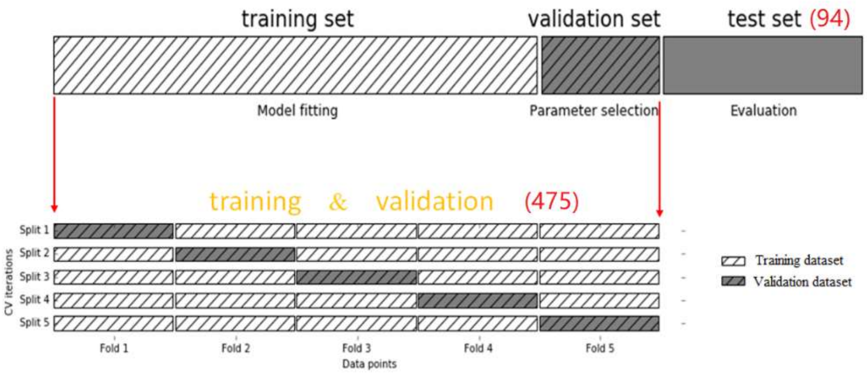
NYCU 2022 FALL 516148
DME Machine Learning HW #5

(1) From your HW4 code, write a Python code to read HW5 data file (”hw5_cancer.csv”)。There are 569 datasets, 1 dataset per line, plus a header line in the CSV file. For each dataset, there are 30 features and 1 classification (0 as malignant and 1 as benign). Data is separated by a comma.
從您的 HW4 代碼，編寫一個 Python3 程序以讀取 HW5 數據文件 (”hw5_cancer.csv”)。整體共計 569 數據集 + header. 除了 header, HW5 數據文件每行是 1 個數據集(dataset). 每個數據集(每行)包含 30 個 features and 1 個 classification (0 是 惡性, 1 是 良性). 每個數據都用逗號分隔.

(2) Randomly pick 94 datasets as the test set, and use the remaining 475 datasets for 5-fold cross validation training.
讀取整體 569 數據集, 隨機選取 94 個數據集作為測試集,使用剩餘的 475 個數據集進行 5 折交叉驗證訓練



(3) Use (a) **LogisticRegression**, (b) **Random Forests**, (c) **Gradient Boosted Regression Trees** methods to train your model with the training and validation dataset (475) and test data set (94).
使用 (a) LogisticRegression、(b) 隨機森林、(c) 梯度提升回歸樹，使用訓練/驗證集 (475)和測試數據 (94)訓練您的模型。

(4) You are to train all three models which all make the test data (94) score above 0.940, and the training/validation data (475) score doesn’t appear overfitting (i.e. close to your test data score). Present your choice of “best” model.
你要訓練這 3 個模型, 使它們測試 (94)分數都高於 0.940，而訓練驗證集(475)分數不出現過度擬合 (接近測試分數). 展示標記您選擇的“最佳”模型。

(5) You can import all corresponding classifiers as shown below.
您可以導入相應的分類器，如下所示。 下面給出了使用 3 個分類器的示例 Python 代碼。

```
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.ensemble import GradientBoostingClassifier
```

(6) Estimate work time: 4-8 hours.
估計所需時間：4-8 小時

(7) Due time: before 12/2/2022 class time. Upload to E3 your Python code (“ yourID_name _HW5.py”) and your running result which includes training and test scores of three methods, such as shell window screen image (“yourID_name_HW5_cancer.jpg”) or shell window printout lines (“yourID_name_HW5_cancer.txt”)
截止時間: 在 2022 年 12 月 2 日上課之前上傳 E3 提交 your python 程序 (“ yourID_name _HW5.py”) 和程序運行結果 (包括三種方法的訓練和測試成績), 和 shell 視窗 print-screen 圖 (“yourID_name_HW5_cancer.jpg”) 或 shell 視窗打印輸出行 (“yourID_name_HW5_cancer.txt”)。