Decision Tree

CART

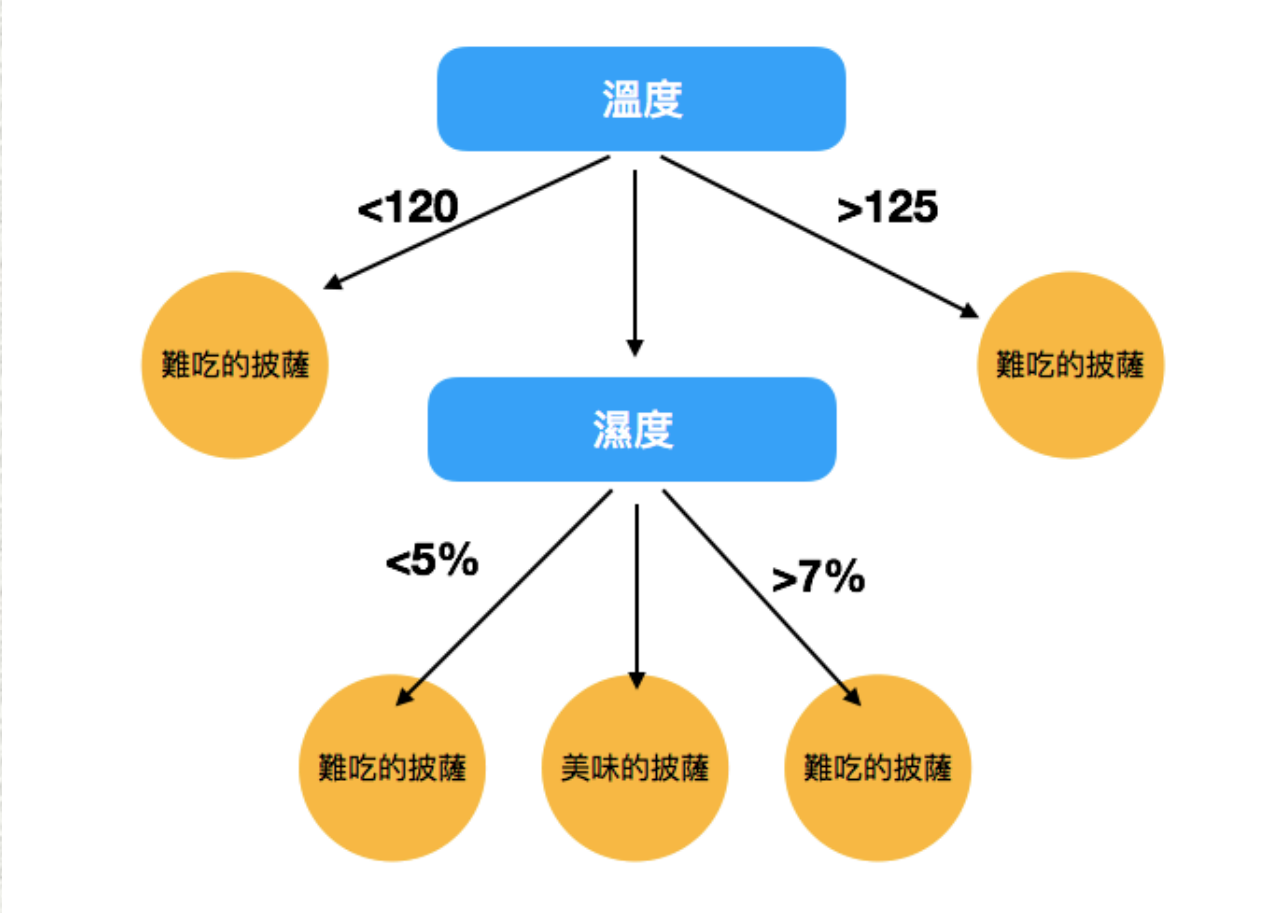
何信賢

什麼是決策樹？

- 用來處理問題的樹狀結構
- 每個內部節點表示一個評估欄位
- 模仿人類決策的過程

特性

- 比較具有解釋力
- 執行速度較快
- 醫藥、商業常用



CART演算法

- Classification & Regression Tree
- 二元樹
- 以吉尼係數(Gini)作為選擇依據（不純度計算）
- 也可以用資訊增益(Information Gain)（用熵計算）

