國立雲林科技大學

機器學習專案作業四

資料降維的應用與實作

成員：

M11021009 吳宥霆

M11021028 劉軒瑋

M11021035 黃堉豪

M11021052 邱守燦

指導教授：

許中川 教授

日期：

2022 年 6 月 27 日

摘要

隨著機器學習的發展，許多先進的降維技術進已逐漸被提出，而資料降維是一個減少資料分析處理難度的核心方法，此法可以進行資料視覺化和觀察資料型態與分布，但又不影響資料原本的特性，透過了解與發掘各變量的特徵，進行特徵選擇與特徵攫取，針對欲進行資料分析的目標更透徹且更有效率執行，本專案利用台灣主要縣市高鐵車站的距離矩陣與飲料資料集進行資料降維的實作；台灣主要縣市高鐵車站的距離矩陣應用 MDS 降維，飲料資料集分別應用到詞嵌入轉換後的相似度矩陣降維與 1-of-k 降維。

關鍵字：資料降維、距離矩陣、詞嵌入、相似度矩陣。

一、 緒論

1.1 研究動機

1.1.1 台灣主要縣市高鐵車站距離資料集

台灣從北到南大約 300 公里，隨著交通的發展，高鐵已取代台鐵，成為西半部交通最快速的選項，東部則因為地形的影響，目前依然為台鐵為主要交通工具。為了理解各車站座標與之間的距離，可透過機器學習的降維方法，將資料投射於 2 維平面上，做為未來政府在交通延伸上可做為設站的參考依據。

1.1.2 飲料資料集

各個國家都有著不同的飲食文化，以美國為例，最具代表性的飲料包含汽水、咖啡，其中更以可口可樂、百事可樂等汽水品牌大廠風靡全球，除了汽水之外，咖啡也深受美國人的喜愛，更研發出了一款獨特的美式咖啡，在眾多不同的飲料中，若想觀察更加細微的特徵，例如：各飲料之間的相似度，可透過機器學習的降維方法，得知不同飲料之間的資訊。

1.2 研究目的

1.2.1 台灣主要縣市高鐵車站距離資料集

為了能將台灣高鐵車站彼此間的距離矩陣標示在 2D 平面上，從而使資料視覺化更加顯而易見，本組選取台灣高鐵車站距離矩陣資料，並使用多維標度(MDS)降維技術進行空間降維，投入 7 個高鐵車站距離資料，以此技術達到在新空間標示各點之距離，並盡可能與原空間的位置保持相等。

1.2.2 飲料資料集

為了讓飲料資料集的名目資料標示在 2D 平面上，並進行資料視覺化顯示降維資訊，本組將資料分別進行詞嵌入與 One-Hot-Encoding 轉換資料，並使用 t-隨機鄰近嵌入法(TSNE)降維技術進行空間降維，投入 7 個不同飲料品牌的資料，以此技術達到在二維空間中標示各點之散佈情形，從而發現各飲料之間的相似性。

二、 資料集

2.1 資料集

2.1.1 台灣主要縣市距離資料集說明

此資料集藉由 Google 地圖標記的各車站經緯度，然後經由經緯度距離計算得出各車站彼此間距離，此表共有 7 個車站，單位為公里。

表 1 台灣各高鐵火車站距離矩陣彙總表

	台北	新竹	台中	台南	高雄	花蓮瑞穗	台東池上
台北	0	54	138	267	289	172	215
新竹	54	0	88	223	247	149	187
台中	138	88	0	136	161	103	125
台南	267	223	136	0	26	128	98
高雄	289	247	161	26	0	141	105
花蓮瑞穗	172	149	103	128	141	0	44
台東池上	215	187	125	98	105	44	0

2.1.2 飲料資料集說明

此飲料資料集欄位包含類別、飲料名稱、等級、各飲料數量母體參數和次數。

表 2 飲料資料集說明欄位資料說明彙總表

Drink Dataset				
Class	Drink	Rank	Amount(N(μ, σ))	Count
A	7Up	7	(100,200)	300
B	Sprite	6	(200,10)	150
C	Pepsi	5	(200,10)	150
D	Coke	4	(400,100)	300
E	Cappuccino	3	(800,10)	150
F	Espresso	2	(800,10)	150
G	Latte	1	(900,400)	300

