

Day 61 非監督式機器學習

降維方法 - t-SNE





知識地圖非監督學習

● 機器學習概論 Introduction of Machine Learning

監督式學習 Introduction of Machine Learning 探索式 特徵 模型 數據分析 前處理 參數調整 集成 選擇 工程 Exploratory Ensemble Fine-tuning Processing Model Feature Data Engineering selection Analysis

非監督式學習

Unsupervised Learning

分群 Clustering K-平均算法 K-Mean 階層分群法 Hierarchical Clustering 降維 Dimension Reduction

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

- 瞭解 PCA 的限制
- t-SNE 概念簡介,及其優劣

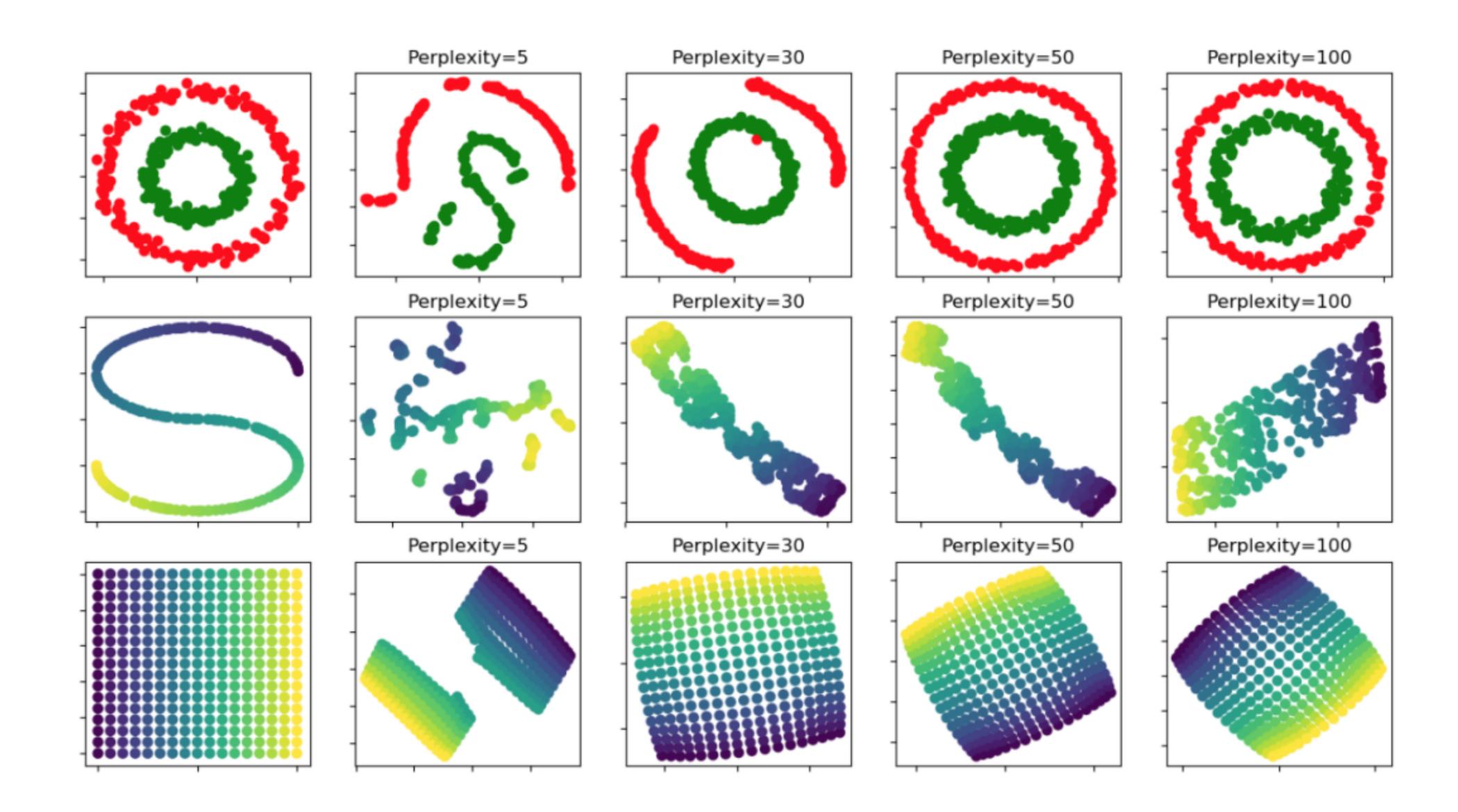
PCA的問題

- 求共變異數矩陣進行奇異值分解,因此會被資料的差異性影響,無法很好的表現相似性及分佈
- PCA 是一種線性降維方式,因此若特徵間是非線性關係,會有 underfitting 的問題

t-SNE

- t-SNE 也是一種降維方式,但它用了更複雜的公式來表達高維和低維之間 的關係
- 主要是將高維的資料用 gaussian distribution 的機率密度函數近似,而低維資料的部分用 t 分佈來近似,在用 KL divergence 計算相似度,再以梯度下降 (gradient descent) 求最佳解

