D25：隨機森林演算法(Random Forest)

重要知識點

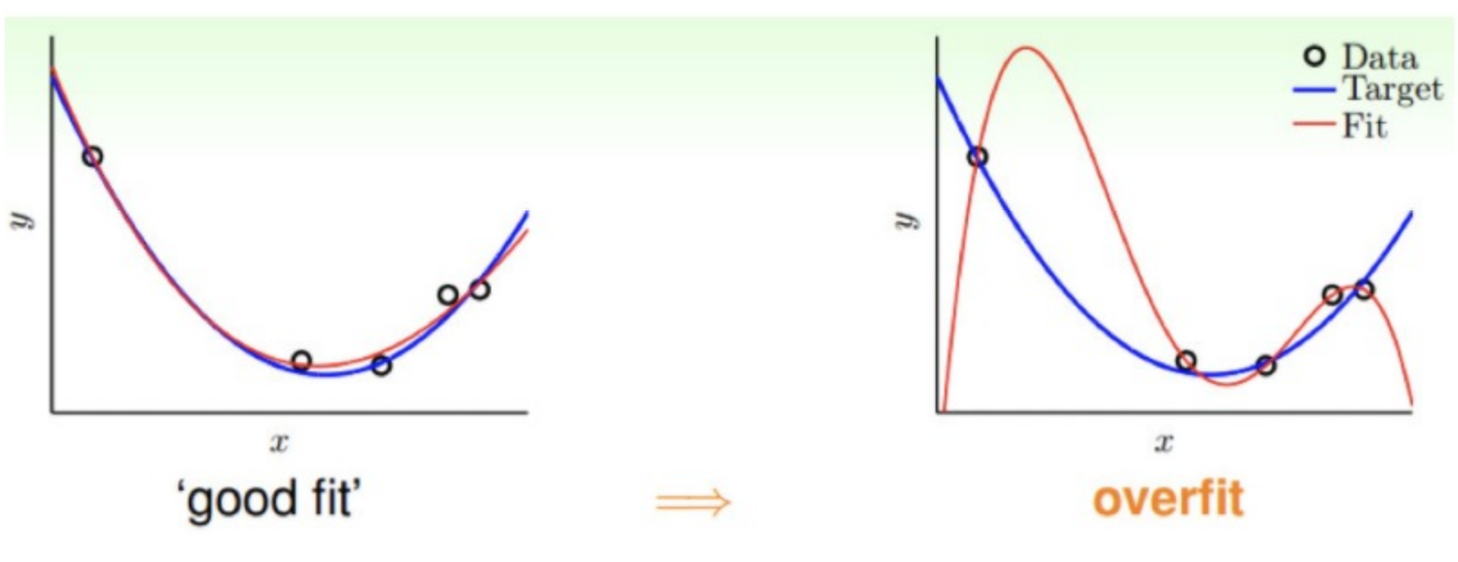
- 了解何謂集成(Ensemble) 中的 Bagging
- 了解何謂隨機森林

決策樹缺點

- 在開始介紹隨機森林前，先來了解一下決策樹的限制，以利我們更加了解為何需要隨機森林的出現。
- 決策樹中若沒有對的樹的成長做限制(樹的深度，每個末端節點 leaf 至少要多少樣本等)，決策樹生長到最後會對每個特徵值創建節點，將所有資料作到 100% 的分類(所有的樣本資料最後都成為一個末端節點 leaf)，進而導致過擬合(overfitting)

何謂過擬合(overfitting)

當模型過度訓練而將訓練資料的分佈記下來，因此在訓練集上可以取得良好得準確度，但當模型遇到訓練集以外的數據時，反而無法得到良好得準確度。類似再考模擬考時透過將答案記下來而得到高分，但因為沒有實際學習到，因此在實際考試時往往無法表現得很好。

