UNIDAD DIDÁCTICA 3: Técnicas de agrupamiento
Tema 2: Análisis Clúster

TÉCNICAS ESTADÍSTICAS PARA EL APRENDIZAJE I

Máster Universitario en Estadística Computacional
y Ciencia de Datos para la Toma de Decisiones

Índice

Métodos jerárquicos

Métodos Jerárquicos Aglomerativos

Agrupamiento que busca construir una jerarquía entre grupos. A diferencia del k-medias no partimos indicando cuántos clústers queremos crear, sino que el algoritmo nos muestra un listado de combinaciones posibles de acuerdo a la jerarquía de las distancias entre puntos.

Este método, de manera iterativa agrupa observaciones basado en sus distancias hasta que cada observación pertenece a un grupo más grande. El método va juntando los elementos más similares y termina en un único gran clúster. Los resultados pueden expresarse en un dendrograma.

Dendrograma

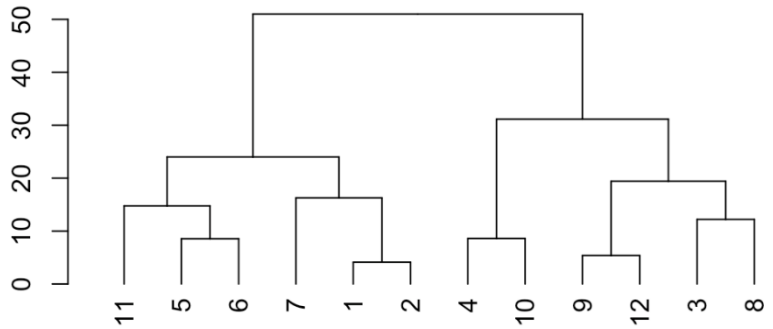
El dendrograma, o árbol jerárquico, es una representación gráfica del resultado del proceso de agrupamiento en forma de árbol.

El **dendrograma se contruye** como sigue:

1. En la parte inferior del gráfico se disponen los n elementos iniciales.

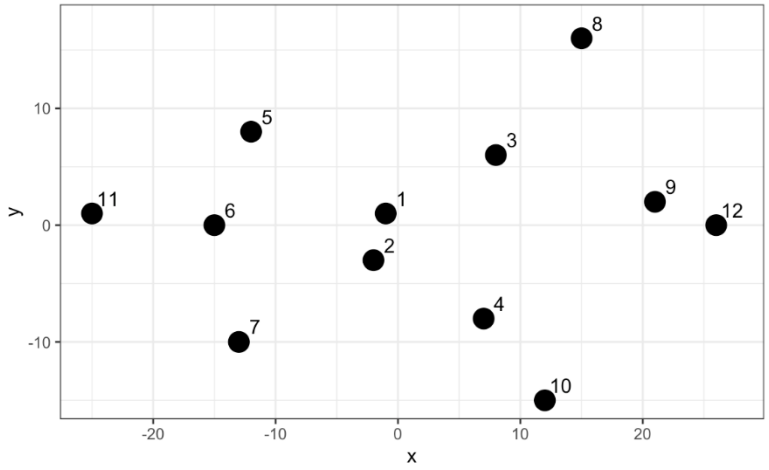
2. Las uniones entre elementos se representan por tres líneas rectas. Dos dirigidas a los elementos que se unen y que son perpendiculares al eje de los elementos y una paralela a este eje que se sitúa al nivel en que se unen.

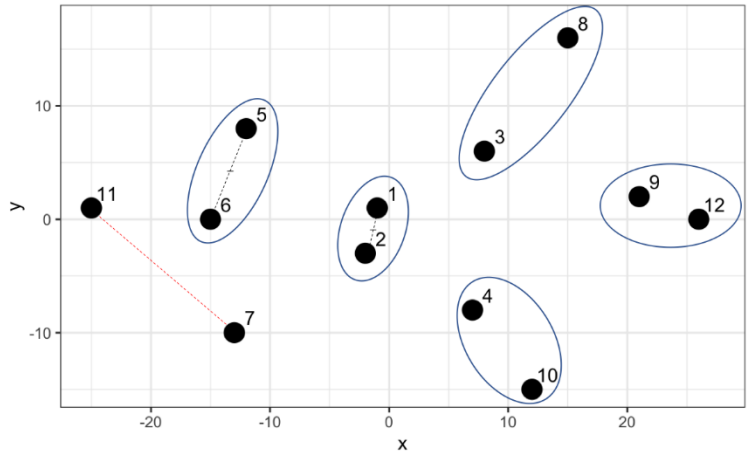
3. El proceso se repite hasta que todos los elementos están conectados por líneas rectas.