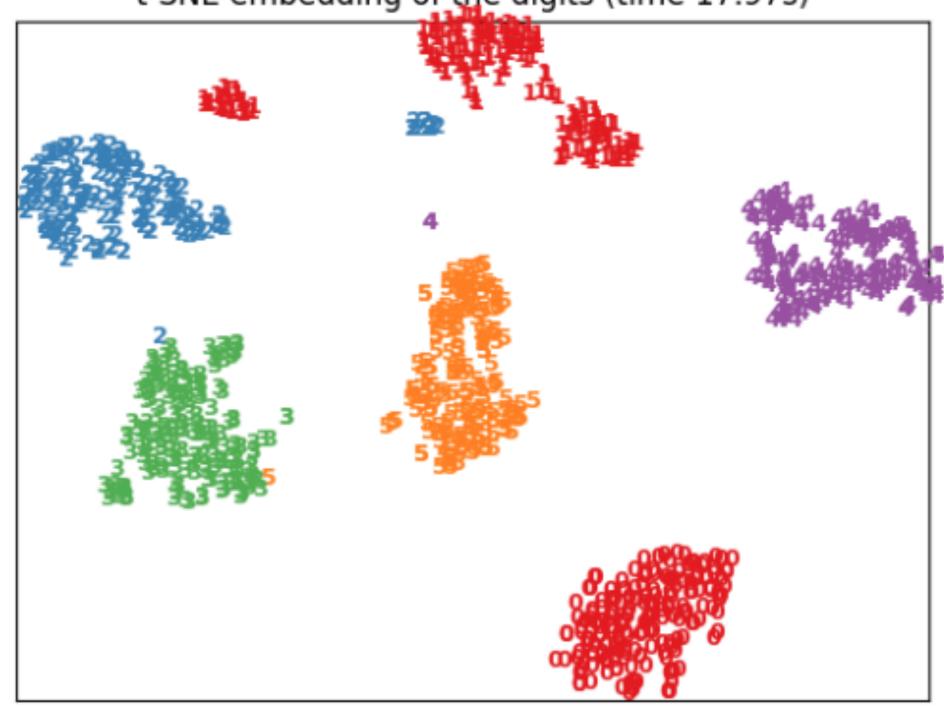
t-SNE 視覺化範例



t-SNE 視覺化範例

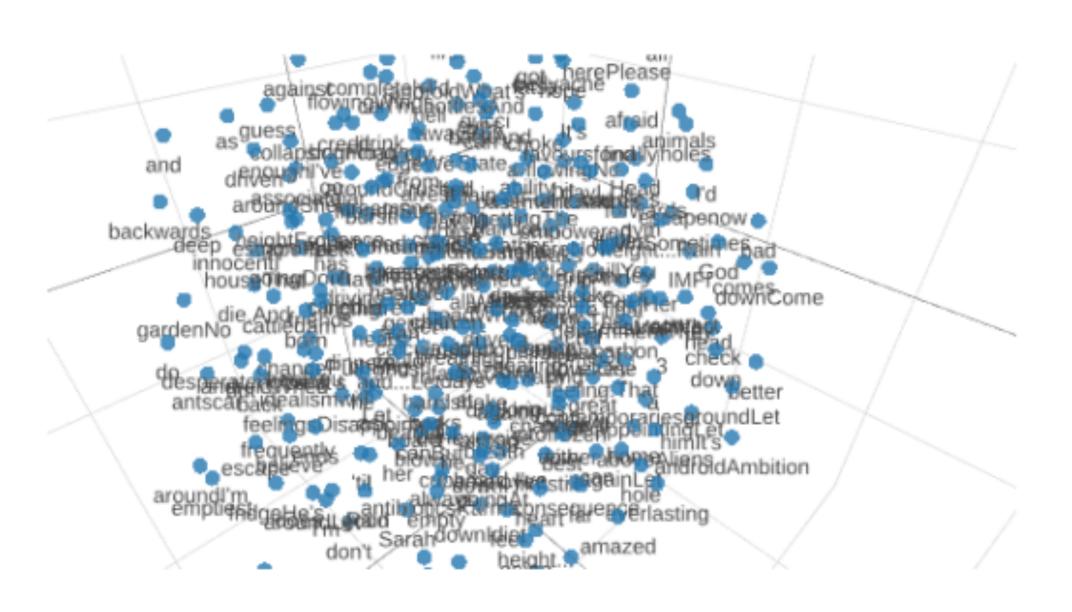
影像資料 (MNIST) 視覺化呈現

t-SNE embedding of the digits (time 17.97s)



文字資料 (MNIST) 視覺化呈現

Ok Computer lyrics



t-SNE 優劣



