國立雲林科技大學

機器學習專案作業三

資料集維度縮減

指導教授：許中川　教授

學生：M10921002 宋沂芸

M10921016 林恩杰

M10921036 童湘庭

M10921038 張珮柔

摘要

隨著時代的科技進步、人口增加，為求便利及不排放過多的二氧化碳，民眾對於大眾運輸交通的需求也日益漸增，台灣於 1999 年開始興建高鐵，2007 年完工通車，直至 2021 年共設立 12 個車站，若未來又增加多個車站，各車站間的距離資料也將變大，。此本組欲使用 MDS 方法將車站間距離資料建維至二維平面，使的民眾更清楚高鐵車站之地理位置。

根據經濟部資料顯示飲料店平均每年成長 8.9%，據財政部統計資料庫資料顯示在 2019 年飲料店數量就將近 2 萬家，在如此競爭激烈的環境下，若是能夠了解消費者的飲料愛好並且對銷售方針進行調整，就能從眾多飲料中脫穎而出。因此本組欲將飲料品項之銷售紀錄轉換至二維平面，且使用 t-SNE 方法降維，並顯示其品項間之相似度。

關鍵字：機器學習、MDS、t-SNE

一、緒論

1.1 動機

1.1.1 高鐵站距離資料集

台灣早期為了運送煤礦於 1876 年興建了第一條輕便軌道，在日治時期為了軍事用途、搬運建材，日軍於 1908 年 3 月完成縱貫線通車，又隨著時代及科技進步，為了配合民眾日益漸增的大眾運輸需求，台灣於 2007 年通車了總長 348.5 公里的高速鐵路，北至新北市，南市高雄市，西部各縣市皆有一高鐵車站，共 12 車站，但是站與站間的距離皆不同，單從相對距離還是難以確認車站真正的位置，若是未來要擴增車站數，民眾會更難從距離表上判斷車站的位置，因此若能直接將距離在二維空間上展示，那將有助民眾更容易清楚高鐵車站之地理位置。

1.1.2 Drink Dataset

根據經濟部調查顯示，飲料店營業額逐年攀升平均每年成長 8.9%，預計在民國 108 年可突破 1000 億元。根據財政部營利事業家數統計，民國 108 年 3 月底飲料店數達 2 萬 2,482 家，較民國 97 年增加了 9079 家(張瑋容，2019)，飲料業因進入門檻較低而蓬勃發展 ，在此激烈競爭之下，若賣方能掌握消費者愛好就能從此競爭中脫穎而出。若能將多品項飲料的銷售記錄將資料進行可視化，銷售者做決策會更加容易判斷。

