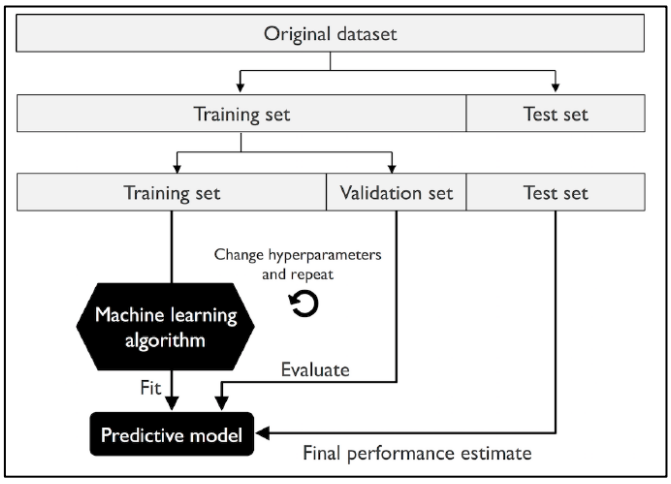
學習模型評估

I. 評估模型有以下幾種標準:

1. **準確率**: 對新的或未知的資料正確判斷或猜測的能力。(low bias)
2. **穩定性 (Robust)**: 給定噪音資料或有空缺值的資料，模型正確預測或判斷的能力。(low variance)
3. **速度**: 產生和使用模型的計算成本之花費。
4. **可度良性**: 對大量資料，有效的建構模型的能力。
5. **可解釋性**: 學習模型提供了解程度。難評量。

II. 針對準確率、穩定性，機器學習會有以下幾種交叉驗證(Cross - validation, CV) 方法來驗證「你設計出來模型」的好壞:

數據庫(database)沒有先切割好「訓練資料(Training data)」和「測試資料(Testing data)」，或是你要從「訓練資料(Training data)」找到一組最合適參數出來，比如 SVM 的懲罰參數(Penalty parameter)，就可以從訓練資料(Training data)做交叉驗證找出來，而不是從「測試資料(Testing data)」得到參數。
機器學習最忌諱把「測試資料(Testing data)」偷偷拿進到模型內訓練或是找參數。在做模型 performance 評估時，要記住一件事情「測試資料(Testing data)」絕對不能進到模型內訓練或是找參數。

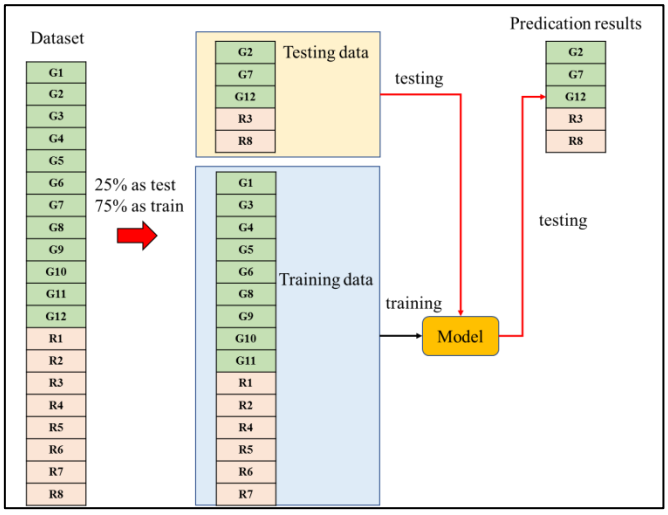


1. Holdout CV:

- Holdout 是指從資料集中隨機取得 p%資料當作「訓練資料(Training data)」和剩下的(1-p)%當做「測試資料(Testing data)」。最後的結果(Predication results)在和測試資料的真實答案(ground truth)進行成效比對(Performance