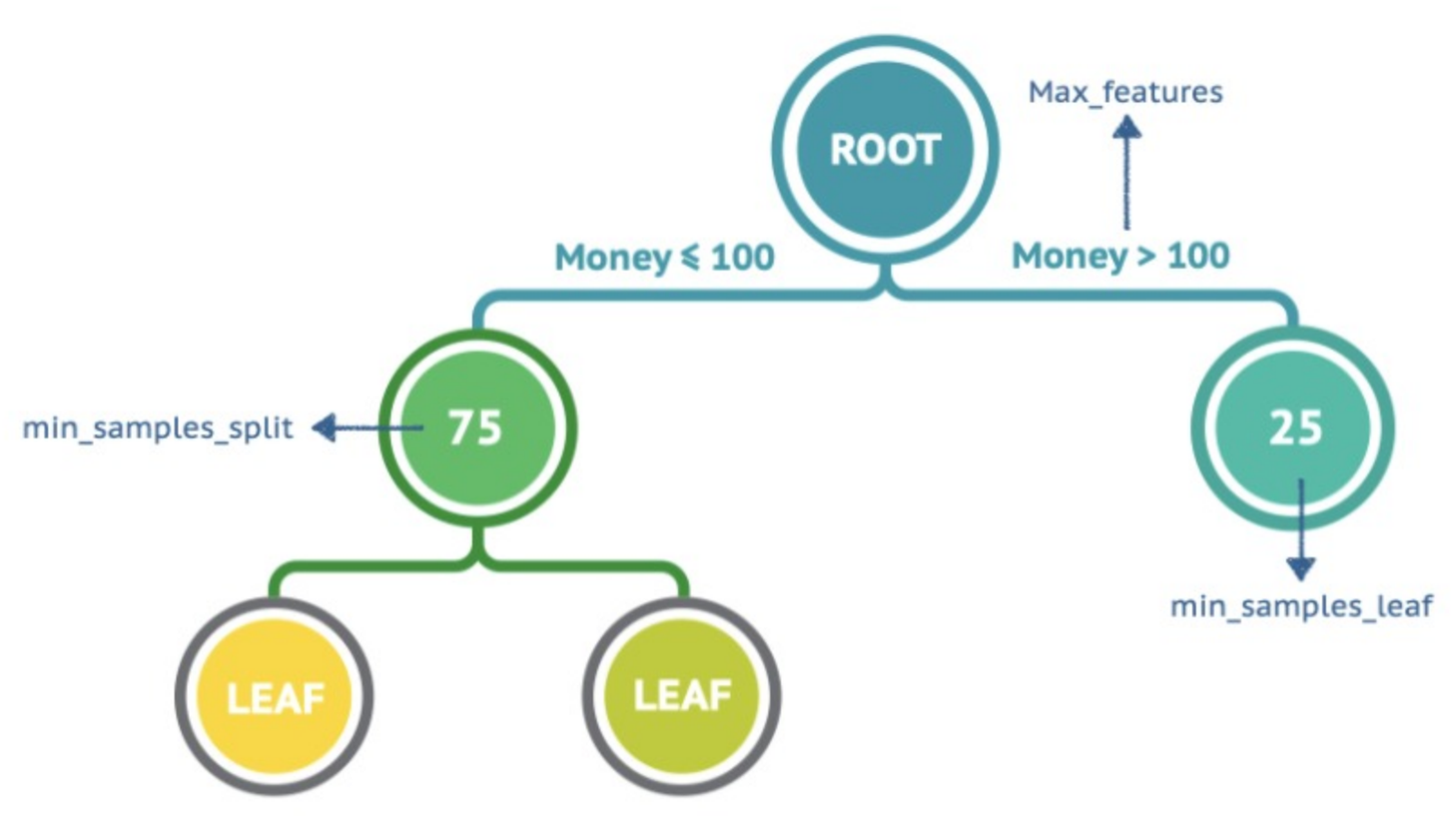


資料來源：解決梯度消失和梯度彌散的方法

設限決策樹

可以透過對決策樹設定限制來預防過擬合(Overfitting) 的現象：

- 最小可分割節點的資料數目 (min_samples_split)
- 最小葉節點(leaf) 的資料數目 (min_samples_leaf)
- 限制樹的高度為幾層 (max_depth)
- 限制最終葉節點(leaf) 個數 (max_leaf_nodes)
- 最多考慮的特徵個數 (max_features)