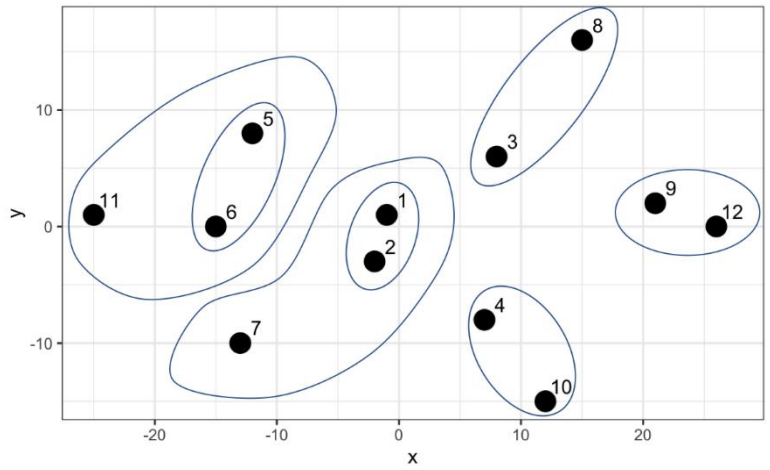


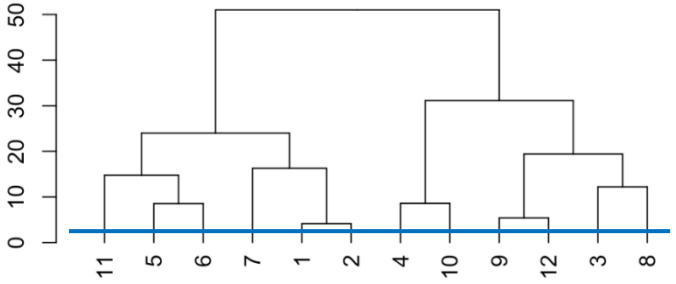
Métodos jerárquicos aglomerativos

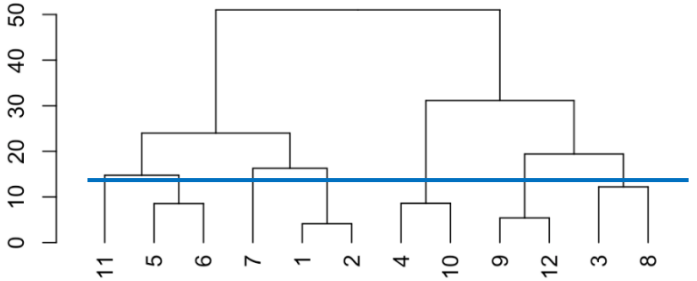
Ejemplo:

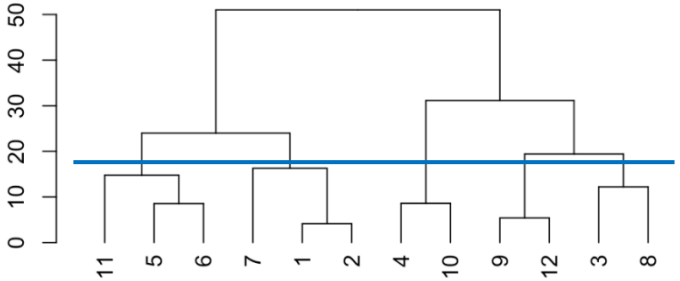












Distancias y similaridades:

Los métodos jerárquicos parten de una matriz de distancias o similaridades entre los elementos de la muestra y construyen una jerarquía basada en una distancia.

Si todas las variables son continuas, la distancia más utilizada es la distancia euclídea entre las variables estandarizadas.

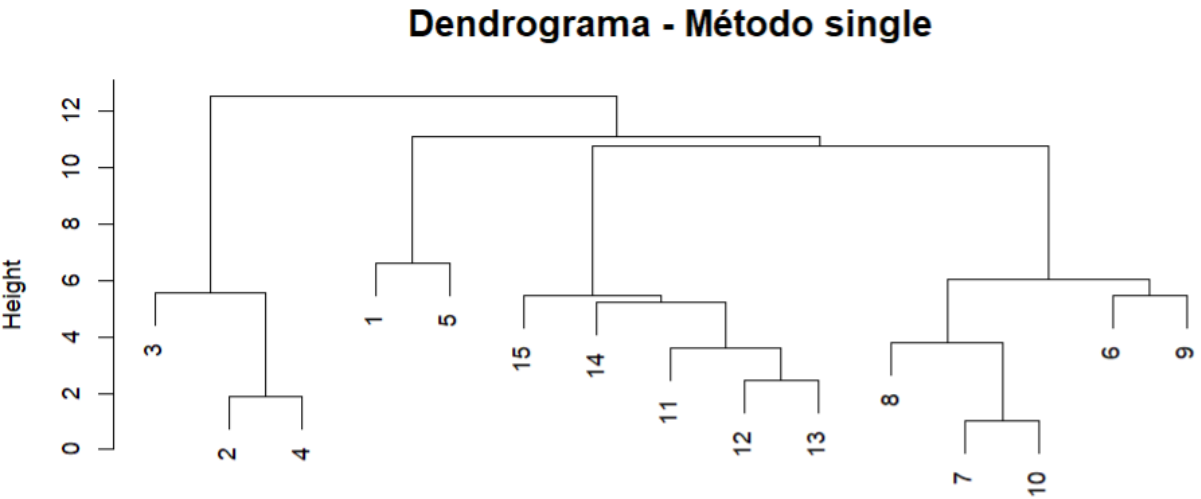
Si no estandarizamos, la distancia euclídea dependerá sobre todo de las variables con valores más grandes.

