
Universidad
Tecmilenio®

Universidad Tecmilenio

Aprendizaje Automático No Supervisado

El algoritmo k-medias aplicado a la clasificación

Introducción

En el mundo real existen diversas situaciones en donde los datos de un problema no están etiquetados, por lo que en esos casos se deben aplicar métodos de aprendizaje no supervisados.

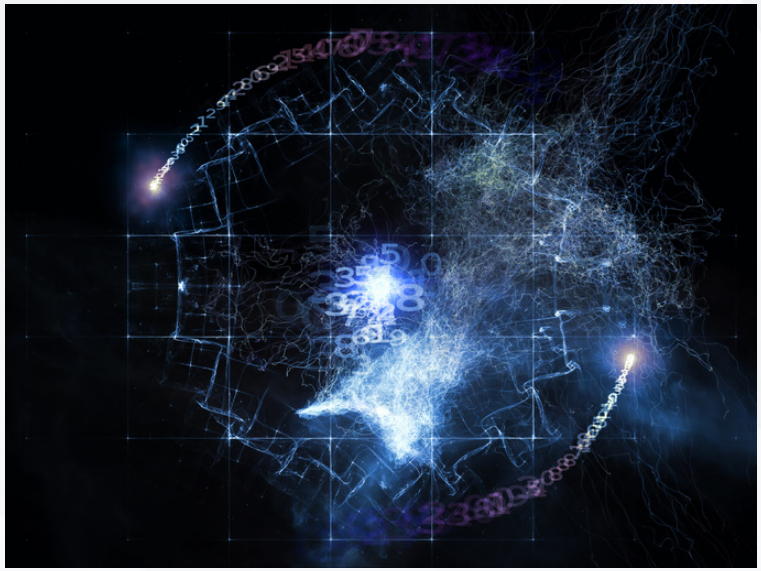
Debido a que no puede obtenerse ninguna forma de retroalimentación supervisada sobre el procesamiento que ocurre, así como por la ausencia de datos etiquetados o categorizados, los métodos supervisados no son capaces de brindar una solución a este tipo de problemas. Otra ventaja de los métodos no supervisados es su costo de etiquetado.

La agrupación en clústeres es uno de los métodos más antiguos en el repertorio del aprendizaje no supervisado, por lo que existen numerosos tipos de este algoritmo en la literatura. En este tema se presentará una de sus formas más simples, pero bastante efectiva y usada para abordar una amplia gama de problemas, a la cual se le conoce como agrupación (clustering) en k-medias. La variable *K* denota el número de conglomerados y el usuario determina el valor de dicha variable antes de comenzar a aplicar el algoritmo.

Explicación

La agrupación (clustering) consiste esencialmente en agregar las muestras en forma de grupos. Los criterios utilizados para decidir la pertenencia a un grupo se determinan mediante el uso de alguna forma de métrica o distancia.

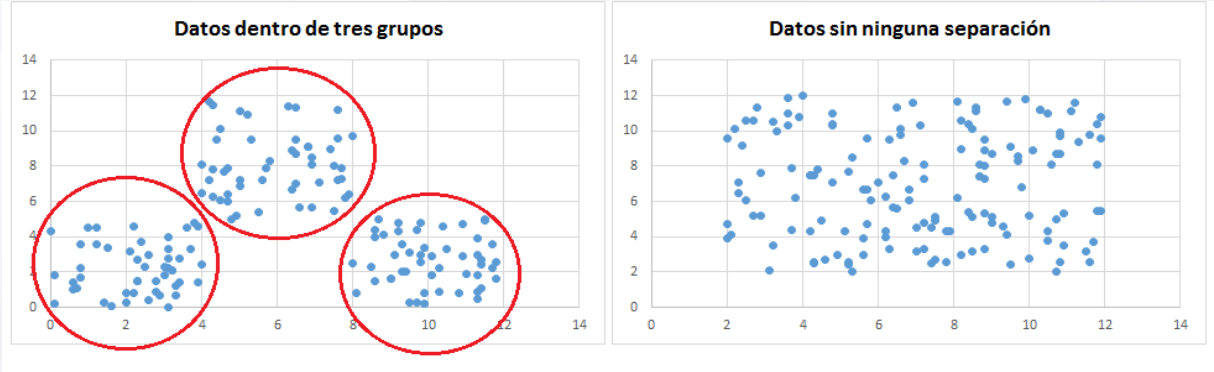
Ahora se presentarán y explicarán los fundamentos del algoritmo más usado bajo este enfoque, es decir, la agrupación de k-medias, por lo que primero se verá de forma general y después en una variante: la agrupación jerárquica de k-medias.