1.2 目的

1.2.1 高鐵站距離資料集

本研究選取目前的台北、桃園、新竹、台中、雲林、台南以及左營高鐵站之間之距離，使用 MDS 方法將各高鐵站之距離降維至二維平面，以展示各車站之相對位置。

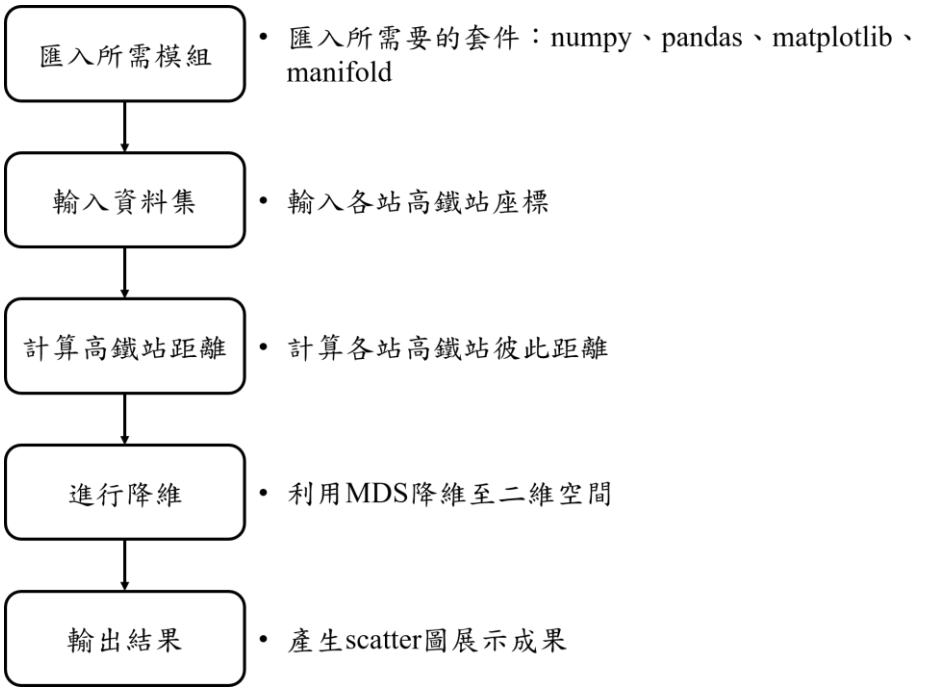
1.2.2 Drink Dataset

在多種品項飲料銷售數據下透過機器學習方法將資料轉為在二維空間中展示，本研究有 7 種不同的飲料品項藉由 t-SNE 方法進行降維，顯示其品項間的相似性。

二、方法

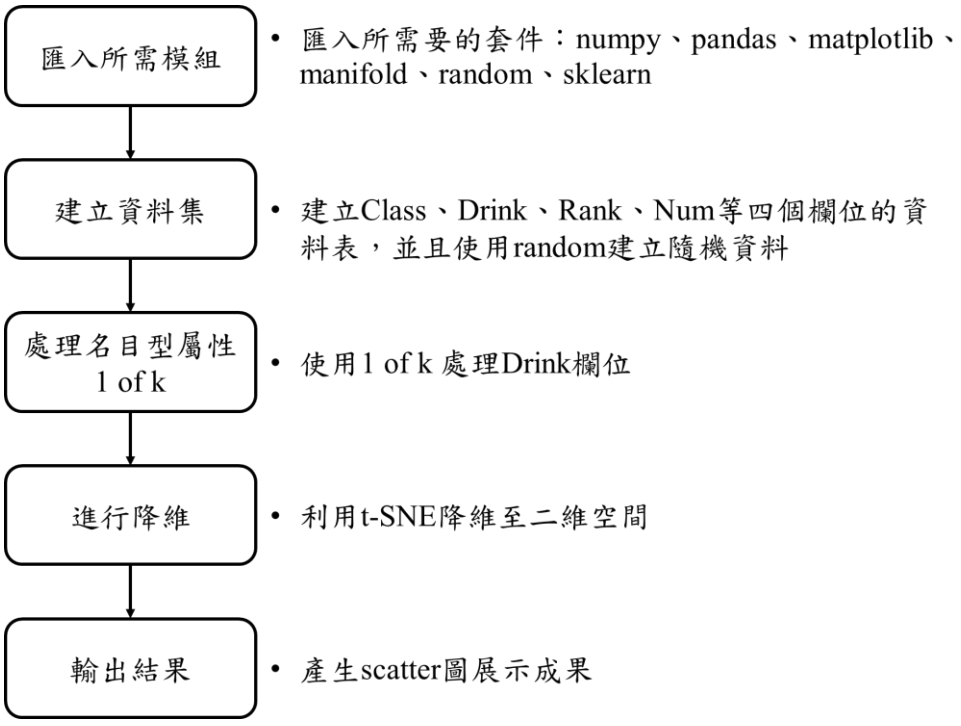
2.1 程式架構

2.1.1 高鐵站距離資料集



圖一高鐵站距離資料集程式架構圖

2.1.2 Drink Dataset



圖二 Drink Dataset 程式架構圖

2.2 執行方法

2.2.1 多維標度法

流形學習是機器學習中一種演算法的統稱，而多維標度法（Multidimensional Scaling, MDS）是將高維度數據在低維度空間中展示並且保留全局屬性的多元分析技術。

2.2.2 t sne-隨機鄰近嵌入法

t-隨機鄰近嵌入法（t-distributed stochastic neighbor embedding,t-SNE）是一種非線性的機器學習降維方法。使用 t 分佈定義低維時的機率分佈來減緩維數災難（curse of dimensionality）造成的擁擠問題（crowding problem），同時對於保持局部資料結構能力十分傑出，因此近年來在學術論文以及各大競賽中常被使用。