Why 隨機森林？

- 上述的限制，雖然可以預防過擬合的現象，但相對的也限制了決策樹的訓練與預測。
- 當一棵樹的表現不好，你有種第二顆嗎？
- 這就是隨機森林的大概念，透過多顆的決策樹增強預測準確度(也可以限制每一棵樹的生長)，而這就是集成學習(Ensemble Learning) 的一種。

集成學習(Ensemble Learning)

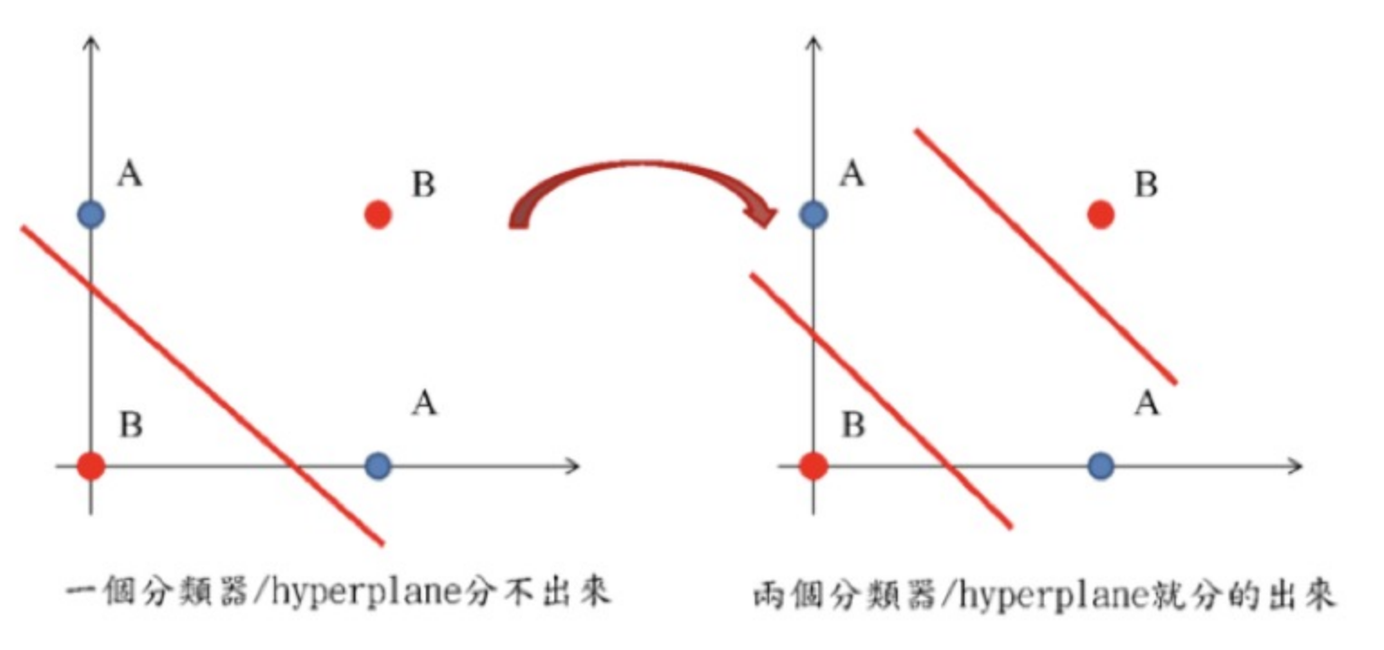
集成學習(Ensemble Learning)的概念是，收集各種不同的分類器，將它們集結起來各司其職，來達到更好的預測效果。(像是無限寶石各有功用，集合起來後的威力更大)

Bagging

隨機森林是集成學習(Ensemble)中的Bagging (Bootstrap aggregating)，其是將多個表現好的分類器，已突破單個分類器的極限，且透過集結多個分類器可降低 Variance，得到更泛化的結果。也就是「三個臭皮匠勝過一個諸葛亮」。