吉尼係數（Gini）

- 假設資料集合S包含n個類別，吉尼係數Gini(S)定義為

$$Gini(S) = 1 - \sum_{j=1}^{n} p_j^2$$

pj為在S中的值組屬於類別j的機率

- 利用屬性A分割資料集合S為S1與S2（二元分割）。則根據此一分割要件的吉尼係數$Gini_A(S)$為

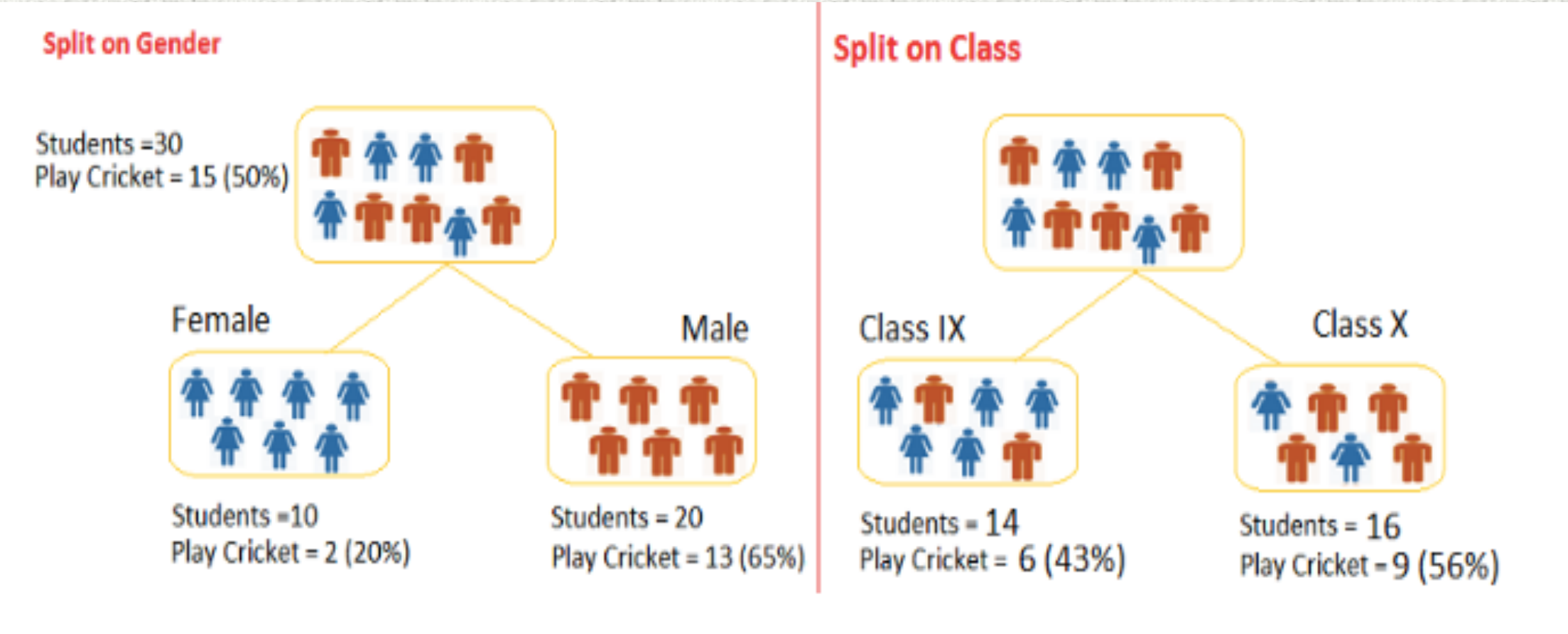
$$Gini_A(S) = \frac{|S_1|}{|S|} Gini(S_1) + \frac{|S_2|}{|S|} Gini(S_2)$$

