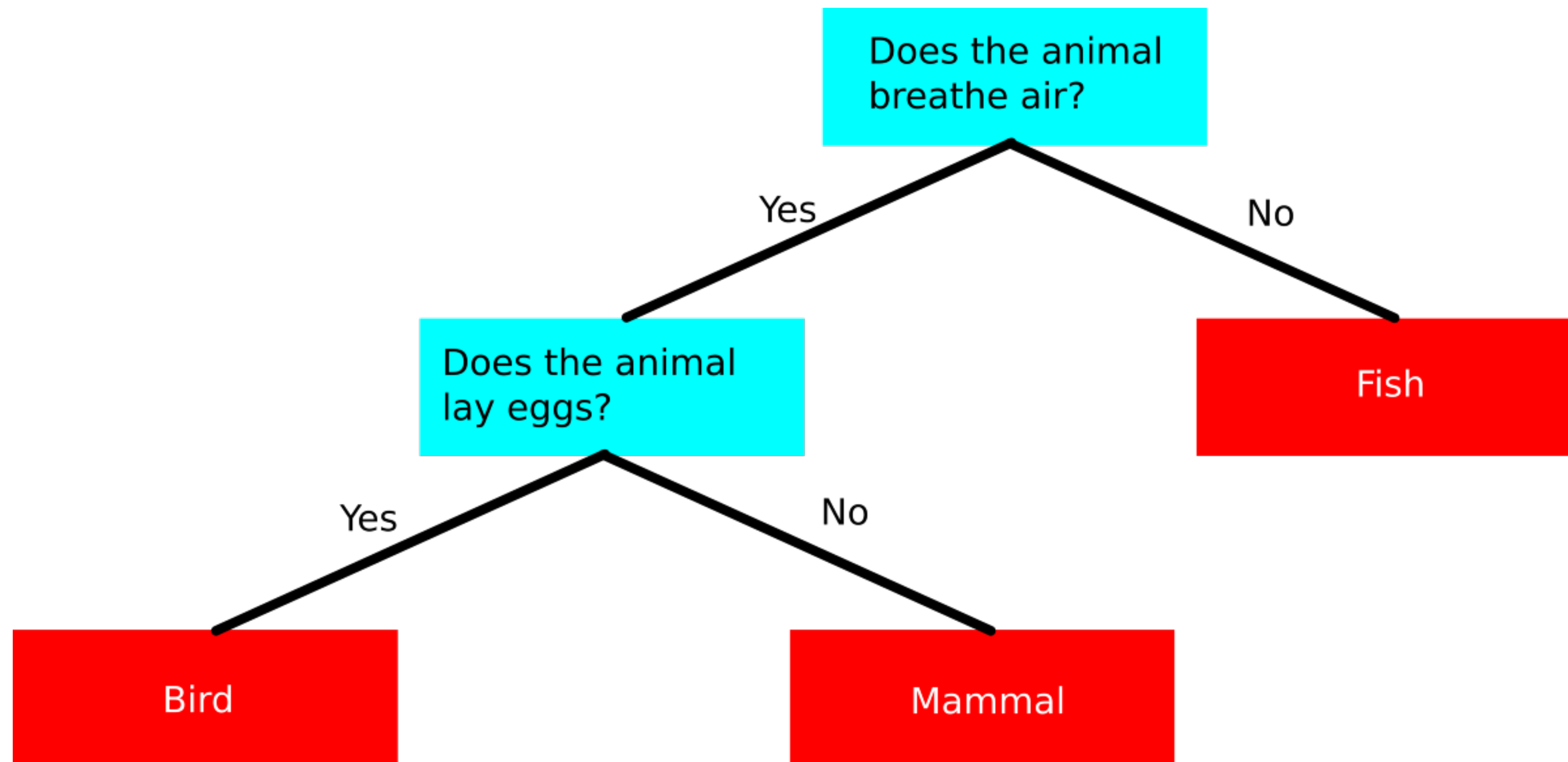
梯度提升機 Gradient Boosting Machine

本日知識點目標

- 了解決策樹的原理、定義與其使用限制
- 如何用 gini-index / entropy 來衡量資料相似程度
- 決策樹是如何對一筆資料做決策

決策樹 (Decision Tree)

