Day 49 集成

混合泛化(Blending)

出題教練 陳明佑

知識地圖 機器學習- 參數調整 - 超參數調整與優化

CUPOY

機器學習概論 Introduction of Machine Learning

監督式學習
Supervised Learning

前處理 Processing → 探索式數據分析 Exploratory Data Analysis → 特徵工程 Feature Engineering → 模型選擇 Model selection → 參數調整 Fine-tuning → 集成 Ensemble

非監督式學習
Unsupervised Learning

- 分群 Clustering
- 降維 Dimension Reduction

參數調整 Fine-tuning

- 混合泛化 Blending
- 堆疊泛化 Stacking

本日知識點目標

- 資料工程中的集成，有哪些常見的內容?
- 混合泛化為什麼能提升預測力，使用上要注意什麼問題?

什麼是集成

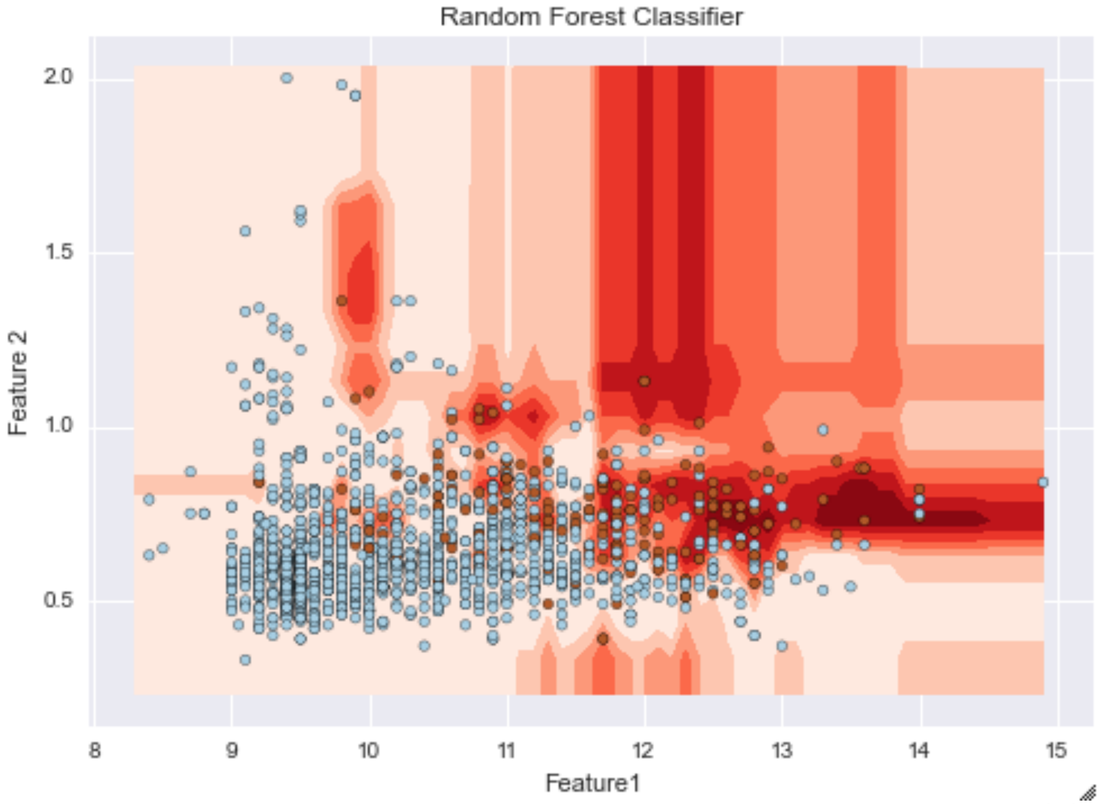
CUPOY

- 集成是使用不同方式，結合多個/多種**不同分類器**，作為**綜合預測**的做法統稱
- 將模型**截長補短**，也可說是機器學習裡的 **和議制 / 多數決**

