CURSO: CMP 5002 - DATA MINING
COLEGIO: POLITÉCNICO
Semestre: 2do Semestre 2023/2024

Proyecto 8: Ejercicio usando agrupaciones (*clustering*)

Problema:
Dado el conjunto de datos " dataset(wq)" enviado por el chat grupal, se desea aplicar los algoritmos de agrupamiento para separar los datos por grupos comunes (***clusters)***. Para la realización de la tarea se exige:

- Cada equipo debe usar dos de los algoritmos presentados a continuación y no pueden repetirse entre equipos:
    - *Affinity Propagation, (Equipo1)*
    - *Agglomerative Hierarchical Clustering, (Equipo2)*
    - *BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies), (Equipo1)*
    - *Mean Shift Clustering, (Equipo2)*
    - *Spectral-clustering, (Equipo3)*
    - *BFR (Equipo3)*
    - *CURE (Clustering Using Representative), (Equipo4)*
    - *DBSCAN (Density-Based Spatial Clustering of Applications with Noise), (Equipo5)*
    - *Expectation-Maximization (Equipo4)*
    - *Gaussian Mixture Models (Equipo5)*
- Cada equipo debe realizar un ***research*** sobre los algoritmos asignados. De forma tal que puedan presentar y discutir sobre la teoría de *clustering* y a su vez de los algoritmos desarrollados.
- Cada equipo debe realizar un ***research*** sobre el método t-SNE (https://en.wikipedia.org/wiki/T-distributed_stochastic_neighbor_embedding), de forma tal que pueda implementarlo para la resolución de literales relacionados a la proyección de los datos y resultados del algoritmo de *clustering* empleado.
- Es obligatorio mostrar la trazabilidad del método de *clustering*:
    - Normalización del *dataset* usando el método *min-max.* **(0.5 punto)**
    - Optimización del valor de *k* (numero de *clusters*) en un intervalo de ***k=2..8.*** **(2 puntos)**
    - Calcular las siguientes métricas para medir el desempeño del modelo y determinar el mejor número de clusters basado en *Rand index, Mutual Information based scores, Homogeneity, completeness and V-measure, Fowlkes-Mallows scores, Silhouette Coefficient, Calinski-Harabasz Index, Davies-Bouldin Index Contingency Matrix, Pair Confusion Matrix*. Usar únicamente las que no dependan de la etiqueta de salida. **(3 puntos)**

- Realizar un plot de desempeño de los algoritmos versus la variación de ***k***. Para los algoritmos que trabajen con el concepto de *centroide o clustroide* pueden usar el *plot* $Dist_{avg}$ *vs k (Ellbow plot)* visto en clase. Para otros algoritmos que no usan este concepto, se debe investigar alguna alternativa de plot que nos permita ver el desempeño por ***k***. **(2 puntos)**
- Imprimir el valor de ***k*** óptimo de acuerdo a su selección. **(0.5 punto)**
- Mostrar el plot t-SNE para el espacio original de los datos normalizados. (**0.5 puntos**)
- Mostrar el plot t-SNE (tres plot en total) después de aplicado el método de *clustering* al *dataset* normalizado para los valores de ***k*** óptimo, ***k-1*** y ***k+1***. De esta forma se podrá visualizar los *clusters* en la vecindad del valor de ***k*** óptimo determinado por el inciso (4). **(1.5 puntos)**

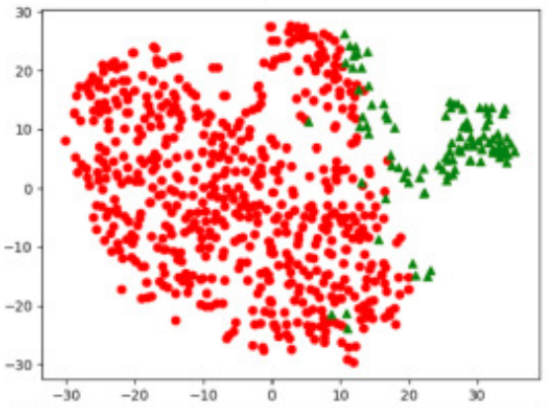

Fig. 1 Ejemplo de representación de dos *clusters* (rojo y verde) usando la técnica t-SNE.

+1 punto: Detectar y eliminar *outliers* usando *clustering*. Se debe demostrar de alguna forma el procedimiento si fuera aplicable.

- Cargar al D2L los códigos implementados dentro del plazo de entrega.