Bagging 假設條件：

- 各個分類器之間具有差異
- 每個分類器的單獨表現夠好(準確度大於 0.5)



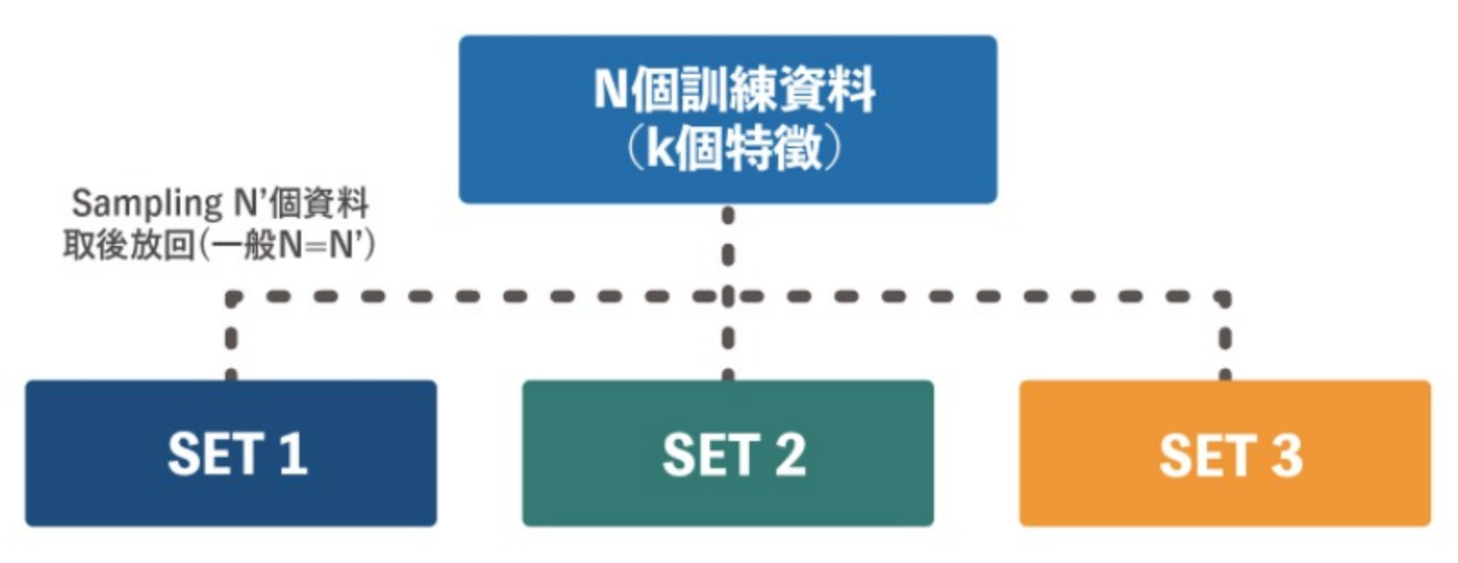
資料來源：Ensemble learning之Bagging、Boosting和AdaBoost

隨機森林(Random Forest)

