

## B10532001 四電資三 黃哲韋上

執行環境: Win10

程式語言: Python 3.6

1. 將 id 欄位刪除
2. 將資料以 k means 方法來分類，經錯測試 12 群應該比較接近真實分群
3. 訓練三組 model，有 2 個以上 model 認為是同一群則輸出 1，相反則為 0

```
import pandas as pd
import numpy as np
from sklearn.cluster import KMeans

# 讀train set做kmeans
data = pd.read_csv("data.csv", header=0, skip_blank_lines=True)
data = data.drop('id', axis = 1) # 拿掉id
kmeans_fit1 = KMeans(n_clusters=12).fit(data)
kmeans_fit2 = KMeans(n_clusters=12).fit(data)
kmeans_fit3 = KMeans(n_clusters=12).fit(data)

# print分群結果
cluster_labels1 = kmeans_fit1.labels_
cluster_labels2 = kmeans_fit2.labels_
cluster_labels3 = kmeans_fit3.labels_

# 讀test set
test = pd.read_csv("test.csv", header=0, skip_blank_lines=True)
test = test.drop('index', axis = 1)

ans = []

for i in range(300):
    if cluster_labels1[test.iloc[i][0]] == cluster_labels1[test.iloc[i][1]]:
        temp1 = 1
    else:
        temp1 = 0

    if cluster_labels2[test.iloc[i][0]] == cluster_labels2[test.iloc[i][1]]:
        temp2 = 1
    else:
        temp2 = 0

    if cluster_labels3[test.iloc[i][0]] == cluster_labels3[test.iloc[i][1]]:
        temp3 = 1
    else:
        temp3 = 0

    if ((temp1+temp2+temp3) >= 2):
        ans.append(1)
    else:
        ans.append(0)

print(ans)

ansWithIndex = pd.read_csv("submission.csv", header=0, skip_blank_lines=True)
ansWithIndex['ans'] = ans
ansWithIndex.to_csv('result1.csv', index=False)
```