

非監督式機器學習

K-mean 觀察：使用輪廓分析

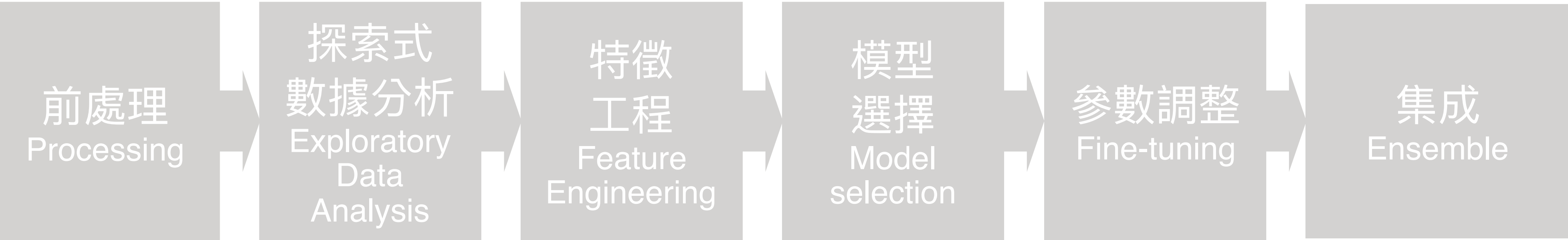


出題教練

陳明佑

機器學習概論 Introduction of Machine Learning

監督式學習
Supervised Learning



非監督式學習
Unsupervised Learning



非監督學習
Unsupervised learning

非監督簡介

| | |
|---------------------------|---|
| 分群 Clustering | K-平均算法 K-Mean |
| | 階層分群法 Hierarchical Clustering |
| 降維 Dimension Deduction | 主成分分析PCA(Principal components analysis) |
| | T 分佈隨機近鄰嵌入 t-SNE |

本日知識點目標

- 大致了解輪廓分析的設計想法與用途
- 如何使用輪廓分析觀察 K-mean 效果

註：因為非監督模型的效果，較難以簡單的範例看出來，所以非監督偶數日提供的檢視工具，僅供觀察非監督模型的效果，與後續其他部分及程式寫作無關，同學只要能感受到這些非監督模型的效果即可，不用執著於完全搞懂該章節所使用的工具

分群模型的評估

- 最大困難

- 與監督模型不同，非監督因為沒有目標值，因此無法使用目標值的預估與實際差距，來評估模型的優劣

- 評估方式類型

- 有目標值的分群
 - 如果資料有目標值，只是先忽略目標值做非監督學習，則只要微調後，就可以使用原本監督的測量函數評估準確性
- 無目標值的分群
 - 但通常沒有目標值/目標值非常少才會用非監督模型，這種情況下，只能使用資料本身的分布資訊，來做模型的評估

輪廓分析(Silhouette analysis) (1 / 3)

不要被名詞嚇到了，其實輪廓分析是一個很直覺的非監督評估方式，讓我們來看看如何計算吧。

歷史

最早由 Peter J. Rousseeuw 於 1986 提出。它同時考慮了群內以及相鄰群的距離，除了可以評估資料點**分群是否得當**，也可以用來**評估**不同分群方式對於資料的**分群效果**

設計精神

同一群的資料點應該很近，不同群的資料點應該很遠，所以設計一種當 **同群資料點越近 / 不同群資料點越遠** 時越大的分數
當資料點在**兩群交界**附近，希望**分數接近 0**

輪廓分析(Silhouette analysis) (2 / 3)