三、方法

3.1 程式架構

3.1.1 台灣主要縣市距離資料集

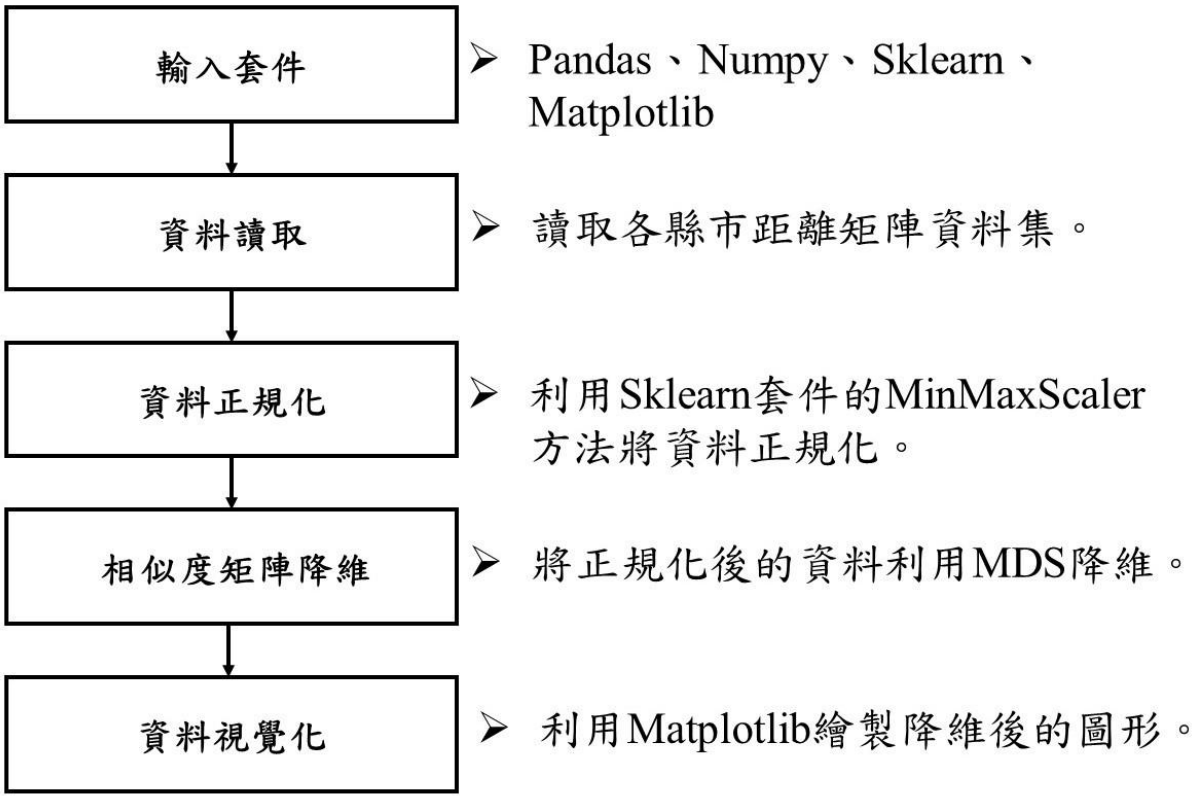

圖 1 台灣主要縣市距離資料集之程式架構流程圖及說明

3.1.2 飲料資料集

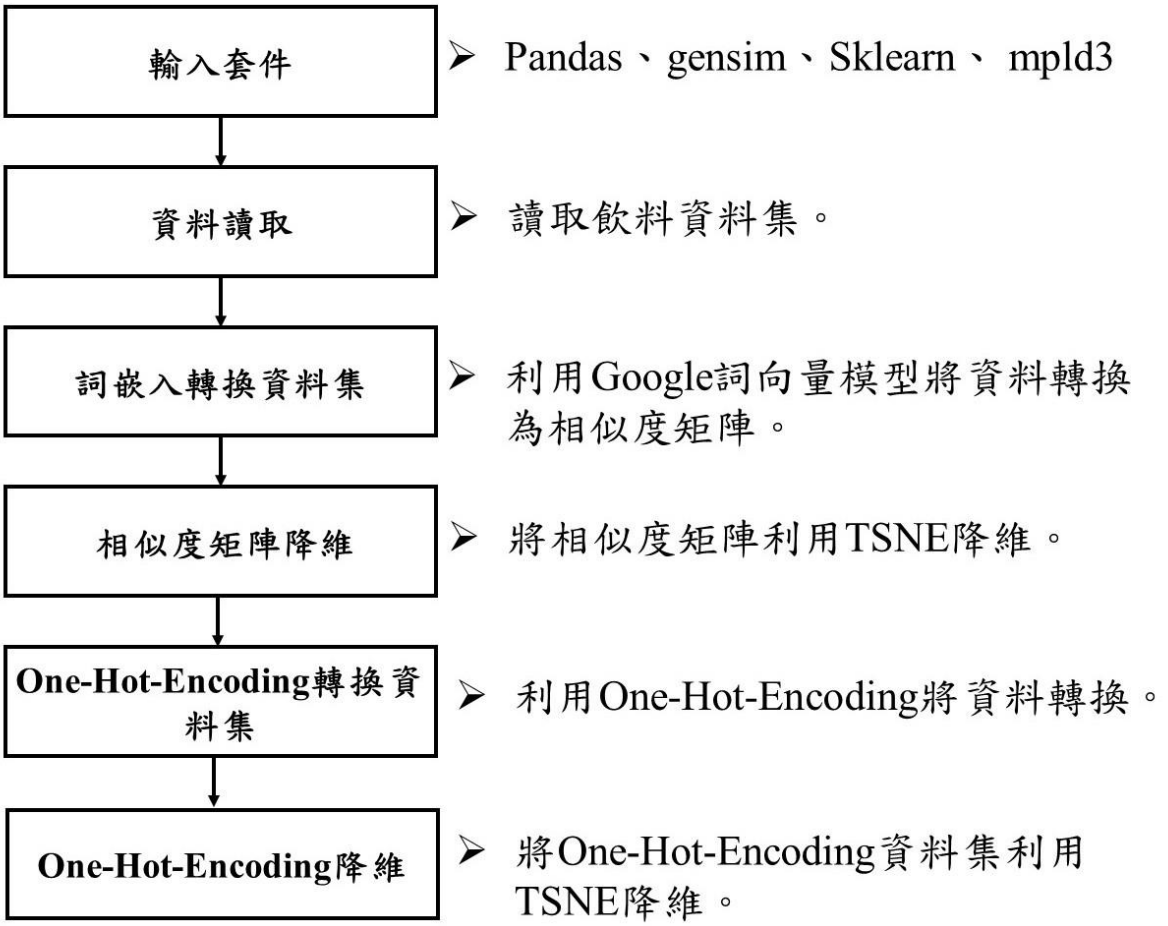

圖 2 飲料資料集之程式架構流程圖及說明

3.2 程式執行方法

3.2.1 多維標度降維技術(MDS)

資料在高緯度情況下進行資料處理將會有極大的資料處理量。為了減少計算量，常常需要緩解這種資料維度災難，這有兩種途徑：降維和特徵選擇。

MDS 演算法要求原始空間中樣本之間的距離在低維空間中得以保持。但是為了有效降維，我們往往只需要降維後的距離與原始空間距離盡可能接近即可。

3.2.2 t-隨機鄰近嵌入法(TSNE)

為非線性的機器學習降維方法，降維時可保持局部結構，在實務應用上，TSNE 常用來將資料投射到 2 維或 3 維的空間做視覺化觀察，透過資料視覺化驗證資料集或演算法的有效性。

三、 實驗

4.1 前置處理

4.1.1 台灣主要車站距離資料集

將台灣主要車站的座標，使用歐式距離計算兩城市之間距離，計算其距離矩陣，並輸出成 CSV 檔。

4.1.2 飲料資料集

將飲料資料集的名目欄位進行詞嵌入轉換為相似度矩陣，利用 gensim 套件導入詞嵌入預訓練詞庫(GoogleNews-vectors-negative300.bin)，該預訓練詞庫由 Google 開發，高達 100 萬個以上的預訓練詞向量被納入該詞庫當中，依照詞嵌入的概念，將各個單詞的相似程度轉換為詞向量建立而成的相似度矩陣。

4.2 實驗設計

4.2.1 台灣主要車站距離資料集

(1)套件載入：載入協助資料分析的 Pandas、Numpy 套件，匯入正規化、距離轉換及 MDS 降維方法的 Sklearn 套件，最後匯入輸出結果可視化的 matplotlib 套件。

