Teoría k means

Jorge Aquino

El algoritmo k-means es un algoritmo de agrupamiento (clustering) que fue propuesto por primera vez por Stuart Lloyd en 1957 como una técnica para la reducción de dimensionalidad en la transmisión de señales. Sin embargo, este algoritmo no se hizo popular hasta la publicación de un artículo de John MacQueen en 1967, donde se describía el algoritmo como una técnica de agrupamiento.

El algoritmo k-means consta de cuatro fases principales: inicialización, clasificación, cálculo de centroides y convergencia:

Inicialización: En esta fase, se seleccionan k centroides de manera aleatoria o utilizando algún método de inicialización específico. Estos centroides se utilizan como representantes iniciales de los k grupos.

Clasificación: En esta fase, cada punto del conjunto de datos se asigna al centroide más cercano. Es decir, se cálcula la distancia entre cada punto y cada centroide, y se asigna cada punto al centroide más cercano. De esta manera, se divide el conjunto de datos en k grupos.

Cálculo de centroides: En esta fase, se actualizan los centroides como la media de todos los puntos asignados a ellos. Es decir, se calcula la media de las coordenadas de todos los puntos asignados a cada centroide, y se utiliza este valor como la nueva posición del centroide. De esta manera, los centroides se mueven hacia el centro de gravedad de los puntos asignados a ellos.

Convergencia: El proceso de clasificación y cálculo de centroides se repite varias veces hasta que los centroides ya no cambien de posición o se alcance un número máximo de iteraciones. Cuando los centroides ya no cambian de posición, se dice que el algoritmo ha convergido y se ha encontrado una solución.

La función objetivo del algoritmo k-means es una medida de la calidad de la solución obtenida por el algoritmo. Esta función se utiliza para evaluar la distancia entre los puntos y los centroides, y para determinar cuán bien los centroides representan a los grupos.

La función objetivo del algoritmo k-means se define como la suma de las distancias cuadradas de cada punto a su centroide correspondiente. Esta medida de distancia se utiliza debido a su simplicidad y su capacidad para capturar la varianza de los datos. La función objetivo se puede expresar de la siguiente manera:

$$J = \sum_{i=1}^{n} \sum_{j=1}^{k} |x_i - \mu_j|^2$$

Donde:

J es la función objetivo.

n es el número de puntos en el conjunto de datos.

k es el número de centroides.

x_i es el punto i-ésimo.

μ_j es el centroide j-ésimo.

El objetivo del algoritmo k-means es minimizar la función objetivo J. Esto se logra mediante la asignación óptima de los puntos a los centroides y la actualización de los centroides. En otras palabras, se busca encontrar los centroides que minimicen la distancia entre los puntos y los centroides.