- 不純度的降低值為: $\Delta Gini(A) = Gini(S) - Gini_A(S)$

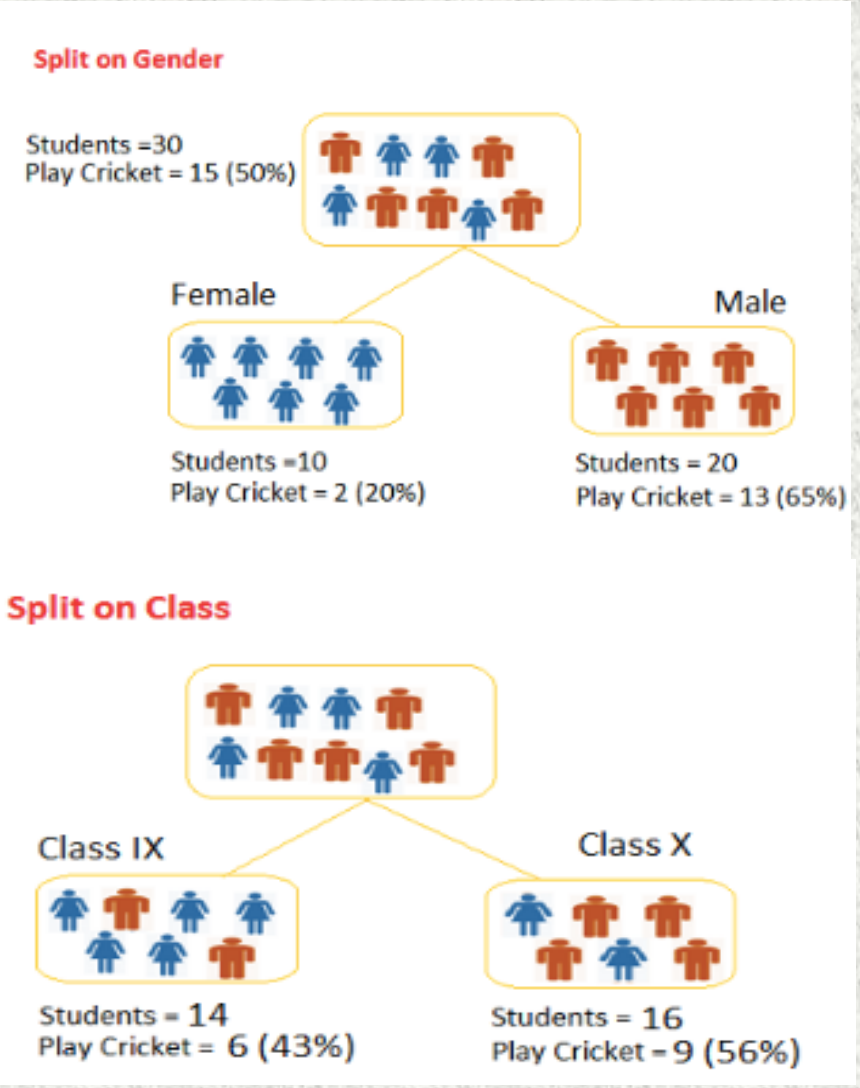
- 挑選擁有最大不純度的降低值、或吉尼係數$Gini_A(S)$最小的屬性作為分割屬性。

舉例

我們想要預測喜歡打板球類型的學生，何者是比較好的分類呢？



Gini怎麼算呢？



性別分類有比較大不純度的降低量

原本資訊量	$1-(15/30)^2-(15/30)^2=0.5$
Female資訊量	$1-(2/10)^2-(8/10)^2=0.32$
Male資訊量	$1-(13/20)^2-(7/20)^2=0.455$
獲得資訊量	$0.5-(10/30)*0.32-(20/30)*0.455=0.09$

原本資訊量	$1-(15/30)^2-(15/30)^2=0.5$
Female資訊量	$1-(6/14)^2-(8/14)^2=0.489$
Male資訊量	$1-(9/16)^2-(7/16)^2=0.492$
獲得資訊量	$0.5-(16/30)*0.489-(14/30)*0.492=0.008$

