Machine Learning No Supervisados

Kmeans - Clustering Jerárquico.

Tomás Fontecilla

31 de agosto de 2023

K-means

Como método de conglomerados, el objetivo de K-means y clustering jerárquico es agrupar las observaciones por sus similitudes, provocando que quienes tengan pocas disimilitudes queden en el mismo cluster y quienes tengan mayores en otros.

Una de las técnicas más conocidas de segmentación es K-means. Este algoritmo está diseñado para cuando todas las variables son cuantitativas la distancia euclideana:

$$d(x_i, x_i') = \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 = \|x_i - x_{i'}\|^2$$

se utiliza como medida de la disimilaridad.

K-means

Tu vecindad te define

K-means - Definiciones

K-means es un procedimiento de clustering que resulta de un problema matemático simple e intuitivo. Sea $C_1, C_2, ..., C_k$ conjuntos que contienen índices de las observaciones en cada cluster. Estos deben cumplir las siguientes propiedades:

- $C_1 \cup C_2 \cup ... \cup C_k = \{1, ..., n\}$. En otras palabras, toda observación debe estar en al menos un cluster K.
- $C_1 \cap C_2 \cap ... \cap C_k = 0 \forall k \neq k'$. O sea, los cluster no se pueden superponer, cada observación puede estar sólo en un cluster.

la idea detrás de KMeans es que un buen agrupamiento es uno en que la variación dentro del cluster sea lo más pequeña posible

K-means - Definiciones

Para poder entender de qué va esto necesitaremos definir algunos criterios:

Dispersión interna

$$W(C) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

Donde $|C_k|$ denota el número de observaciones en el K-ésimo cluster.

Asignación de Cluster

$$C^* = \min_{C_i,\ldots,C_k} \sum_{k=1}^{K} W(C_k)$$

Cada x se iría a un cluster más cercano, la optimización queda como

K-means - Algoritmo

Todo bien y tranquilo, pero, ¿Cómo opera el modelo?

Algoritmo

1. Para un cluster C, la varianza total de cluster es minimizada con respecto a $\{m_1, ..., m_k\}$ con respecto a la media del cluster asignado ($\bar{x}_S = \underset{m}{\text{argmin}} \sum_{i \in S} \|x_i - m\|^2$).
2. Dado el conjunto de medias $\{m_1, ..., m_k\}$ C^* es minimizado al asignar las variables *actualmente* más cercanas a la media del cluster, esto es:

$$C(i) = \underset{1 \leq k \leq K}{argmin} \|x_i - m_k\|^2$$

3. pasos 1 y 2 se repiten hasta que las asignaciones no cambian.

K-means - Potenciales problemas

Desventajas

- puede converger a un óptimo local
- Puede converger a diferentes puntos dependiendo de los puntos iniciales.
- El cómputo de los centroides puede no ser robusto ante outliers.
- Aunque el dato esté entre 2 clústeres, el método lo asignará con probabilidad 1 al más cercano. Es más apropiado cuando los atributos son valores continuos pues usa el cuadrado de la distancia euclidiana.

Aún así muchas veces los elementos son soslayables cambiando metodologías y con el avance de la ciencia. Ejemplo, cambiar la distancia euclidiana por manhattan permite mejor aproximación a datos discretos.

K-means - Escoger el número de clusters

Como la idea de K-means es agrupar minimizando la varianza intra grupo y maximizando la varianza entre grupos, podemos usar la el screeplot igual que en componentes principales pero:

- No hace referencia a la varianza explicada, sino que a la varianza intragrupo
- debemos ejecutar varios modelos de kmeans para obtener las distintas varianzas intragrupo de cada uno de ellos
- puede ser costoso determinar de antemano cuántos clusters hacer debido a que tiene que hacer matrices de distancia en cada iteración para cada número de grupos por cada observación.

Clustering Jerárquico

K-medias tiene un componente de extrema importancia que es el número de clusters que va a realizar. La selección del número de clusters es efectivamente uno de los grandes detrimentos del algoritmo, ya que es necesario segmentar antes de saber si sería necesario u óptimo el número.

Los modelos jerárquicos no requieren de esta especificación. En su lugar este método ve las disimilaridades entre grupos disjuntos de observaciones, basado en las disimilaridades en apres entre las observaciones en los dos grupos.

Existen distintas estrategias que se dividen en dos paradigmas: aglomerativo (agrupando por similitud) o divisivos (separando por disimilitud) En resumen, esta es la técnica de clustering de árboles, sea

desde las hojas hasta la raiz o de la raiz a las hojas.