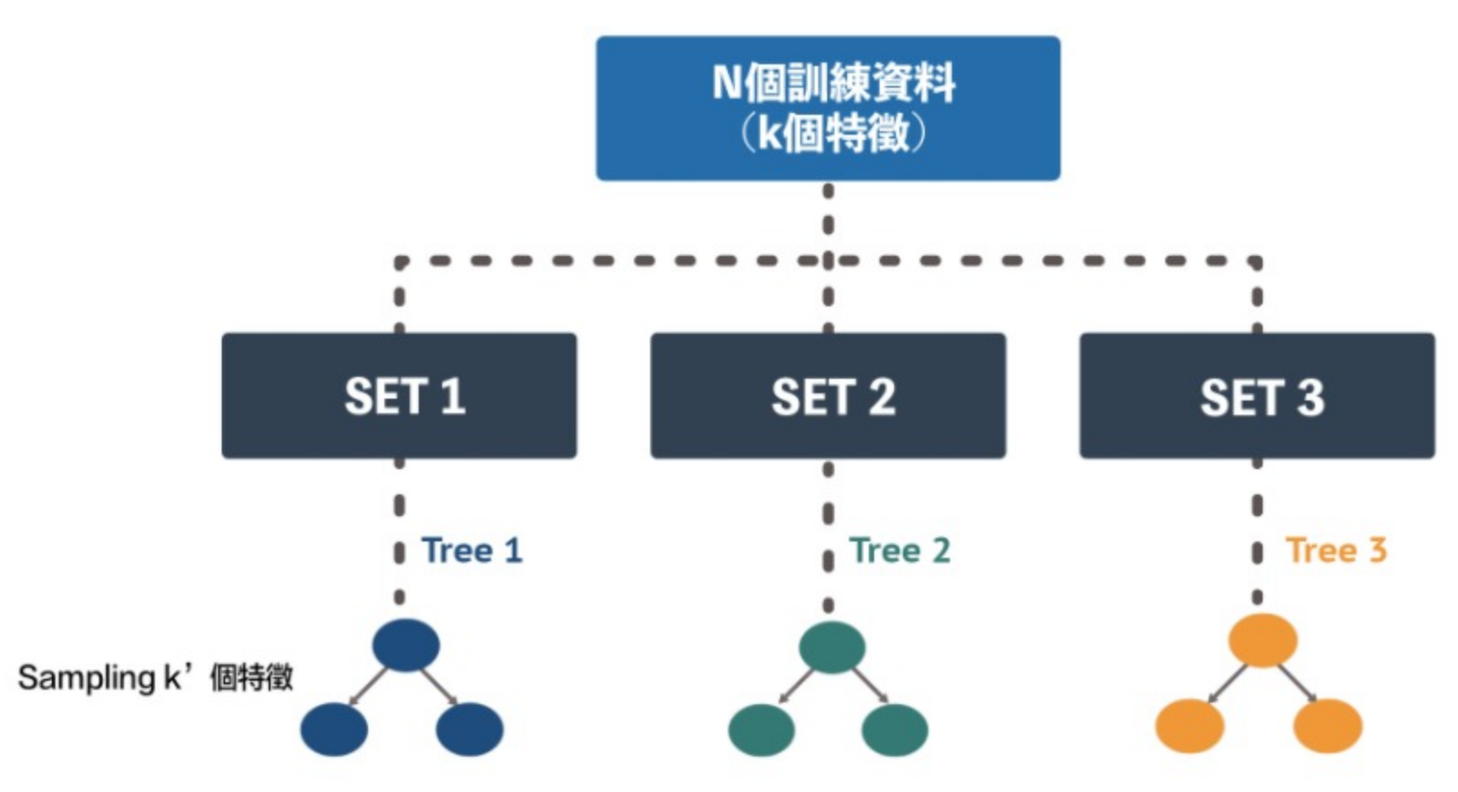
如何取得具有差異的決策樹組成隨機森林呢？可以透過以下方法來得到不同的決策樹：

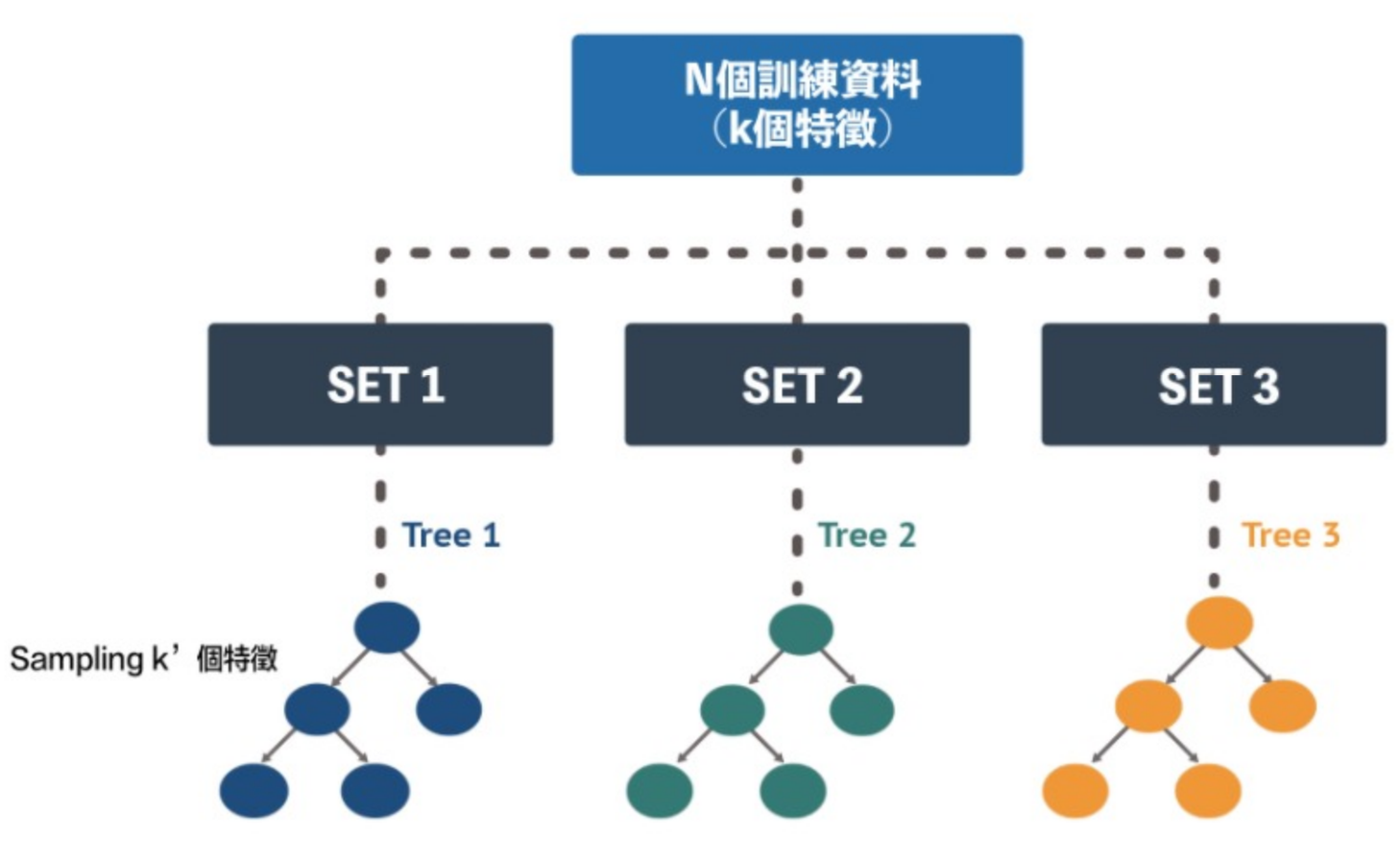
1. 從訓練集中隨機選取資料來進行模型訓練 (一般為取後放回)
2. 從資料特徵中隨機選取部分特徵來進行模型訓練

1. 從訓練集中隨機選取資料來進行模型訓練 (一般為取後放回)