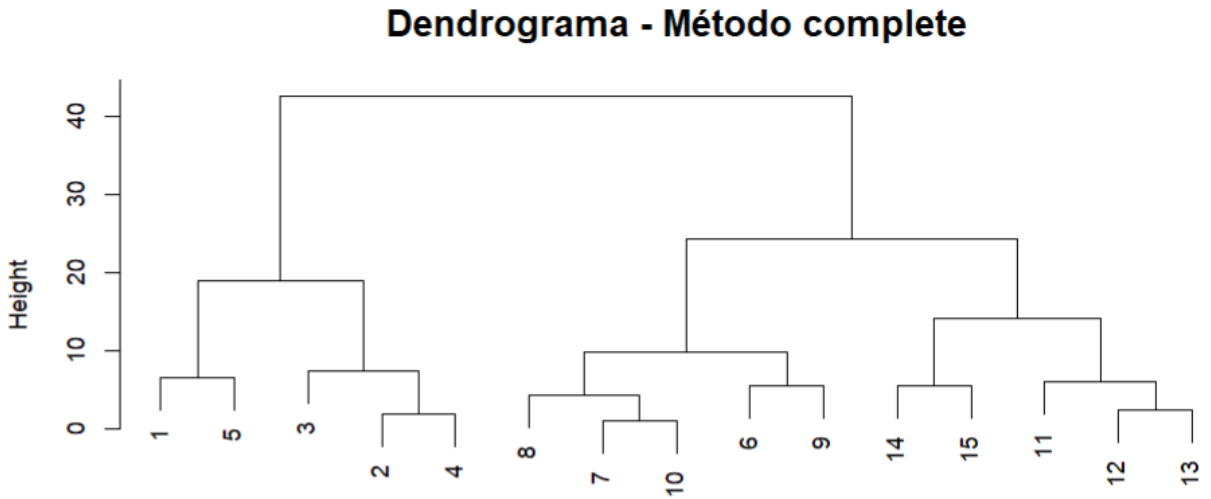
Criterios para definir distancias entre grupos:

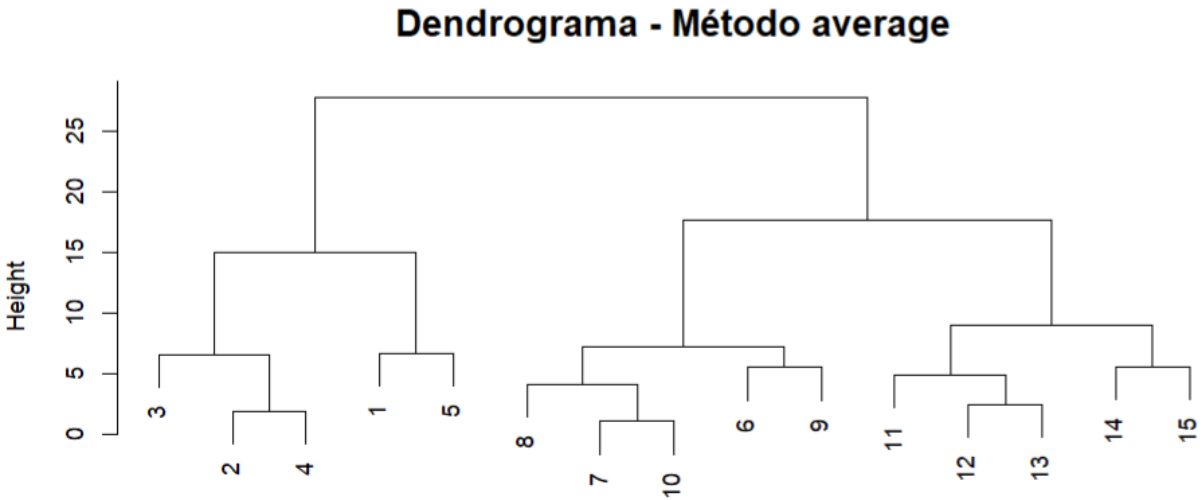
Distancia entre pares de grupos, cada uno formado por varias observaciones, esta distancia se realiza habitualmente por alguna de las cinco reglas siguientes (*linkage)*:

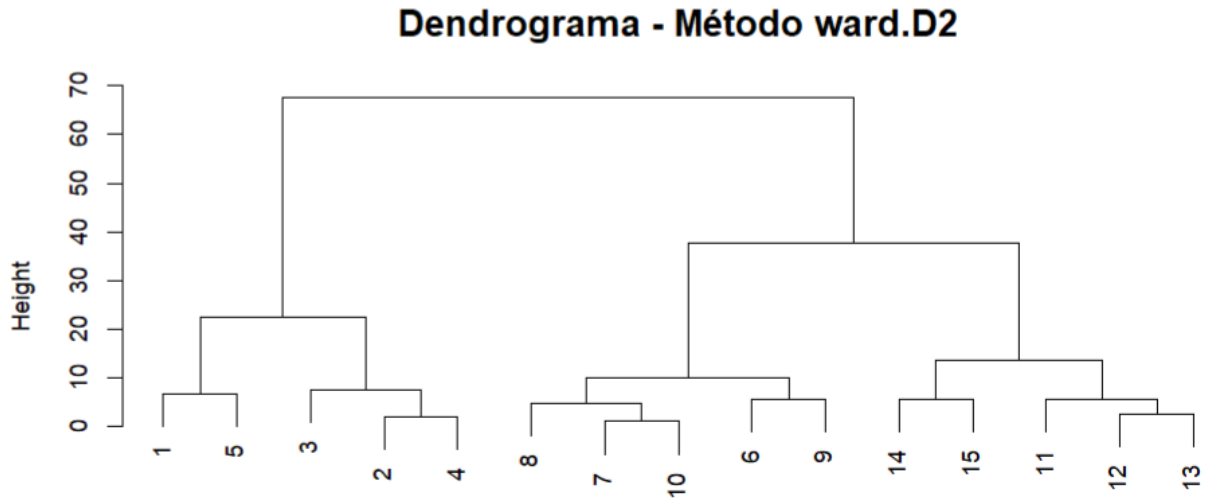
1. **Mínimo (Single or Minimum):** Se calcula la distancia entre todos los posibles pares formados por una observación del clúster A y una del clúster B. La menor de todas ellas se selecciona como la distancia entre los dos clústers.
2. **Máximo (Complete or Maximum):** Se calcula la distancia entre todos los posibles pares formados por una observación del clúster A y una del clúster B. La mayor de todas ellas se selecciona como la distancia entre los dos clústers
3. **Media (Average):** Se calcula la distancia entre todos los posibles pares formados por una observación del cluster A y una del cluster B. El valor promedio de todas ellas se selecciona como la distancia entre los dos clusters
4. **Centroide (Centroid):** La distancia entre dos grupos se hace igual a la distancia euclídea entre sus centros, donde se toman como centros los vectores de medias de las observaciones que pertenecen al grupo.
5. **Método de Ward.** El conocido método Ward's minimum variance. En cada paso, se identifican aquellos 2 clústers cuya fusión conlleva menor incremento de la varianza total intra-cluster.

Criterios para definir distancias entre grupos:





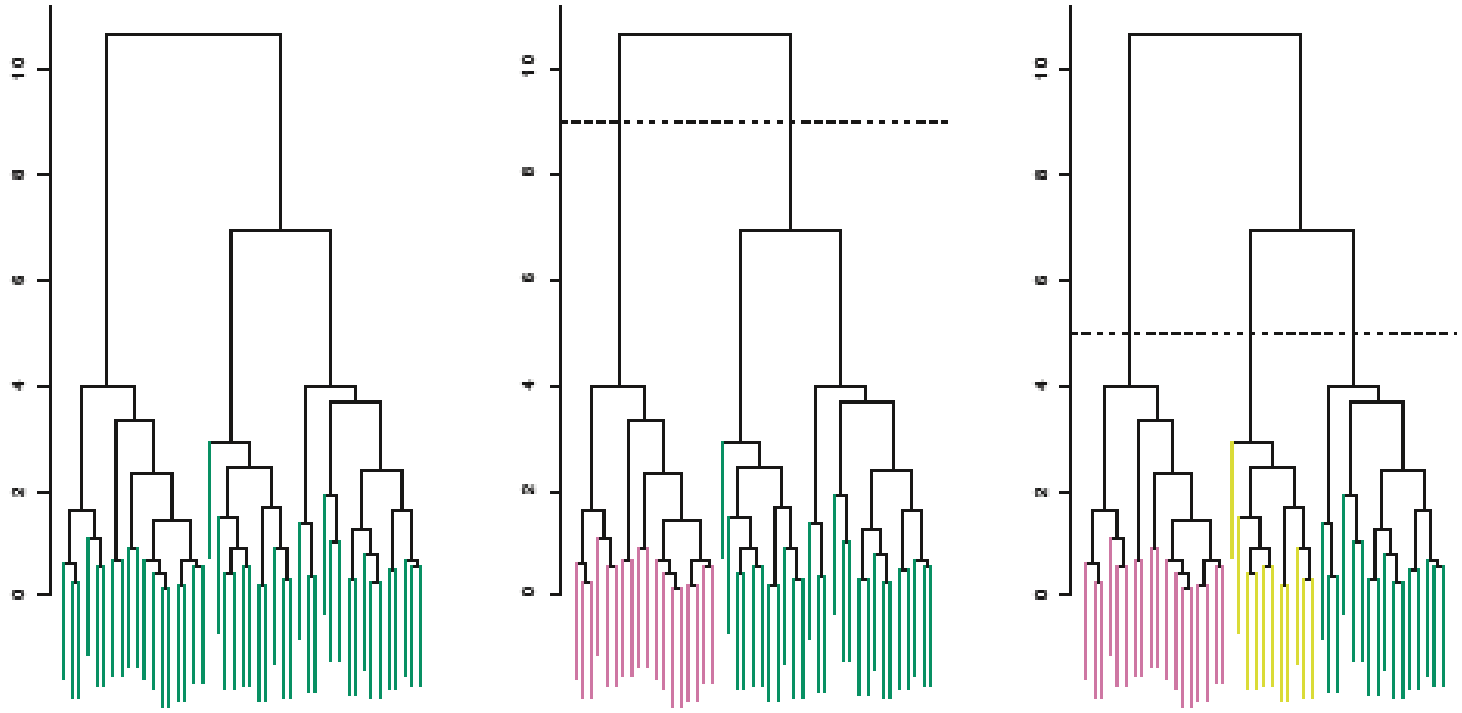




Dendrograma

Si cortamos el dendrograma a un nivel de distancia dado, obtenemos una clasificación del número de grupos existentes a ese nivel y los elementos que los forman.

