Gaussian Mixture Model

WEI-CHIH CHANG Q36111281
December 6, 2022

Abstract

應用不同參數下的 Gaussian Mixture Model 應用在三種不同 data 上，分析資料的特性以及不同參數對 Gaussian Mixture Model 分群的影響。

1. Related Work

這次報告的資料格式是 json 檔(JavaScript Object Notation)，是一種輕量級的資料交換格式，其內容由屬性和值所組成，類似於 Python 的 Dictionary 資料格式。報告中因為有大量作圖的需求，因此在資料處理時，預先將 json 檔轉換成 csv 檔儲存，再用 pandas 讀出，以利接下來作圖的流程。

2 Model Analysis

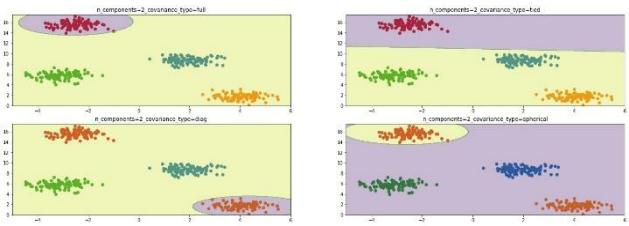
資料分群上使用 GMM(Gaussian Mixture Model)作為分群的演算法，實作上使用的是 sklearn 預建好的 mixture.GaussianMixture()函式，透過調整 n_components 和 covariance_type 兩種參數，分析不同模型在不同資料上分群結果的不同。

n_components 表示分群的群數，covariance_type 表示共變異數的類型，分為‘full’, ‘tied’, ‘diag’, ‘spherical’四種， ‘full’指每個分量具有各自不同的標準共變異數矩陣；‘tied’ 指每個分量具有相同的標準共變異數矩陣；‘diag’ 指每個分量具有各自不同的對角共變異數矩陣；‘spherical’指每個分量具有各自的單一變異數。

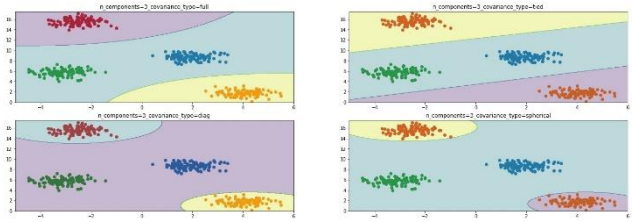
2.1 Data 1

從圖中可以看出，Data 1 分為 4 個類別，彼此之間分散，其中 x 軸分量為權重較重之量。因此我設定 n_components 從 2 到 6，並且將不同 n_components 下的四種不同共變異數類個別合成一張圖。

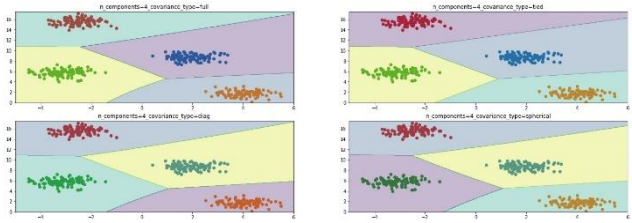
從圖表 1 可以看出‘tied’的分類結果明顯與其他三者不同，分群結果類似於線性分割。圖表 2 則可以看出‘full’對 x 軸分量較為敏感。圖表 3 可以看出在適當的分群數下，四種不同的共變異數類型的表現結果趨向一致。圖表 4 中，除了‘diag’外，其他三種共變異數類型皆將左上角紅色類別分為兩群，從圖中也可以看出雖然紅色的點皆為同一類別，但卻明顯有條直線可以將其線性分割成兩類。