Explicación

En la agrupación de k-medias se especifica el número de grupos (clústeres) *k* deseados, por lo que a su vez el algoritmo asigna exactamente cada observación a cada uno de ellos. Se optimizan los grupos minimizando la variación tal del agrupamiento (también conocida como inercia), de modo que la suma de las variaciones sea lo más pequeña posible.

Por lo general, el algoritmo de k-medias realiza varias ejecuciones y elige la que tiene la mejor separación, que es definida como la suma total más baja de variaciones dentro del clúster para todos los *k* clústeres.



La figura anterior muestra dos ejemplos de casos extremos posibles en el proceso de agrupamiento.

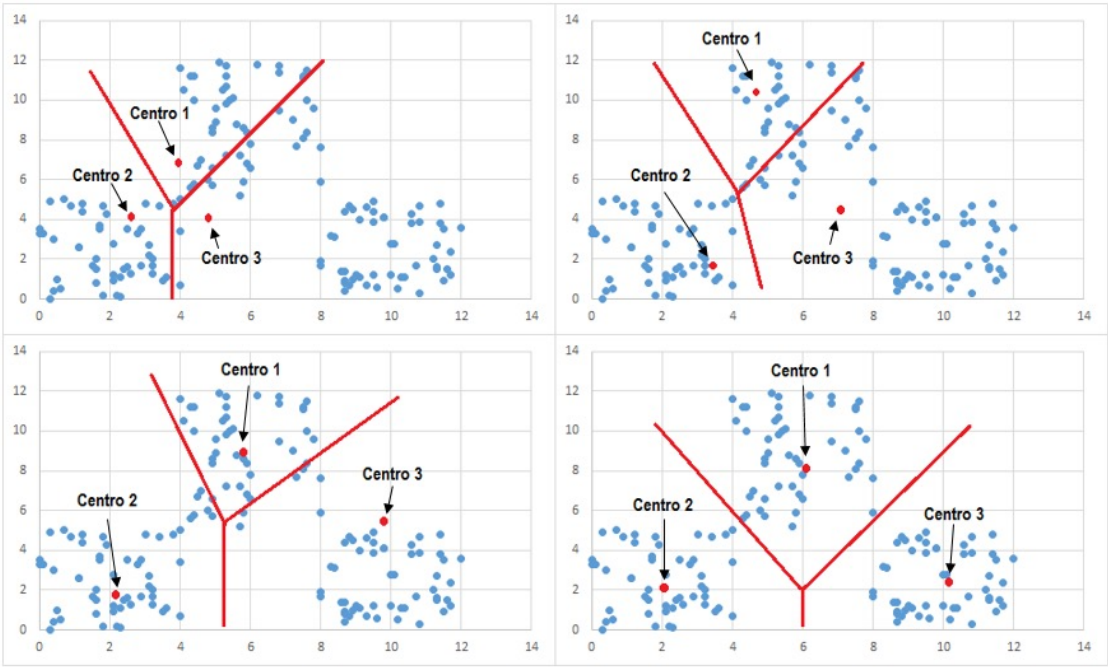
Explicación

Algunas de **las mejoras u optimizaciones** que se aplican comúnmente al algoritmo son las siguientes:

- Restringir el número de iteraciones a un número máximo de estas.
- Encontrar el número de muestras en cada grupo. Sin embargo, si es menor que un cierto umbral habrá que eliminar ese grupo y repetir el proceso.
- Encontrar la distancia intragrupo vs. la distancia entre grupos. No obstante, si dos están demasiado cerca entre sí (en relación con otros) habrá que combinarlos y repetir el proceso.
- Si algunos conglomerados crecen demasiado habrá que aplicar un umbral de número máximo de muestras en un conglomerado y dividirlo en dos o más, repitiendo el proceso.

Explicación

La siguiente figura muestra el algoritmo en etapas intermedias a medida que converge a los grupos deseados (este es un caso ideal). En la mayoría de las situaciones prácticas, donde los conglomerados no están bien separados o el número de conglomerados naturales es diferente al valor inicializado de *k*, es posible que el algoritmo no converja.



Explicación

Entonces, el algoritmo de agrupación en clústeres de k-medias se puede resumir de la siguiente manera:

1. Empezar con un valor predeterminado de *k*, que es el número de conglomerados a encontrar en los datos dados.
2. Inicializar aleatoriamente los *k* centros de conglomerados, así como las *k* muestras en los datos de entrenamiento, de modo que no haya duplicados.
3. Asignar cada una de las muestras de entrenamiento a uno de los *k* centros de clúster según la métrica de distancia elegida.
4. Una vez creadas las clases, actualizar los centros de cada clase como media de todas las muestras de esa clase.
5. Repetir los pasos 2 al 4 hasta que no haya cambios en los centros de los grupos.