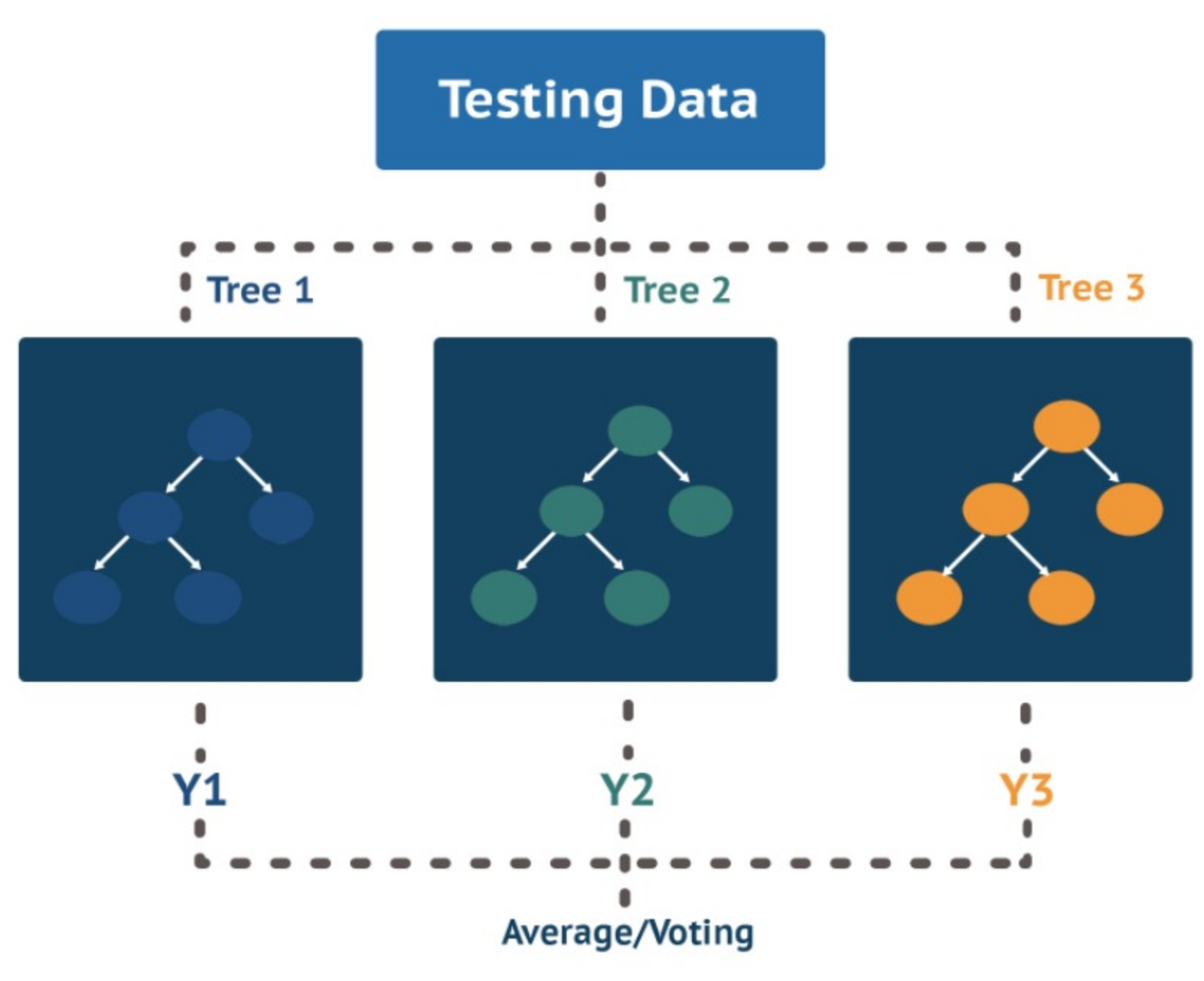


2. 從資料特徵中隨機選取部分特徵來進行模型訓練





預測時將資料輸入所有的決策樹，將其輸出進行平均(Regression)或投票(Classification)來決定最後預測結果。



在 Bagging 演算法中(隨機森林)，最重要的部分就是隨機選取樣本與特徵，這可以使每個子模型(決策樹)，擁有不一樣的特性。且若原始資料存在噪音(noise)，透過抽樣有機會使沒有噪音的資料被選取到，增加模型的泛化能力。