資訊獲利（IG）

- 以熵 (Entropy) 為基礎
- 熵（亂度），可當作資訊量的凌亂程度（不確定性）指標，當熵值愈大，則代表資訊的凌亂程度愈高。

Entropy = -p * log2 p - q * log2q

p：成功的機率（或true的機率） q：失敗的機率（或false的機率）

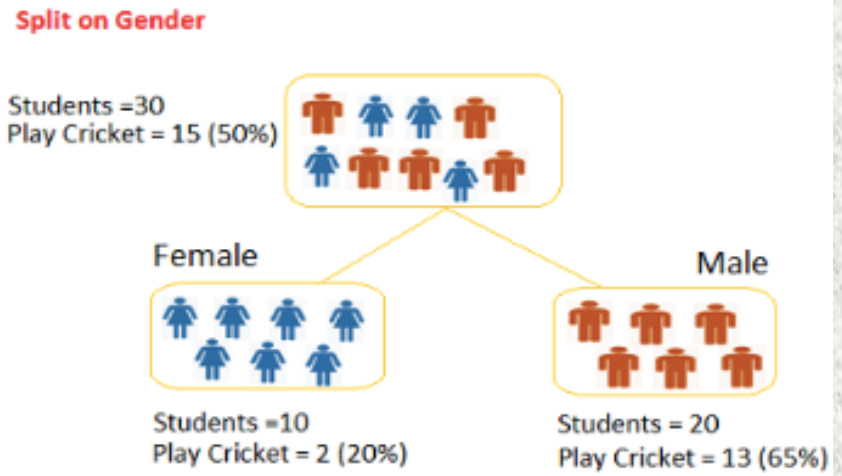
【範例】丟公平銅板，則丟出正面與反面的機率是一樣的（最凌亂）
若不公平銅板，則丟出正面與反面的機率不會是一樣的（愈不凌亂）

- 若丟了14次銅板，出現了9個正面與5個反面(記為[9+, 5-])，則這個 範例的熵為:

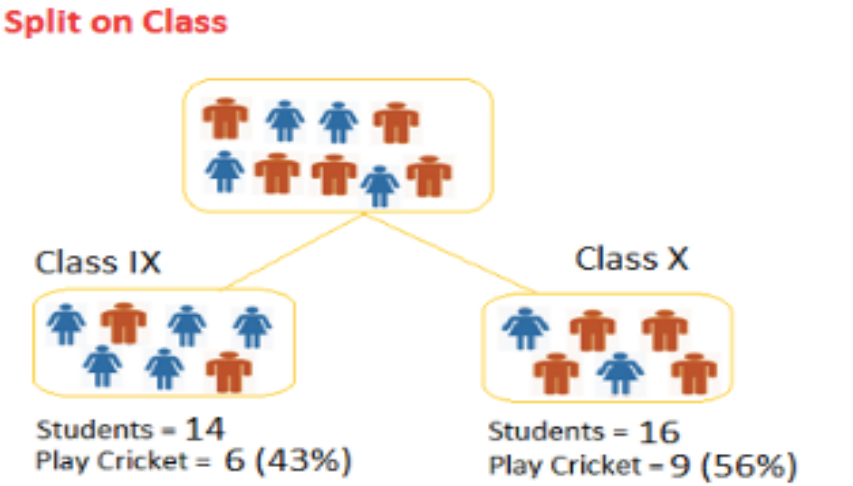
$$\text{Entropy}([9+, 5-]) = -(9/14)\log_2(9/14) - (5/14)\log_2(5/14) = 0.94$$