- 其中又分為資料面的集成：如裝袋法(Bagging) / 提升法(Boosting)
- 以及模型與特徵的集成：如混合泛化(Blending) / 堆疊泛化(Stacking)

資料面集成：裝袋法 (Bagging)

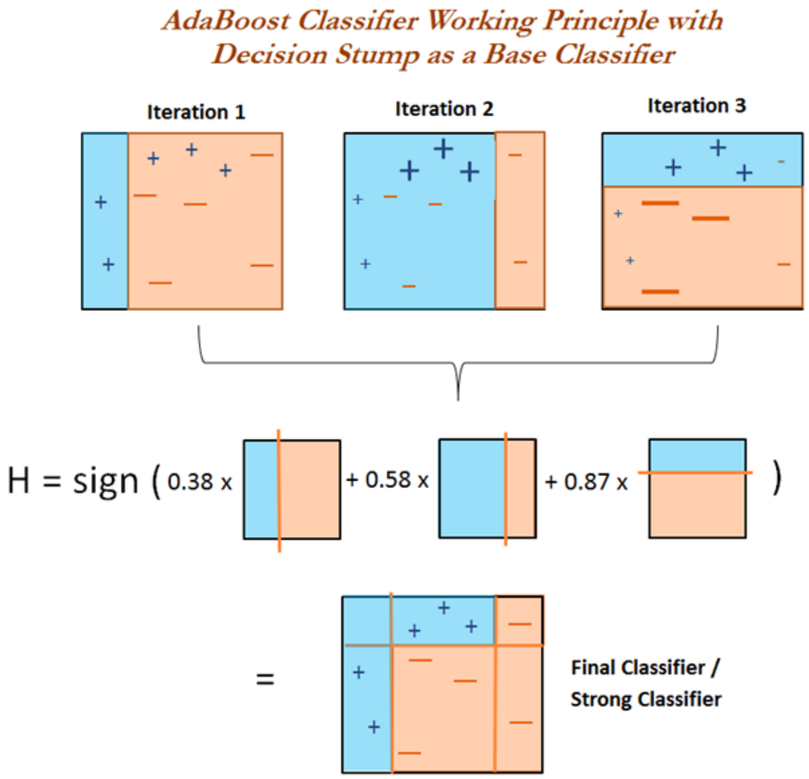
- 裝袋法顧名思義，是將資料放入袋中抽取，每回合結束後全部**放回袋中重抽**
- 再搭配弱分類器取平均/多數決結果，最有名的就是前面學過的**隨機森林**



圖片來源：stackexchange.

資料面集成：提升法 (Boosting)

- 提升法則是由之前模型的預測結果，去改變資料被抽到的權重或目標值
- 將錯判資料被抽中的機率放大，正確的縮小，就是**自適應提升** (AdaBoost, Adaptive Boosting)
- 如果是依照估計誤差的殘差項調整新目標值，則就是**梯度提升機** (Gradient Boosting Machine) 的作法，只是梯度提升機還加上用梯度來選擇決策樹分支