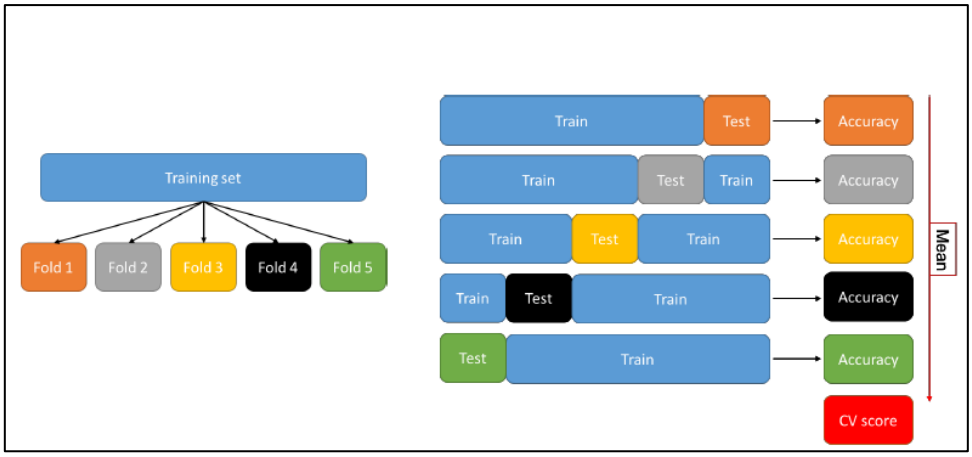
Comparison)。

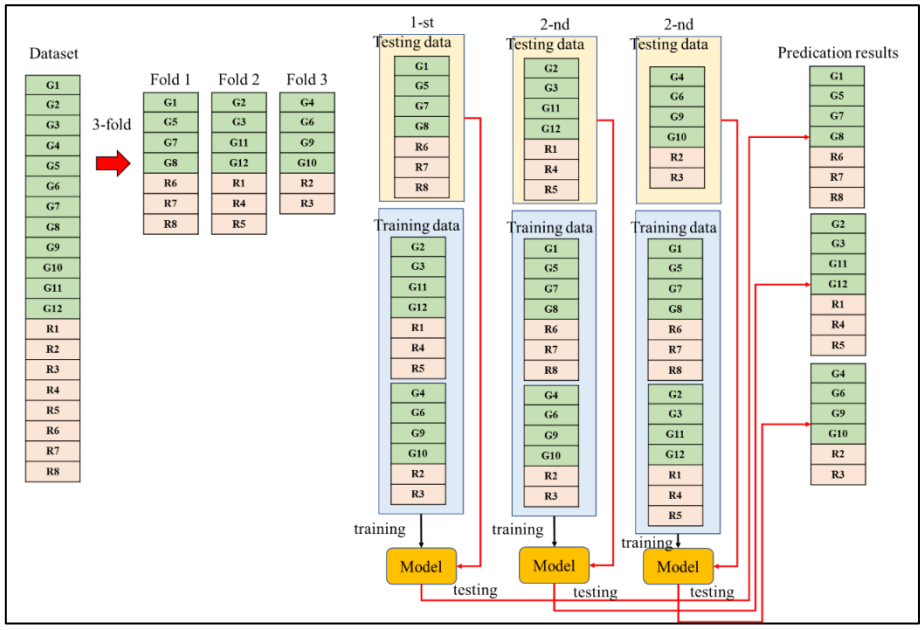
- 這個隨機不是不考慮資料的類別，也就是說 Holdout 是從每一類都取 p%資料當作「訓練資料(Training data)」，從每一類剩下的(1-p)%當做「測試資料(Testing data)」。 (Stratified)



2. K-fold CV:

- K-fold 是比較常用的交叉驗證方法。做法是將資料隨機平均分成 k 個集合，然後將某一個集合當做「測試資料(Testing data)」，剩下的 k-1 個集合做為「訓練資料(Training data)」，如此重複進行直到每一個集合都被當做「測試資料(Testing data)」為止。最後的結果(Predication results)在和真實答案(ground truth)進行成效比對(Performance Comparison)。
- 看這 n 次的 performance 最後的平均和標準差，最後會去 report 這個平均值當做模型的 performance。而標準差用來表示模型的穩定程度，如果標準差太大代表模型的穩定度不夠好，在做 generalized model 的時候這個方法可能就不太合適。
- k = 10 為最普遍。
- 這個隨機分 k 個集合也是要考慮資料的類別，也就是說 K-fold 是從每一類都隨機分割成 k 個集合。(分層 k 折交叉驗證 Stratified k-fold cross-validation)



