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)

```
[2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 2 2 2 2 2 2 2 2 2 2 2 2 2 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 1 1 1 0 1 1 1 1
 1 1 0 0 1 1 1 1 0 1 0 1 0 1 1 0 0 1 1 1 1 1 0 1 1 1 1 0 1 1 1 0 1 1 1 0 1
 1 0]
[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 2 2 2 2 1 2 2 2 2
 2 2 1 1 2 2 2 2 1 2 1 2 1 2 2 1 1 2 2 2 2 1 2 2 2 2 1 2 2 2 1 2 2 2 1 2
 2 1]
```

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。

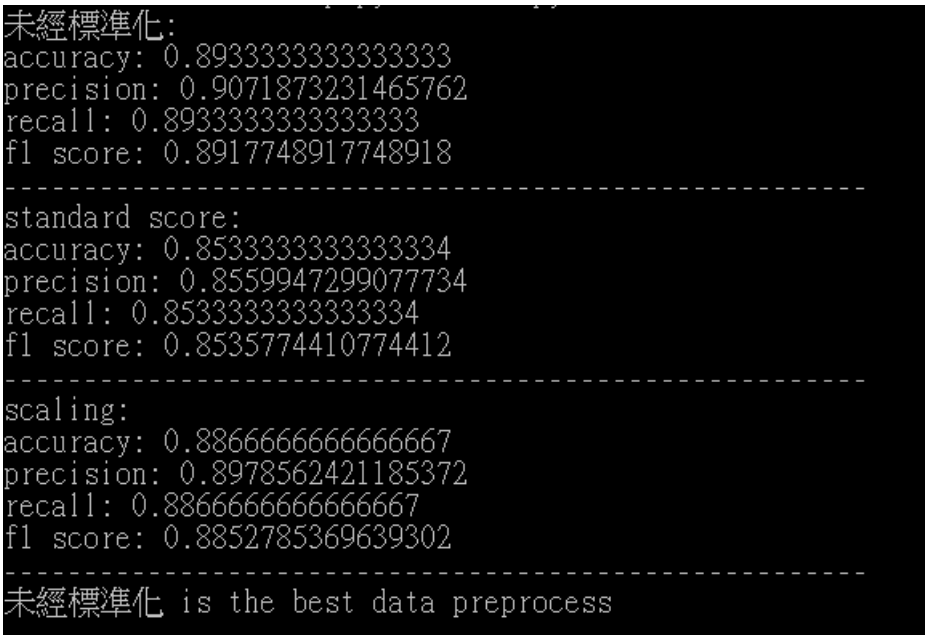
recall：(TP) / (P)，真實為 true 而 predict 也是 true 的比例，體現了分

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

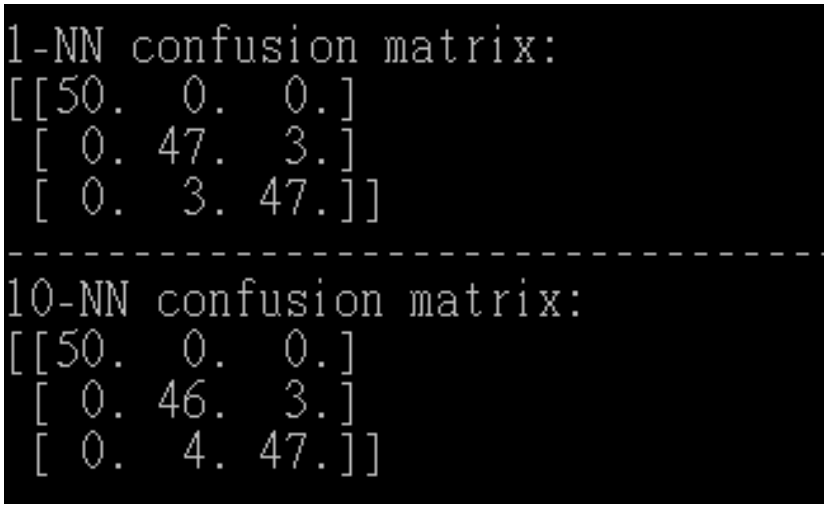
F1-score：2 * (precision * recall) / (precision + recall)，是兩者的綜合。F1-score 越高，說明分類模型越穩健。

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

2. 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