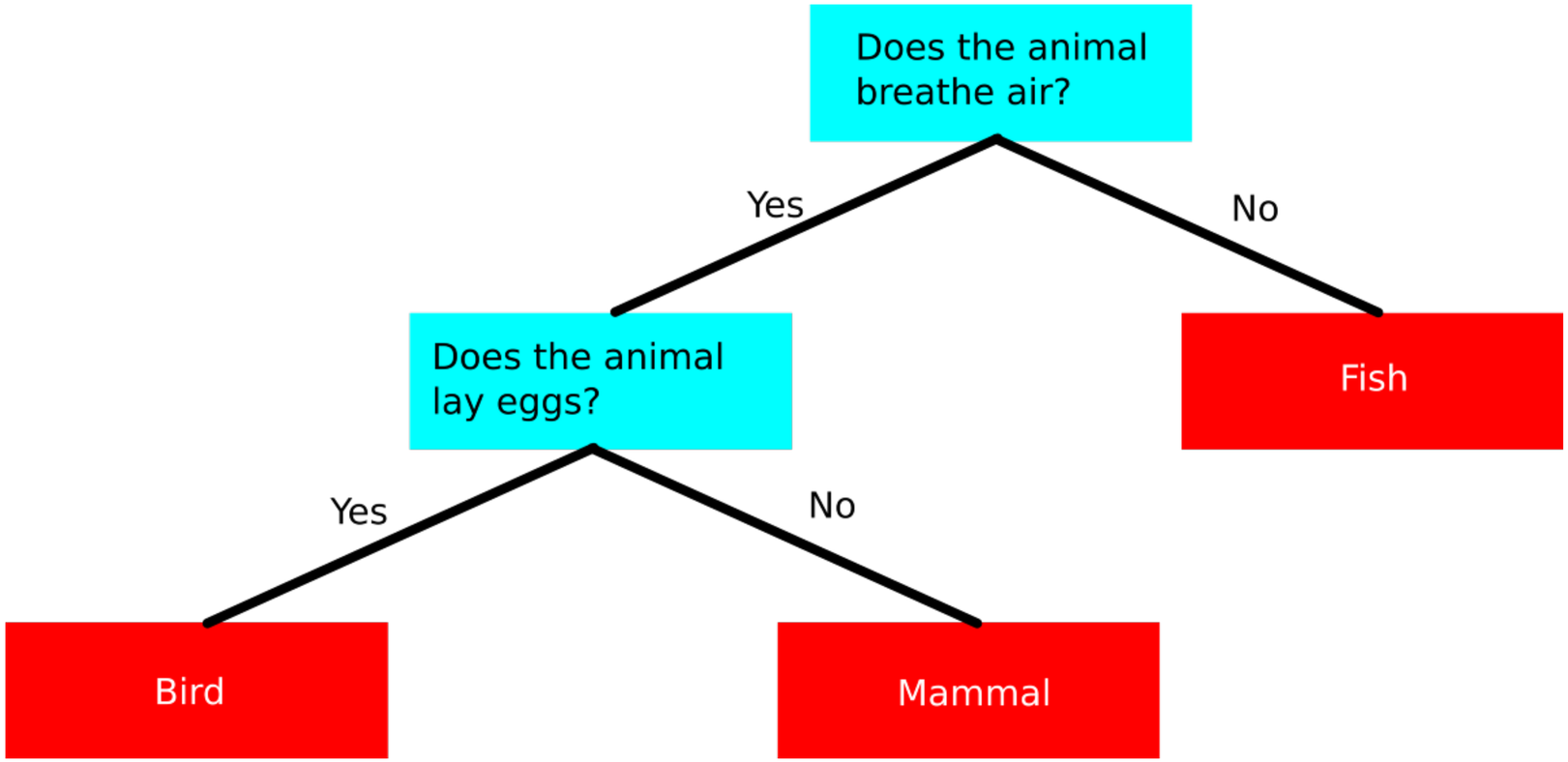
Day 41

機器學習

決策樹

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



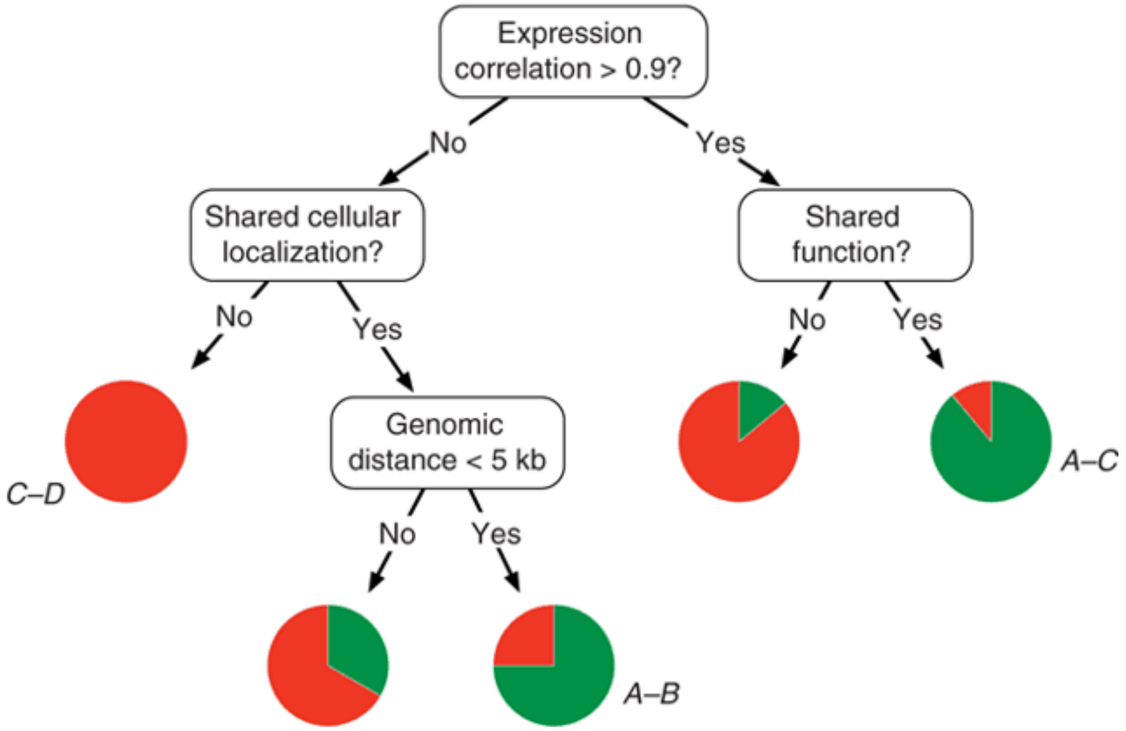
圖片來源：towardsdatascience

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。



圖片來源：nature