Actividad

Después de haber estudiado el tema puedes abordar las siguientes cuestiones:

- ¿Qué otro tipo de aplicaciones podrías darle al algoritmo de agrupación en k-medias?
- Busca y estudia un ejemplo de aplicación que trabaje aplicando el algoritmo de agrupamiento por k-medias.

Cierre

En este tema se abordó una de las técnicas principales básicas en el aprendizaje no supervisado. Esto se debe a que la manera en que cada clúster o grupo finaliza es aleatoria hasta cierto punto, lo cual es útil para formar una idea de lo mencionado al inicio del tema sobre los principios del aprendizaje sin supervisión.

Por otra parte, la agrupación en clústeres tiene muchas aplicaciones, por ejemplo, en la detección de fraudes con tarjetas de crédito, ya que algunos algoritmos son capaces de agrupar las transacciones fraudulentas, separándolas de las transacciones normales, asimismo, si solo se tuvieran unas pocas etiquetas para las observaciones en el conjunto de datos, podría usarse *clustering* para agrupar las observaciones primero (sin etiquetas) para después transferir la información de las pocas observaciones etiquetadas al resto de las mismas dentro del grupo. Esta es una forma de aprendizaje por transferencia, lo cual es un campo en rápido crecimiento en el aprendizaje automático.

Finalmente, su impacto también podría apreciarse en áreas como, por ejemplo, compras minoristas y en línea, marketing, redes sociales, sistemas de recomendación de películas, música, libros, citas, etc., los algoritmos de agrupación pueden acomodar a personas similares en función de su comportamiento. Una vez establecidos estos grupos, los usuarios comerciales tendrán una mejor percepción de su base de usuarios, por lo que podrán diseñar estrategias comerciales específicas para cada uno de los distintos grupos.

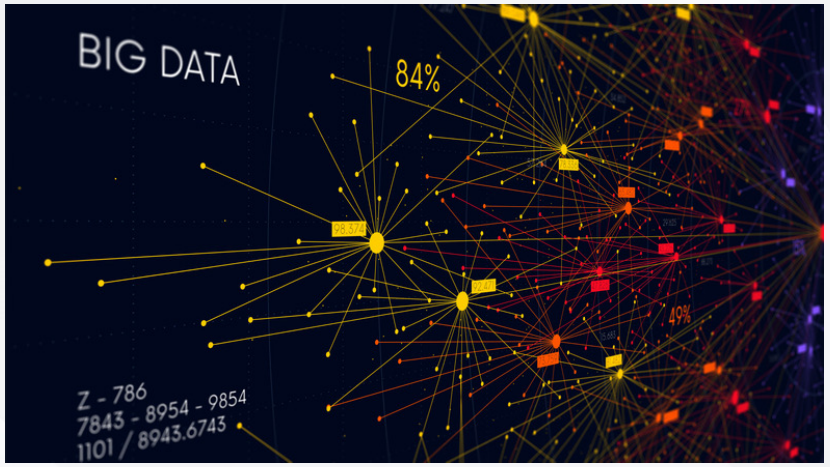
Universidad Tecmilenio

Aprendizaje Automático No Supervisado

Técnicas mejoradas de agrupamiento

Introducción

De forma general, el agrupamiento es una herramienta muy utilizada para identificar segmentos o patrones de características en conjuntos de datos variados. Según las propiedades de los datos y las necesidades del problema a resolver, se puede optar por una de las principales implementaciones: el clusterizado rígido (fuerte) o el suave (débil).



En la agrupación rígida, cada muestra de los datos se agrupa en un clúster de manera estricta. Una observación puede pertenecer completamente a un clúster o no pertenecer en lo absoluto. El algoritmo de agrupamiento k-means es un ejemplo clásico de los pertenecientes a esta familia. Por otro lado, en el caso de la agrupación suave, en lugar de poner cada muestra en clústeres separados, se le asigna una probabilidad de pertenencia ese grupo, por lo cual en el agrupamiento suave o difuso cada observación puede pertenecer a uno o varios clústeres en función de una puntuación de probabilidad o verosimilitud. El ejemplo más significativo de este tipo de agrupamiento es el algoritmo en clúster de c-means difuso (FCM).

En el desarrollo de este tema se verán algunas técnicas avanzadas que se utilizan para incorporarle mejoras al método de agrupamiento k-means, las cuales están relacionadas con las dos formas de clusterizado que se mencionaron anteriormente.