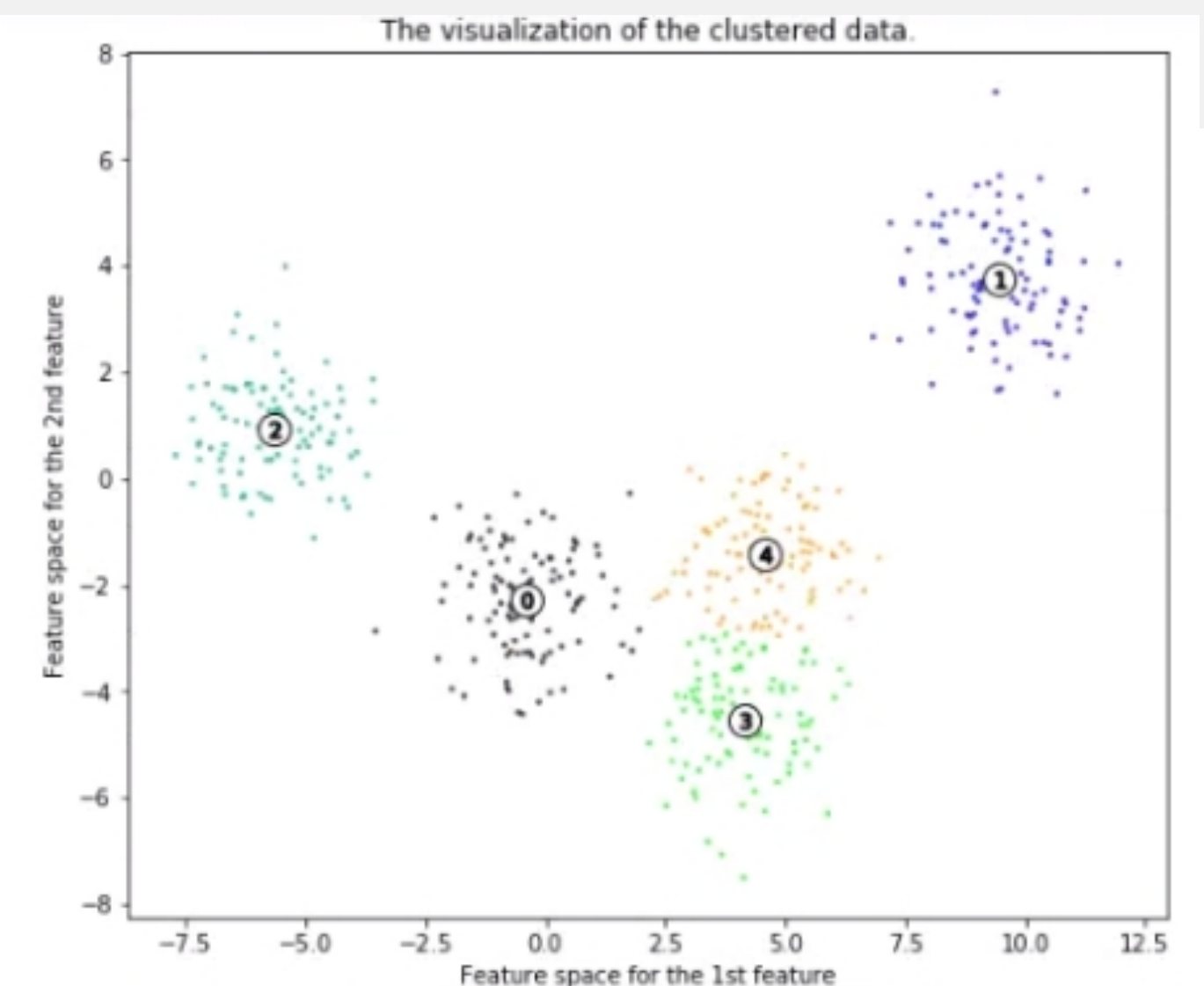