El dendrograma es útil cuando los puntos tienen claramente una estructura jerárquica, pero puede ser engañoso cuando se aplica ciegamente, ya que dos puntos pueden parecer próximos cuando no lo están, y pueden aparecer alejados cuando están próximos.



Métodos Jerárquicos Divisivos (DIANA)

El algoritmo más conocido de divisive hierarchical clustering es **DIANA (DIvisive ANAlysis Clustering)**.

Este algoritmo se inicia con un único clúster que contiene todas las observaciones, a continuación, se van sucediendo divisiones hasta que cada observación forma un clúster independiente. En cada iteración, se selecciona el clúster con mayor diámetro, entendiendo por diámetro de un clúster la mayor de las diferencias entre dos de sus observaciones.

Una vez seleccionado el clúster, se identifica la observación más dispar, que es aquella con mayor distancia promedio respecto al resto de observaciones que forman el clúster, esta observación inicia el nuevo clúster. Se reasignan las observaciones en función de si están más próximas al nuevo clúster o al resto de la partición, dividiendo así el clúster seleccionando en dos nuevos clúster. Esto se repite recursivamente en cada grupo hasta que haya un grupo para cada observación.

A diferencia del método aglomerativo, en el que hay que elegir un tipo de distancia y un método de *linkage*, en el método divisivo solo hay que elegir la distancia, no hay *linkage*.

Análisis Clúster Jerárquico k-medias

K-medias es uno de los métodos de *Análisis Clúster* más utilizados y cuyos resultados son satisfactorios en muchos escenarios, sin embargo, sufre las limitaciones de necesitar que se **especifique el número de *clusters* de antemano** y de que sus resultados puedan variar en función de la iniciación aleatoria. Una forma de contrarrestar estos dos problemas es **combinando** el ***K-medias*** con el ***Análisis Clúster Jerárquico***.

Los pasos a seguir son los siguientes:

1. Aplicar ***Análisis Clúster Jerárquico*** a los datos y **cortar el árbol en *k clusters***. El número óptimo puede elegirse de forma visual o con cualquiera de los métodos explicados en la sección ***Número de clusters***.

2. Calcular el **centro** (por ejemplo, la media) de cada ***clúster***.

3. Aplicar ***k-medias*** empleando como **centroides iniciales los centros calculados** en el paso anterior.

El algoritmo de K-medias tratará de mejorar la agrupación hecha por el Análisis Clúster Jerárquico en el paso 1, de ahí que las agrupaciones finales puedan variar respecto a las iniciales.

Fuzzy clustering

Los métodos de *Análisis Clúster* descritos hasta ahora (*K-medias, jerárquico, DIANA,…*) asignan cada observación únicamente a un *clúster*. Los métodos de ***fuzzy clustering o soft clustering*** se caracterizan porque, cada **observación**, puede **pertenecer potencialmente a varios clusters**, en concreto, cada observación tiene asignado un grado de pertenencia a cada uno de los *clúster*.

Fuzzy c-means (FCM) es uno de los algoritmos más empleado para generar *fuzzy clustering*. Se asemeja en gran medida al algoritmo de *k-medias* pero con dos diferencias:

- El cálculo de los **centroides de los *clusters***. La definición de centroide empleada por *c-means* es: **la media de todas las observaciones del conjunto de datos ponderada por la probabilidad de pertenecer al *clúster***.

- Devuelve para cada observación **la probabilidad de pertenecer a cada *clúster***.