Explicación

El agrupamiento jerárquico es una variante de los métodos de clusterizado fuerte que no necesita de la definición previa del número de clústeres. Según la estrategia que se utilice para conformar los grupos, se pueden clasificar en dos tipos :

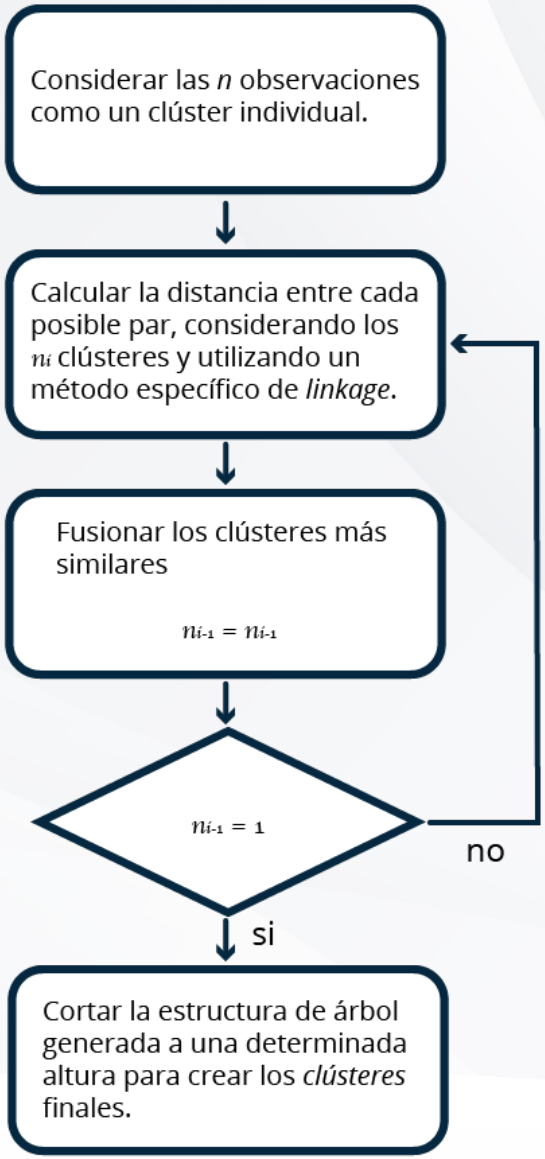
- **Agrupamiento aglomerativo (*agglomerative clustering*)**: en este tipo de agrupamiento se comienza separando todas las observaciones, considerándolas como un clúster individual. De manera progresiva se van creando combinaciones de estos clústeres primarios, haciendo crecer la estructura hasta que converja en una sola entidad.

- **Agrupamiento divisivo (*divisive clustering*):** este tipo de agrupamiento realiza la operación inversa al aglomerativo, ya que comienza agrupando todas las observaciones en un clúster de orden superior, para posteriormente realizar divisiones de esta hasta que cada muestra queda independiente del resto.

Explicación

Para comprender el algoritmo que se aplica en el **agrupamiento aglomerativo**, se puede utilizar el diagrama de la figura.

Es importante destacar que, para la adecuada realización de este proceso, es necesario establecer un criterio para cuantificar la semejanza entre los dos grupos. Por lo general, para realizar esta operación se utiliza un parámetro denominado ***linkage***.



Explicación

El **linkage** extiende el concepto de distancia entre observaciones individuales para que pueda ser aplicable a las parejas de grupos integradas por varias observaciones.

Los métodos de linkage, completo, promedio y Ward, suelen ser los más utilizados por los analistas de datos, debido a que generan dendrogramas más compensados.

En un caso real, la decisión final sobre cuál de estos métodos se debe utilizar dependerá en gran medida de la problemática que se esté intentando resolver y de la calidad de los datos con los que se esté trabajando.

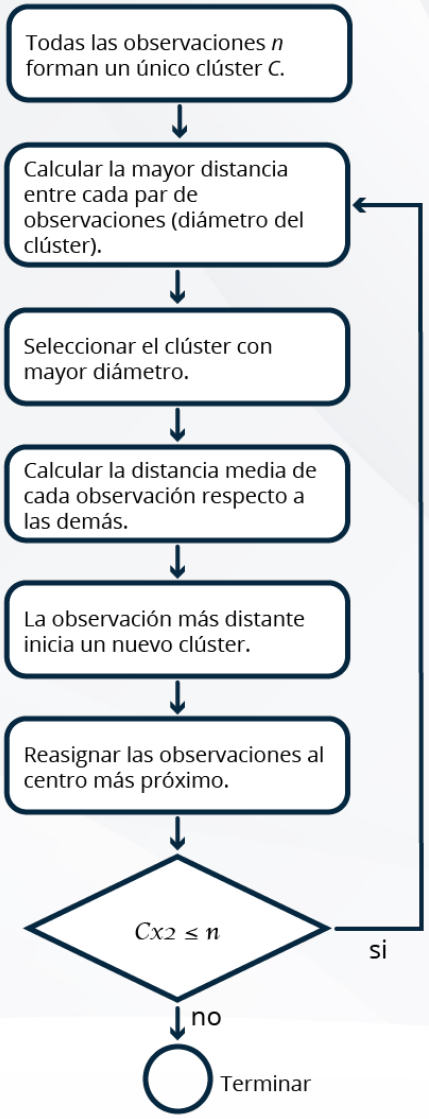
Tipo de linkage	Descripción
Completo *(Complete or Maximum)*	Calcula la distancia entre todos los posibles pares formados por una observación del primer clúster y una del segundo. La mayor de todas se selecciona como la distancia válida entre los dos clústeres. Se considera uno de los tipos de linkage más conservadores.
Sencillo *(Single or Minimum)*	Calcula la distancia entre todos los posibles pares formados por una observación del primer clúster y una del segundo. La menor de todas se selecciona como la distancia válida entre los dos clústeres. Se considera uno de los tipos de linkage más conservadores.
Promedio *(Average)*	Calcula la distancia entre todos los posibles pares formados por una observación del primer clúster y una del segundo. El valor promedio de todas se selecciona como la distancia válida entre los dos clústeres.
Centroide *(Centroid)*	Se calcula el centroide de cada uno de los clústeres y selecciona la distancia entre ellos como la distancia valida entre los dos grupos.
Método Ward *(Ward)*	Es un método general donde la selección del par a fusionar en cada iteración del algoritmo se basa en el valor óptimo de una función objetivo definida por el experto. Un ejemplo de función es el método *Ward's minimum variance*, cuyo objetivo es minimizar la suma total de la varianza entre los clústeres. Esta es la misma métrica que se minimiza en el algoritmo k-means tradicional.

