Machine Learning No Supervisados

Componentes Principales - Kmeans - Clustering Jerárquico.

Tomás Fontecilla

29 de septiembre de 2022

Machine Learning No Supervisado

Hasta ahora hemos visto modelos donde disponemos de una variable que nos permite revisar si acaso lo que hacemos "es correcto". Así, nos hemos ocupado en predecir si la salida del modelo está cerca (caso valores numéricos) o igual (caso valores categóricos) al real.

En esta última sección veremos modelos que permitan determinar las asociaciones entre variables

Machine Learning No Supervisado

En general,

- si el problema es de baja dimensionalidad ($p \leq 3$) existen varios métodos no paramétricos que estiman las propiedades del conjunto de datos y son representables gráficamente.
- Estos modelos no paramétricos fallan rápidamente en más dimensiones debido a la maldición de la dimensionalidad.
- Debido a esto, uno debe conformarse con realizar modelos globales algo crudos, como mezclas de gaussianas o estadísticas descriptivos que puedan caracterizar los datos.

Machine Learning No Supervisado

Para estas estadísticas descriptivas

- Componentes principales (escalamiento multidimensional, mapas auto organizados entre otros) intenta identificar colectores de baja dimensión dentro del espacio X que representan una alta densidad de los datos. Esto provee información sobre las asociaciones que permite considerar o no si son funciones de una variable menor "latente".
- K-means (y otros métodos de clustering) intenta encontrar regiones en el espacio X que contengan distintas modas de X, para determinar si X puede ser representado como una mezcla de densidades simples representadas por los tipos de clases de observaciones.
- Los modelos de mezclas tienen un objetivo similar al anterior. A través de reglas de asociación intenta construir descripciones simplesque describan regiones de alta densidad en caso especial de datos evaluados binariamente de muy alta dimensionalidad.

Machine Learning - Motivación

Componentes Principales

- Para poder eliminar la maldición de dimensionalidad de los datos, podemos tomar el enfoque de proyectar p variables en q rangos de tal manera que $q \leq p$. Con componentes principales intentamos perder la menor cantidad de información posible.
- Para no perder mucha información haremos algo que hasta ahora hemos intentado evitar: realizaremos las proyecciones *maximizando la varianza*. Dicho de otra forma, *minimizaremos la distancia* entre los datos originales y las proyecciones.
- Para simplificar, asumiremos que los datos están "centrados", lo que hace que cada variable tenga media 0. Así,

$$x^T x = \nu v$$

donde v es la matriz de covarianzas de los datos.

Componentes Principales - Minimizando los residuales de la proyección

Empecemos con una proyección unidimensional.

- La proyección de un vector de datos $\vec{x_i}$ sobre el vector unitario $\vec{\omega}$ (que es una linea que pasa por el origen) como $\vec{x_i} \cdot \vec{\omega}$ lo que lo hace escalar.
- Esta es la distancia de la proyección al origen.
- la coordenada realmente es $(\vec{x_i}\vec{\omega})\vec{\omega}$. La proyección media será cero, porque el promedio de los vectores $\vec{x_i}$ es cero:

$$\frac{1}{n}\sum_{i=1}^{n}(\vec{x_i}\vec{\omega})\vec{\omega} = ((\frac{1}{n}\sum_{i=1}^{n}(\vec{x_i})\vec{\omega})\vec{\omega})$$

- si intentamos usar los vectores proyectados o imagen en lugar de los vectores originales habrá errores, porque las imágenes no coincidirán con el vector original. Esta diferencia se llama el **residual** de la proyección.
- Su tamaño está dado por:
$\|\vec{x_i} - (\vec{x_i} \cdot \vec{\omega})\vec{\omega}\|^2 = \vec{x_i} \cdot \vec{x_i} - (\vec{x_i} \cdot \vec{\omega})^2$ dado que $\vec{\omega} \cdot \vec{\omega} = \|\vec{\omega}\|^2 = 1$

