**Aglomerative Clustering**

**INTRODUCCIÓN**

Tema: Aglomerative Clustering

Objetivos:

* Definir el clustering jerárquico aglomerativo
* Explicar los diferentes métodos de aglomeración
* Describir las ventajas y desventajas del clustering jerárquico aglomerativo

Introducción

El clustering jerárquico aglomerativo es un método de aprendizaje automático no supervisado que se utiliza para agrupar datos en función de su similitud. El algoritmo comienza con cada dato en su propio grupo y luego va fusionando grupos de forma sucesiva hasta que todos los datos se encuentran en un solo grupo.

Definición

El clustering jerárquico aglomerativo es un método de clustering que se basa en la idea de construir una jerarquía de grupos. El algoritmo comienza con cada dato en su propio grupo y luego va fusionando grupos de forma sucesiva hasta que todos los datos se encuentran en un solo grupo.

* El análisis de agrupación jerárquico es un método de aprendizaje automático no supervisado, lo que significa que no se proporciona información sobre los grupos correctos.
* El análisis de agrupación jerárquico es un método flexible que puede utilizarse para agrupar datos de cualquier tipo.
* El análisis de agrupación jerárquico puede ser utilizado para una variedad de propósitos, como la segmentación de clientes, el análisis de datos clínicos y la detección de fraude.

**DIAPO 2**

* **El aglomerative clustering es un método iterativo.** Esto significa que el algoritmo comienza con cada dato en su propio cluster y luego va fusionando clusters de forma sucesiva hasta que solo queda un cluster.
* **El aglomerative clustering utiliza una medida de similitud o distancia para determinar la proximidad entre los clusters.**La medida de similitud o distancia se utiliza para calcular la distancia entre dos clusters.
* **La elección de la medida de similitud o distancia puede afectar a los resultados del clustering.** Es importante elegir una medida de similitud o distancia que sea adecuada para el conjunto de datos y la aplicación.

Ejemplos de uso

* **Segmentación de clientes:** El aglomerative clustering se puede utilizar para agrupar a los clientes en función de sus características demográficas, sus preferencias de compra o su comportamiento.
* **Análisis de datos clínicos:** El aglomerative clustering se puede utilizar para agrupar a los pacientes en función de sus síntomas, su historial médico o su respuesta al tratamiento.
* **Detección de fraude:** El aglomerative clustering se puede utilizar para identificar transacciones fraudulentas.

**DIAPO 3**

1. Construye una jerarquía de clusters

El análisis de agrupación jerárquico construye una jerarquía de clusters, es decir, una estructura de tipo árbol basada en la jerarquía. La jerarquía comienza con cada punto de datos en su propio cluster. Luego, los clusters se fusionan de forma sucesiva hasta que solo queda un cluster.

Explicación concisa pero expresiva:

El análisis de agrupación jerárquico crea una estructura de árbol que representa la relación entre los clusters. Los clusters se fusionan de forma sucesiva hasta que solo queda uno.

2. Agrupa puntos de datos cercanos

El análisis de agrupación jerárquico agrupa los puntos de datos cercanos entre sí en función de la medida de similitud o distancia. La medida de similitud o distancia se utiliza para determinar la proximidad entre los puntos de datos.

Explicación concisa pero expresiva:

Los puntos de datos que son similares se agrupan juntos. La similitud se puede medir de diferentes maneras, como la distancia euclidiana o la similitud coseno.

3. Se supone que los puntos de datos cercanos son más similares

El análisis de agrupación jerárquico se basa en la suposición de que los puntos de datos cercanos son más similares. Esta suposición se basa en la idea de que los puntos de datos que están cerca en el espacio de datos también son similares o están relacionados.

Explicación concisa pero expresiva:

Los puntos de datos que están cerca en el espacio de datos también son similares. Esta suposición es la base del análisis de agrupación jerárquico.

**DIAPO 4**

Aglomerative Clustering y Divisive Clustering son dos métodos de clustering jerárquico que se basan en la idea de construir una jerarquía de clusters. Sin embargo, los dos métodos tienen enfoques diferentes para construir la jerarquía.

Aglomerative Clustering comienza con cada dato en su propio cluster y luego va fusionando grupos de forma sucesiva. Divisive Clustering comienza con todos los datos en un solo cluster y luego va dividiendo los clusters en subclusters.

En general, Aglomerative Clustering es un método más flexible que Divisive Clustering. Aglomerative Clustering puede utilizarse para agrupar datos de cualquier tipo, mientras que Divisive Clustering puede ser más adecuado para datos que se pueden dividir de forma natural en subclusters.

Por otro lado, Divisive Clustering puede ser más eficiente que Aglomerative Clustering para grandes conjuntos de datos. Divisive Clustering comienza con un solo cluster, lo que puede reducir el número de operaciones que deben realizarse.

**Aglomerative Clustering**

* **Principio:** comienza con cada dato en su propio cluster y luego va fusionando grupos de forma sucesiva hasta que todos los datos se encuentran en un solo grupo.
* **Ventajas:**
  + Es un método flexible que puede utilizarse para agrupar datos de cualquier tipo.
  + Es fácil de entender y de implementar.
* **Desventajas:**
  + Puede ser computacionalmente costoso para grandes conjuntos de datos.
  + Puede ser difícil determinar el número óptimo de clusters.