Explicación

Por otra parte, el algoritmo más conocido dentro de la familia de **agrupamiento jerárquico divisivo** es el llamado DIANA (*Divisive Analysis Clustering*). Su proceso se ilustra en el diagrama de flujo.

En este caso, se parte de un único clúster que contiene todas las observaciones, por lo que posteriormente se van realizando divisiones hasta que cada observación queda de forma independiente. Por ende, las observaciones se reasignan en función del valor de referencia más cercano, dando lugar a la división del clúster original en dos nuevos grupos.

A diferencia del agrupamiento aglomerativo, en donde hay que elegir un tipo de distancia y un método de *linkage*, en este tipo de agrupamiento solo hay que elegir la primera de estas métricas.



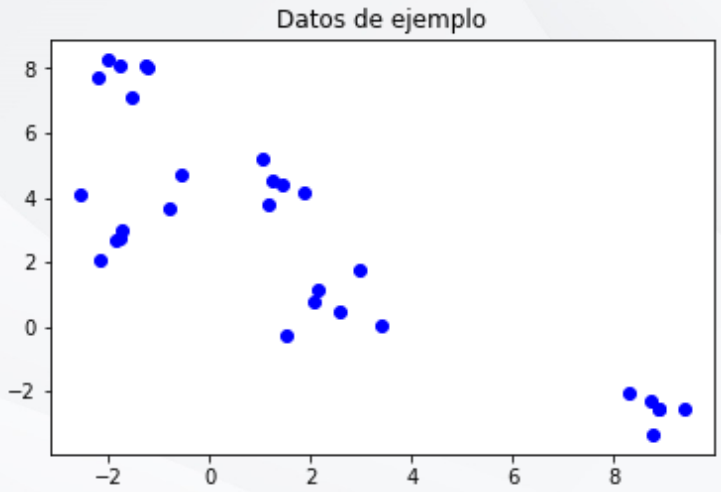
Explicación

A continuación, se verá cómo se realiza el agrupamiento jerárquico utilizando el lenguaje de programación Python y varias librerías utilitarias que facilitan este proceso.

Supongamos que se cuenta con un conjunto de datos como el que se muestra en la figura.

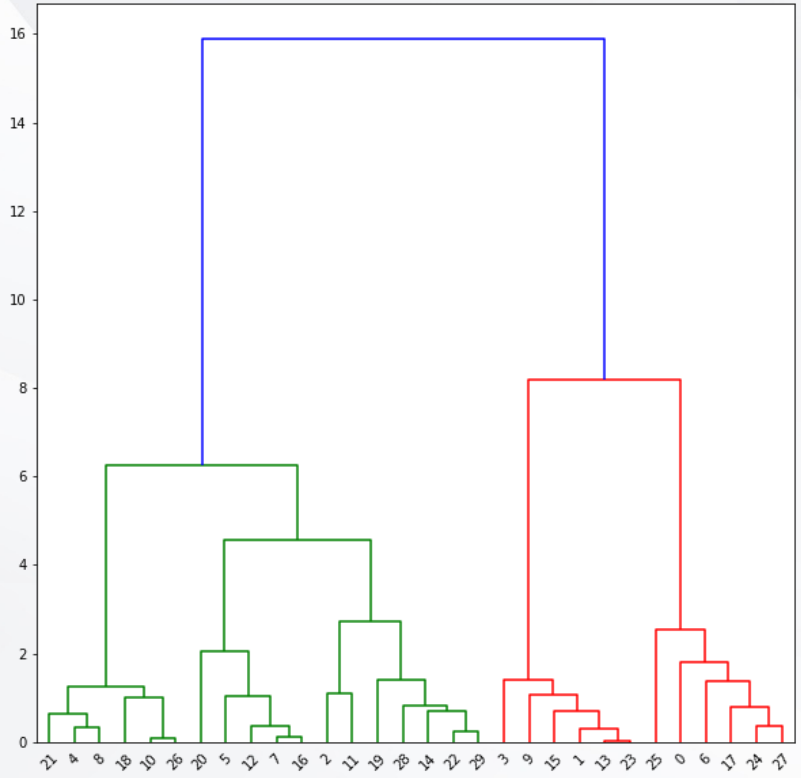
Para implementar el agrupamiento jerárquico y construir el dendrograma, Python cuenta con un método específico denominado *cluster.hierarchy()*, que se encuentra en la librería especializada Scipy.