(2)正規化：使用 Maxmin 方法，將資料轉換成 0～1 之間。

(3)歐式距離：使用 Sklearn 套件 pairwise_distances 發訪，進行城市距離的轉換。

(4)資料降維：使用 MDS 方法，將原始資料從 3 維降至 2 維。

(5)模型輸出：最後使用 matplotlib 套件，將降維結果可視化輸出。

4.2.2 飲料資料集

(1)套件載入:載入協助資料分析的 Pandas、載入 sklearn.manifold 套件的 TSNE 降維方法，並載入繪圖（matplotlib）與詞嵌入（gensim）套件。

(2)詞嵌入:利用 gensim 套件導入詞嵌入預訓練詞庫（GoogleNews-vectors-negative300.bin），將各個單詞的相似程度轉換為詞向量建立而成的相似度矩陣。

(3)相似度矩陣資料降維：使用 TSNE 對相似度矩陣降維並視覺化呈現。

(4)1-of-k 資料降維：利用 sklearn.preprocessing 的 One-Hot-Encoding 方法轉換原始資料並利用 TSNE 進行 1-of-k 降維並視覺化呈現。

4.3 實驗結果

4.3.1 台灣主要車站距離集降維結果

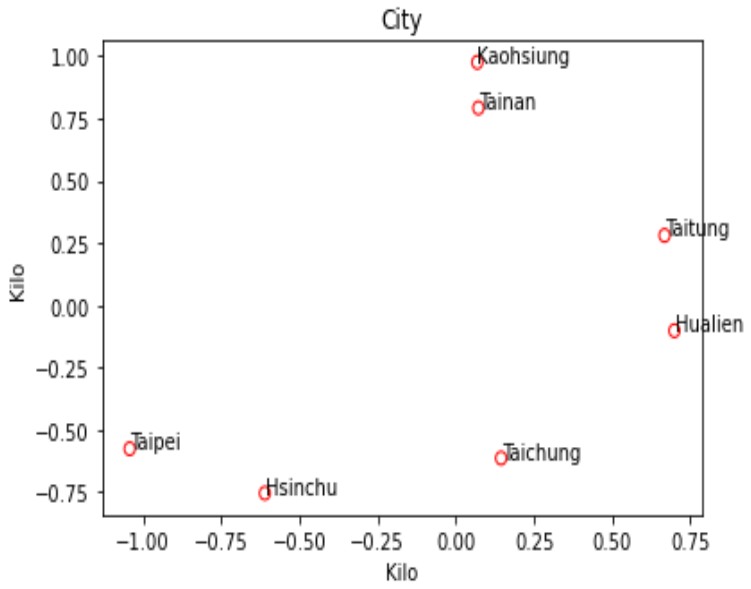

圖 3 台灣主要車站距離集降維結果\

4.3.2 飲料資料集降維結果

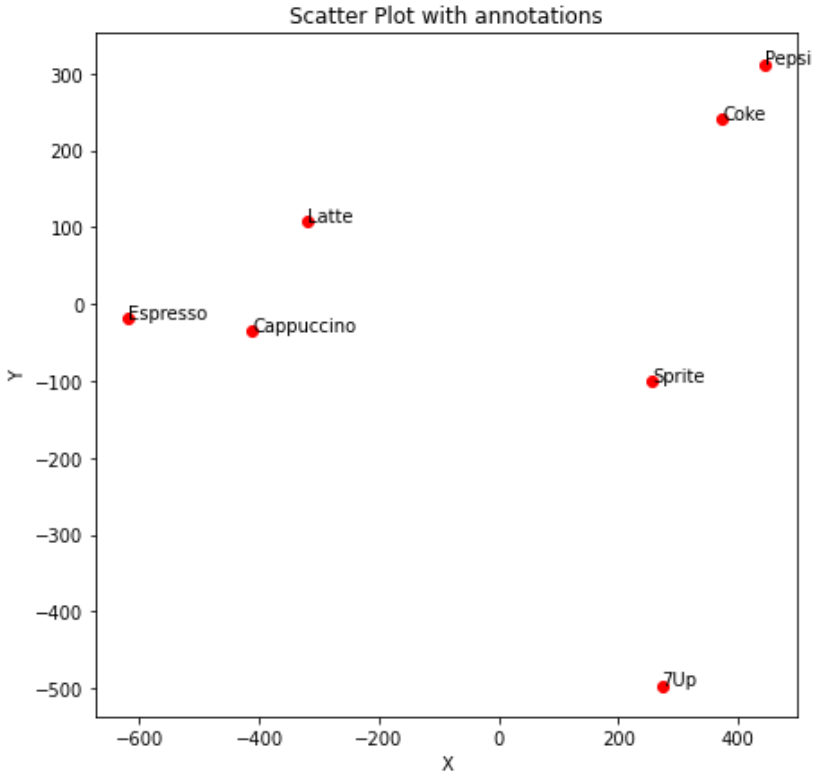

圖 4 相似度矩陣的降維結果

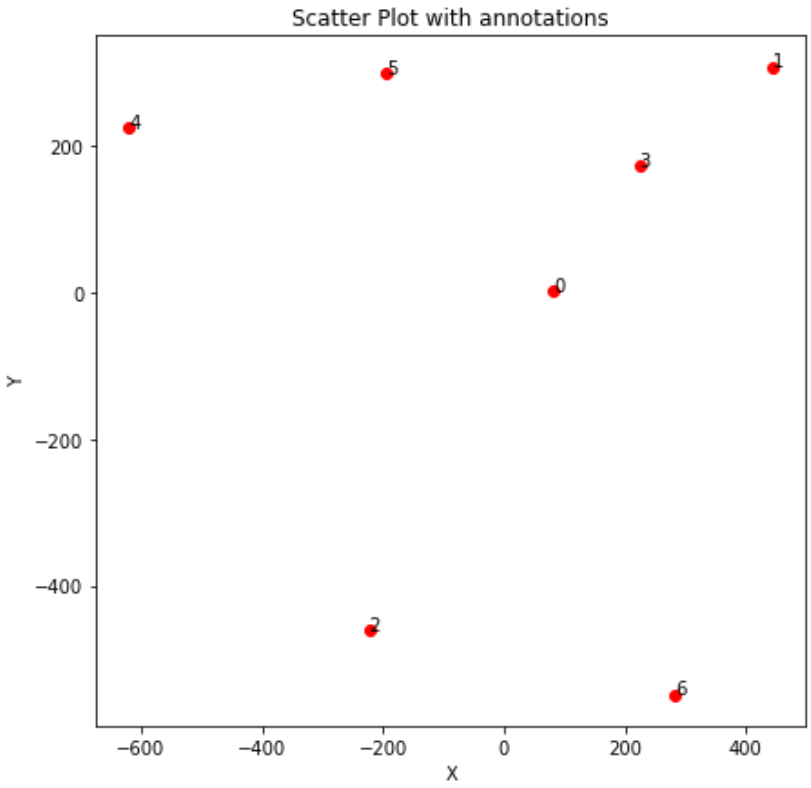

圖 5 1-of-k 的降維結果

四、 結論

根據台灣主要車站距離資料集的結果，透過 MDS 方法可有效進行資料降維，可將城市定位於 2D 平面上，或許因為經緯度差距不大，導致只能明確定位兩城市之間的距離，但在相對位置上還有改進的空間。

