隨機森林
Random Forest

雲科財金系 張子溥 2018.11.15

ALLPPT.com _ Free PowerPoint Templates, Diagrams and Charts

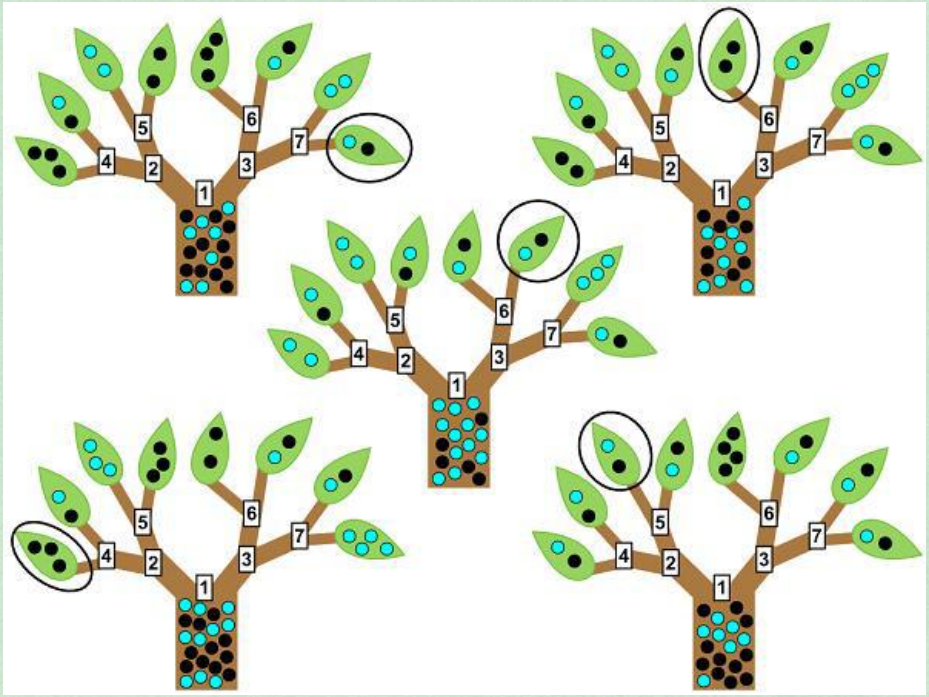
何謂森林？？

字詞	【森林】
注音	ㄙㄣ ㄌㄧㄣˊ
漢語拼音	sēn lín
相似詞	叢林
釋義	樹木密生的寬廣地區。如：「森林兼具木材供應、水土保持、觀光休憩等功能，是重要的天然資源之一。」

資料來源：教育部重編國語辭典修訂本

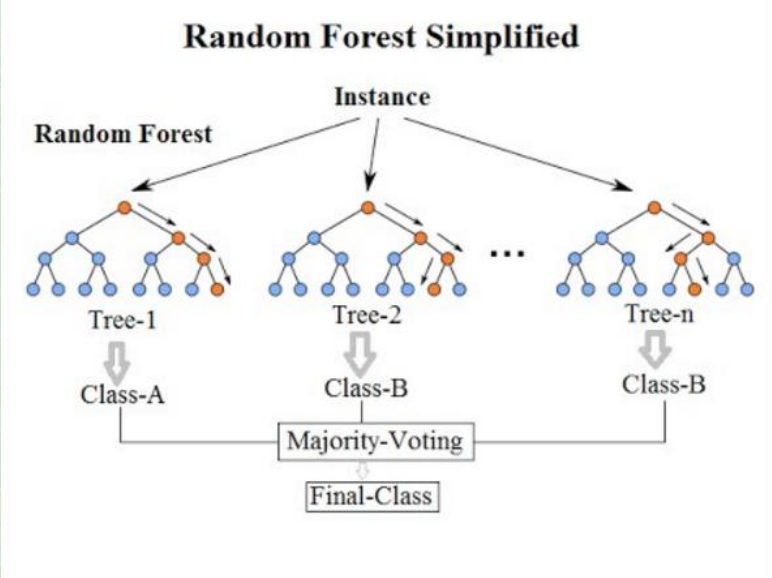
增強學習的概念

- 單一分類結果(決策樹)不一定可靠，而且沒有一個單一分類結果可以以一敵百。因此，一個民主的取向就是增強學習(ensemble learning，或稱集成學習)，也就是將多種結果綜合起來，增強單一結果。