- 透過一系列的是非問題，幫助我們將資料進行切分
- 可視覺化每個決策的過程，是個具有非常高解釋性的模型

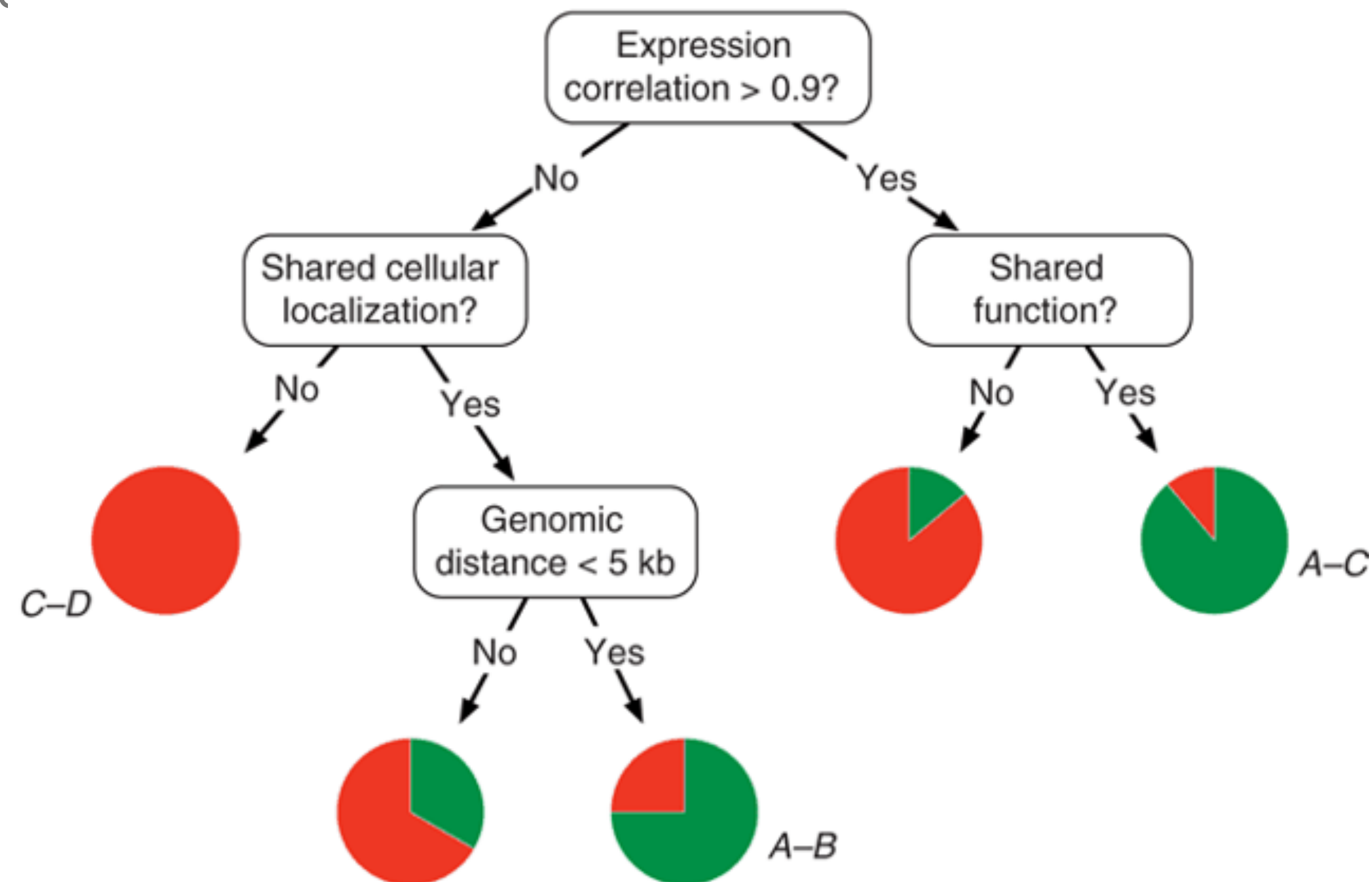


決策樹 (Decision Tree)