Componentes Principales - MSE

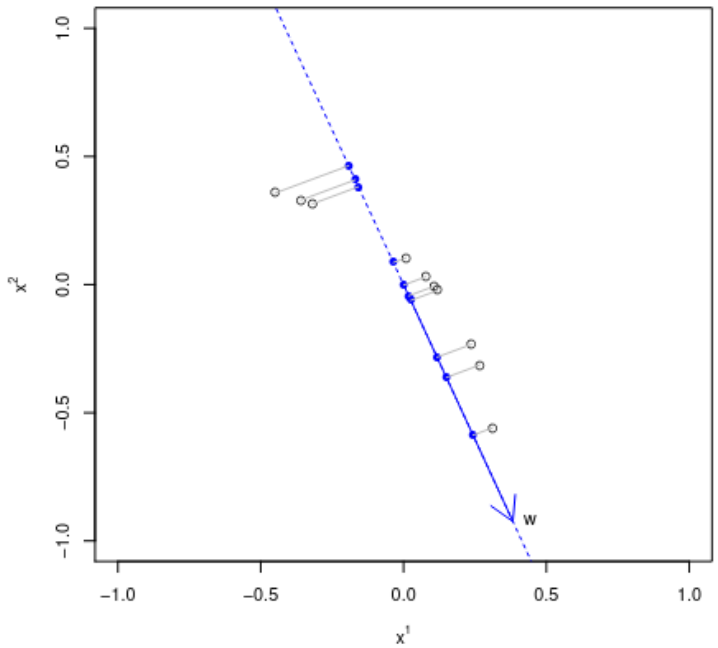
Al summar los residuos, se obtiene:

$$\begin{aligned} MSE(\vec{\omega}) &= \frac{1}{n}\sum_{i=1}^{n}\|\vec{x_i}\|^2 - (\vec{\omega}\vec{x_i})^2 \\ &= \frac{1}{n}\left(\sum_{i=1}^{n}\|\vec{x_i}\|^2 - \sum_{i=1}^{n}(\vec{\omega}\vec{x_i})^2\right) \end{aligned}$$

La suma no depende de $\vec{\omega}$ asi que no importa para minimizar la media cuadrática de los residuos.

Componentes Principales - Resumiendo la historia... y la matemática detrás

Todo esto es conocido!



- Queremos maximizar varianza
- queremos proyectar con respecto a ejes
- queremos minimizar los errores

Muy parecido a LDA!

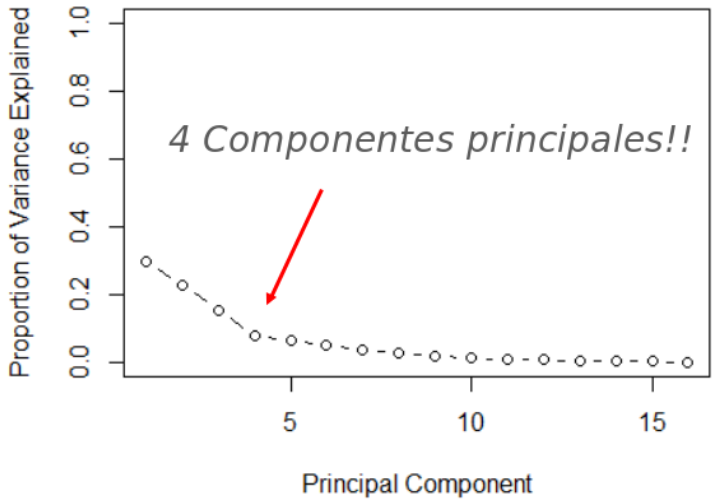
¿En qué se diferencia? No tenemos la regla de clasificación, la estamos intentando deducir.

Componentes Principales - ¿Cuántos Componentes?

utilizando la varianza maximizada y la proporción de esta explicada por cada componente, podemos realizar un gráfico scree

Sea λ la varianza explicada por el componente principal. Luego, la proporción está dada por:

$$\frac{\lambda_p}{\sum_{i=1} \lambda_i} = \frac{(u^p)^T S u^p}{\sum_{i=1} (u^i)^T S u^i}$$



Componentes Principales - Recordatorio

Debido a que este algoritmo es no supervisado, no tenemos una forma real para saber cómo validar directamente estos modelos, por lo tanto no existe una forma pura de determinar si están bien ajustados y suele hacerse combinaciones de modelos no supervisados e interpretaciones gráficas para determinar si hay cambios o no en los resultados.
En el caso de PCA, es la distribución de la varianza la que manda, y la acumulación de ella en los componentes para determinar si es suficiente o mejor seguir con otro método.

