Más sobre al algoritmo K-medias

El algoritmo K-medias se basa en un esquema iterativo con dos pasos básicos, como puedes ver en la Fig. 1. En el primero, se asigna cada dato al prototipo más cercano para crear las particiones. Para realizar esta tarea se calcula la distancia de todos los datos a cada prototipo, usualmente utilizando la distancia Euclídea. El segundo paso tiene como objetivo actualizar el prototipo con el punto medio de todos los datos del grupo. A continuación, veremos algunos aspectos que caracterizan este algoritmo.

Algoritmo K-medias

Entrada = *{D: conjunto de datos, K: número de grupos}*

Comienzo_procedimiento

Seleccionar K datos de D como prototipos iniciales $\mathbf{c}_i$, $i=1...K$

Repetir

Crear las particiones G_i *asignando cada dato de D a su prototipo más cercano*

Actualizar cada prototipo $\mathbf{c}_i$ *como la media de los datos de la partición* G_i

hasta *alcanzar la convergencia*

Fin-procedimiento

Salida = *Conjunto de prototipos* $\mathbf{c}_i$, $i=1..K$

Fig. 1. Algoritmo K-medias.

Procedimiento de optimización

El algoritmo K-medias emplea una función objetivo para guiar la formación de los grupos. Esta función evalúa la variación interna de los grupos y se define como la suma de las distancias cuadráticas entre cada dato $\boldsymbol{x}$ y su prototipo $\boldsymbol{c}$ (conocida también como criterio de inercia), como se detalla a continuación:

$$suma\ de\ las\ distancias\ al\ cuadrado\ =\ SSE = \sum_{i=1}^{k} \sum_{\boldsymbol{x} \in G_i} \|\boldsymbol{x} - \boldsymbol{c}_i\|^2$$

$$donde\ \begin{cases} \boldsymbol{x} \in D = \{\boldsymbol{x}_j\}_{j=1\ldots N} \\ G_i = grupo\ asociado\ al\ prototipo\ \boldsymbol{c}_i, i = 1..k \end{cases}$$

Al calcular un nuevo prototipo en cada iteración como el punto medio de los datos del grupo, se mejora la variación dentro de los grupos e, implícitamente, se está minimizando la función objetivo. En este proceso se intenta obtener grupos tan compactos como sea posible.

Inicialización de los centros

Un aspecto que caracteriza al algoritmo K-medias es que los resultados dependen de los valores de los prototipos que utiliza para comenzar a construir los grupos. Diferentes valores iniciales conducen a agrupaciones diferentes. La recomendación es entonces ejecutar el algoritmo con diferentes valores de centros iniciales y seleccionar la agrupación que mejores resultados ofrezca con base en alguna métrica de evaluación (por ejemplo, SSE). También se han propuesto esquemas de inicialización para mejorar la convergencia, como K-means++.

Otras características

El algoritmo K-medias, al utilizar una métrica de distancia como la Euclídea para determinar la similitud entre instancias, requiere de un escalado de los datos con el fin de no suministrar una importancia o ponderación de variables de manera implícita, donde atributos de mayor rango tendrán un mayor peso en el resultado. Por otra parte, si se utiliza la distancia Euclídea, el algoritmo tendrá una mayor sensibilidad a los *outliers*.

¿Qué pasa si hay datos categóricos?

Otro aspecto para resaltar en K-medias es que los prototipos finales no se encuentran en el conjunto de datos, lo cual podría ser una desventaja a la hora de interpretar los resultados. Por otra parte, si

el conjunto de datos tiene atributos categóricos, estos deberían ser transformados en numéricos para poder aplicar el algoritmo. En este contexto no tendría sentido derivar una media para estos atributos. Una solución está en el algoritmo K-medoides. En este, el centro de cada grupo siempre es un dato del conjunto de datos. Este algoritmo es una variación de K-medias que utiliza como criterio de optimización la suma de las distancias de cada punto a su prototipo, es más interpretable y menos sensible a los *outliers*, aunque tiene en contra un mayor costo computacional.

Bibliografía

Géron, A. (2022). Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow. Capítulo 9. Third edition. O'Reilly.

Jin, X., Han, J. (2011). K-Medoids Clustering. In: Sammut, C., Webb, G.I. (eds) Encyclopedia of Machine Learning. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-30164-8_426.

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.