UNIVERSIDAD AUSTRAL | INGENIERÍA

Introducción a Data Mining – 2022

Trabajo Práctico Nº 4 – Análisis de Conglomerados

Nota: Realizar los ejercicios sin recurrir a software de data mining

1) Computar la distancia Minkowski entre {1,2} y {3,4} para h=1, 2, ∞.

2) Computar la similitud del coseno y el coeficiente de Jaccard entre los conjuntos {A, B, C} y {A, C, D, E}.

3) Divida las muestras del dataset en 2 clústeres utilizando el algoritmo K-means. Como centroides tome las dos observaciones más alejadas.

   ¿Es mejor dividir la muestra en 3 *clústeres*? Justifique.

Observación	Sueldo en miles	Antigüedad Laboral
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

4) Dado el dataset {1, 5, 8, 10, 2} utilice el Algoritmo Aglomerativo de Clustering con *link complete* y distancia Euclidiana para establecer las relaciones jerárquicas entre los grupos. ¿Cuántos Clústeres encuentra? ¿Cuál es la relación entre cada uno?

5) Clustering Jerárquico es utilizado a veces para generar K clústeres, K> 1 tomando los clústeres al nivel K^{th} del dendograma. (La raíz es el nivel 1). Mirando los clústeres producidos de este modo, podemos evaluar el comportamiento de Clustering Jerárquico en diferentes tipos de datos y clústeres, y también comparar Clustering Jerárquico con K-means.

   El siguientes un conjunto unidimensional de puntos {6, 12, 18, 24,30, 42, 48}

a. Para cada uno de los siguientes conjuntos de centroides iniciales, crear dos clústeres asignando cada punto al centroide más cercano, y luego calcular el error cuadrático total para cada uno de los dos clústeres. Muestre los dos clústeres y luego el error total cuadrático para cada uno de los centroides.

   i. {18, 45}

   ii. {15, 40}

b. ¿Los dos centroides dados representan una solución estable, por ejemplo, si el algoritmo K-means fuera corrido en este conjunto de puntos utilizando los centroides dados como los centroides iniciales, habría algún cambio en los clústeres generados?

c. ¿Cuáles con los clústeres generados utilizando *single link*?

d. ¿Cuál técnica, K-means o singlelink, parece producir los clústeres "más naturales" (Para K-means considere los clústeres con menor error cuadrático)?

6) Nuestra tarea de minería de datos consiste en conglomerar los siguientes ocho puntos (x,y) que representan una ubicación en el espacio R^2, en tres conglomerados.

   A1 (2,10), A2 (2,5), A3 (8,4), B1 (5,8), B2 (7,5), B3 (6,4), C1 (1,2), C2 (4,9)

   Emplee la función de distancia *Euclidea y* suponga que inicialmente asignamos A1, B1 y C1 como los centros de cada conglomerado. Emplee *K-medias* para mostrar *solamente*:

   a) Los tres centros de los conglomerados luego de la primera ronda de movimientos y

   b) Los tres conglomerados finales.