El algoritmo Mean Shift. Densidad más centroides

Mean Shift es un algoritmo basado en centroides, que considera todo el espacio de características como una función de densidad de probabilidad. En este contexto, los grupos se corresponden con los máximos de esta distribución, es decir, con regiones de alta densidad. El objetivo es entonces ubicar los centroides en estos "picos". Para realizar esta tarea, se basa en un procedimiento conocido como "desplazamiento de media" el cual, de manera iterativa, "mueve" estos puntos hacia la "media" de regiones más densas hasta alcanzar la convergencia. A continuación, se describen los pasos que realiza el algoritmo:

1. Definición del ancho de banda (*bandwidth*): este parámetro determina el tamaño de la ventana de búsqueda alrededor de cada punto para calcular el desplazamiento de media.

2. Selección. El algoritmo comienza seleccionando aleatoriamente un punto del conjunto de datos o, en algunos casos, puede utilizar una estrategia de muestreo para seleccionar varios puntos iniciales. Este punto es el centroide inicial.

3. Cálculo del desplazamiento de media. Para el centroide inicial se calcula el desplazamiento de media. Este desplazamiento indica la dirección hacia la que se mueve el punto para dirigirse a una región más densa (ver Fig. 1). Para ello, se realiza lo siguiente:

- Cálculo de la media ponderada de todos los puntos dentro de la ventana de búsqueda, donde cada punto tiene un peso que se determina a través de una función Kernel. La más comúnmente utilizada para este propósito es la función gaussiana, que asigna los pesos con base en la cercanía al centroide inicial: pesos más altos mientras más cercanos, y más bajos mientras más alejados.
- Determinación de la dirección de desplazamiento desde el centroide inicial hacia el punto medio ponderado.
- Actualización de la posición del centroide. El punto se mueve hacia la dirección del desplazamiento de media calculado en el paso anterior. Este proceso se repite iterativamente

hasta que el centroide converja hacia una región de alta densidad. Para esto, se verifica si el desplazamiento de media es menor que un umbral determinado (por ejemplo, si la distancia entre el punto actual y el punto después del desplazamiento es menor que un valor de tolerancia).

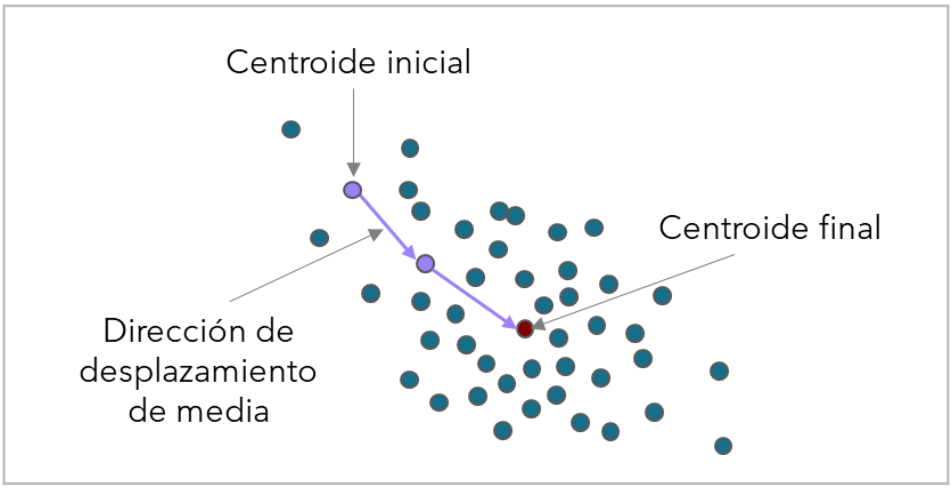

Fig. 1. Desplazamiento de media en el algoritmo Mean Shift.

4. Iteración. Repetir los pasos 2 y 3 para cada punto del conjunto de datos.

5. Fusión de grupos. Después de la convergencia, se pueden fusionar grupos si están lo suficientemente cercanos entre sí, lo que puede ayudar a reducir el número final de clústeres.

¿Cuáles son sus hiperparámetros?

- Ancho de banda (*bandwidth*). Representa el tamaño de la ventana de búsqueda alrededor de cada punto en el que se busca calcular el desplazamiento de media. Si el ancho de banda es demasiado pequeño el algoritmo puede terminar convergiendo a muchas regiones locales, lo que daría lugar a una segmentación excesiva. Por otro lado, si el ancho de banda es demasiado grande puede ocurrir lo contrario y terminar en una sola región global.
- Función Kernel. Juega un papel importante en la ponderación de los puntos dentro de la ventana de búsqueda durante el cálculo del desplazamiento de media. Es una función que asigna pesos a los puntos basándose en la distancia respecto al centroide. La función Kernel

más comúnmente utilizada es la gaussiana, pero también se pueden emplear otras. Su elección puede afectar significativamente los resultados del algoritmo, ya que la diferencia entre funciones se refleja en la importancia otorgada a los puntos cercanos y lejanos.

- Criterio de convergencia. Se refiere a la condición que determina cuándo detener el proceso de desplazamiento del centroide. Puede ser un número fijo de iteraciones, un umbral de cambio mínimo en la posición del centroide o una combinación de ambos.

Para hallar la mejor combinación de estos hiperparámetros para un problema dado se puede construir una "grilla" de posibles valores y generar una agrupación para cada combinación, con su respectivo valor de la métrica de evaluación que haya sido seleccionada. Al final, se escogería la combinación que arroje el mejor valor para esta. El proceso es similar al método del codo o el de la silueta para determinar el valor de K para un algoritmo como K-medias y asimismo es aplicable a otros algoritmos de agrupación.

En resumen

- Aunque Mean Shift utiliza la densidad de los datos para determinar la dirección del desplazamiento de media, en sí mismo es un algoritmo basado en centroides.
- No requiere que le especifiquemos el número de grupos a priori ya que es capaz de encontrar este número con base en la estructura de densidad de los datos.
- Es robusto frente a formas y tamaños de grupos irregulares, lo que lo hace especialmente útil cuando los datos no tienen estructuras esféricas o elípticas.
- Es muy utilizado en procesamiento de imágenes.

Bibliografía

Aggarwal, Ch., Reddy, Ch. (2013). Data *clustering*. O'Reilly.

Gallatin, K., Albon, C. (2023). *Machine Learning with Python Cookbook, Capítulo 19.* O'Reilly, 2nd Edition.

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.