了解隨機森林的運作後，來看看隨機森林的特性與優缺點：

優點：

- 透過 Bagging，取得的精度較單個模型算法好
- 引入隨機性(隨機樣本、特徵)，不容易陷入過擬合
- 能處理數值型與類別型的資料

缺點：

- 所需的訓練時間與空間(複雜度)較大
- 對於小資料或特徵較少的資料，效果較不好
- 相較於決策樹，可解釋性較不足

知識點回顧

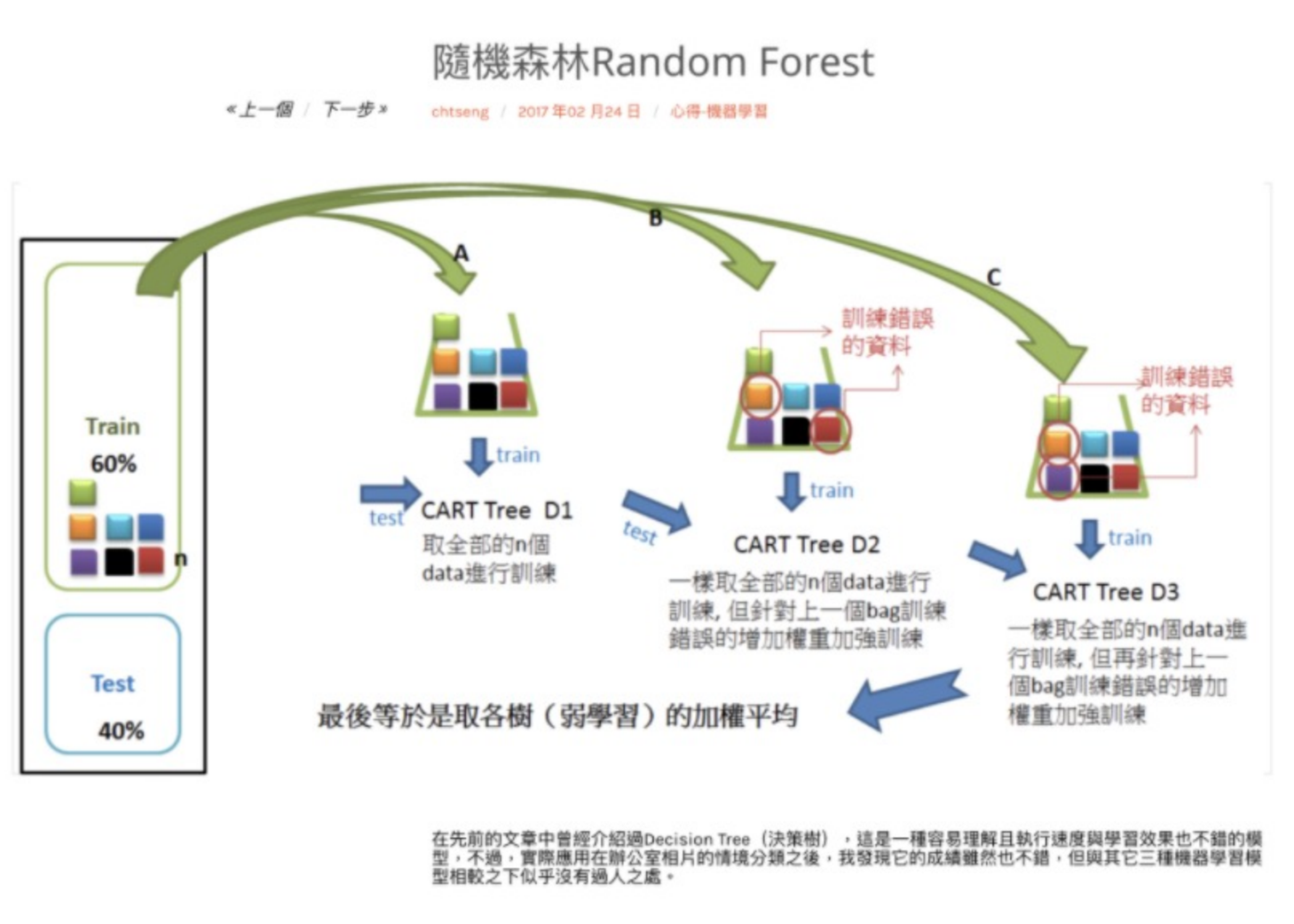
在這章節我們學習到了

- 了解集成(Ensemble)中的Bagging
- 了解隨機森林與其運作原理

延伸閱讀

網站：Bagging介紹
Bagging 介紹 YouTube 影片

網站：更多隨機森林
更多關於隨機森林的介紹



下一步：完成作業