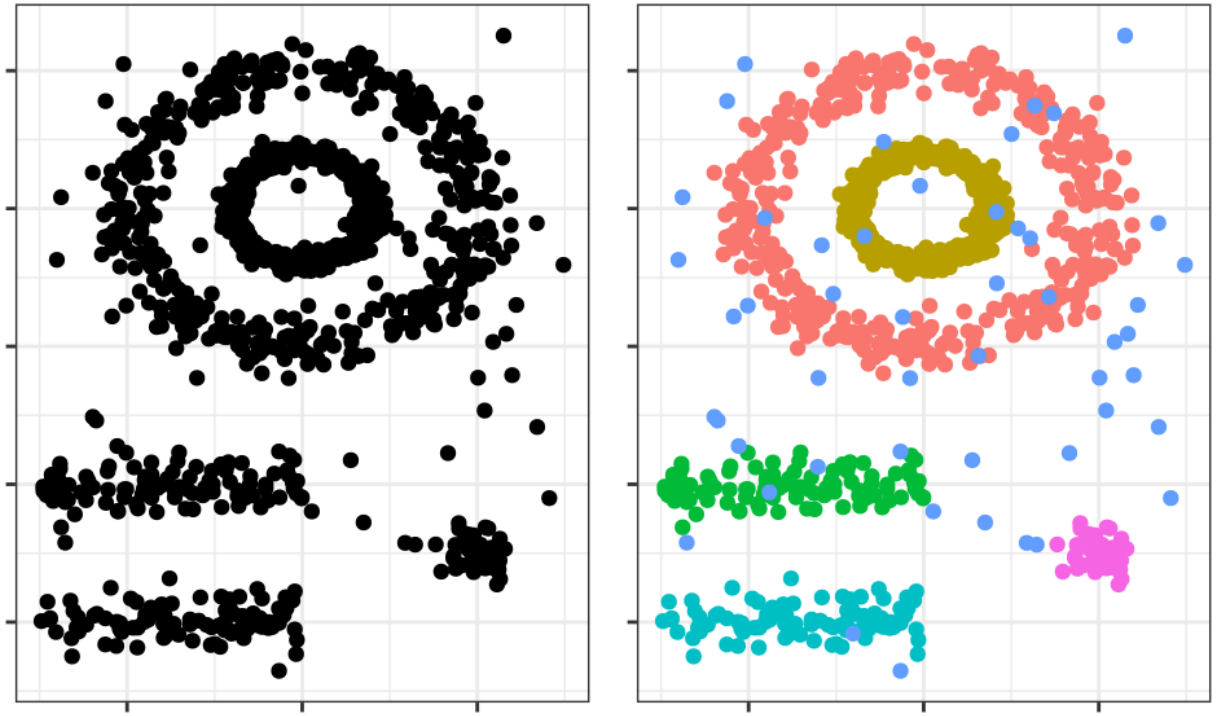
Elección del mejor algoritmo de clustering

Decidir cuál es el método de clustering más adecuado para un determinado conjunto de datos es un proceso complejo ya que se tienen que analizar uno a uno múltiples índices, estadísticos y parámetros (número de clusters, homogeneidad, separación,…). El **paquete clValid** agiliza el proceso ofreciendo la posibilidad de comparar, de forma simultánea, múltiples algoritmos de clustering en una única función.

Análisis Clúster Basado en las densidades

Análisis Clúster Espacial Basado en Densidades (DBSCAN)

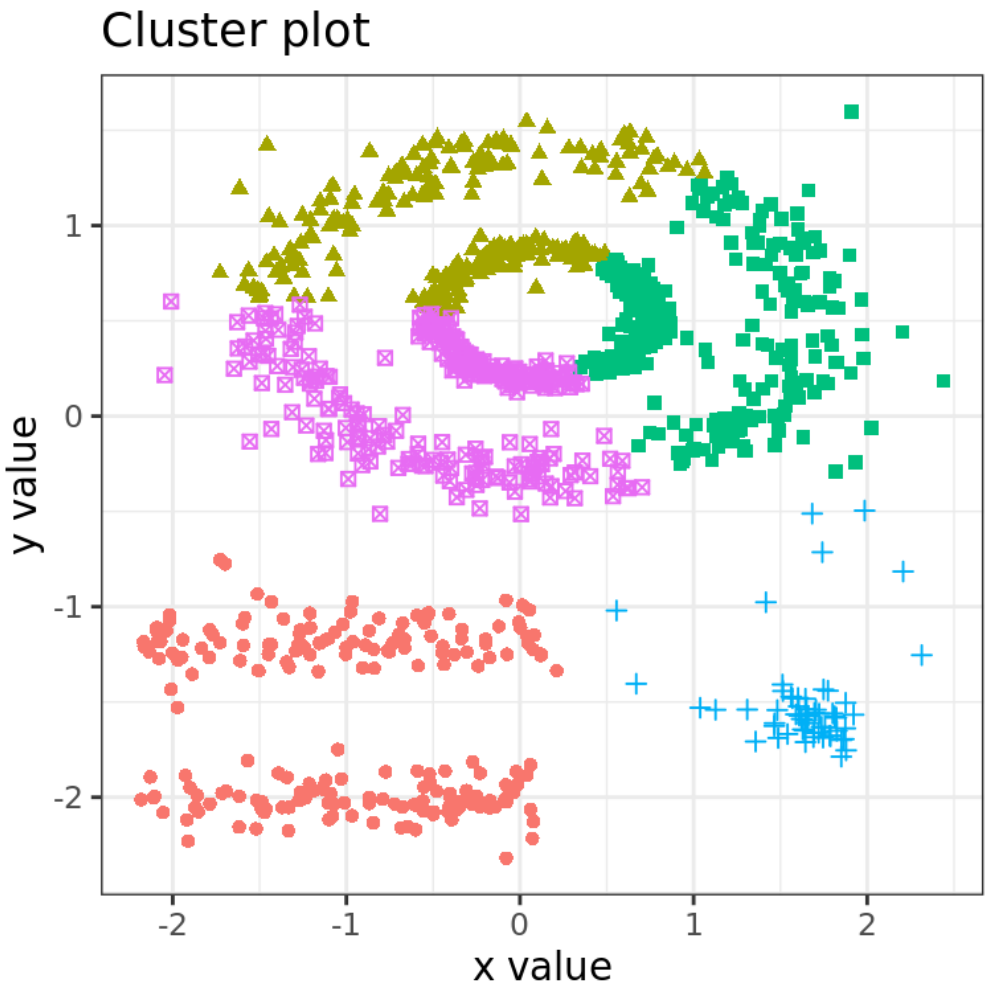
Análisis Clúster Espacial basado en las densidades *(Density-based spatial clustering of applications with noise, DBSCAN))* fue presentado en 1996 por Ester et al. (1996) como una forma de identificar *clusters* siguiendo el modo intuitivo en el que lo hace el cerebro humano, identificando regiones con alta densidad de observaciones separadas por regiones de baja densidad.