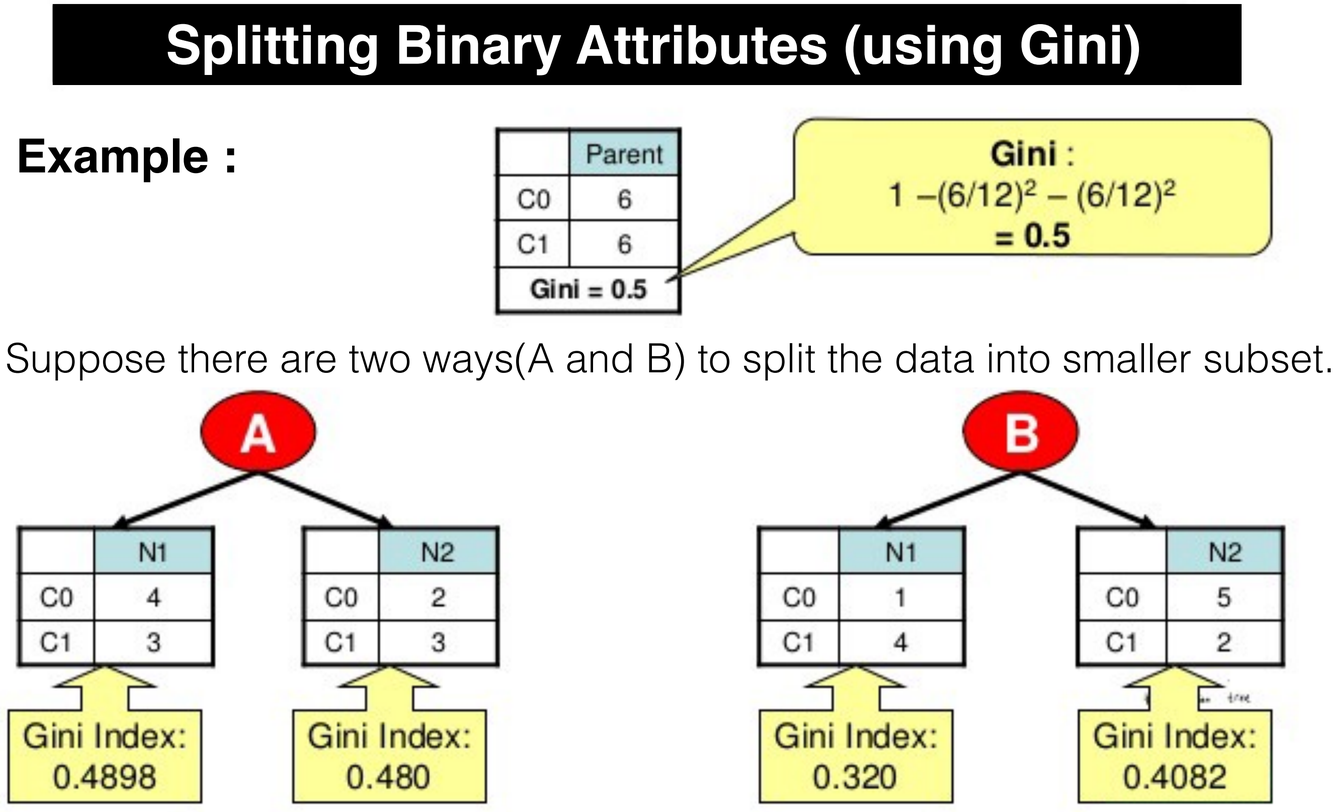
衡量資料相似: Gini vs. Entropy

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

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

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

訊息增益 (Information Gain)