那Entropy怎麼算呢？

性別分類
熵較低



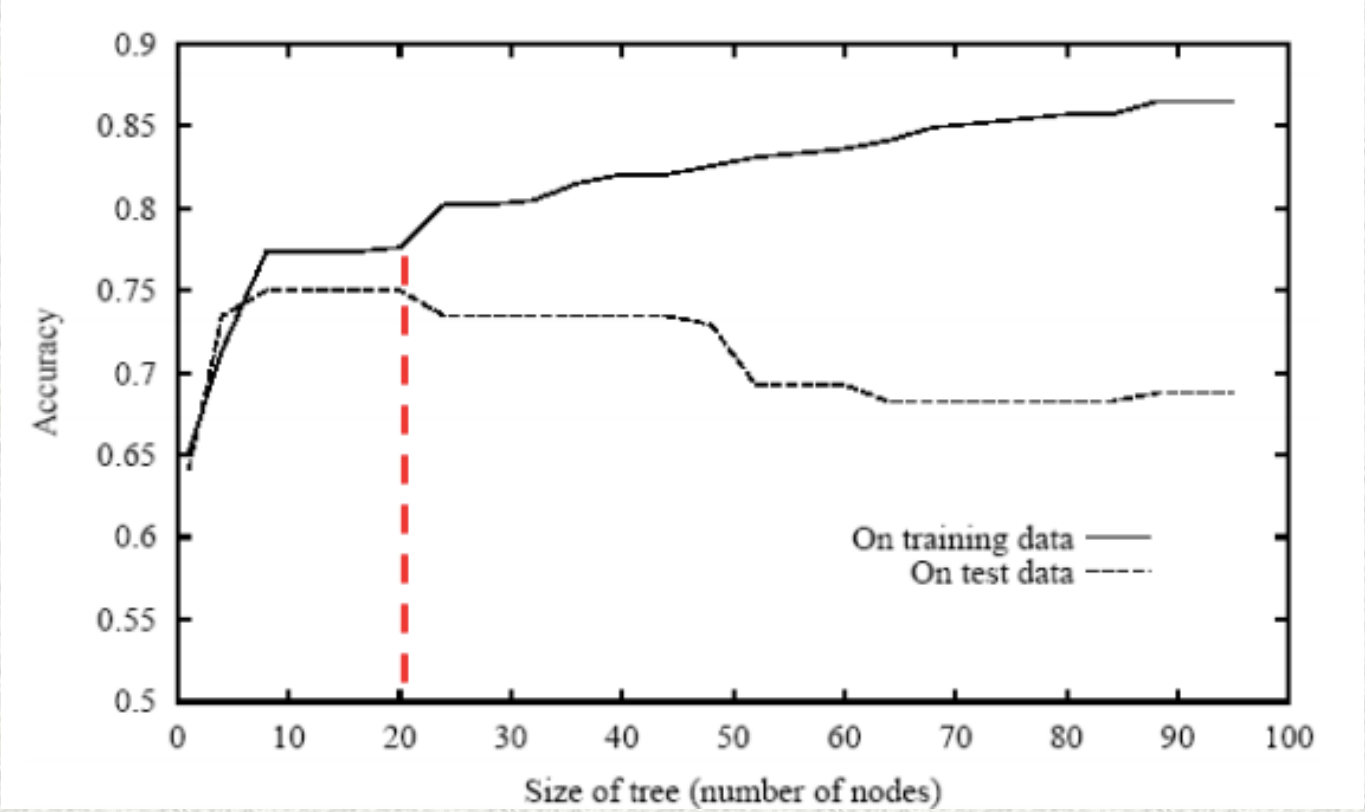
原本熵	$-(15/30)\log_2(15/30)-(15/30)\log_2(15/30)=1$
Female熵	$-(2/10)\ \log_2(2/10)\ -\ (8/10)\ \log_2(8/10)\ =0.72$
Male 熵	$-(13/20)\ \log_2(13/20)\ -\ (7/20)\ \log_2(7/20)\ =0.93$
加權平均	$(10/30)*0.72\ +(20/30)*0.93\ =0.86$



原本熵	$-(15/30)\log_2(15/30)-(15/30)\log_2(15/30)=1$
Female熵	$-(6/14)\ \log_2(6/14)\ -\ (8/14)\ \log_2(8/14)\ =0.99$
Male 熵	$-(9/16)\ \log_2(9/16)\ -\ (7/16)\ \log_2(7/16)\ =0.99$
加權平均	$(14/30)*0.99\ +(16/30)*0.99\ =0.99$

決策樹學習常見問題

- 決策樹學習可能遭遇模型過度配適（overfitting）的問題
- 因此樹的階層越少比較好



修剪

- 事前修剪

運用統計門檻值(Significance Level)加以衡量，譬如卡方值或資訊獲得值等技術，評估是否該繼續分割某內部節點成數個子分支或是應該立刻停止。

- 事後修剪

允許決策樹過度配適情形的合理存在，當完成決策樹的建立之後，再進行修剪的程序。

衡量複雜度

need a **regularizer**, say, $\Omega(G) = \text{NumberOfLeaves}(G)$

want **regularized** decision tree:

