

2.4 模型評估和調參



交叉驗證 Cross-Validation

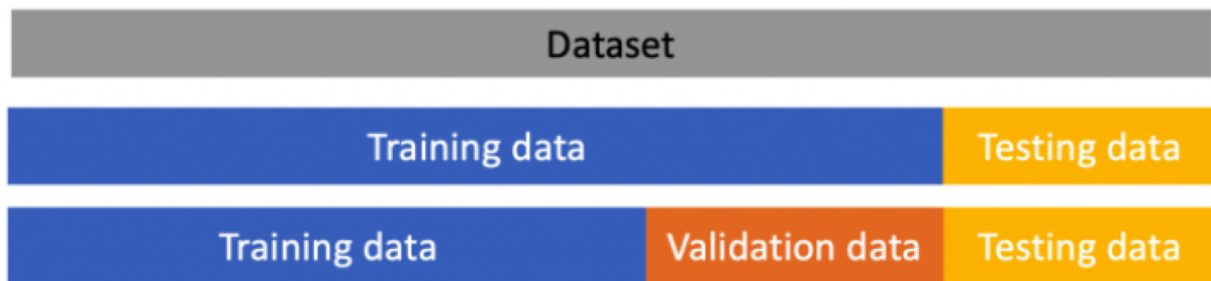


為什麼要做交叉驗證(Cross-Validation)

為了避免模型訓練發生過度擬合，會從訓練集切一小部分資料出來進行驗證。

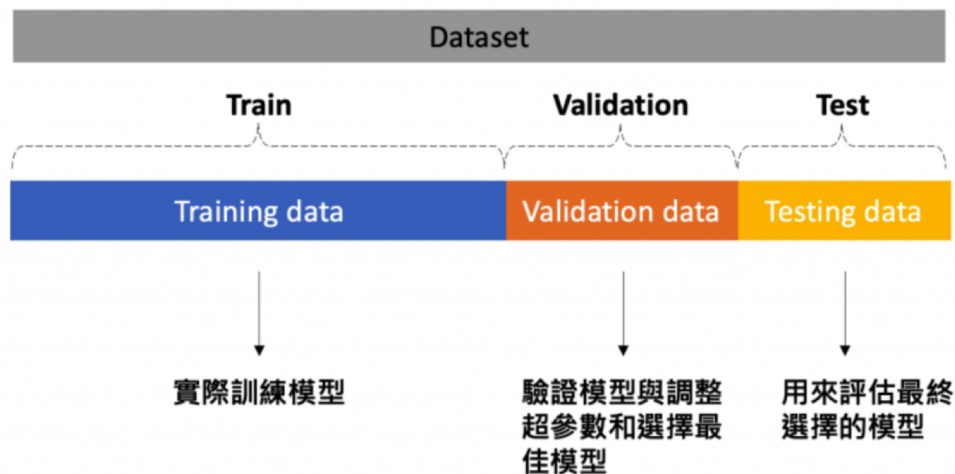
驗證集是用來檢視模型在訓練過程中每次的迭代結果訓練的好不好。

為確認模型泛化能力，以有效的切割驗證集來評估模型，可採用交叉驗證 Cross-Validation 的技巧來獲得最佳驗證。



- Holdout
- K-fold
- Bootstrap

Holdout Method



優點：

1. 簡單實作。
2. 驗證集可以被拿來評估模型在訓練過程中的學習成果。
3. 測試集可以評估模型泛化能力。

缺點：

1. 當資料集變異量較大時，驗證集與測試集可能無法足以評估模型。
2. 不適合用在資料不平衡的資料集。

K-fold Cross-Validation

K-Fold 的方法中 K 是可自由調控的，在每次的迭代中會選擇一組作為驗證集，其餘 (k-1) 組作為訓練集。以不同分組訓練的結果進行平均來減少變異，因此模型的性能對數據的劃分就不會那麼敏感。



優點:

1. 降低模型訓練對於資料集的偏差。
2. 訓練集與驗證集完整被充分利用與學習。

缺點:

1. 不適合用於資料不平衡的資料集。
2. 如果要簡單的 K-fold 來尋找超參數會有資料洩漏問題導致訓練結果有偏差，因為在每個 Fold 中都會使用同一組資料進行驗證。
3. 在相同的驗證集計算模型的誤差，當找到了最佳的超參數。可能會導致重大偏差，有過擬合疑慮。

Bootstrapping (自助抽樣法)

從給定訓練集中有放回的均勻抽樣，每當選中一個樣本，它等可能地被再次選中並被再次添加到訓練集中。假設每次訓練都採樣十個樣本，在這十筆資料中很有可能會再次被隨機抽到。剩下沒有抽到的資料則都變成測試集，用來評估訓練完的模型。



StratifiedKFold

StratifiedKFold 是 KFold 的加強版，特別設計來處理類別不平衡的情況。

它會確保每個摺疊中各類別的樣本比例與原始資料集中的比例相同。

保證每一個摺疊中的類別分佈大致相同，有助於減少因為資料不平衡而導致的偏差。

- 資料分割：在 StratifiedKFold 中，分割資料的方式是按類別進行**分層抽樣** (stratified sampling)，確保每一個摺疊中的類別比例與原始資料集相同。避免某些摺疊中某些類別樣本過少，從而影響模型訓練的穩定性。
- 適用情況：當資料集存在類別不平衡的情況時（例如，某個類別的樣本數量遠小於其他類別），使用 StratifiedKFold 可以更好地反映每個類別在每一折中的分佈。

超參數調參



- Grid Search 網格搜索/窮舉搜索
- Random Search 隨機搜索

Grid Search 網格搜索/窮舉搜索

在所有可能的參數中，透過排列組合嘗試每一種可能性。
將表現最好的參數最為最終的超參數搜尋結果。

缺點：

有許多超參數要尋找時，排列組合就會變非常多，導致搜索的時間變長花費的資源也變大。
因此這種暴力式的搜索方法適合在小的資料集上被採用

Sklearn 中有提供 GridSearchCV 方法，可設定參數列表，並透過所有可能的參數組合一個個嘗試找到最合適的參數。

Random Search 隨機搜索

在所有可能的候選參數中隨機挑選一個數值並嘗試。

優點：

如需調的參數較多時，使用隨機搜索可以降低搜索時間，同時又能確保一定的模型準確性。

Sklearn 有提供 `RandomizedSearchCV` 方法可以呼叫，與網格搜索的差別在於可將欲搜尋的超參數設定一個期望的範圍。該方法會在此範圍中隨機抽一個數值並進行模型訓練並驗證模型。並找出所有隨機組合中表現最好的一組超參數。



THANKS

