¿Cómo pasamos de la Reducción de Dimensiones a una segmentación de clusters jerárquicos?

Diego Figueroa

2025-05-15

Conexión con la Generación de Clusters Jerárquicos: De Reducción de Dimensiones a Segmentación

El PCA, al ser una técnica de reducción de dimensiones, toma un conjunto de variables correlacionadas y las transforma en un número menor de variables no correlacionadas (los componentes principales) que capturan la mayor parte de la varianza en los datos.

¿Cómo pasamos de la Reducción de Dimensiones a una segmentación de clusters jerárquicos?

1- **Aplicar PCA**: Primero, aplicamos PCA a nuestras variables activas para reducir la dimensionalidad del conjunto de datos. En lugar de trabajar con muchas variables potencialmente correlacionadas,ahora tenemos un número menor de componentes principales que explican la mayor parte de la variabilidad.

2- **Usar los Componentes Principales para el Clustering**: Los componentes principales resultantes son combinaciones lineales de las variables originales y representan las dimensiones más importantes de variabilidad en los datos. Estos componentes principales pueden ser utilizados como las nuevas "características" para realizar un clustering jerárquico.

3- **Clustering Jerárquico**: El clustering jerárquico es un método de clustering que construye una jerarquía de clusters. Puede ser:

```
* **Aglomerativo (bottom-up)**: Comienza con cada individuo en su propio cluster y fusiona los clusters más cercanos iterativamente hasta que todos los individuos pertenecen a un solo cluster.

* **Divisivo (top-down)**: Comienza con todos los individuos en un solo cluster y lo divide recursivamente en clusters más pequeños.

Al aplicar el clustering jerárquico a las puntuaciones de los individuos en los componentes principales, estamos agrupando los individuos que son similares en las dimensiones que explican la mayor parte de la varianza en los datos. La reducción de dimensionalidad previa con PCA ayuda a:

* Simplificar el espacio de datos: Trabajar con menos dimensiones (los componentes principales) hace que el cálculo de distancias entre individuos sea más eficiente y menos propenso a la "maldición de la dimensionalidad".

* Enfocarse en la variabilidad importante: Los componentes principales capturan la mayor parte
```

de la varianza, por lo que el clustering se basa en las dimensiones que más distinguen a los individuos.

4- **Visualizar el Dendrograma**: El resultado del clustering jerárquico se visualiza mediante un dendrograma, un diagrama de árbol que muestra la secuencia de fusiones o divisiones de los clusters. Cortar el dendrograma a diferentes niveles nos permite obtener diferentes números de clusters.

5- **Interpretar los Clusters**: Una vez que hemos identificado los clusters, podemos volver a las variables originales (e incluso a las variables suplementarias) para caracterizar y entender las diferencias entre los grupos encontrados. Por ejemplo, si encontramos tres clusters de autos basados en los componentes principales, podemos analizar si estos clusters difieren significativamente en su precio promedio (variable cuantitativa suplementaria) o en la proporción de autos de lujo de cada tipo (variable cualitativa suplementaria).