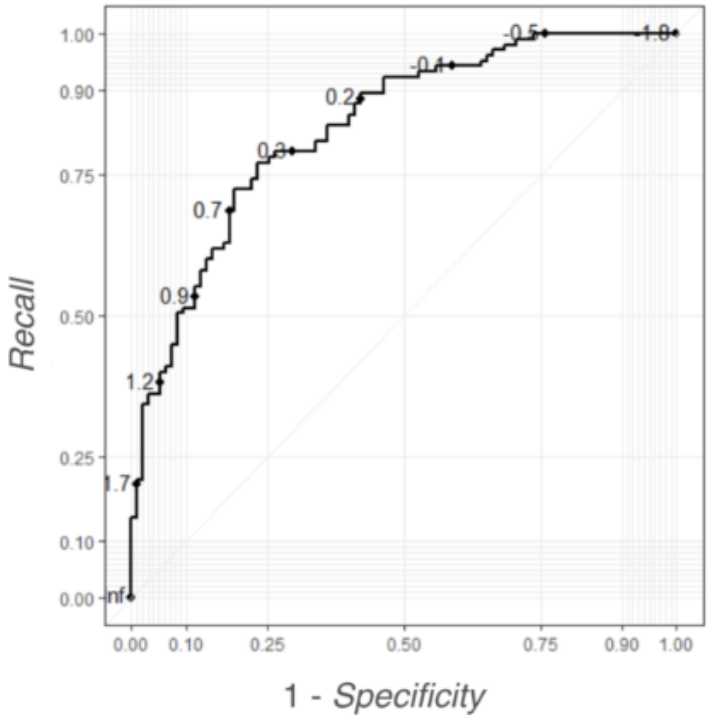


1. 分類效能度量指標 ：ROC 曲線、AUC 值、正確率、召回率、敏感度、特異度

預測類別 / 實際類別	Class 1	Class 2
Class 1	TP（true positive）	FN（false negative）
Class 2	FP（false positive）	TN（true negative）

- T/F 表示 true/false　表示預測的是不是對的
- P/N 表示 positive/negative　表示预测数据是正样本还是负样本

- True Positive （真正, TP）被模型预测为正样本，实际为正样本
  False Positive （假正, FP）被模型预测为正样本， 实际为负样本
  True Negative（真负 , TN）被模型预测为负样本，实际为负样本
  False Negative（假负 , FN）被模型预测为负样本，实际为正样本
- **True Positive Rate（真正率 , TPR）或靈敏度（sensitivity）**
  TPR = TP /（TP + FN）
  True Negative Rate（真負率 , TNR）或特指度（specificity）
  TNR = TN /（TN + FP）
  False Positive Rate （假正率, FPR）
  FPR = FP /（FP + TN）
  False Negative Rate（假負率 , FNR）
  FNR = FN /（TP + FN）
- 準確率 precision： TP/ (TP+FP) 預測為正樣本中有多少是真正的正樣本
  召回率 recall： TP/ (TP +FN) 正樣本有多少被成功預測為正樣本
  F1: 2 / (1/p + 1/r)
- ROC 曲線:
  ROC 曲線指受試者工作特徵曲線 / 接收器操作特性曲線(receiver operating characteristic curve), 是反映敏感性和特異性連續變數的綜合指標,是用構圖法揭示敏感性和特異性的相互關係，它通過將連續變數設定出多個不同的臨界值，從而計算出一系列敏感性和特異性，再以敏感性為縱座標、(1-特異性）為橫座標繪製成曲線，曲線下面積越大，診斷準確性越高。ROC 曲線和它相關的比率。
  理想情況下，TPR 應該接近 1，FPR 應該接近 0。ROC 曲線上的每一個點對應於一個 threshold，對於一個分類器，每個 threshold 下會有一個 TPR 和 FPR。比如 Threshold 最大時，TP=FP=0，對應於原點；Threshold 最小時，TN=FN=0，對應於右上角的點(1,1)。
  隨著閾值 theta 增加，TP 和 FP 都減小，TPR 和 FPR 也減小，ROC 點向左下移動。
  實際中需要根據實際場景進行合理選擇閾值,比如在人臉識別支付的時候，對 FPR 比較敏感，FPR 越小，錯誤接收的使用者可能性越小，用戶的錢財越安全， 這個時候，可以提高閾值，降低 FPR，TPR 也會下降(用戶體驗會下降)。在如精准行銷領域的商品推薦模型，模型目的是儘量將商品推薦給感興趣的用戶，若用戶對推薦的商品不感興趣，也不會有很大損失，因此此時 TPR 相對 FPR 更重要，這個時候可以降低閾值。



參考資料

1. Class Handout, Lee, Chia Jung professor, MDM64001, School of Big Data Management, Soochow University
2. 交叉驗證(Cross-validation, CV) https://medium.com/@chih.sheng.huang821/%E4%BA%A4%E5%8F%89%E9%A9%97%E8%AD%89-cross-validation-cv-3b2c714b18db
3. 【機器學習】分類效能度量指標：ROC 曲線、AUC 值、正確率、召回率、敏感度、特異度 https://www.itread01.com/content/1547130433.html