- 從訓練資料中找出規則，讓每一次決策能使訊息增益 (Information Gain) 最大化
- 訊息增益越大代表切分後的兩群資料，群內相似程度越高
- 例如使用健檢資料來預測性別，若使用頭髮長度超過 50 公分來切分，則切分後兩群資料很有可能多數都為男生或女生 (相似程度高) 這樣頭髮長度就是個很好的 feature。

訊息增益 (Information Gain)

- 決策樹模型會用 features 切分資料，該選用哪個 feature 來切分則是由訊息增益的大小決定的。希望切分後的資料相似程度很高，通常使用吉尼係數來衡量相似程度。



衡量資料相似: Gini vs. Entropy

該怎麼衡量資料相似程度？通常使用吉尼係數 (gini-index) 或熵 (entropy) 來衡量，兩者都可使用，更詳細可參考[Stack Exchange](#)

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

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

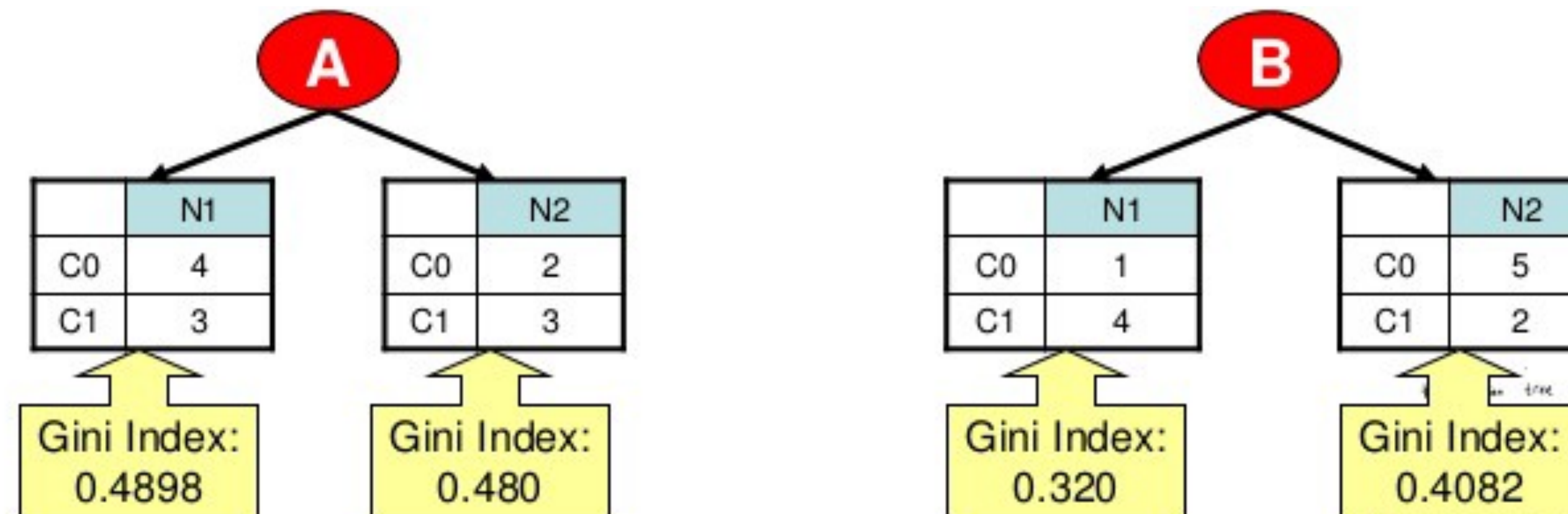
Splitting Binary Attributes (using Gini)

Example :

	Parent
C0	6
C1	6
Gini = 0.5	

$$\begin{aligned}\text{Gini :} \\ 1 - (6/12)^2 - (6/12)^2 \\ = 0.5\end{aligned}$$