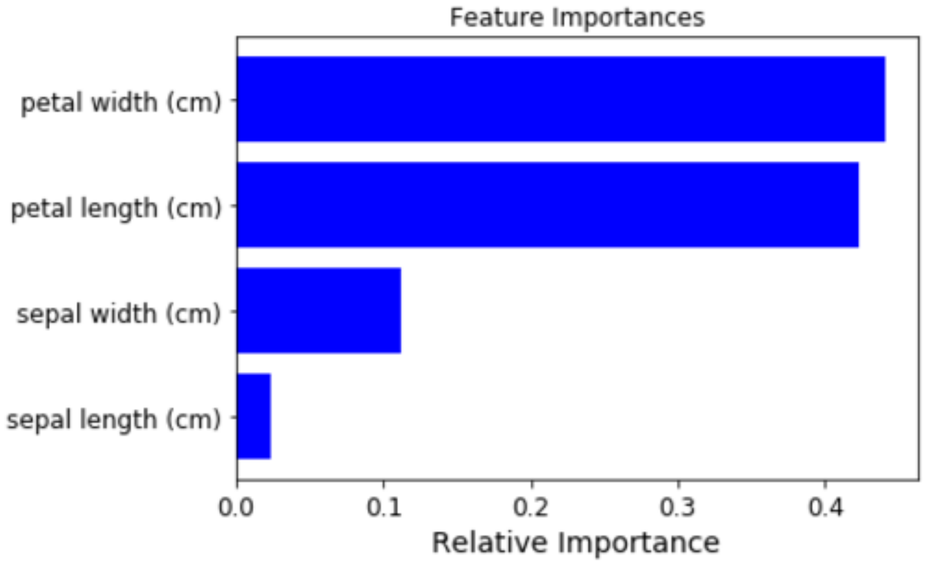


By : Mohd.Noor Abdul Hamid,Ph.D
(universiti Utara Malaysia)

圖片來源：Medium

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



圖片來源：Stack Overflow

解題時間
It’s Your Turn

請跳出PDF至官網Sample Code＆作業
開始解題