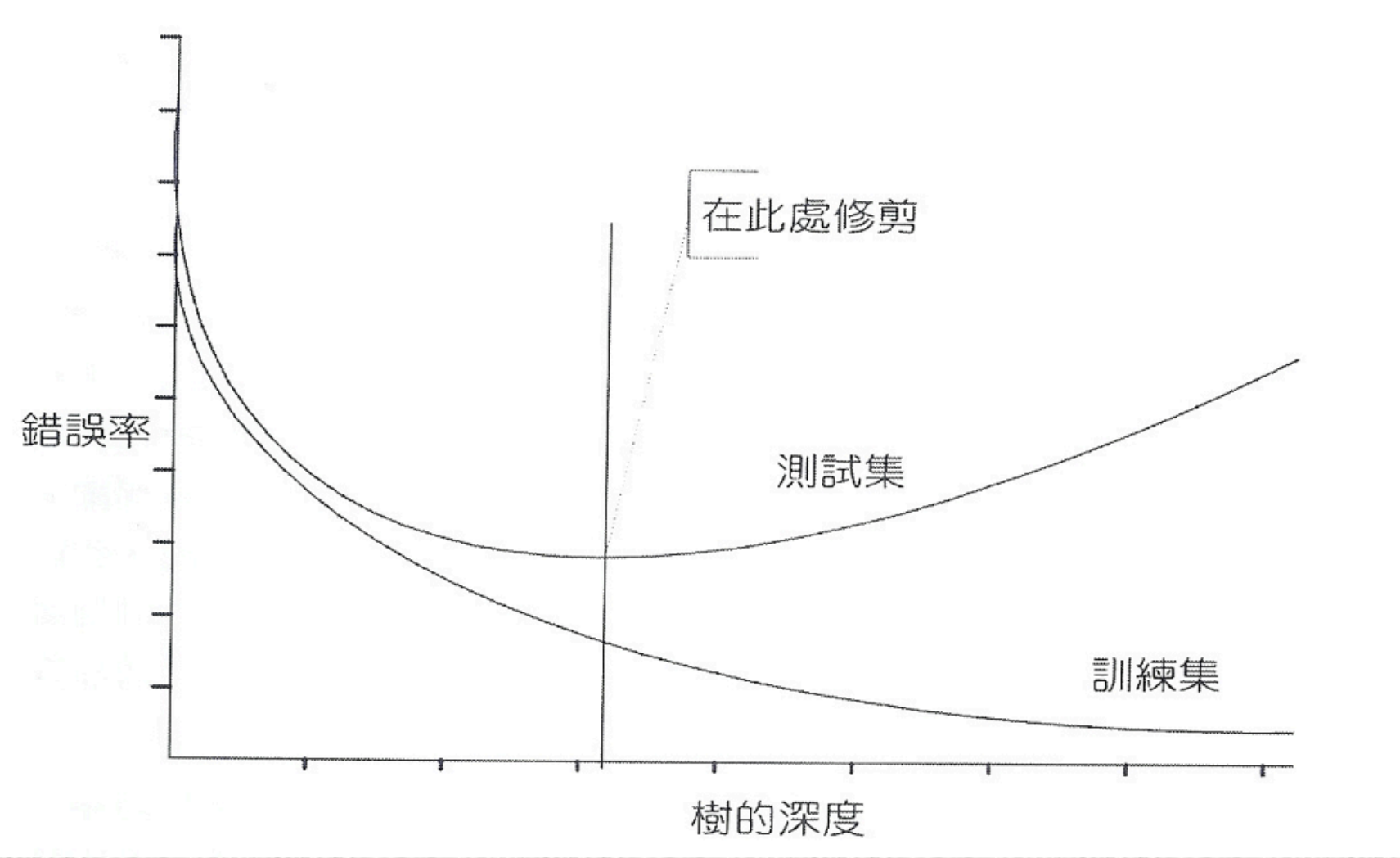
$$\underset{\text{all possible } G}{\text{argmin}} \; E_{\text{in}}(G) + \lambda\Omega(G)$$

—called **pruned** decision tree

- cannot enumerate all possible G computationally:
  —often consider only
  - $G^{(0)}$ = fully-grown tree
  - $G^{(i)} = \text{argmin}_G E_{\text{in}}(G)$ such that G is **one-leaf removed** from $G^{(i-1)}$

systematic choice of λ? **validation**

如何修剪



Demo

資料來源： https://www.kaggle.com/mrisdal/exploring-survival-on-the-titanic/data

- 資料匯入
- 資料整理
- 特徵工程
- 建模型
- 修剪
- 決策樹視覺化
- 上傳Kaggle