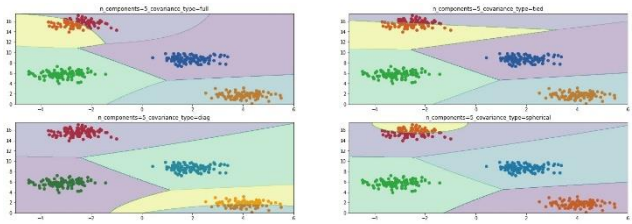
圖表 1_n=2



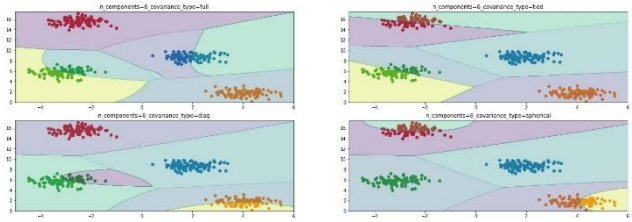
圖表 2_n=3



圖表 3_n=4



圖表 4_n=5



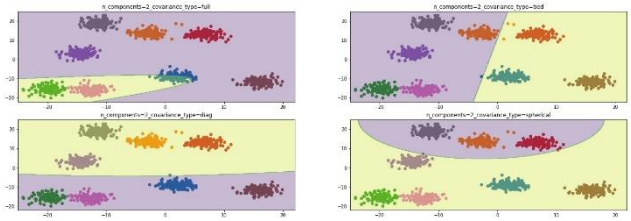
圖表 5_n=6

2.2 Data 2

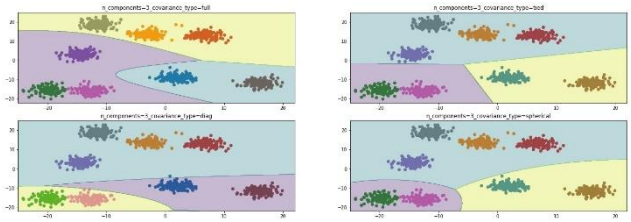
從圖中可以看出，Data 2 分為 8 個類別，彼此之除左下角粉色和綠色類別外，類別間分布均勻。我設定 n_components 從 2 到 9，並且將不同 n_components 下的四種不同共變異數類別個別合成一張圖。

從圖表 6 可以看出‘tied’的分類結果為線性分割，‘diag’的分類結果也接近線性分割，‘full’將左下角過於接近的兩個類別視為同一類。從圖表 7 可以發現所有共變異數類別皆將下角過於接近的兩個類別視為同一類 X，同時‘full’ 分群結果是四類之中表現最不為線性的。在圖表 8 中，’diag’的結果最不線性，‘spherical’的結果為各自的橢圓形。圖表 9 中值得注意的是‘diag’將左下角過於接近的兩個類別從錯誤的地方進行線性分割，可能是 y 軸分量在分群的過程中有較大的權重，才導致‘diag’選擇 y 軸分量作為分類標準導致錯誤的分類。圖表 11 中，‘spherical’成功將

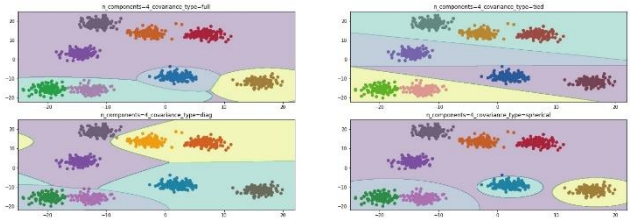
綠色與粉色類別區分。圖表 12 除了‘diag’外，其他共變異數類型皆能夠將資料正確分群。