圖片來源：subscription

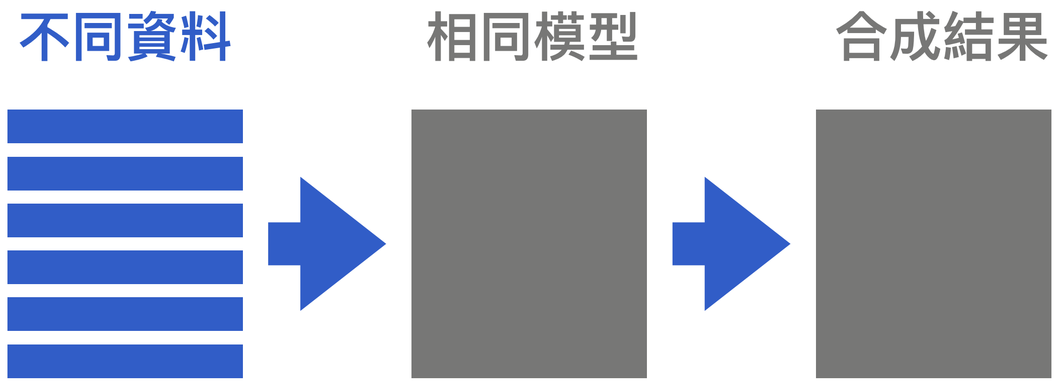
資料集成 v.s. 模型與特徵集成

CUPOY

- 兩者雖然都稱為集成，其實適用範圍差異很大，通常不會一起提及
- 這裡為了避免同學混淆，在這邊將兩者做個對比
- 資料集成

  Bagging / Boosting

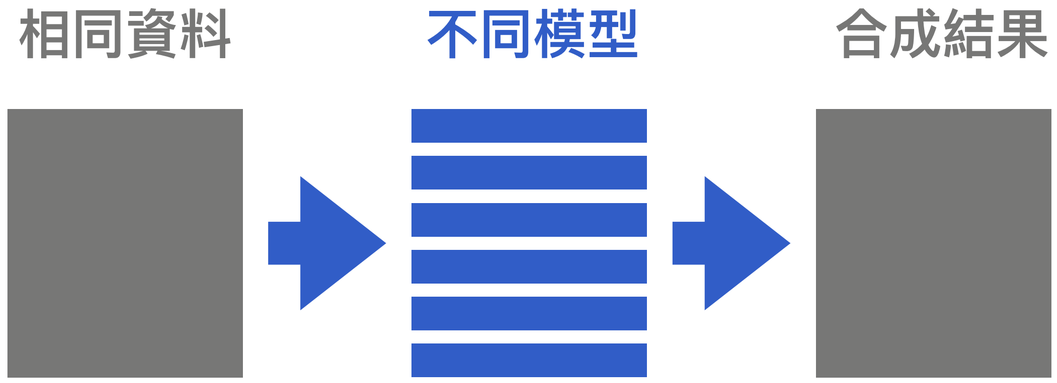
  - 使用不同訓練資料 + 同一種模型，多次估計的結果合成最終預測



- 模型與特徵集成

  Voting / Blending / Stacking

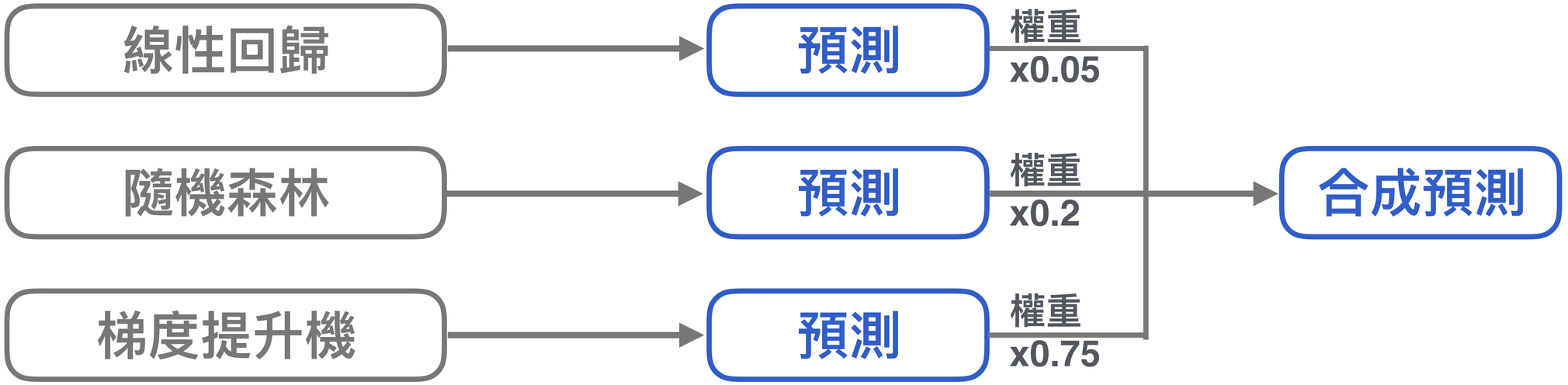
  - 使用同一資料 + 不同模型，合成出不同預測結果



混合泛化 (Blending) (1 / 3)

