

機器學習作業指示 - Homework #1

目標：

1. 透過散佈圖(Scatter plot)視覺化分析鳶尾花資料集(Iris data set)特徵的鑑別度。
2. 利用程式自主實現K-NN分類器，並利用其進行Iris data set之訓練及測試，求出分類率。

資料描述：

- 安德森鳶尾花卉數據集(Anderson's Iris data set)為機器學習領域中，常被用來驗證演算法效能的資料庫。數據庫中包含三種不同鳶尾花標籤(Label)：山鳶尾(Setosa)、變色鳶尾(Versicolor)以及維吉尼亞鳶尾(Virginica)，且每種鳶尾花皆有50筆樣本。而每筆樣本以四種屬性作為特徵(單位：cm)：花萼長度(Sepal length)、花萼寬度(Sepal width)、花瓣長度(Petal length)以及花瓣寬度(Petal width)等四種屬性作為定量分析之數據。
- 讀取鳶尾花資料後會產生150×5的陣列，其中第5行為資料的類別標籤。

作業內容（程式相關部分請使用Python或MATLAB實現）：

1. 於E3的HW1中取得Iris data set(iris.txt)。
2. 將特徵兩兩組合後繪製scatter plot（共6張圖）。
3. 利用程式實現K-NN分類器（請勿使用開源程式或套件）。
4. 利用Iris dataset測試於第3.點完成的K-NN分類器，其中K=1，列出所有可能之特徵組合（共15種組合）的分類率（分類率計算方式請參考Note），並以**表格**呈現(分類率以百分率呈現，並四捨五入至小數第二位，例如：94.67%)。
5. 根據4.之流程，將最近鄰居數K改為3，列出所有可能之特徵組合（共15種組合）的分類率，並以**表格**呈現。
6. 詳細討論分類率與scatter plot之關係，並附上心得。

Note：

- 分類率計算方式：
首先，將各類別資料中的前半資料當作測試資料(Training data)，剩下的後半資料當作測試資料(Test data)，求得第一個分類率；之後，將 Training data 和 Test data 互換，求得第二個分類率；最後，將兩分類率平均，如此便可得到最終分類率。

作業繳交注意事項(遲交視同該次作業0分)：

- 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。
- 壓縮檔名請符合下列格式：「作業_學號_姓名」(如：HW1_309511001_王小明)，否則作業成績打8折。
- 請將壓縮檔上傳至E3教學平台。

Deadline：2024/09/23 11:59 P.M. (Two weeks from now)