隨機森林的邏輯

- 在機器學習中，隨機森林是一個**包含多個決策樹**的分類器，並且其輸出的類別是由個別樹輸出的類別的**眾數**而定。
- 其實從直觀角度來解釋，每棵決策樹都是一個分類器，那麼對於一個輸入樣本，N棵樹會有N個分類結果。而隨機森林集成了所有的分類投票結果，將投票次數最多的類別指定為最終的輸出



隨機森林的「隨機」

- **對於樣本與變數的隨機抽樣**
- 如果訓練集大小為N，對於每棵樹而言，隨機且有放回地從訓練集中的抽取N個訓練樣本（這種採樣方式成為bootstrap sample），作為該樹的訓練集
- 如果每個樣本的特徵維度為M，指定一個常數m<<M，隨機地從M個特徵中選取m個特徵子集，每次樹進行分裂時，從這m個特徵中選擇最優的；
- 兩個隨機性的引入對隨機森林的分類性能至關重要。由於它們的引入，使得隨機森林不容易陷入過擬合，並且具有很好得抗噪能力

隨機森林的生成

- 對採樣之後的資料使用完全分裂的方式建立出決策樹，這樣決策樹的某一個葉子節點要麼是無法繼續分裂的，要麼裡面的所有樣本的都是指向的同一個分類。分裂的辦法是：採用上面說的列採樣的過程從這m個屬性中採用某種策略（比如說資訊增益）來選擇1個屬性作為該節點的分裂屬性。
- 決策樹形成過程中每個節點都要按完全分裂的方式來分裂，一直到不能夠再分裂為止。

隨機森林分類效果（錯誤率）

- 與兩個因素有關：
  - 森林中任意兩棵樹的相關性：相關性越大，錯誤率越大；
  - 森林中每棵樹的分類能力：每棵樹的分類能力越強，整個森林的錯誤率越低。
- 減小特徵選擇個數m，樹的相關性和分類能力也會相應的降低；增大m，兩者也會隨之增大。所以關鍵問題是如何選擇最優的m（或者是範圍），這也是隨機森林唯一的一個參數。

袋外錯誤率（oob error）

- 構建隨機森林的關鍵問題就是如何選擇最優的m，這裡要依據袋外錯誤率oob error（out-of-bag error）
- 隨機森林有一個重要的優點就是，沒有必要對它進行交叉驗證或者用一個獨立的測試集來獲得誤差的一個無偏估計。它可以在內部進行評估，也就是說在生成的過程中就可以對誤差建立一個無偏估計。
- 我們知道，在構建每棵樹時，我們對訓練集使用了不同的bootstrap sample（隨機且有放回地抽取）。所以對於每棵樹而言（假設對於第k棵樹），大約有1/3的訓練實例沒有參與第k棵樹的生成，它們稱為第k棵樹的oob樣本。

袋外錯誤率（oob error）

- 而這樣的採樣特點就允許我們進行oob估計，它的計算方式如下：（note：以樣本為單位）
- 1）對每個樣本，計算它作為oob樣本的樹對它的分類情況（約1/3的樹）；
- 2）然後以簡單多數投票作為該樣本的分類結果；
- 3）最後用誤分個數占樣本總數的比率作為隨機森林的oob誤分率。

- oob誤分率是隨機森林泛化誤差的一個無偏估計，它的結果近似於需要大量計算的k折交叉驗證。