Hola! Soy Jesús Gutiérrez, y en este video te hablaré sobre el mini batch K-means. Si alguien ya habló de este método antes que mí pues qué pena, a ver si así engancha mejor.

Todos conocemos al K-medias. Es el método de agrupación más popular. K centroides son ubicados aleatoriamente en el conjunto de datos, y se forman k grupos que asocian cada dato a su centroide más cercano, luego el centroide se recalcula, y se iteran estos dos últimos pasos hasta llegar a la convergencia. La forma en la que se recalculan los centroides sigue la función objetivo



Donde x es cada dato en el conjunto X, y C es el centroide. Lo que quiere decir es que quiere minimizar la suma de las distancias euclidianas entre cada dato y el centroide del grupo en el que está.

No obstante, K means deja de ser tan atractivo cuando tenemos muchos datos, pues necesita de todo el conjunto en la memoria y se vuelve lento.

Para solucionarlo, llega a la puerta de nuestras casas mini batch K-means, K-medias con mini lotes en español, o como a mi me gustó traducirlo K-medias con lotecitos. Es una variante de K-means que reduce el tiempo de cómputo mientras busca optimizar la misma función objetivo.

Para enfrentarse al problema de la memoria, lo que hace es trabajar con lotecitos, o sea, muestras aleatorias de los datos en cada iteración. Pero todo tiene un costo, y el sacrificio es tener un resultados tantito peores que K-means.

Vamos a ver con un poco más de detalle el algoritmo:

La entrada es el conjunto de datos, el número de grupos k, el tamaño del lotecito b, y las iteraciones t y la salida son los grupos.

Comienza con k centroides inicializados en algún dato x del conjunto D. Inicializa los clústeres Ci y el contador de cada centroide.

Aquí sigue el for grandote:

De 1 hasta donde t le permita, lo que hará será

Primero tomar una muestra aleatoria M de tamaño b del conjunto de datos.

Y ahora sigue el primer for chiquito, que va desde 1 hasta b:

Este ubica el centroide más cercano al dato xm

Una vez que termina prosigue el siguiente for chiquito, que también va de 1 hasta b

Toma el centro que le corresponde a xm

Actualiza el contador.

Y esta es la parte que ayuda a que el algoritmo converja, la tasa de aprendizaje. Y con lo de aquí, llamado “dar el paso del gradiente” actualiza el centroide.

Termina el for chiquito y el grandote.

Si no es con el número de iteraciones predeterminadas, el algoritmo acaba cuando converge.