Data Mining Homework 2

Due: 2021/12/2

➢ Clustering: K-means

Dataset description:
# of data: 600
# of attribute: 60
# of cluster: 6

公式一、 cost of clustering using Euclidean distance

$$\phi = \sum_{x \in \mathcal{X}} \min_{c \in \mathcal{C}} ||x - c||^2$$

公式二、 cost of clustering using Manhattan distance

$$\psi = \sum_{x \in \mathcal{X}} \min_{c \in \mathcal{C}} |x - c|$$

(a) 請使用上課所教的 Euclidean distance 作為計算資料點間相似度的依據，並探討以下兩種 centroid initialization 的方法，分別執行 Simple K means 演算法。

1. 以作業所提供的 c1.txt 中的六個座標點，作為演算法起始狀態的 initial cluster centroids (若是使用 weka 作為分析工具，請設定 initialization method 為 "random"，並把 seed 參數調整為: 456)，然後對作業提供的資料集，執行 k-means 分群演算法，並記錄下每個 iteration 所得到的 Within cluster sum of squared error (SSE)(公式一)，直到演算法執行至停止條件，請將得到的結果，繪製成 x-y 折線圖，x 軸標記由 1 開始的 iteration 數量，y 軸為該 iteration 得到的 SSE。

2. 同上題的題目要求，但改用 c2.txt 檔案提供的 initial centroids，作為演算法起始條件(使用 weka 分析的同學請改用 "farthest first" 為 initialization method，並一樣使用 456 作為 seed)，請將本題得到的 iteration-error 折線圖，**繪製到和上題同一個圖中，並在圖例上標記清楚，兩條折線圖各自的名稱 (Euclidean-random / Euclidean-farthest)**。

3. 請分別對兩種情況下的分群結果，以多至少的方式，列出最終 cluster 包含的資料點數量，舉例來說:
   Euclidean-random: 150, 120, 100, 90, 80, 60
   Euclidean-farthest: 180, 140, 100, 100, 40, 40

(b) 請使用上課所教的 Manhattan distance 作為計算資料點間相似度的依據，並探討以下兩種 centroid initialization 的方法，分別執行 Simple K means 演算法。

1. 以作業所提供的 c1.txt 中的六個座標點，作為演算法起始狀態的 initial cluster centroids (若是使用 weka 作為分析工具，請設定 initialization method 為 "random"，並把 seed 參數調整為: 456)，然後對作業提供的資料集，執行 k-means 分群演算法，並記錄下每個 iteration 所得到的 Sum of within cluster distances(公式二)，直到演算法執行至停止條件，請將得到的結果，繪製成 x-y 折線圖，x 軸標記由 1 開始的 iteration 數量，y 軸為該 iteration 得到的 Sum of within cluster distances。
2. 同上題的題目要求，但改用 c2.txt 檔案提供的 initial centroids，作為演算法起始條件(使用 weka 分析的同學請改用 "farthest first" 為 initialization method，並一樣使用 456 作為 seed)，請將本題得到的 iteration-error 折線圖，**繪製到和上題同一個圖中，並在圖例上標記清楚，兩條折線圖各自的名稱 (Manhattan-random / Manhattan-farthest)。**

3. 請分別對兩種情況下的分群結果，以多至少的方式，列出最終 cluster 包含的資料點數量，舉例來說:
   Manhattan-random: 150, 120, 100, 90, 80, 60
   Manhattan-farthest: 180, 140, 100, 100, 40, 40

(c) 試比較 (a) 和 (b) 的結果，有沒有觀察到什麼值得討論的現象?並請根據學習到的 k-means 知識，試著解釋你的發現。

(d) 請繪製出 number of clusters(k)-to-SSE 的折線圖。實驗條件為: 以 Euclidean distance 為 similarity function、初始 centroids 採用 random initialization、maximum iteration 數量設為 20。探討 number of clusters (k)，從 1,2,3…依序漸增到 12 時，相對應的 Within cluster sum of squared errors 分別為多少。

(e) 根據 (d) 的結果，並設想一個資料探勘的應用場景，是否我們會傾向使用越大的 cluster 數量來跑 k-means 分析我們的資料集會越好?請說明你回答 yes or no 的理由。

(f) 若不是使用 weka 完成作業的同學，請簡介你使用的工具，並於報告附上自行實作的程式碼，於本題提供能夠再現你結果的必要資訊(例如: 如何調整參數、用什麼工具繪製圖表等等)

➢ Multi-class classification

Dataset description:

# of Training data : 1000

# of Testing data: 200

# of Attribute: 14

# of class : 5

Rules:

限使用 Tree-based 分類演算法，可以上課提過或是投影片有列出的方法為主。

可以對 Data 作任何你覺得能增進效能的前處理，並請在報告中討論原因。

請在 Kaggle 平台上，上傳你對 testing dataset 的預測結果，上傳格式請參考 prediction_sample.csv，public 與 private 的比例為 33% 比 66%。

Hint: 避免 classifier overfitting。

Baseline accuracy: 0.90909

結果超過 baseline(private) 可以拿到基本分，其餘根據報告內容評分

https://www.kaggle.com/t/148b9fdf660d49a2890c13434e67857f

每日上傳次數限制: 3 次。

名稱請用學號_姓名。

➢ 繳交項目

1. Clustering 報告 (pdf)
2. Classification 報告(pdf)，開頭請先列出你是用什麼 Tool 完成，並簡介你所使用的演算法工作原理，需要調整哪些參數來優化你的模型?以及使用什麼方式進行 validation 以避免 overfitting?
3. 擇一:

若是用 WEKA 完成的同學請附上自己進行完資料前處理後的檔案。

若是用其他 Tool 包含自行實作的同學請附上 code。

最後所有項目包裝成壓縮檔上傳 moodle，檔名為學號 +_DM_HW2

EX: M12345678_DM_HW2