El código correspondiente a la construcción del modelo y a la construcción del dendrograma se ha implementado empleando tales herramientas.



Explicación

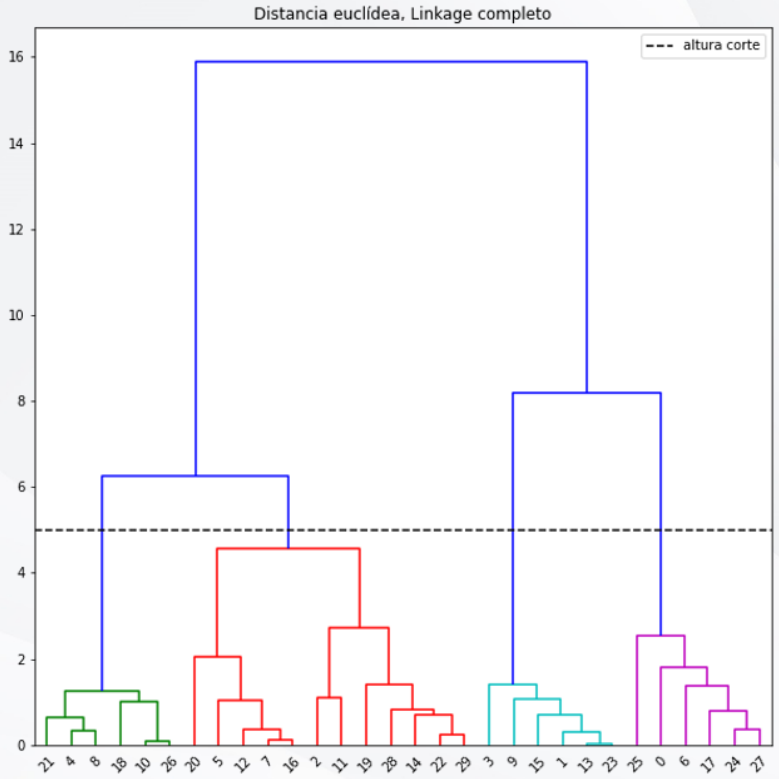
Como se muestra en el código anterior, el linkage seleccionado fue el completo. Por tanto, el dendrograma generado por este programa se muestra en la siguiente figura.



Explicación

Una forma simple que se utiliza para determinar el número adecuado de clústeres es examinar visualmente el dendrograma y realizar un corte horizontal. La cantidad de ramas que se interceptan con la línea de corte corresponden a la cantidad de clústeres en los que se dividen los datos.

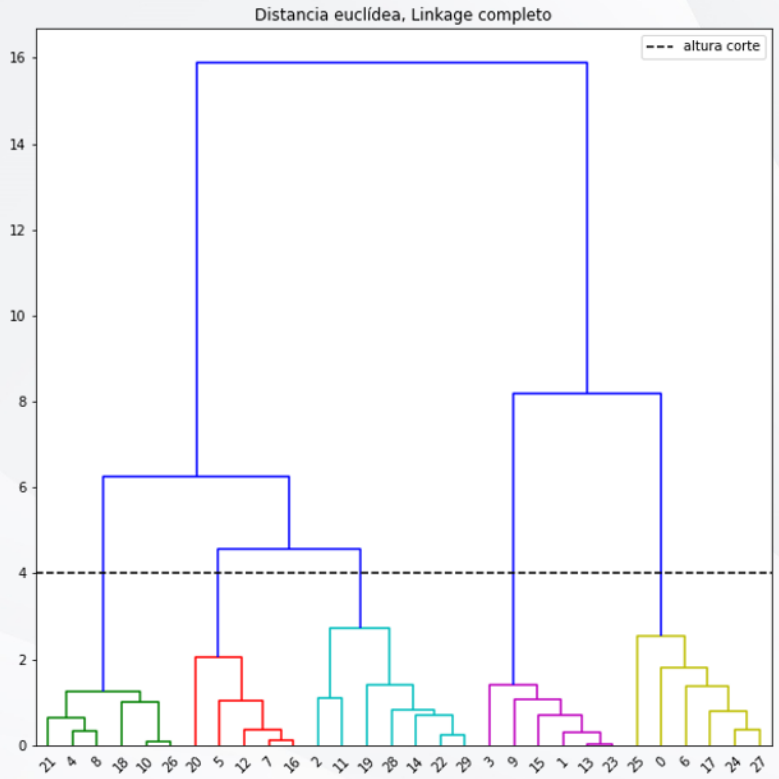
Por ejemplo, en la siguiente figura el corte está a una altura entre 5 y 6, lo cual corresponde a la selección de 4 clústeres.