K-means

Como metodo 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 - Definiciones

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

Dispersión interna

$$W(C) = \frac{1}{2}\sum_{k=1}^{K}\sum_{C(i)=k}\sum_{C(i')=k}\|x_i - x_{i'}\|^2 = \sum_{k=1}^{K} N_k \sum_{C(i)=k}\|x_i - \bar{x}_k\|^2$$

Donde $\bar{x}_k = (\bar{x}_{1k}, ..., \bar{x}_{pk})$ es el vector de medias asociadas al k-ésimo cluster y $N_k = \sum_{i=1}^{N} I(C(i) = k)$

Asignación de Cluster

$$C^* = \min_{C} \sum_{k=1}^{K} N_k \sum_{C(i)=k}\|x_i - \bar{x}_k\|^2$$

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

$$C^* = \min_{C, \{m_k\}_1^K} \sum_{k=1}^{K} N_k \sum_{C(i)=k}\|x_i - m_k\|^2$$

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

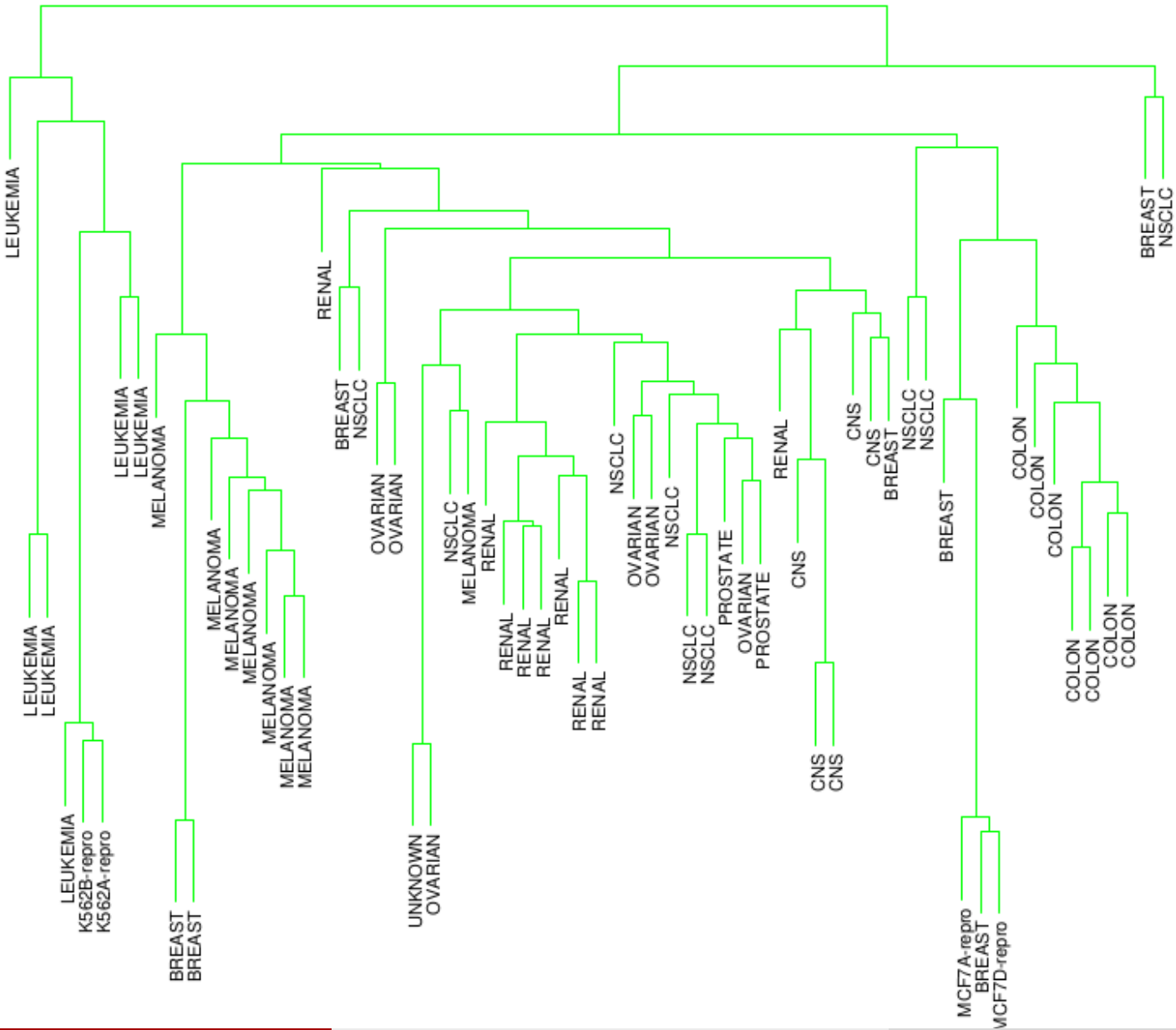
$$d_{CL}(G,H) = \max_{\substack{i \in G \\ i' \in H}} d_{ii'}$$