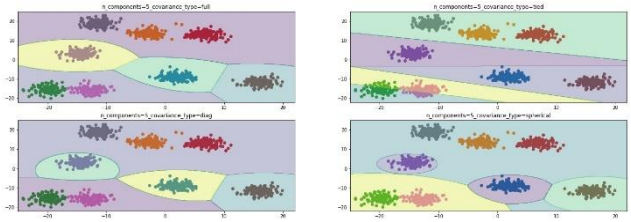
圖表 6_n=2



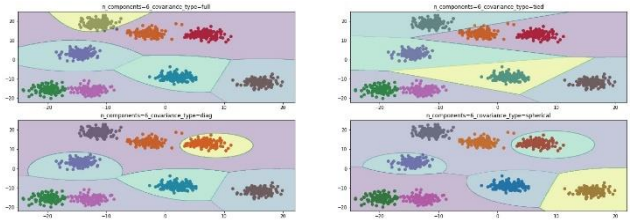
圖表 7_n=3



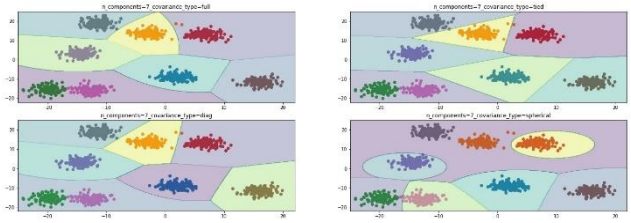
圖表 8_n=4



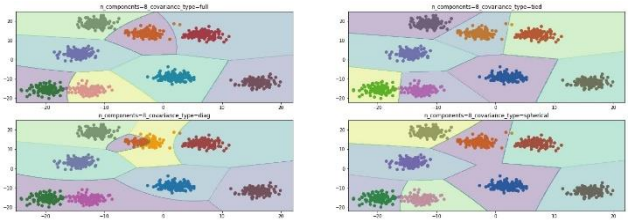
圖表 9_n=5



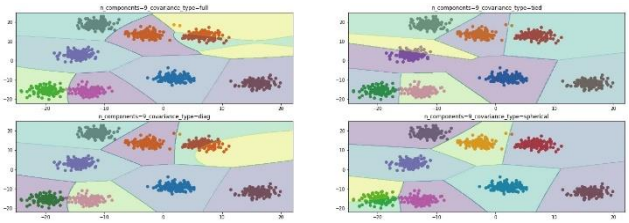
圖表 10_n=6



圖表 11_n=7



圖表 12_n=8

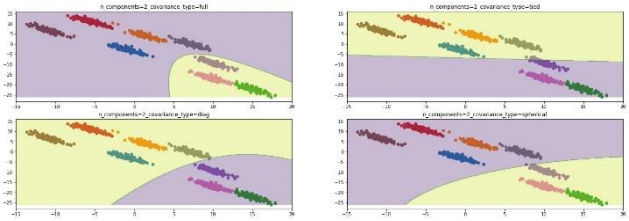


圖表 13_n=9

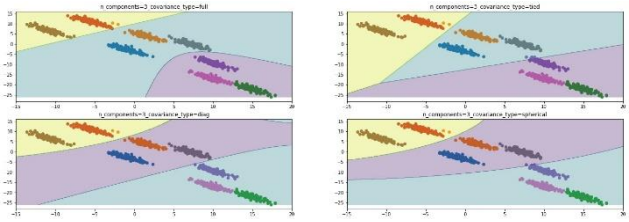
2.3 Data 3

從圖中可以看出，Data 3 分為 8 個類別，彼此之間呈負相關，但人工仍可正確辨識其類別。我設定 n_components 從 2 到 9，並且將不同 n_components 下的四種不同共變異數類別個別合成一張圖。

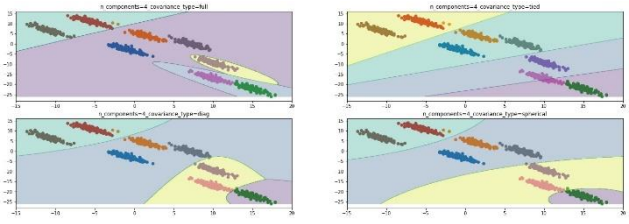
在圖表 14、圖表 15 與圖表 16 中，四種共變異數類型的分類結果皆相同。有趣的是儘管分類結果相同，但是四種不同策略的分群方式卻大不相同。觀察圖表 17，可以看出‘full’將 y 軸分量視為重要的分類依據，與其他三者相反。從圖表 20 可以發現，除了’tied’外，另外三者皆無法有效區別離散的資料，例如在紅色類別上方的兩個紅點，尤其‘diag’與‘spherical’對區分離散資料的效果最差(因為‘full’仍可區分橘色商方的兩個灰點)。



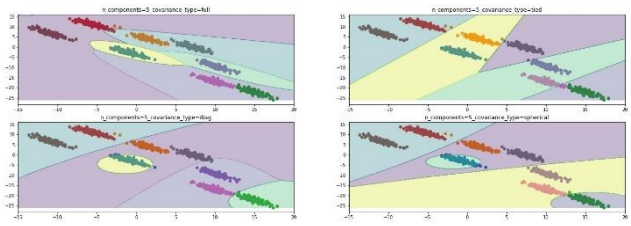
圖表 14_n=2



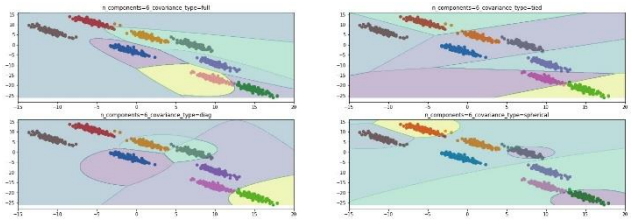
圖表 15_n=3



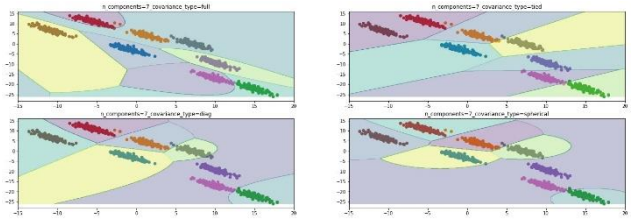
圖表 16_n=4



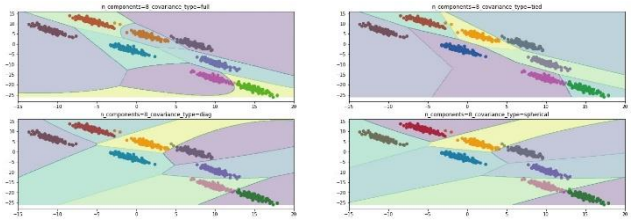
圖表 17_n=5



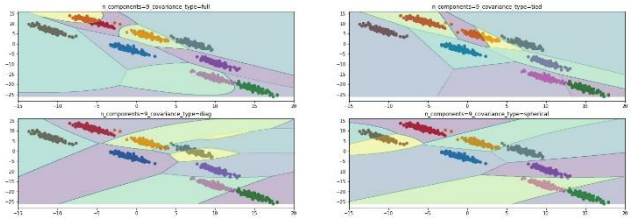
圖表 18_n=6



圖表 19_n=7



圖表 20_n=8



圖表 21_n=9

3. Conclusion

因為‘full’指每個分量具有各自不同的標準共變異數矩陣，這使得聚類的結果較大程度地受到較重要的分量影響，在 data3 中該特性尤其明顯；‘tied’ 指每個分量具有相同的標準共變異數矩陣，這使得聚類的結果像是多次線性分割後的結果；‘diag’ 指每個分量具有各自不同的對角共變異數矩陣，這使得不同分量對資料聚類的結果較為一致；‘spherical’指每個分量具有各自的單一變異數，因此聚類的結果多為各自的圓形。

4. Reference

(1)https://scikit-learn.org/stable/modules/generated/sklearn.mixture.GaussianMixture.html