**Divisive Clustering**

* **Principio:** comienza con todos los datos en un solo cluster y luego va dividiendo los clusters en subclusters hasta que cada dato se encuentra en su propio cluster.
* **Ventajas:**
  + Puede ser más eficiente que el clustering aglomerativo para grandes conjuntos de datos.
  + Puede ser más fácil determinar el número óptimo de clusters.
* **Desventajas:**
  + Puede ser menos flexible que el clustering aglomerativo.
  + Puede ser más difícil de implementar.

**DIAPO 5**

* ¿Qué es un dendograma?
* ¿Cuáles son los componentes de un dendograma?
* ¿Cómo se interpreta un dendograma?

Espero que esto te ayude. ¡Buena suerte con tu presentación!

Aquí hay algunos comentarios adicionales que puedes incluir en tu presentación:

* **El dendograma se puede utilizar para visualizar la jerarquía de clusters.**
* **El dendograma se puede utilizar para determinar el número óptimo de clusters.**
* **El dendograma se puede utilizar para comparar diferentes métodos de clustering.**

También puedes mencionar que el dendograma se utiliza en una variedad de aplicaciones, como la segmentación de clientes, el análisis de datos clínicos y la detección de fraude.

Un dendograma se puede crear utilizando un algoritmo de clustering jerárquico. Los algoritmos de clustering jerárquico son un tipo de algoritmo de aprendizaje automático no supervisado que se utilizan para agrupar datos en función de su similitud.

* **Libro:** "Clustering: Theory, Applications, and Software" por Kaufman, Rousseeuw, y Rousseeuw.
* **Artículo de Wikipedia:** "Dendrogram"

**DIAPO 6**

El aglomerative clustering funciona de la siguiente manera:

1. **Inicialización:** Cada dato se coloca en su propio cluster.
2. **Iteración:** Se fusionan los clusters más similares.
3. **Terminación:** El proceso se detiene cuando solo queda un cluster.

En cada iteración, se calcula la similitud entre todos los clusters. Los clusters más similares se fusionan para formar un nuevo cluster. Este proceso continúa hasta que solo queda un cluster.

**Medidas de similitud**

Hay muchas medidas de similitud que se pueden utilizar en el aglomerative clustering. Algunas de las medidas de similitud más comunes incluyen:

* **Distancia euclidiana:** La distancia euclidiana es la distancia entre dos puntos en un espacio euclidiano.
* **Distancia Manhattan:** La distancia Manhattan es la suma de las distancias entre dos puntos en un espacio euclidiano.
* **Distancia coseno:** La distancia coseno es una medida de similitud que se basa en el ángulo entre dos vectores.

**Métodos de fusión**

Hay dos métodos principales de fusión que se utilizan en el aglomerative clustering:

* **Aglomeración de enlace completo:** Este método fusiona los dos clusters más cercanos.
* **Aglomeración de enlace simple:** Este método fusiona los dos clusters que tienen el punto de datos más cercano en común.

**DIAPO 7**

**entajas:**

* **Versátil y aplicable a una amplia gama de problemas:** El aglomerative clustering se puede utilizar para una amplia gama de aplicaciones, como la segmentación de clientes, el análisis de datos clínicos y la detección de fraude.
* **Robusto y puede manejar datos ruidosos y no lineales:** El aglomerative clustering puede manejar datos ruidosos y no lineales, lo que significa que puede funcionar bien incluso cuando los datos no cumplen con las suposiciones de otros métodos de clustering.
* **Outliers pueden ser absorbidos por clusters más grandes:** Los outliers, que son datos que se encuentran muy alejados de la mayoría de los datos, pueden ser absorbidos por clusters más grandes, lo que puede ayudar a mejorar la precisión del clustering.

**Desventajas:**

* **Computacionalmente costoso para grandes conjuntos de datos:** El aglomerative clustering puede ser computacionalmente costoso para grandes conjuntos de datos, lo que significa que puede tardar mucho tiempo en ejecutarse.
* **Puede ser difícil determinar el número óptimo de clusters:** El número óptimo de clusters es un problema abierto en el aprendizaje automático, lo que significa que no hay una respuesta única que funcione para todos los conjuntos de datos.
* Existen programas especializados para el seguimiento del tiempo y la gestión de empleados. Estos programas suelen ser más eficientes y fáciles de usar que el aglomerative clustering para estas tareas específicas.

**DIAPO 8**

Aplicaciones.  
**Segmentación de clientes:** El aglomerative clustering se puede utilizar para segmentar clientes en grupos con características similares. Esto puede ayudar a las empresas a desarrollar estrategias de marketing más efectivas. Por ejemplo, una empresa de telecomunicaciones podría utilizar el aglomerative clustering para segmentar a sus clientes en función de su historial de uso, su ubicación y sus preferencias. Esto le permitiría a la empresa desarrollar ofertas y mensajes de marketing personalizados para cada segmento.

