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

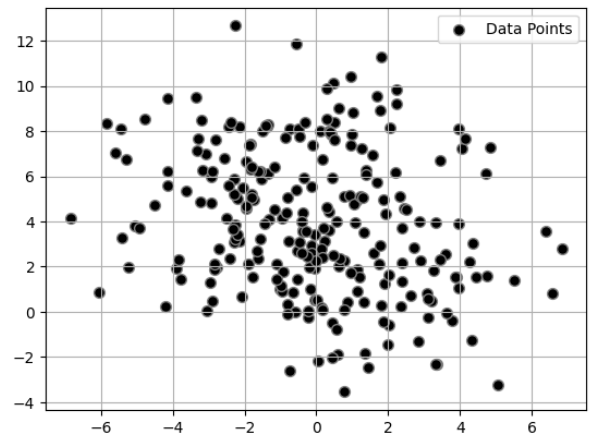
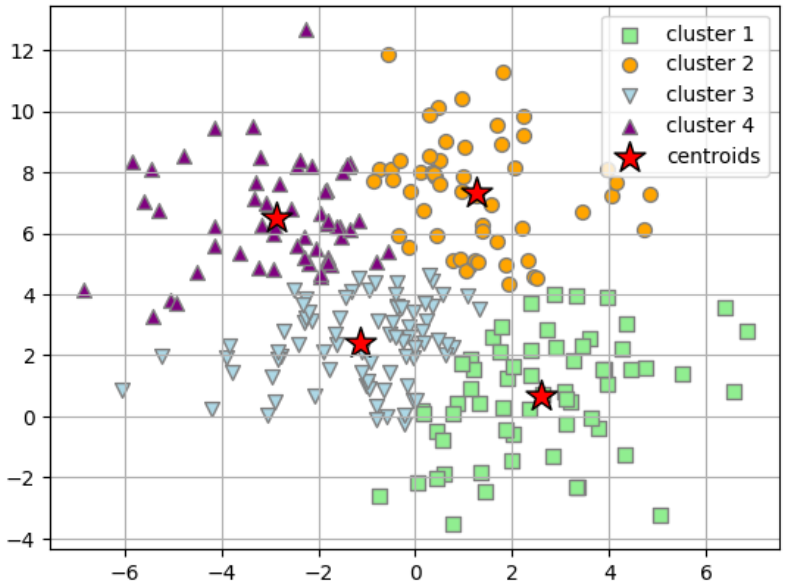
使用 find_closest_centroids 方法找出 X 的所有點目前最近的 centroid，儲存在 closest_cluster_ids

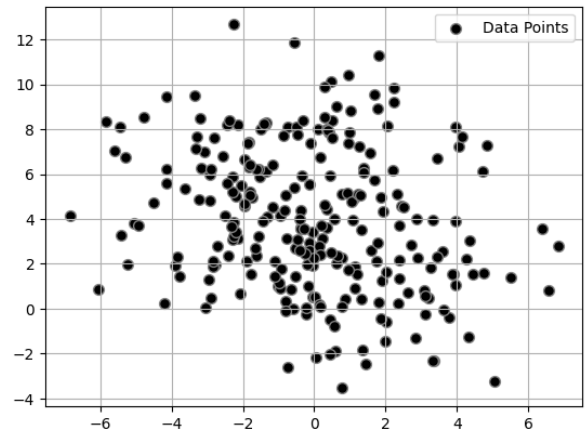
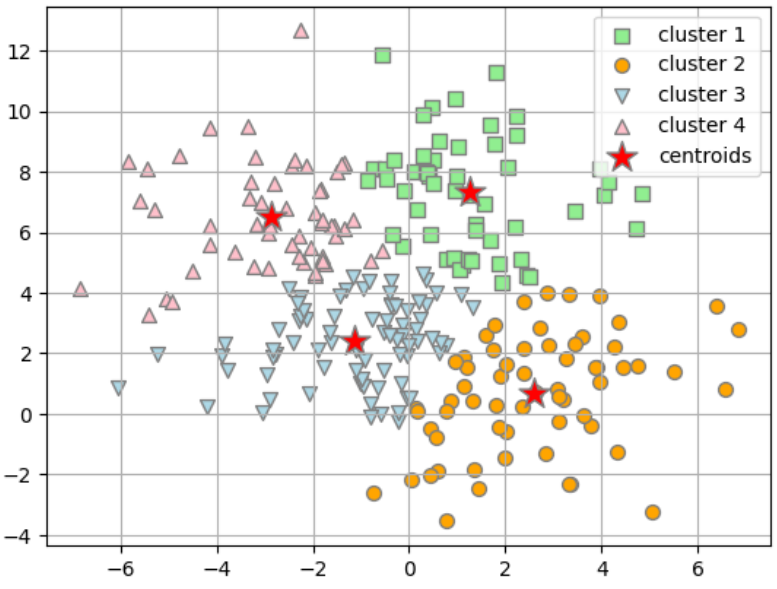
Step 3: 疊代 n 次，每次都更新 cluster，並更新 cluster 的 centroid

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。