Clustering Jerárquico - Árboles Aglomerativos

El clustering aglomerativo toma las disimilitudes entre observaciones y las va agrupando a través de una función de enlace. Esta puede ser de tres tipos:

Single linkage - vecino más cercano

$$d_{SL}(G, H) = \min_{\substack{i \in G \\ i' \in H}} d_{ii'}$$

Complete linkage - vecino más lejano

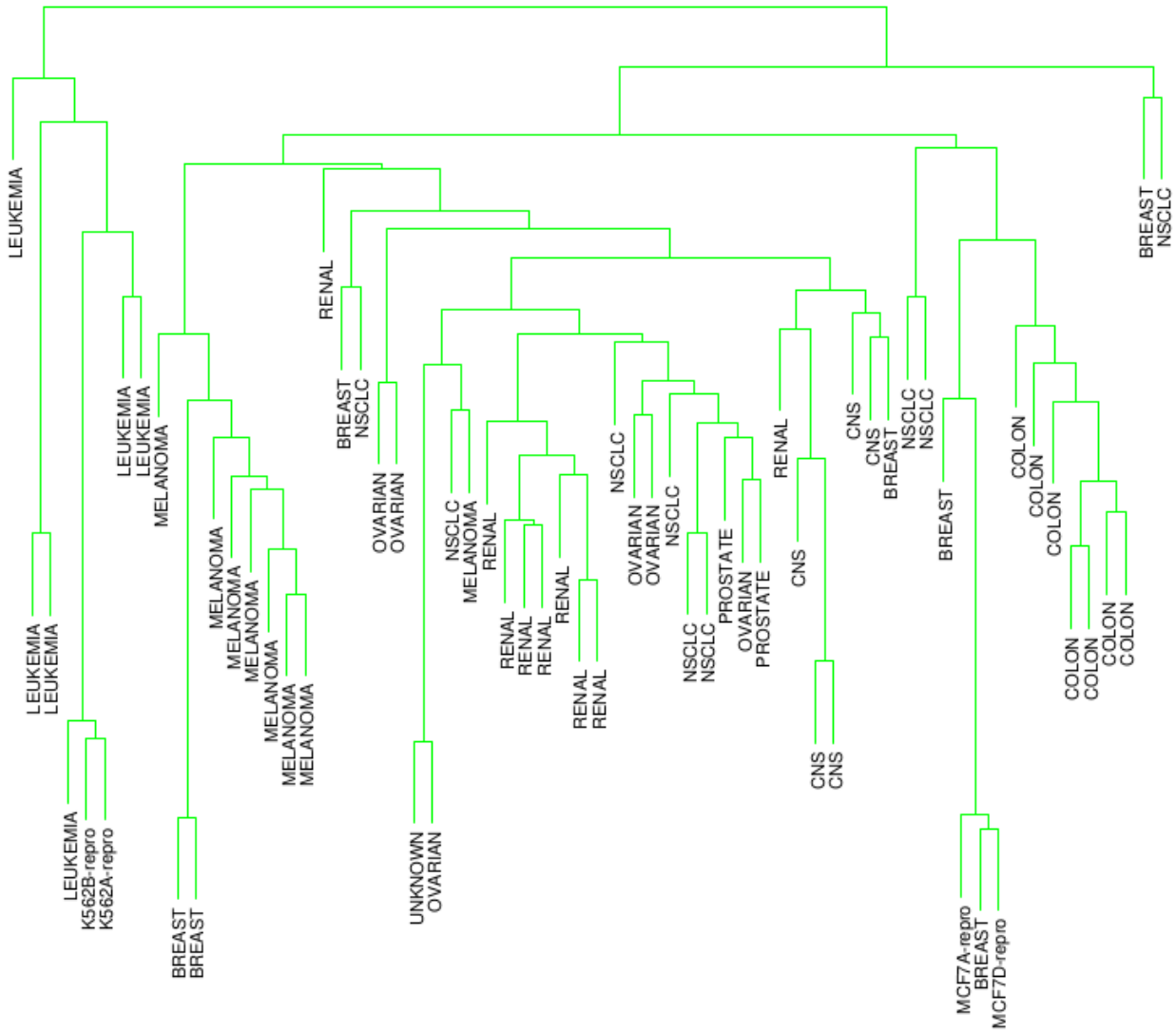
$$d_{CL}(G, H) = \max_{\substack{i \in G \\ i' \in H}} d_{ii'}$$