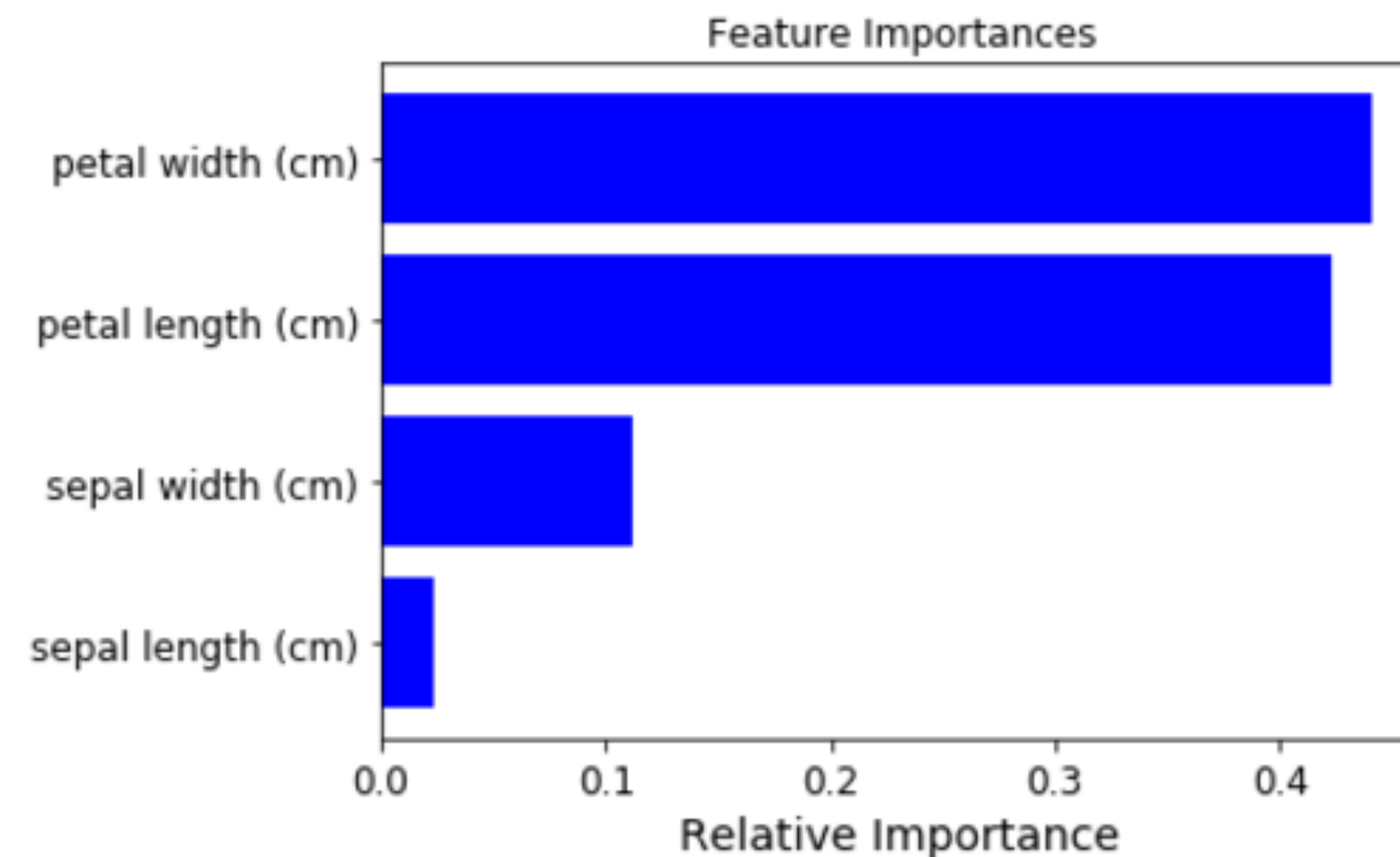
Suppose there are two ways(A and B) to split the data into smaller subset.



By : Mohd.Noor Abdul Hamid,Ph.D
(universiti Utara Malaysia)

決策樹的特徵重要性 (Feature importance)

- 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 的總和為 1
- 實務上可以使用 feature importance 來了解模型如何進行分類



解題時間 It's Your Turn

請跳出PDF至官網Sample Code & 作業
開始解題

