**人工智慧與金融科技實務HW6**

**0616098 黃秉茂**

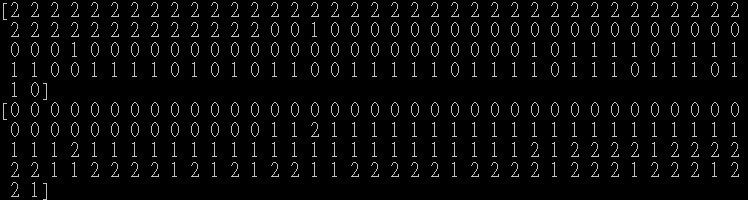
1. 請用下列程式讀取iris資料庫

|  |
| --- |
| From sklearn import datasets  iris = datasets.load\_iris() |

iris.data紀錄150朵鳶尾花萼片及花瓣的長及寬(共4個數值)，iris.target紀錄3種不同的花

請用k-means演算法將資料分成三種類別，並實作上課所述之標準化方法，因這些資料已有類別資訊，請設計一種計算分類準確與否的評估方式，試解釋你的設計理念，並分別針對未經標準化、standard score方法及scaling方法三種作法計算準確度（你設計的評估方式）。

因為Kmeans只會分群，label僅代表誰是同一群，為了好和真正的label做比較，會對label做處理，讓label大致上是一群0再一群1再一群2。改變label但同一label的仍是同一label。 (如下圖所示)



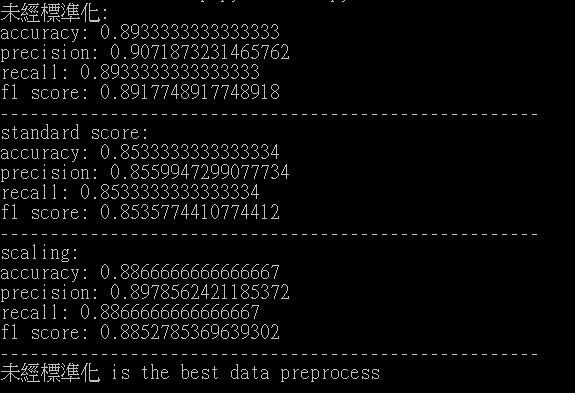
accuracy：(TP + TN) / (P + N)，預測正確的比例，是所有分類正確得百分比，體現了分類模型對樣本的識別能力，accuracy越高，說明模型對樣本的識別能力越強。

recall：(TP) / (P)，真實為true而predict也是true的比例，體現了分類模型對正樣本的識別能力，recall越高，說明模型對正樣本的識別能力越強

precision：(TP) / (TP + FP)，predict為true而真的也是true的比例，體現了模型對負樣本的區分能力，precision越高，說明模型對負樣本的區分能力越強

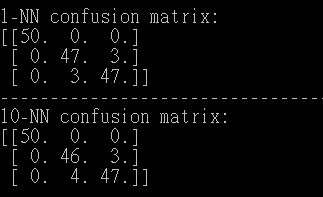
F1-score：2 \* (precision \* recall) / (precision + recall)，是兩者的綜合。F1-score 越高，說明分類模型越穩健。

recall, precision, F1-score有用到average='weighted'，因為多個label不是只有true / false，所以讓各個label輪流當true其他的當false，再把結果依據真實的label數量做平均。



未經標準化 is the best data preprocess 因為他的分數大致上都最高

1. 請用K-nearest neighbors (KNN)演算法對iris資料分群，計算leave one out cross validation，每次拿一筆資料當作test資料，剩下當作train資料，印出1-NN及10-NN的兩個confusion matrix



X軸為true 0 / 1 / 2

Y軸為predcit 0/ 1 / 2