El algoritmo K-means agrupa los datos tratando de separar las muestras en n grupos de varianza igual, minimizando un criterio conocido como *inercia* o suma de cuadrados dentro del grupo (ver más abajo). Este algoritmo requiere que se especifique el número de clústeres. Se adapta bien a una gran cantidad de muestras y se ha utilizado en una amplia gama de áreas de aplicación en muchos campos diferentes.

El algoritmo de k-medias divide un conjunto de N muestras X dentro K clústeres disjuntos C, cada uno descrito por la media μj de las muestras en el conglomerado. Los medios se denominan comúnmente "centroides" del grupo; tenga en cuenta que no son, en general, puntos de X, aunque viven en el mismo espacio.

El algoritmo de K-medias tiene como objetivo elegir centroides que minimicen la inercia , o el criterio de suma de cuadrados dentro del clúster :

$$\sum_{i=0}^{n} \min_{\mu_j \in C} (||x_i - \mu_j||^2)$$

La inercia se puede reconocer como una medida de cuán coherentes son los clústeres internamente.

En términos básicos, el algoritmo tiene tres pasos. El primer paso elige los centroides iniciales, siendo el método más básico elegir k muestras del conjunto de datos X. Después de la inicialización, K-means consiste en recorrer los otros dos pasos. El primer paso asigna cada muestra a su centroide más cercano. El segundo paso crea nuevos centroides tomando el valor medio de todas las muestras asignadas a cada centroide anterior. Se calcula la diferencia entre el antiguo y el nuevo centroide y el algoritmo repite estos dos últimos pasos hasta que este valor es menor que un umbral. En otras palabras, se repite hasta que los centroides no se mueven significativamente.

Método del codo

El método del codo ayuda a elegir el valor óptimo de 'k' (número de grupos) ajustando el modelo con un rango de valores de 'k'.

Este método utiliza los valores de la inercia obtenidos tras aplicar el K-means a diferente número de Clusters (desde 1 a N Clusters), siendo la inercia la suma de las distancias al cuadrado de cada objeto del Cluster a su centroide:

$$Inercia = \sum_{i=0}^{N} \|x_i - \mu\|^2$$

Una vez obtenidos los valores de la inercia tras aplicar el K-means de 1 a N Clusters, representamos en una gráfica lineal la inercia respecto del número de Clusters. En esta gráfica se debería de apreciar un cambio brusco en la evolución de la inercia, teniendo la línea representada una forma similar a la de un brazo y su codo. El punto en el que se observa ese cambio brusco en la inercia nos dirá el número óptimo de Clusters a seleccionar para ese dataset; o, dicho de otra manera: el punto que representaría al codo del brazo será el número óptimo de Clusters para ese data set.

Métodos

fit(X [, y, sample_weight])	Calcula la agrupación en clústeres de k-medias.
fit_predict(X [, y, sample_weight])	Calcula los centros de los conglomerados y prediga el índice del conglomerado para cada muestra.
fit_transform(X [, y, sample_weight])	Calcula la agrupación en clústeres y transforme X en un espacio de distancia de clústeres.
get_params([deep])	Obtenga parámetros para este estimador.
predict(X [,sample_weight])	Predecir el conglomerado más cercano al que pertenece cada muestra de X.
score(X [, y, sample_weight])	Lo contrario del valor de X en el objetivo de K-medias.
set_params(** parámetros)	Establece los parámetros de este estimador.
transform(X)	Transforma X en un espacio de distancia de grupo.