機器學習作業指示 - Homework #4

目標：

利用鳶尾花資料(Iris data set)來訓練Support Vector Machine (SVM)，採用one-against-one strategy來處理三類別分類問題，並以grid search來最佳化SVM之參數C與sigma，其過程再以two-fold cross validation使所得到的最佳參數組具有較佳的泛化性。

作業內容：

Step1. 將Iris data set的山鳶尾(Setosa)、變色鳶尾(Versicolor)以及維吉尼亞鳶尾(Virginica)各別取前25筆data設為training data，剩餘的75筆設為test data，所有資料皆採用全部4個特徵，並採用同一組SVM參數(C與sigma)進行以下訓練及測試流程。

Step2. 將山鳶尾(Setosa)以及變色鳶尾(Versicolor)分別設為positive class與negative class，並以這兩類training data訓練SVM_{12}。

Step3. 將山鳶尾(Setosa) 以及維吉尼亞鳶尾(Virginica)分別設為positive class與negative class，並以這兩類training data訓練SVM_{13}。

Step4. 將變色鳶尾(Versicolor)以及維吉尼亞鳶尾(Virginica)分別設為positive class與negative class，並以這兩類training data訓練SVM_{23}。

Step5. 將test data分別輸入至**Step2 – Step4**所訓練的3組SVM decision function，並根據3組SVM的決策結果採多數決進行分類，若同票則該筆資料視為分類錯誤，並於最後得到第一個測試分類率。

Step6. 將**Step1**的training data與test data互換，並重複**Step1 – Step5**進行交叉驗證得到第二個測試分類率。

Step7. 將**Step5**與**Step6**求出的兩個分類率平均，得到平均分類率 (數值請列到小數點後第2位，in %)，並**記錄**下來。

Step8. 改用下一組SVM參數(C與sigma)來重複**Step1 – Step7**，直到所有參數組合(7×41組)皆完成測試，並將每組參數的分類率記錄下來，並標註最佳化後的參數組及分類率。

其中，本次作業皆採用RBF-kernel之SVM，且grid search範圍如下：

C: 1, 5, 10, 50, 100, 500, 1000

sigma: 1.05^{-100}, 1.05^{-95}, …, 1.05^{-10}, 1.05^{-5}, 1.05^{0}, 1.05^{5}, 1.05^{10}, …, 1.05^{95}, 1.05^{100}

討論：

1. 請問在grid search的結果中，C的大小與分類率高低有何關係?
2. sigma的大小的改變與分類率是否有關係? 若有，請探討sigma的差異與特徵的數值有什麼關聯性?
3. 若分析過程不採用two-fold cross validation，則分類率是否會更高? 請探討之。

將上列實驗之結果(包括一些step中要記錄的結果)整理後做詳細的討論，並以書面報告呈現。

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_HW4」(如：309511001_王小明_HW4)，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：2024/11/25 11:59 P.M. (Two weeks from now)