El cerebro humano identifica fácilmente 5 agrupaciones y algunas observaciones aisladas.

Análisis Clúster Espacial Basado en Densidades (DBSCAN)

Clusters que se obtienen si se aplica, por ejemplo, *K-medias*.



Los *clusters* generados distan mucho de representar las verdaderas agrupaciones. Esto es así porque los métodos vistos (k-medias, clúster jerárquico, clúster k-medias jerárquico), son buenos encontrando agrupaciones con forma esférica o convexa que no contengan un exceso de *outliers*, pero fallan al tratar de identificar formas arbitrarias.

Análisis Clúster Espacial Basado en Densidades (DBSCAN)

DBSCAN evita este problema siguiendo la idea de que, para que una **observación** forme **parte de un *clúster*,** tiene que haber un **mínimo de observaciones vecinas** dentro de un radio de proximidad y de que los *clusters* están separados por regiones vacías o con pocas observaciones.

El algoritmo *DBSCAN* necesita dos parámetros:

- ***Epsilon*** (ϵ): radio que define la región vecina a una observación, también llamada ϵ-*neighborhood*.

- ***Minimum points* (*minPts*):** número mínimo de observaciones dentro de la región *epsilon*.

Análisis Clúster Espacial Basado en Densidades (DBSCAN)

Empleando los parámetros (**ϵ** y ***minPts)*** , cada observación del conjunto de datos se puede clasificar en una de las siguientes tres categorías:

- ***Puntos núcleo (Core point)*:** observación que tiene en su ϵ-*neighborhood* un número de observaciones vecinas igual o mayor a *minPts*.

- ***Puntos frontera (Border point)*:** observación no satisface el mínimo de observaciones vecinas para ser *core point* pero que pertenece al ϵ-*neighborhood* de otra observación que sí es *un punto núcleo*.

- ***Ruido (Noise* u *outlier)*:** observación que no es *ni punto núcleo ni punto frontera*.

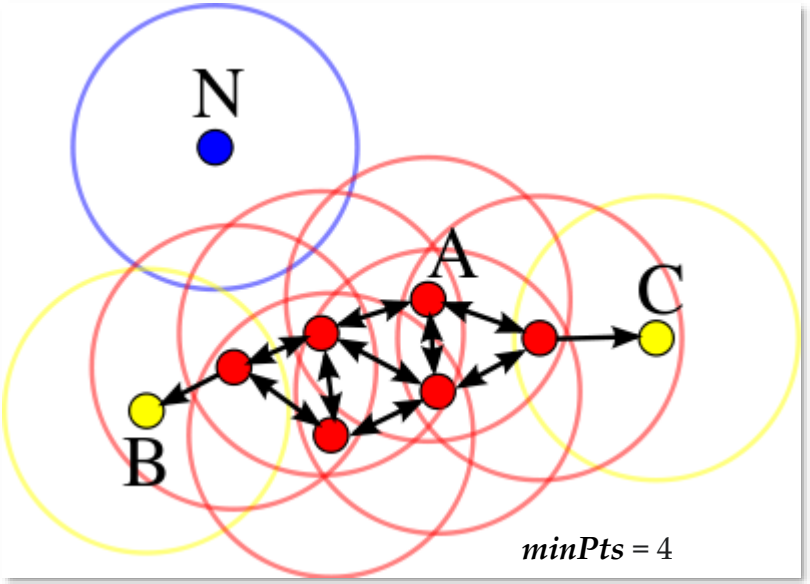


Fig: Wikipedia

Análisis Clúster Espacial Basado en Densidades (DBSCAN)

Algoritmo:

1. Para cada observación x_i, calcular la distancia entre ella y el resto de observaciones. Si en su ϵ hay un número de observaciones $\geq$ *minPts* marcar la observación como punto núcleo, de lo contrario marcarla como visitada.

2. Para cada observación x_i marcada como punto núcleo, si todavía no ha sido asignada a ningún clúster, crear uno nuevo y asignarla a él. Encontrar recursivamente todas las observaciones conectadas a ella y asignarlas al mismo clúster.