Explicación

Sin embargo, considerando la altura de corte con un valor ubicado entre las distancias 3 y 4, el resultado obtenido sería 5 clústeres como se muestra en la figura.

La altura de corte controla el número de clústeres obtenidos, lo cual corresponde a la misma función que tiene que establecer el valor de *K* en el algoritmo k-means tradicional.



Explicación

A partir del análisis anterior se pueden establecer dos propiedades adicionales que complementan este criterio de selección:

- Debido a que las ramas tienen una longitud variable, existe un intervalo de altura donde cualquier corte corresponde a la misma cantidad de clústeres.
- Un dendrograma permite seleccionar desde 1 hasta *n* clústeres (donde *n* es la cantidad de observaciones), por lo cual la selección del número óptimo depende de la valoración del observador. Una sugerencia es considerar la relación entre las ramas principales y la altura donde se efectúan las intersecciones.

Explicación

Entre los estudios que se han enfocado en mejorar el algoritmo k-means clásico está la utilización combinada de este con las técnicas de agrupamiento jerárquico. Entre las diferentes opciones que más se han destacado podemos mencionar las siguientes:

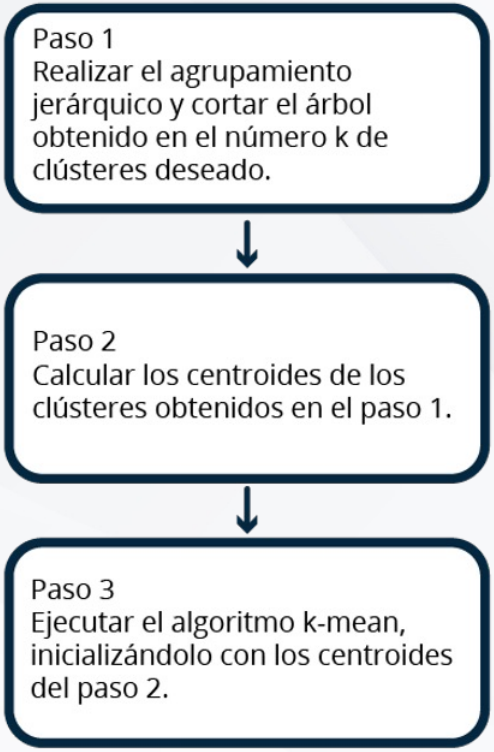
- **Agrupación jerárquica de k-mean:** consiste en utilizar el algoritmo k-means clásico de forma recursiva (ver la figura). Después de cada finalización exitosa de la agrupación se inicializa un nuevo conjunto de agrupaciones aleatorias dentro de cada clúster obtenido, por lo que se repite el mismo algoritmo para encontrar todos los subgrupos.

Explicación

- **Agrupación jerárquica de k-mean híbrida (hkmeans):** una parte de los fundamentos de k-means radica en la selección inicial del número de clústeres, los cuales son ubicados en el espacio definido por los datos de forma aleatoria (STHDA, s.f.).

Esta condición provoca que la solución obtenida sea muy sensible a esta elección inicial, por tanto, los resultados pueden ser relativamente diferentes en cada ejecución del algoritmo.

Para evitar esta situación, se suele utilizar una combinación entre las técnicas de agrupamiento jerárquico y el método clásico de k-means. El procedimiento que describe este tipo de agrupación se muestra en la figura.