CUPOY

- 其實混合泛化非常單純，就是將不同模型的預測值**加權合成**，權重和為 1
  如果取預測的平均 or 一人一票多數決(每個模型**權重相同**)，則又稱為 **投票泛化(Voting)**



- 雖然單純，但因為**最容易使用**且**有效**，至今仍然是競賽中常見的作法

混合泛化 (Blending) (2 / 3)

CUPOY

容易使用

- 不只在一般機器學習中有用，影像處理或自然語言處理等深度學習，也一樣可以使用
- 因為只要有預測值(Submit 檔案)就可以使用，許多跨國隊伍就是靠這個方式合作
- 另一方面也因為只要用預測值就能計算，在競賽中可以快速合成多種比例的答案，妥善消耗掉每一天剩餘的 Submit 次數

混合泛化 (Blending) (3 / 3)

CUPOY

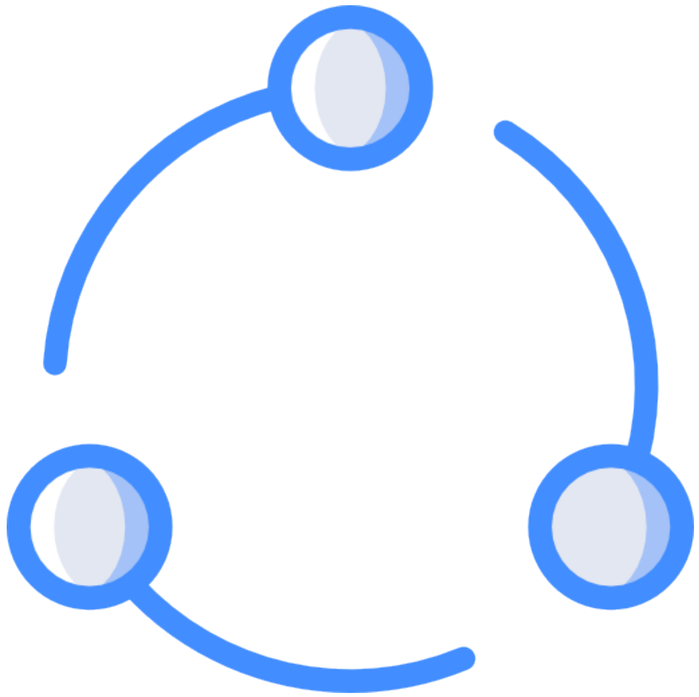
效果顯著

- Kaggle 競賽截止日前的 Kernel，有許多只是對其他人的輸出結果做 Blending，但是因為分數較高，因此也有許多人樂於推薦與發表
- 在2015年前的大賽中，Blending 仍是主流，例如林軒田老師也曾在課程中提及：有競賽的送出結果，是上百個模型的 Blending

注意事項

- Blending 的前提是：個別**單模效果都很好**(有調參)並且**模型差異大**，單模要好尤其重要，如果單模效果差異太大，Blending 的效果提升就相當有限

重要知識點複習



- 資料工程中的集成，包含了資料面的集成 - **裝袋法(Bagging)** / **提升法(Boosting)**，以及模型與特徵的集成 - **混合泛化(Blending)** / **堆疊泛化(Stacking)**
- 混合泛化提升預測力的原因是基於**模型差異度大**，在預測細節上能互補，因此預測模型只要各自調參優化過且原理不同，通常都能使用混合泛化集成

解題時間
It’s Your Turn

請跳出PDF至官網Sample Code＆作業
開始解題