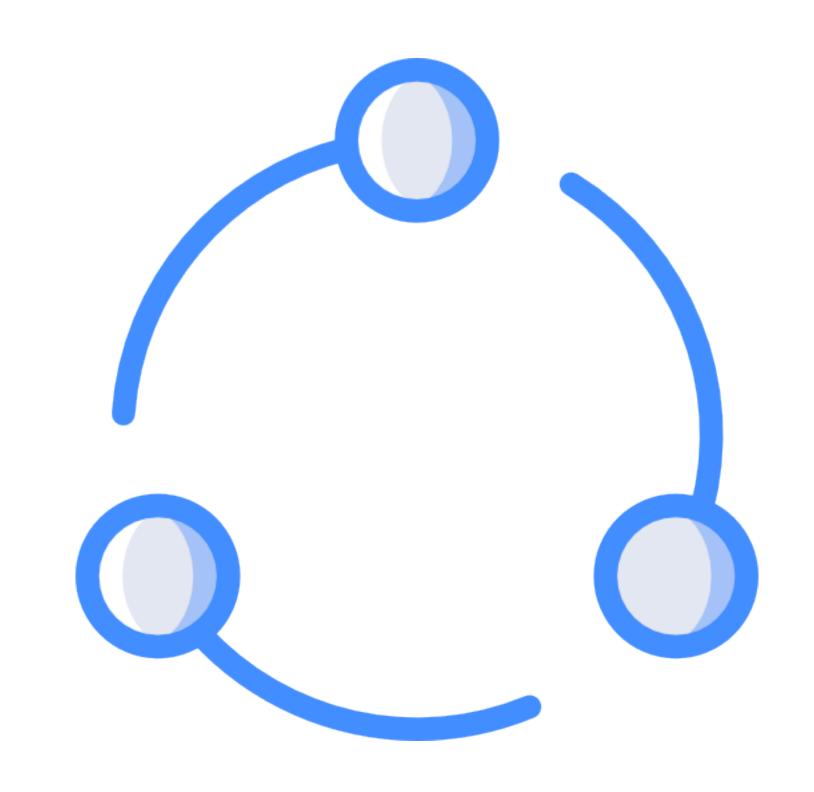
當特徵數量過多時,使用 PCA 可能會造成降維後的 underfitting,這時可以考慮使用 t-SNE 來降維



t-SNE 的需要比較多的時間執行

重要知識點複習

- 特徵間爲非線性關係時 (e.g. 文字、影像資料)PCA很容易 underfitting
- t-SNE 對於特徵非線性資料有更好的降維呈 現能力





請跳出PDF至官網Sample Code&作業 開始解題