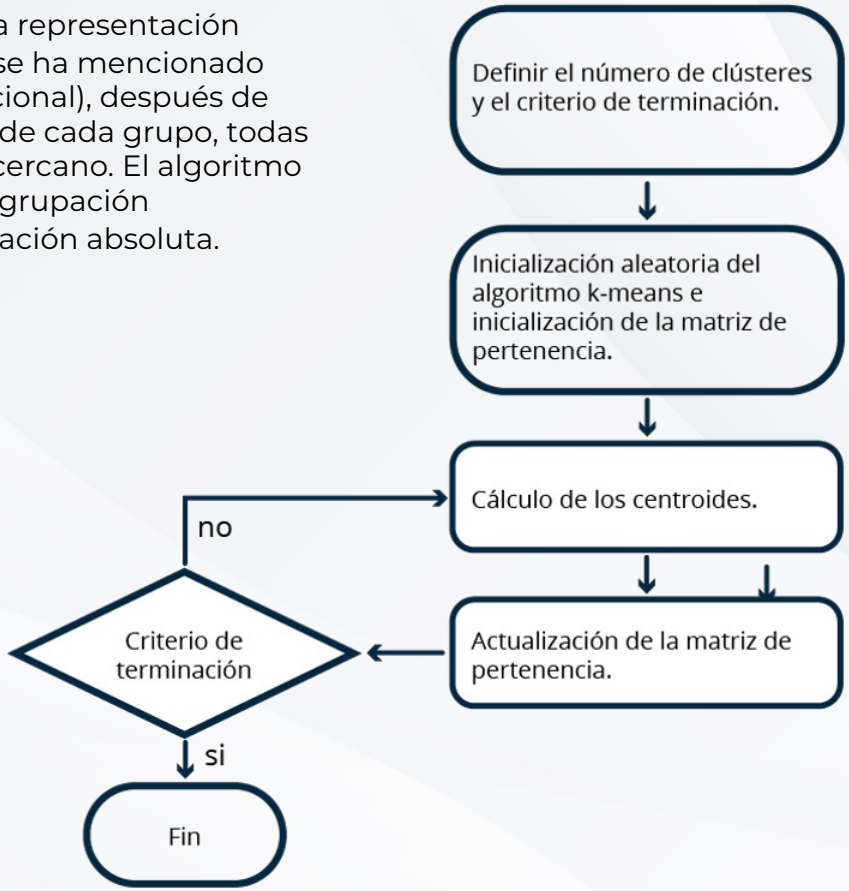


Explicación

El agrupamiento suave también tiene una representación relacionada al algoritmo k-means. Como se ha mencionado anteriormente (en el agrupamiento tradicional), después de que se ubican o actualizan los centroides de cada grupo, todas las muestras se juntan en su clúster más cercano. El algoritmo de **k-means difuso** sugiere un uso de la agrupación probabilística en lugar de utilizar la agrupación absoluta.

En este caso, cada muestra tiene una probabilidad distinta de cero de pertenecer a varios clústeres al mismo tiempo, por tanto, cuanto más cerca esté la observación del centro del grupo, mayor será la probabilidad de pertenecer a este. Entre los representantes más significativos de este tipo de propuesta se encuentra el algoritmo de agrupación en clúster c-means difuso (FCM).

Este diagrama se puede utilizar para comprender el funcionamiento del algoritmo FCM.



Actividad

Piensa en qué tipo de problemas podrías resolver aplicando las versiones mejoradas del algoritmo k-medias.

¿Consideras que el método del dendrograma es efectivo para elegir el número óptimo de clústeres? ¿Por qué? ¿Conoces algún otro método para tomar tal decisión?

¿En qué casos consideras necesario aplicar el agrupamiento divisivo y en cuáles el aglomerativo?

Cierre

En este tema diste un recorrido por las diversas técnicas de agrupamiento, por tanto, aprendiste que las muestras pueden estar claramente separadas o solapadas entre sí. Ante estas problemáticas se puede optar por la agrupación rígida, pero esta decisión garantiza que cada muestra quede asignada a un clúster específico, lo cual puede ser adecuado, pero no totalmente óptimo. Por el contrario, si se utiliza el agrupamiento suave se obtendrá un resultado más flexible y determinado, esto a partir de la probabilidad de pertenencia de la observación a un grupo determinado.

Por su parte, también existe otra forma de implementar el agrupamiento mediante el método jerárquico, el cual puede ser aglomerativo o divisivo. En ambos casos el proceso de separación se representa mediante una estructura en forma de árbol conocida como dendrograma, la cual es una herramienta muy útil para determinar de forma visual el valor adecuado de clústeres en donde dividiremos los datos.

En algunas ocasiones es necesario considerar que una observación puede pertenecer a uno o más grupos. No obstante, si el objetivo es, por ejemplo, construir un sistema de recomendación de productos, considerar una probabilidad de pertenencia no es una idea tan descabellada, por lo que una adecuada implementación del algoritmo k-means difuso pudiera ser la opción más adecuada.

Tecmilenio no guarda relación alguna con las marcas mencionadas como ejemplo. Las marcas son propiedad de sus titulares conforme a la legislación aplicable, estas se utilizan con fines académicos y didácticos, por lo que no existen fines de lucro, relación publicitaria o de patrocinio.