三、實驗

3.1 資料集

3.1.1 高鐵站距離資料集

以下共有 7 站高鐵站經緯度座標，分別為：台北、桃園、新竹、台中、雲林、台南和高雄。

表一 高鐵站距離表

高鐵站	經緯度座標
台北	(25.048,121.517)
桃園	(25.013,121.215)
新竹	(24.808,121.040)
台中	(24.112,120.616)
雲林	(23.734,120.417)
台南	(22.926,120.286)
高雄	(22.687,120.308)

3.1.2 Drink Dataset

此資料集共有三個特徵欄位：Amount、Drink、Rank 和一個類別欄位：Class。特徵欄位中 Drink 為名目型態，其餘皆為數值型態。

表二 Drink dataset 說明表

Class	Drink	Rank	Amount($N(\mu,\sigma)$)	Count
A	Coke	7	(100, 200)	200
B	Pepsi	6	(200, 10)	100
C	7Up	5	(200, 10)	100
D	Sprite	4	(400, 100)	200
E	Latte	3	(800, 10)	100
F	Espresso	2	(800, 10)	100
G	Cappuccino	1	(900, 400)	200

3.2 前置處理

3.2.1 高鐵站距離資料集

找尋各大高鐵站經緯度座標後，為研究進行統一四捨五入取小數點後三位，將其資料列為 array 方便後續程式進行。

3.2.2 Drink Dataset

依照常態分配隨機產生資料，轉換生成後資料的型態。依照題目要求針對 Drink 欄位進行 1-of-k 轉換，使用 One hot 編碼進行轉換，最後將其資料轉換成 array。

3.3 實驗設計

3.3.1 高鐵站距離資料集

利用 MDS 將高鐵站各站座標投射至二維空間內部，參數 n_components 設為 2，dissimilarity 選用歐基里德距離，random_state 設為 1。下表為高鐵站各站距離。

表 高鐵站各站距離矩陣

	台北	桃園	新竹	台中	雲林	台南	高雄
台北	0	0.304	0.534	1.30	1.714	2.453	2.653
桃園	0.304	0	0.270	1.082	1.508	2.284	2.497
新竹	0.534	0.270	0	0.815	1.242	2.027	2.244
台中	1.30	1.082	0.815	0	0.427	1.231	1.458
雲林	1.714	1.508	1.242	0.427	0	0.819	1.053
台南	2.453	2.284	2.027	1.231	0.819	0	0.240
高雄	2.653	2.497	2.244	1.458	1.053	0.240	0

3.3.2 Drink Dataset

使用 sklearn 中的套件 TSNE，將 Drink 的資料集降為到二維空間內，參數 n_components 設定為 2。

3.4 實驗結果

3.4.1 高鐵站距離資料集

將 MDS 輸出的結果利用 matplotlib 套件產生散佈圖，並且將點進行標記，使每個點都有相對應的名稱方便觀察。降維後 X 軸座標介於-0.732 至 0.52 之間，而 Y 軸介於-1.159 至 1.476 之間。