Promedio de grupo

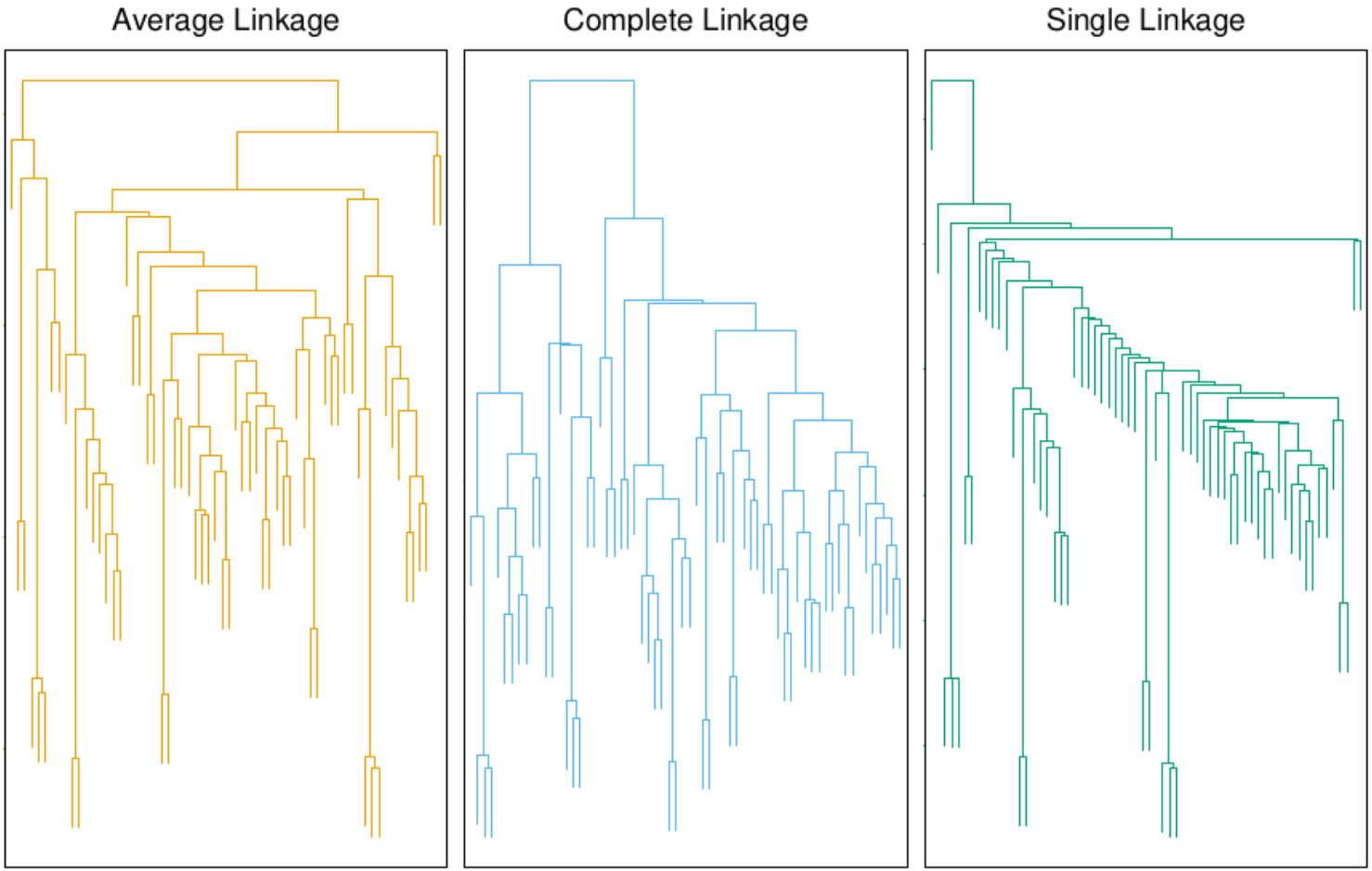
$$d_{GA}(G,H) = \frac{1}{N_G N_H} \sum_{i \in G} \sum_{i' \in H} d_{ii'}$$

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 - 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.