Promedio de grupo

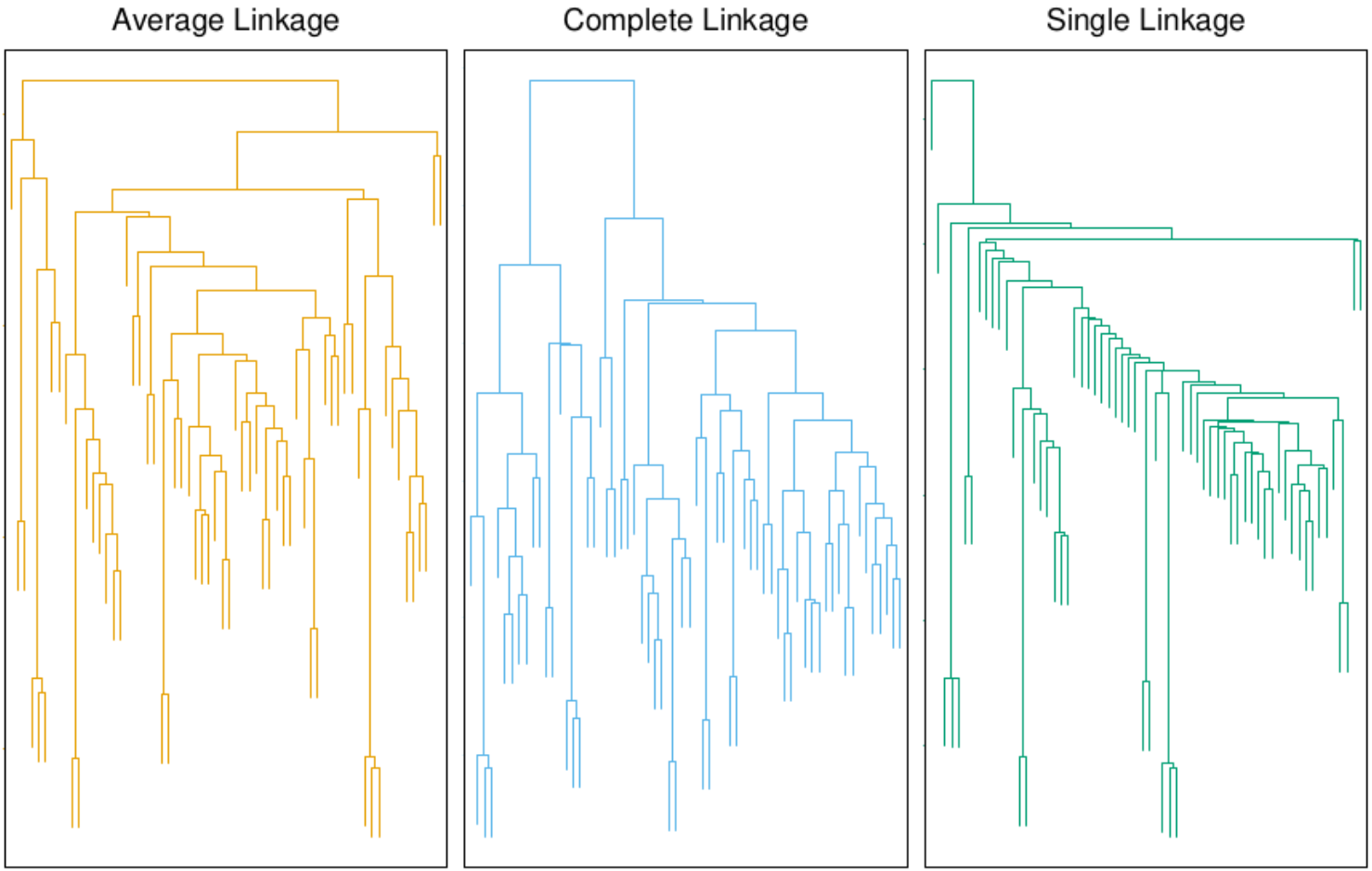
$$d_{GA}(G, H) = \frac{1}{N_G N_H} \sum_{i \in G} \sum_{i' \in H} d_{ii'}$$

donde N_G y N_H son los números de observaciones en cada grupo.

