機器學習 專案作業四

本作業主要目的為使用Python語言實作資料集維度縮減。

作業內容包括：

* + 利用Python API練習及比較不同降維度的方法。
    - 資料集一：請找出**台北高鐵站、苗栗高鐵站、雲林高鐵站、台南高鐵站、高雄高鐵站、花蓮豐濱、台東鹿野**彼此之間的距離，然後用MDS畫在2D平面上。也請使用Google地圖標記其位置。
    - 資料集二：Drink Dataset共有四個特徵欄位(Drink, Rank, Amount, Quantity)及一個類別欄位(Class)：Drink為名目型欄位，Rank、Amount、Quantity為數值型欄位。針對每一個類別(Class)，依照Amount的常態分配及亂數分配Quantity，隨機產生Count數量的資料筆數。以tSNE將資料呈現在2D平面。
    - tSNE方法處理名目型屬性時，請採用兩種方法，並比較結果：1-of-*k*及考慮屬性值的相似度。類似ML04.1 Dimensionality Reduction (MDS t-SNE)投影片第19頁的結果。
    - 當考慮名目型屬性值相似度時，**同類品項的相似度值應該較高**，例如，碳酸飲料彼此之間的相似度應該高於與咖啡類飲料之間的相似度。
    - 名目型屬性值的不相似度值可以人工設定，例如，同類品項的不相似度值為1，不同類品項的不相似度值為2。或者可以Word2Vec的方式計算名目型屬性值之間的距離。或者以其他機器學習的方法，從Drink Dataset本身學到名目型屬性值之間的不相似度值。提醒：計算任兩筆資料距離之前，需進行屬性值正規化。

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | | **Table 1. Drink Dataset** | | | | |
| **Class** | **Drink** | | **Rank** | **Amount(*N*(*μ*,*σ*))** | **Quantity** | **Count** |
| A | 7Up | | 7 | (100, 200) | Random(500, 1000) | 300 |
| B | Sprite | | 6 | (200, 10) | Random(500, 1000) | 150 |
| C | Pepsi | | 5 | (200, 10) | Random(500, 1000) | 150 |
| D | Coke | | 4 | (400, 100) | Random(500, 1000) | 300 |
| E | Cappuccino | | 3 | (800, 10) | Random(1, 500) | 150 |
| F | Espresso | | 2 | (800, 10) | Random(1, 500) | 150 |
| G | Latte | | 1 | (900, 400) | Random(1, 500) | 300 |
|  | | *N*(*μ*,*σ*)：Normal Distribution | | | | |

* + 請利用Linked brushes或者dash的技術，於螢幕互動式的圈選資料群集，取得該群組的資料點。參考資料：<https://stackoverflow.com/questions/44531241/python-how-to-get-data-from-linked-brushes-in-mlpd3-bokeh-plotly>
  + 專案程式直接上傳Github，專案名稱：hw4+組長學號，並新增助教為專案成員，以利下載評分。
  + 請遵循學術論文寫作格式撰寫專案報告，內容至少必須包括下列要項：

**摘要** (學術論文摘要通常不分段，內容簡要地涵蓋動機、目的、方法與實驗結果)

**一、緒論**

**1.1動機** (探勘所選用的資料集之動機)

**1.2目的** (想進行哪種探勘，期望了解甚麼？)

**二、方法**

(1. 簡單說明程式架構，2. 執行程式的方法，無法執行程式，視為不正確)

**三、實驗**

**3.1資料集** (簡介資料集：名稱、筆數、欄位、型態…、並以資料表格呈現部份資料內容)

**3.2前置處理** (例如，欄位值正規化至0~1之間等。另外，須說明名目型屬性值的處理方法)

**3.3實驗設計** (實驗如何進行、若有參數的話，參數如何設定等）

**3.4實驗結果** (呈現結果)

**四、結論**

* 評分標準

內容(難度、豐富性、方法正確性)：60%，論文格式：20%，寫作技巧(文筆通順)：20%

* **任何複製剪貼網站資料或學長姐以前作業，皆屬抄襲，嚴重者零分。**
* 此份報告最多不超過15頁，請加註頁碼。請遵循投稿論文寫作格式，使用一欄式，無需製作論文目錄。可以觀摩 ACM Transactions on Knowledge Discovery from Data 期刊論文的寫作技巧與格式。
* 文件請以電子檔(轉成PDF格式)上傳到Tronclass。沒有上傳電子檔視為未完成繳交作業。電子檔命名：HW4-組長學號。
* 每遲交一天，扣該次作業成績10分。
* 繳交日期：112/06/26（一）