**KMeans報告書**

**學號:M123040018 姓名:陳彥宇**

**程式說明**

我建立一個class KMeans，並用此類別的實例來對輸入資料進行clustering，下表是程式設計的詳細資訊

|  |  |  |
| --- | --- | --- |
| **屬性** | **用途** | **預設值** |
| n\_clusters | cluster數量 | 4 |
| max\_iter | 最高疊代次數 | 400 |
| centroids | 儲存所有centroids | None |
| tol | centroid變化幅度的筏值 | 1e-4 |
| **方法** | **用途** | **輸出** |
| fit\_predict | 分群 | X的分群結果 |
| find\_closest\_centroids | 更新centroid座標 | 所有點最近的centroid |
| update\_centroids | 更新最近的centroid | 最近的centroid |

KMeans其實只有兩個功能

1. 找出目前的cluster
2. 更新cluster的centroid座標

我分別將這兩個功能建立為類別的方法，所以分群的方法只需要呼叫這兩個方法即可

**Step 1: 隨機初始化centroids**

首先必須先初始化centroids，從X中random sample出n\_clusters個點作為centroids，之所以是sample資料而非隨機生成是因為如果生成的centroid離得太遠可能會造成形成不存在的cluster

**Step 2: 初始化最近的centroid**

使用find\_closest\_centroids方法找出X的所有點目前最近的centroid，儲存在closest\_cluster\_ids

**Step 3: 疊代n次，每次都更新cluster，並更新cluster的centroid**

1. 如果centroid的更新幅度太小就視為分群完成
2. 如果某個cluster的點數為0，則直接sample一個資料點當作centroid

**<補充>**: 程式碼檔案裡有註解對程式碼做竹行說明，報告僅說明程式邏輯以及特別處理的部分

**執行結果**

|  |  |
| --- | --- |
| 執行前的散點圖 | 用我寫的KMeans分群完後的圖 |
|  |  |

|  |  |
| --- | --- |
| 執行前的散點圖 | 用sklearn的KMeans分群完後的圖 |
|  |  |

**結論**

Sklearn的KMeans在初始centroid時如果分到一個離所有點都很遠的座標，那麼就會成為一個沒有點的cluster，同時centroid也不會更新，就是因為害怕這種情況，我才決定要在初始centroid時質接sample資料點，而且在cluster為空時也會sample資料點作為centroid。