機器學習作業指示 - Homework #5

目標：

實做兩種特徵篩選方法(Sequential Forward Selection 和 Fisher's Criterion)；比較 Filter-based 和 Wrapper-based 特徵篩選法的異同；並利用**乳癌資料集**，搭配 LDA 分類器和 2-Fold CV 完成分類任務，並使用平衡分類率以評估分類器效能。

乳癌資料集, Breast Cancer dataset：

https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic

資料描述：

- 乳癌資料集如同Iris dataset，是在機器學習領域中常被用作演算法效能驗證的開放資料集
- 包含兩個類別：惡性腫瘤和良性腫瘤，前者標籤為 M = malignant 而後者為 B = benign
- 共包含569筆資料，而每筆資料皆以30種特徵($N_s = 30$)進行描述
- 取得資料集：可以透過上述網址直接下載，而在 Python 上可以使用scikit-learn套件載入資料集(sklearn.datasets.load_breast_cancer)

Part1: Sequential Forward Selection, SFS

Step1: 透過 LDA 和 2-Fold CV ，在全部N_s種特徵中找出能達到最高之交叉驗證平衡分類率(Highest validated balanced accuracy)的單一特徵(Single feature)。(紀錄底線標記之結果)

Hint：如出現兩個以上的 Highest validated balanced accuracy，選擇排序最靠前者即可。

Step2: 將餘下的特徵($N_s - 1$種)各自與 **Step1** 找到的最佳之單一特徵進行組合，形成多組特徵對($N_s - 1$組)，並選擇能達到 Highest validated balanced accuracy 的特徵子集合。(紀錄底線標記之結果)

Hint：在此步驟中選出的特徵子集合包含 2 種特徵

Step3: 基於與 **Step2** 相同的邏輯，將餘下的特徵($N_s - 2$種)各自與 **Step2** 找到的特徵子集合進行組合，形成多組三維特徵子集合($N_s - 2$組)，並選擇能達到 Highest validated balanced accuracy 的特徵子集合。(紀錄底線標記之結果)

Hint：在此步驟中選出的特徵子集合包含 3 種特徵

Step4: 重複相同的程序直到選出的特徵子集合中的特徵數相等於N_s，並記錄 Highest validated balanced accuracy。(紀錄底線標記之結果)

Step5: 在前述步驟所記錄的所有 Highest validated balanced accuracy 中，最高者的特徵子集合即作為最佳特徵子集合(Optimal feature subset)；並將最佳特徵子集合(Optimal feature subset)中所包含的特徵和特徵數記錄下來

Part2: Fisher's Criterion

Step1: 實現 Fisher's Criterion 演算法（**請勿直接使用開源的** Fisher's Criterion **套件**）

Step2: 計算全部 N_s 種特徵的 Fisher's score

Step3: 根據 **Step2** 所計算出的結果，對特徵進行降序排列(Rank in descending order)；並初始化 $N = 0$

Step4: $N = N + 1$；並透過 LDA 和 2-Fold CV 計算 Fisher's score 最高 N 筆特徵 (Top-N-ranked features) 之 Validated balanced accuracy。(紀錄底線標記之結果)

Step5: 重複 **Step4**，直到 $N = N_s$

Step6: 在前述步驟所記錄的 Validated balanced accuracy 中，最高者的特徵子集合即作為最佳特徵子集合(Optimal feature subset)；並將最佳特徵子集合(Optimal feature subset)中所包含的特徵和特徵數記錄下來

Part3：Discussion and results presenting

整理上述結果並加以呈現，以圖表或表格等方式，重點在於使其清晰易讀。

另外，請試著討論以下問題：

1. Sequential Forward Selection 和 Fisher's Criterion 分別屬於 Filter-based 和 Wrapper-based 中的何種特徵篩選方法？
2. 一般來說 Filter-based 和 Wrapper-based 各有什麼性質或優缺點？
3. 在本次作業的結果中是否有展現出跟上一題你的回答有一致的現象呢? 不管是否一致皆請你試著討論與分析原因。

作業繳交注意事項(遲交一周該次作業打8折，遲繳超過一周視同該次作業0分)：

1. 作業報告請以書面(pdf or Markdown)呈現，並將程式碼一併壓縮在一個壓縮檔中。
2. 壓縮檔名請符合下列格式：「學號_姓名_HW5」(如：309511001_王小明_HW5)，否則作業成績打8折。
3. 請將壓縮檔上傳至E3數位教學平台。

Deadline：2024/12/16 11:59 P.M. (About Two weeks from now)