從飲料資料集的降維結果可以發現，利用詞嵌入的降維方式效果優於 1-of-k 降維，詞嵌入可保留單詞的涵義，可發現降維後的資料集區分出汽水類與咖啡類，然而 1-of-k 僅有表示量化的功能，忽略了單詞的語意和順序性，讓降維效果不佳。

參考文獻

[1] [Day 7] 非監督式學習-降維

https://ithelp.ithome.com.tw/m/articles/10267685

[2] 降維- 維基百科，自由的百科全書

https://zh.wikipedia.org/zh-tw/%E9%99%8D%E7%BB%B4

[3] 機器學習-降維演算法(MDS 演算法)

https://www.796t.com/content/1550606051.html

[4] 數據降維-多維尺度縮放（MDS）

https://www.twblogs.net/a/5efd970741b2036d50955e9b

[5] 機器學習降維技術（PCA，ICA 和流形學習）及醫學中流形學習的應用

https://kknews.cc/health/jk6lox6.html

[6] 十種方法實作影像資料集降維

https://www.uj5u.com/qita/277459.html

[7] 資料降維與視覺化：t-SNE 理論與應用

https://mropengate.blogspot.com/2019/06/t-sne.html

[8] 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

[9] Python - 如何使用 t-SNE 進行降維 - Mortis

https://mortis.tech/2019/11/program_note/664/

[10] [Day 25] tSNE - dimension reduction / 非線性降維方法與視覺化

https://ithelp.ithome.com.tw/articles/10219137