**Clasificación:** El aglomerative clustering se puede utilizar para clasificar datos en grupos. Esto puede ayudar a las empresas a identificar patrones en los datos. Por ejemplo, un hospital podría utilizar el aglomerative clustering para clasificar a los pacientes en función de su historial médico. Esto le permitiría al hospital identificar grupos de pacientes con necesidades similares, lo que podría ayudar a mejorar la atención al paciente.

**Minería de datos:** El aglomerative clustering se puede utilizar para descubrir patrones ocultos en los datos. Esto puede ayudar a las empresas a tomar mejores decisiones. Por ejemplo, una empresa de retail podría utilizar el aglomerative clustering para descubrir patrones en las compras de sus clientes. Esto le permitiría a la empresa identificar tendencias de consumo y desarrollar nuevos productos y servicios que satisfagan las necesidades de sus clientes.

Aquí hay algunos ejemplos específicos de cómo se ha utilizado el aglomerative clustering en la práctica:

* **Netflix utiliza el aglomerative clustering para recomendar películas y programas de televisión a sus usuarios.**
* **Amazon utiliza el aglomerative clustering para agrupar productos similares.**
* **Google utiliza el aglomerative clustering para clasificar imágenes en línea.**

El aglomerative clustering es un método poderoso que se puede utilizar para una amplia gama de aplicaciones. Es importante comprender los principios básicos del aglomerative clustering para poder utilizarlo de forma eficaz.

**DIAPO 9**

**Ward linkage**

Ward linkage minimiza la varianza dentro de los clusters. Esto significa que los clusters se fusionan de manera que se minimicen las diferencias entre los datos dentro de cada cluster.

En el mapa de puntos de Ward, los clusters se fusionan de manera que se minimiza la distancia euclidiana entre los centroides de los clusters. Esto se puede ver en el hecho de que los clusters tienden a ser más compactos en el mapa de puntos de Ward que en los mapas de puntos de los otros enlaces.

**Average linkage**

Average linkage minimiza la distancia media entre los puntos de datos de dos clusters diferentes. Esto significa que los clusters se fusionan de manera que se minimicen las diferencias entre los datos de los dos clusters.

En el mapa de puntos de average linkage, los clusters se fusionan de manera que se minimice la distancia media entre los puntos de datos de los dos clusters. Esto se puede ver en el hecho de que los clusters tienden a ser más irregulares en el mapa de puntos de average linkage que en los mapas de puntos de los otros enlaces.

**Complete linkage**

Complete linkage minimiza la distancia máxima entre los puntos de datos de dos clusters diferentes. Esto significa que los clusters se fusionan de manera que se maximicen las diferencias entre los datos de los dos clusters.

En el mapa de puntos de complete linkage, los clusters se fusionan de manera que se minimice la distancia máxima entre los puntos de datos de los dos clusters. Esto se puede ver en el hecho de que los clusters tienden a ser más dispersos en el mapa de puntos de complete linkage que en los mapas de puntos de los otros enlaces.

**Single linkage**

Single linkage minimiza la distancia mínima entre los puntos de datos de dos clusters diferentes. Esto significa que los clusters se fusionan de manera que se minimicen las diferencias entre los datos más cercanos de los dos clusters.

En el mapa de puntos de single linkage, los clusters se fusionan de manera que se minimice la distancia mínima entre los puntos de datos de los dos clusters. Esto se puede ver en el hecho de que los clusters tienden a ser más alargados en el mapa de puntos de single linkage que en los mapas de puntos de los otros enlaces.

En general, el tipo de enlace que se utilice puede afectar a los resultados del clustering. Ward linkage tiende a producir clusters compactos, mientras que average linkage tiende a producir clusters irregulares. Complete linkage tiende a producir clusters dispersos, mientras que single linkage tiende a producir clusters alargados.

**CLUSTERIZACIÓN DE DATOS**

La clusterización de datos es un proceso de aprendizaje automático no supervisado que se utiliza para agrupar datos en función de su similitud. Los datos se agrupan en grupos, denominados clusters, de modo que los datos dentro de un mismo cluster sean similares entre sí y los datos de clusters diferentes sean diferentes entre sí.

La clusterización de datos se utiliza en una variedad de aplicaciones, como la segmentación de clientes, el análisis de datos clínicos y la detección de fraude.

**Definición breve:**

La clusterización de datos es un proceso de agrupar datos en función de su similitud.

**Ejemplo:**

Podemos imaginar un conjunto de datos de clientes que consta de información sobre la edad, el sexo, los ingresos y el gasto en compras. Podemos utilizar la clusterización de datos para agrupar a los clientes en función de su similitud en términos de estos factores. Por ejemplo, podemos agrupar a los clientes jóvenes y con ingresos elevados en un cluster, a los clientes mayores y con ingresos bajos en otro cluster, y así sucesivamente.

**Ventajas:**

* Puede ayudar a descubrir patrones ocultos en los datos.
* Puede ayudar a identificar grupos de datos con características similares.
* Puede ser utilizado para una variedad de aplicaciones.

**Desventajas:**

* Puede ser difícil determinar el número óptimo de clusters.
* Los resultados de la clusterización pueden ser subjetivos.