El algoritmo K-medoides

El algoritmo K-medoides (*K-medoids*) es una variante del algoritmo K-medias que se utiliza también para dividir un conjunto de datos en K grupos exclusivos y al mismo nivel. Sin embargo, hay algunas diferencias entre ellos que puedes observar en la Tabla 1. Veamos:

Tabla. 1. Diferencias entre K-medias y K-medoides.

Aspecto	K-medias	K-medoides
Prototipos	Cada grupo está representado por su centroide, que es el punto promedio de todos los puntos en el grupo.	Cada grupo está definido por un punto real de los datos, conocido como medoide, que es el dato más representativo del grupo.
Paso de actualización	Determina los nuevos prototipos como el promedio de todos los puntos del grupo.	Considera cada dato $\boldsymbol{x}_i$ como candidato para sustituir a un prototipo. La decisión de si sustituir o no un centroide $\boldsymbol{c}_j$ por $\boldsymbol{x}_i$ se basa en un criterio que cuantifica el costo de reemplazo.
Criterio para optimizar	Suma de las distancias al cuadrado (inercia): $$SSE = \sum_{i=1}^{k} \sum_{\boldsymbol{x} \in G_i} \|\boldsymbol{x} - \boldsymbol{c}_i\|^2$$	Suma de las distancias absolutas: $$SSE = \sum_{i=1}^{k} \sum_{\boldsymbol{x} \in G_i} \|\boldsymbol{x} - \boldsymbol{c}_i\|$$
Sensibilidad a los valores atípicos (*outliers*)	Utiliza la distancia euclidiana para medir la similitud entre los puntos y los centroides, la cual puede verse afectada por los valores atípicos o outliers.	Al utilizar la suma de las distancias absolutas resulta menos sensible a los *outliers*. Por otra parte, emplear un dato real como medoide conlleva mayor robustez ante la presencia de valores atípicos

Interpretabilidad	Los centroides son puntos abstractos que pueden no estar presentes en los datos reales. Esto puede dificultar la interpretación y comprensión de los grupos cuando se tiene variables categóricas.	Los medoides son puntos reales de los datos. Esto facilita la interpretación y análisis de los resultados, sobre todo cuando se tienen variables categóricas.

Para finalizar, resaltar que ambos algoritmos seleccionan K instancias del conjunto de datos, de manera aleatoria, como prototipos iniciales para generar la primera partición, mediante la asignación de cada dato a su prototipo más cercano.

Bibliografía

Bonaccorso, G. (2019). Hands-On Unsupervised Learning with Python. O'Reilley.

Han, J., Kamber, K., Pei, J. (2011). Cluster Analysis. Data mining: Concepts and Techniques. Capítulo 10. O'Reilley.

Universidad de los Andes | Vigilada Mineducación Reconocimiento como
Universidad: Decreto 1297 del 30 de mayo de 1964.
Reconocimiento personería jurídica: Resolución 28 del 23 de febrero de 1949 Minjusticia.