3. Iterar el mismo proceso para todas las observaciones que no hayan sido visitadas.

4. Aquellas observaciones que tras haber sido visitadas no pertenecen a ningún clúster se marcan como outliers.

Como resultado, todo **clúster cumple dos propiedades**:

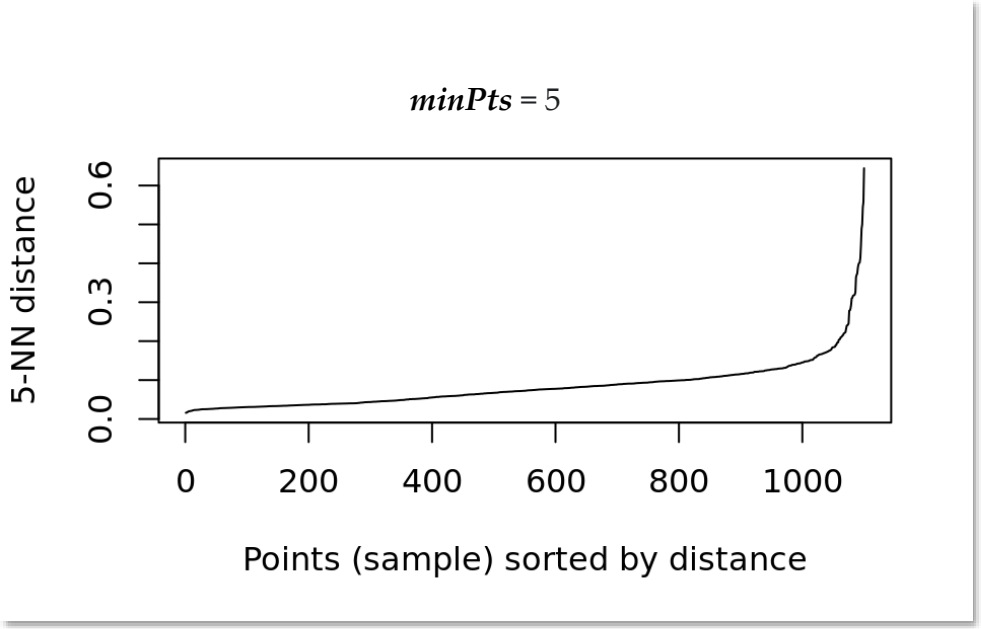
Todos los puntos que forman parte de un mismo clúster están densamente conectados entre ellos y, si una observación A es densamente alcanzable desde cualquier otra observación de un clúster, entonces A también pertenece al clúster.

Selección de parámetros:

No existe una forma única y exacta de encontrar el valor adecuado de epsilon (ϵ) y minPts. A **modo orientativo** se pueden seguir las siguientes premisas:

minPts*:** cuanto mayor sea el tamaño del conjunto de datos, mayor debe ser el valor mínimo de observaciones vecinas. Se recomienda no bajar nunca de 3. Si los datos contienen niveles altos de ruido, aumentar ***minPts favorecerá la creación de clústers significativos menos influenciados por outliers.

Epsilon **(ϵ):** una buena forma de escoger el valor de ϵ es estudiar las distancias promedio entre las *k=minPts* observaciones más próximas. Al representar las distancias promedios (k-NN distancia) ordenadas, en función del número de puntos, el punto de inflexión de la curva suele ser un valor óptimo para ϵ.



Limitaciones Análisis Clúster

El *Análisis Clúster* puede ser una herramienta muy útil para encontrar agrupaciones en los datos, sin embargo, posee limitaciones o problemas que pueden surgir al aplicarlo.

Pequeñas decisiones pueden tener grandes consecuencias: A la hora de utilizar el método se tienen que tomar decisiones que influyen en gran medida en los resultados obtenidos.

- **Escalado y centrado de las variables**
- Qué medida de **distancia/similitud** emplear
- **Número de *clusters***
- Tipo de ***linkage*** empleado en *los métodos jerárquicos*
- A qué altura establecer el **corte de un dendrograma**

Validación de los clusters obtenidos: No es fácil comprobar la validez de los resultados ya que en la mayoría de escenarios se desconoce la verdadera agrupación.

Falta de robustez: Los métodos de *K-medias* e *Clúster Jerárquico* asignan obligatoriamente cada observación a un grupo. Si existe en la muestra algún *outlier*, a pesar de que realmente no pertenezca a ningún grupo, el algoritmo lo asignará a uno de ellos provocando una distorsión significativa del *clúster* en cuestión.
Algunas alternativas también pertenecientes al Análisis Clúster, *k-medoids (PAM), CLARA* y *HDBSCAN*.

Bibliografía

Bibliografía

- Aldás Manzano, J., & Uriel Jimenez, E. (2017). Análisis multivariante aplicado con R. Ediciones Paraninfo, SA.
- Amat Rodrigo, J. Clustering y heatmaps: aprendizaje no supervisado by Joaquín, available under a Attribution 4.0 International (CC BY 4.0) https://www.cienciadedatos.net/documentos/37_clustering_y_heatmaps
- Ester, Martin; Kriegel, Hans-Peter; Sander, Jörg; Xu, Xiaowei (1996). Simoudis, Evangelos; Han, Jiawei; Fayyad, Usama M. (eds.). A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. AAAI Press. pp. 226–231.
- Guy Brock, Vasyl Pihur, Susmita Datta, and Somnath Datta. clValid, an R package for cluster validation. Department of Bioinformatics and Biostatistics, University of Louisville
- Husson, F., Lê, S., & Pagès, J. (2011). Exploratory multivariate analysis by example using R (Vol. 15). Boca Raton: CRC press.
- Peña, D. (2002). Análisis de datos multivariantes. McGraw-Hill.