

Data Science HW2

B10815057 廖聖郝

1. 執行程式:

(1) pip install pandas

(2) pip install sklearn

(3) python clustering.py

2. 程式架構流程:

(1) import 必要函式庫

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

(2) 讀取檔案，包括資料集與測試資料

```
data_df = pd.read_csv('data.csv')
test_df = pd.read_csv('test.csv')
```

(3) 用 kmeans 演算法將資料分成 5 類

```
kmeans = KMeans(n_clusters=5)
kmeans = kmeans.fit(data_df)
```

(4) 取得分群結果

```
kmeans_label = kmeans.labels_
```

(5) 創建空的 array，size 預設為測試資料數，用於存放判斷結果

```
test_amount = test_df['0'].size
ans = np.empty(test_amount, dtype=int)
```

(6) for 迴圈跑過每筆測試資料判斷是否同一群，將結果輸出到 ans

```
for i in range(test_amount):
    print('test case:%d' %(i))
    if kmeans_label[test_df.at[i, '0']]==kmeans_label[test_df.at[i, '1']]:
        ans[i] =1
    else :
        ans[i] =0
    print('answer:%d' %(ans[i]))
```

(7) 將 ans 這個 array 打包成 dataframe

```
ans=pd.DataFrame(ans)
```

(8) 設定 index 欄位名稱為"id"，index 從 int 轉為 float (繳交規定)

```
ans.index.name = 'id'
ans.index = ans.index.astype(float)
```

(9) 輸出到 csv 檔，並設定結果欄位名稱為"ans"

```
ans.to_csv('submit.csv',header=[ 'ans' ])
```

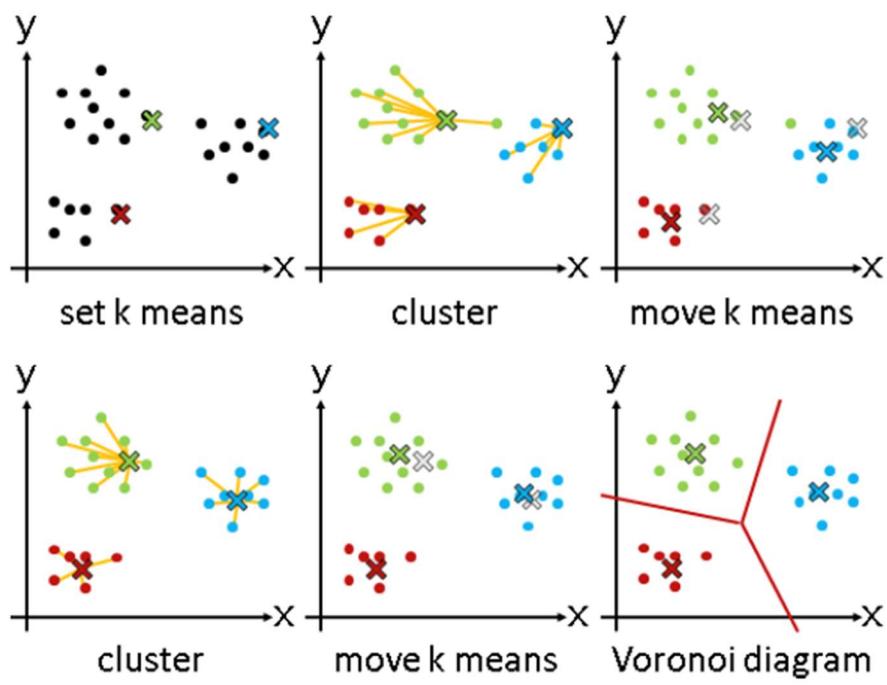
3. Kmeans 解釋

kmeans 是一種非監督式學習，且是一種 clustering 常用的方法。

流程:

- 要先決定要分 k 群，接著隨機選 k 個點作為每群中心點。
- 將資料集中的資料，分群到最靠近自己中心點。
- 重新計算每群的中心點。 重複 2.和 3.，直到中心點不再移動時，

kmeans 結束。



(圖片出處：<http://www.csie.ntnu.edu.tw/~u91029/Fitting.html>)