Clustering Jerárquico - Modelo Aglomerativo



Clustering Jerárquico - Funciones de enlace

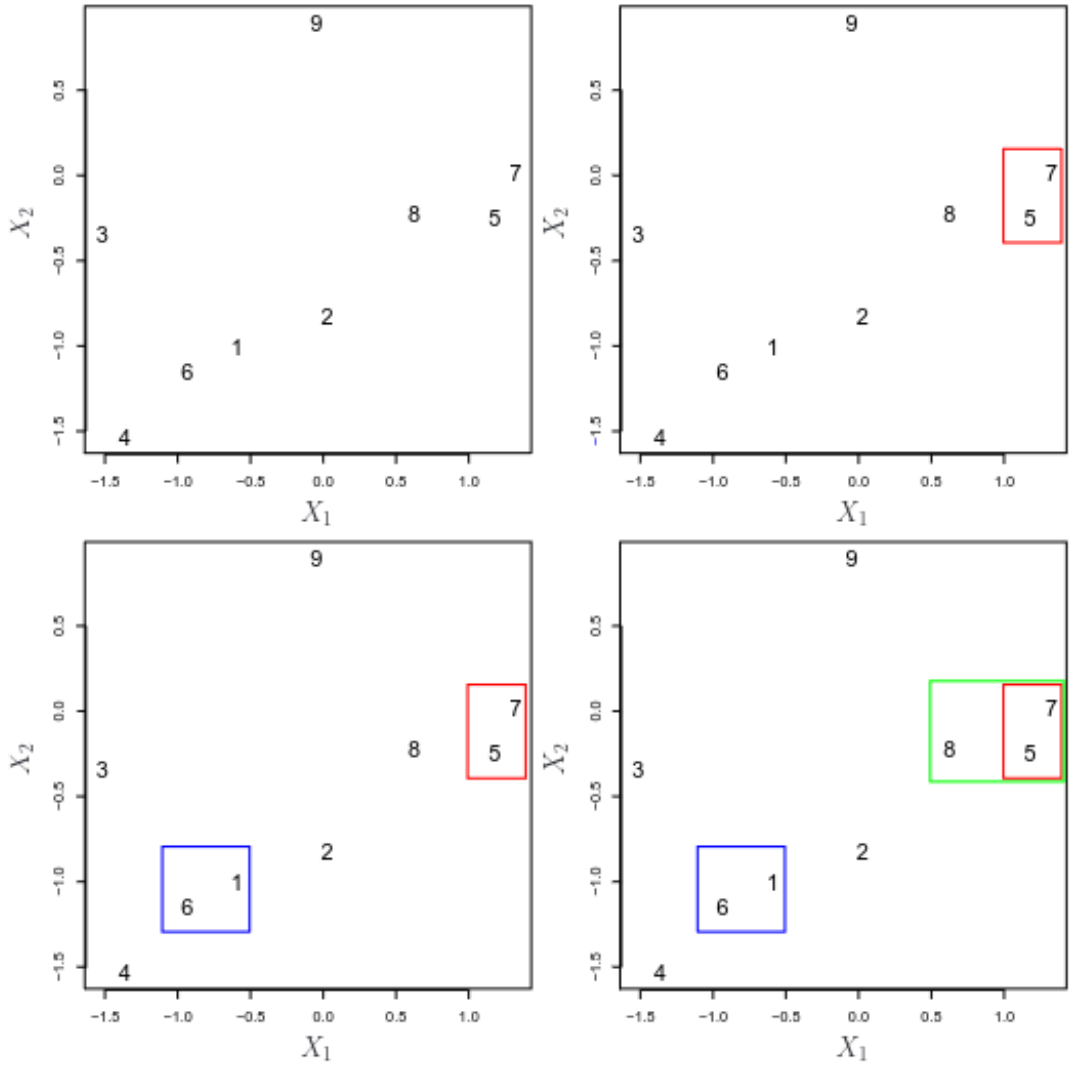


Clustering Jerárquico - Algoritmo

Algoritmo

1. Empezar con la medida de distancia para los $\binom{n}{2} = n(n-1)/2$ disimilaridades. Tratar cada observación como su propio cluster.
2. Para $i = n, n-1, ..., 2$:
   a) Examinar *todas* las disimilaridades entre pares de *inter-cluster* entre los i clusters e identificar los pares que son menos disimilares. Ésta representa la altura en el dendograma
   b) Computar las nuevas disimilaridades entre los i-1 cluster restantes.

Clustering Jerárquico - Funciones de enlace



Clustering Jerárquico - árboles divisivos

A la inversa del algoritmo anterior, el clustering divisivo considera todos los datos como un todo, y va separando la información para generar los grupos. En general y por su comportamiento, se utiliza cuando se quiere agrupar en pocos grupos los datos, lo que llevaría pocas iteraciones.
La forma en que se realiza es tomando la distancia promedio de las observaciones y separando la distancia más lejana. A ese lo apartan.
Esto se repite iterativamente hasta que cada observación quede singularizada o se llegue al límite deseado.