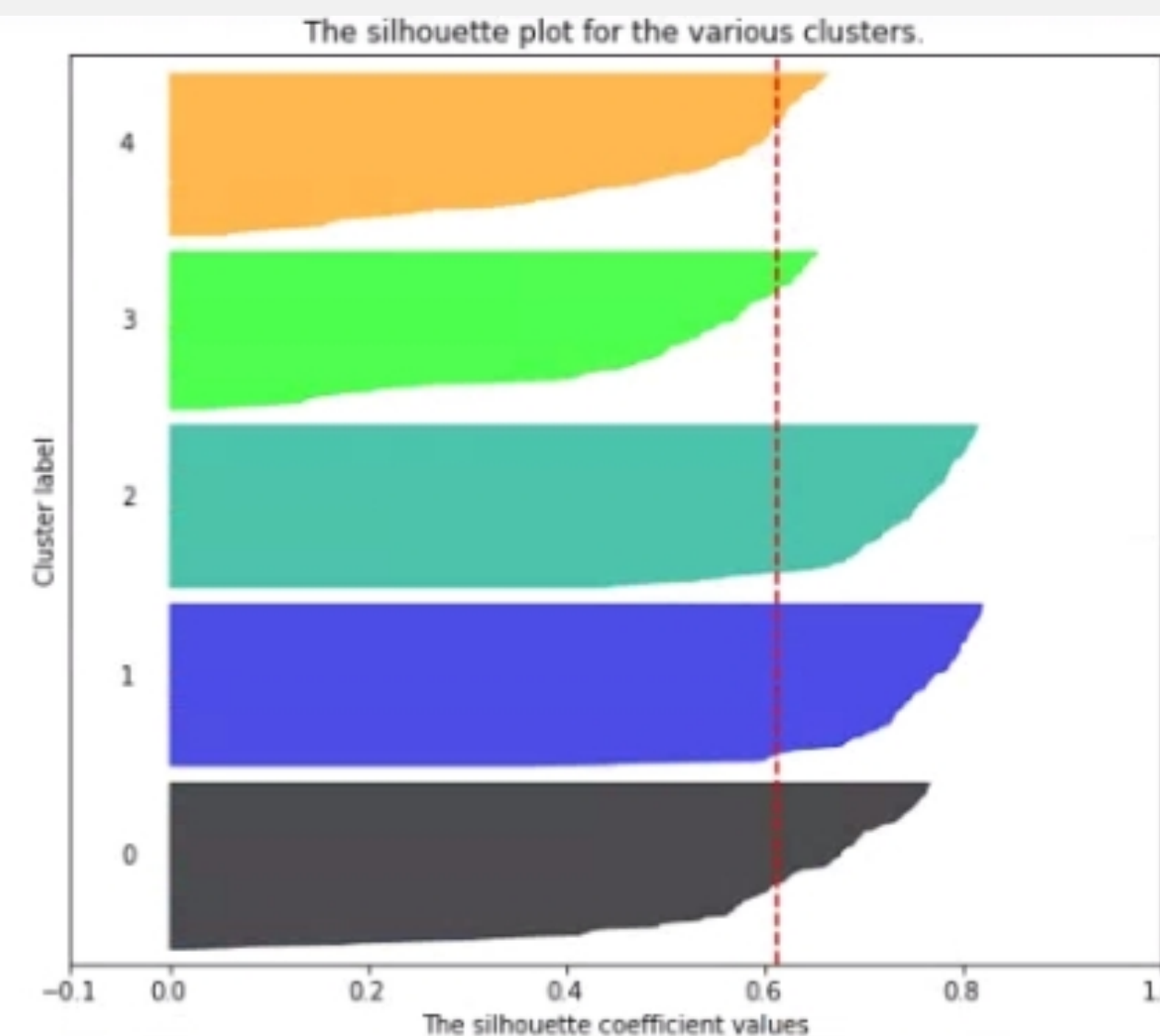
單點 輪廓值