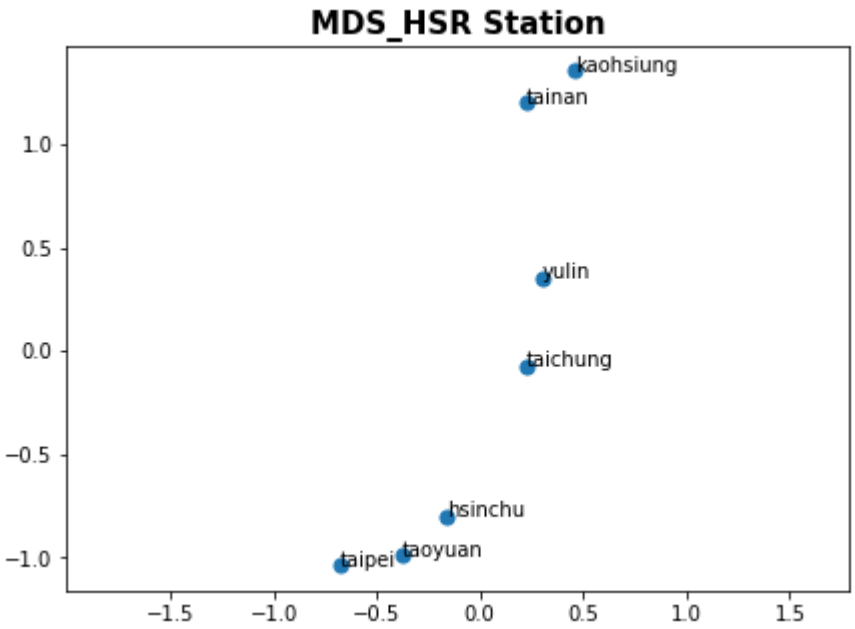


圖 MDS_高鐵站座標圖

3.4.2 Drink Dataset

我們使用不同文字編碼方式進行實驗，將 T-SNE 的結果輸出產生圖片如下圖，得到兩種形狀較不一樣的圖形，但可以發現線條重疊部分以及有涵蓋到的但大致上都一致。

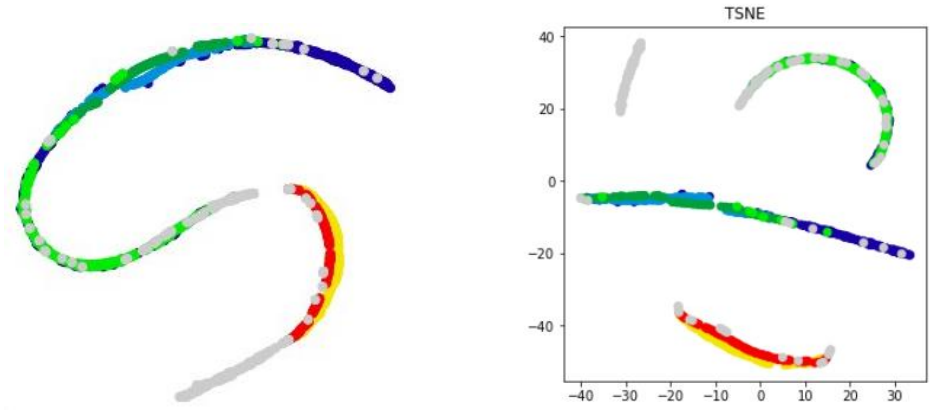


圖 T-SNE_Drink

四、結論

4.1 高鐵站距離資料集

將 n_components 參數設置為想要的維度，並且要注意 random_state 設置否則會造成每次結果不同。利用 MDS 可直接進行降為並且將資料視覺化，透過研究發現 MDS 可以保留資料間原本彼此的關係。

4.2 Drink Dataset

本實驗將 t-SNE 用來將資料投影到 2 維的空間作定性的視覺化觀察，通過視覺化直觀的驗證某資料集或演算法的有效性。t-SNE 的演算法優化是基於機率的方法，在判讀圖像時有許多要注意的地方。在資料處理上的不同也會導致圖像出

來的不一樣，但從圖中可以發現資料點的關係大致上不會改變。t-SNE 降維時保持局部結構的能力十分傑出，也因為這樣成為近年來學術論文與模型比賽中資料視覺化的常客。

五、參考文獻

[1] 張瑋容(2019)。飲料店營業額連續 14 年正成長。取自 https://www.moea.gov.tw/Mns/dos/bulletin/Bulletin.aspx?kind=9&html=1&menu_id=18808&bull_id=6099

[2] 流形學習 https://kknews.cc/zh-tw/health/jk6lox6.html

[3] 資料降維與視覺化：t-SNE 理論與應用 https://mropengate.blogspot.com/2019/06/t-sne.html

[4] Python – 如何使用 t-SNE 進行降維 https://mortis.tech/2019/11/program_note/664/

[5] 淺談降維方法中的 PCA 與 t-SNE https://medium.com/d-d-mag/%E6%B7%BA%E8%AB%87%E5%85%A9%E7%A8%AE%E9%99%8D%E7%B6%AD%E6%96%B9%E6%B3%95-pca-%E8%88%87-t-sne-d4254916925b

[6] 資料降維與視覺化：t-SNE 理論與應用 https://mropengate.blogspot.com/2019/06/t-sne.html

[7] [Python]資料視覺化 M01—運用 matplotlib 完成散布圖(scatter) https://ithelp.ithome.com.tw/articles/10211370

[8] Visualising high-dimensional datasets using PCA and t-SNE in Python https://towardsdatascience.com/visualising-high-dimensional-datasets-using-pca-and-t-sne-in-python-8ef87e7915b