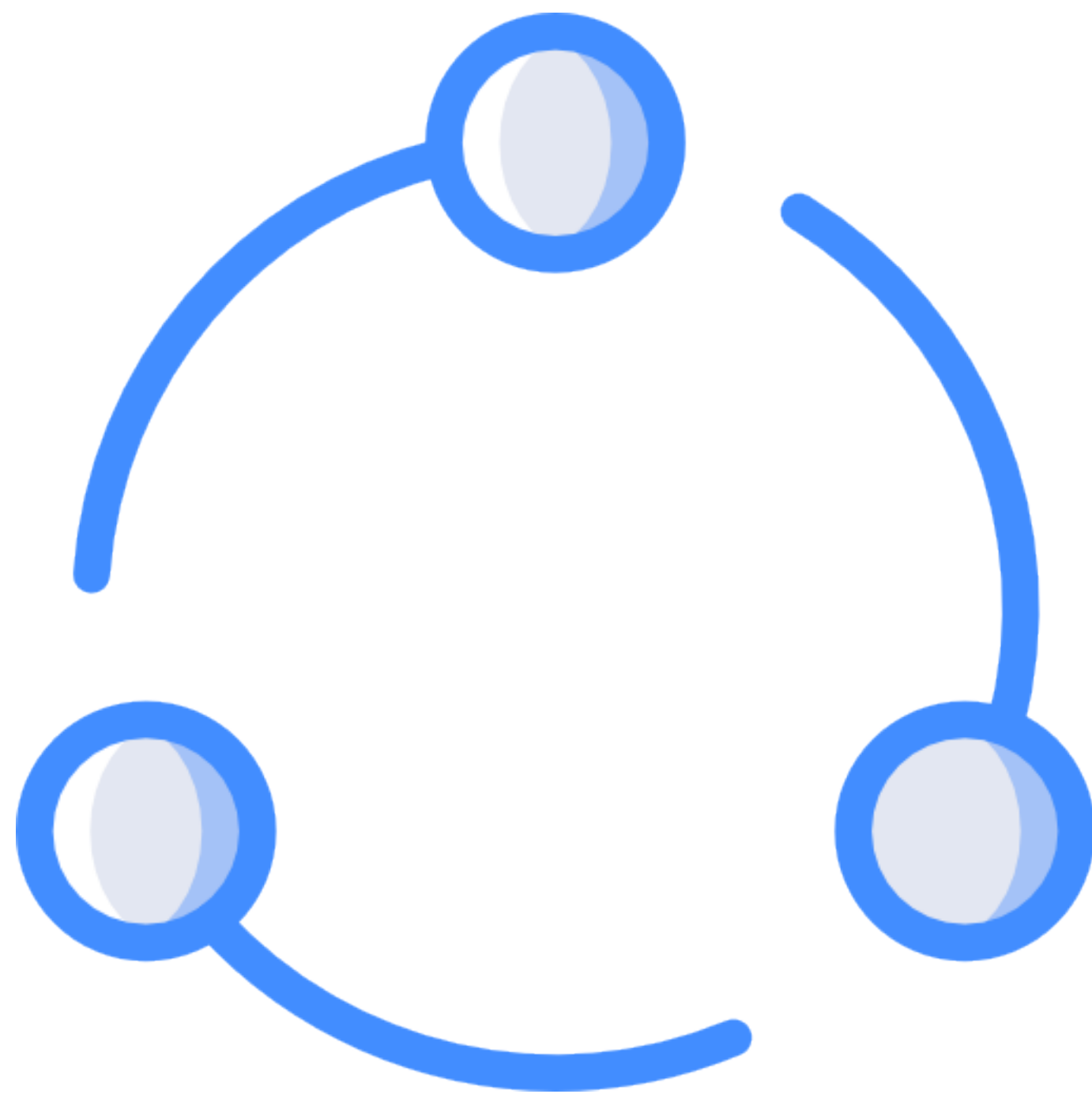
- 對任意單一資料點 i ，「與 i 同一群」的資料點，距離 i 的平均稱為 a_i
- 「與 i 不同群」的資料點中，不同群距離 i 平均中，最近的稱為 b_i (其實就是要取第二靠近 i 的那一群平均，滿足交界上分數為 0 的設計)
- i 點的輪廓分數 $s_i : (b_i - a_i) / \max\{b_i, a_i\}$
- 其實只要不是刻意分錯， b_i 通常會大於等於 a_i ，所以上述公式在此條件下可以化簡為 $1 - a_i / b_i$

輪廓分析(Silhouette analysis) (3 / 3)

整體的 輪廓分析

- 分組觀察 如下圖，左圖依照不同的類別，將同類別的輪廓分數排序後顯示，可以發現黃綠兩組的輪廓值大多在平均以下，且比例上接近 0 的點也比較多，這些情況都表示這兩組似乎沒分得那麼開 (可對照右圖)
- 平均值觀察 計算分群的輪廓分數總平均，分的群數越多應該分數越小，如果總平均值沒有隨著分群數增加而變小，就說明了那些分數數較不恰當





- 輪廓分數是一種 同群資料點越近 / 不同群資料點越遠 時會越大的分數，除了可以評估資料點分群是否得當，也可以用來評估分群效果
- 要以輪廓分析觀察 K -mean，除了可以將每個資料點分組觀察以評估資料點分群是否得當，也可用平均值觀察評估不同 K 值的分群效果

解題時間 Coding Time

請跳出PDF至官網Sample Code & 作業
開始解題

