DIPLOMADO EN CIENCIA DE DATOS

Módulo: Minería de Datos

Métodos de clusterización

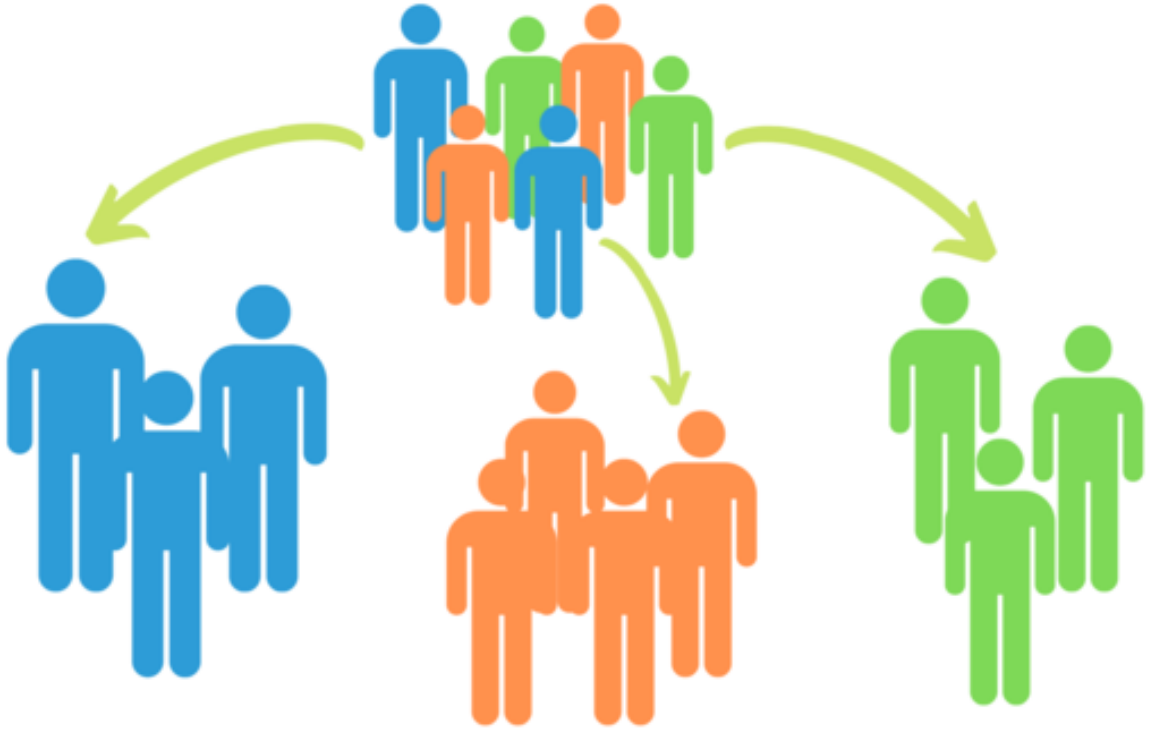
Universidad Nacional de Colombia

Contenido

- Introducción
- Tipos de clusterización
- Clusterización Jerárquica
  - Aglomerativo y divisivo
- Clusterización Particional
  - K-means
- Clusterización basada en densidades
  - DBSCAN

Agrupamiento

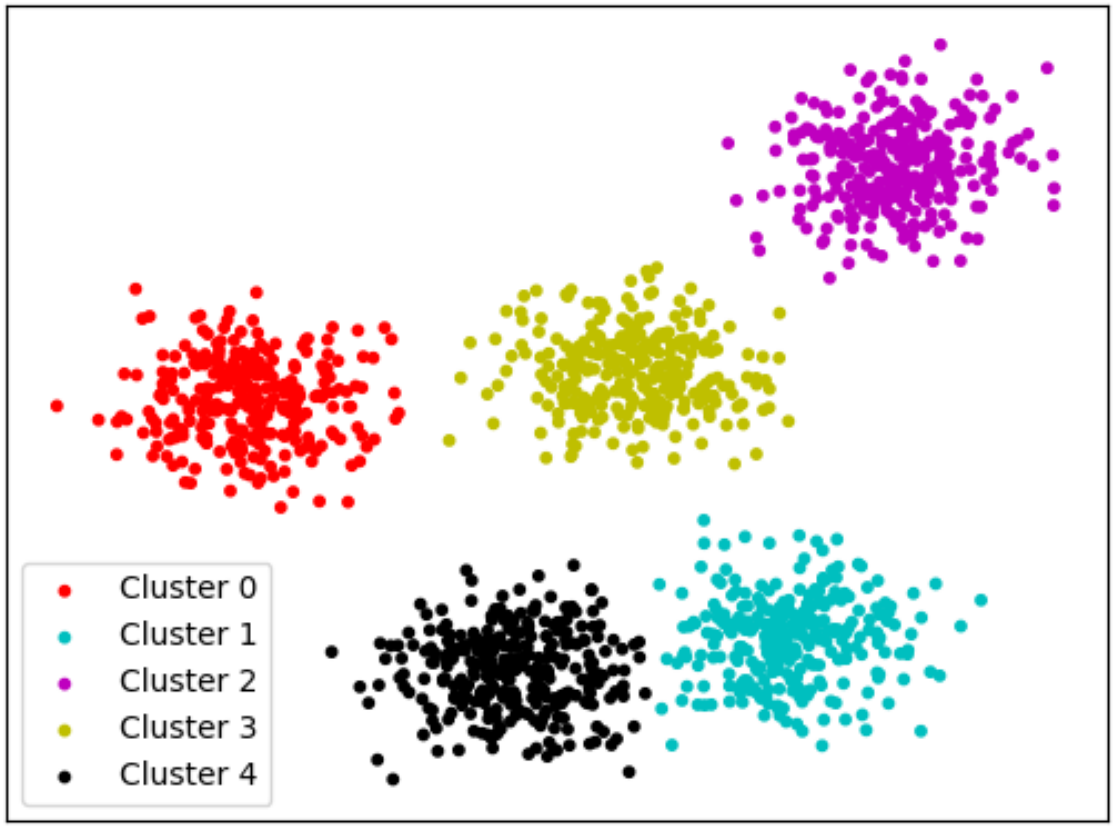
El análisis de clusters (agrupamiento) es una metodología donde agrupamos objetos/ observaciones en subgrupos



La idea es que los objetos en un mismo grupo (cluster) sean mas similares entre ellos que objetos en otros grupos. Los *clusters* deberían ser homogéneos dentro y heterogéneos entre ellos.

Agrupamiento

Es uno de los análisis más relevantes en la minería de datos (muy útil), y una técnica común de análisis estadístico de datos



Aplicaciones: machine learning, reconocimiento de patrones, análisis de imágenes, bioinformática, segmentación de clientes, y más

Agrupamiento - ¿por qué?

Organizar los datos en clusters, muestra cuál es la estructura interna de la data para encontrar patrones

A veces la meta es particionar/segmentar para entender mejor los datos (segmentación de mercados)

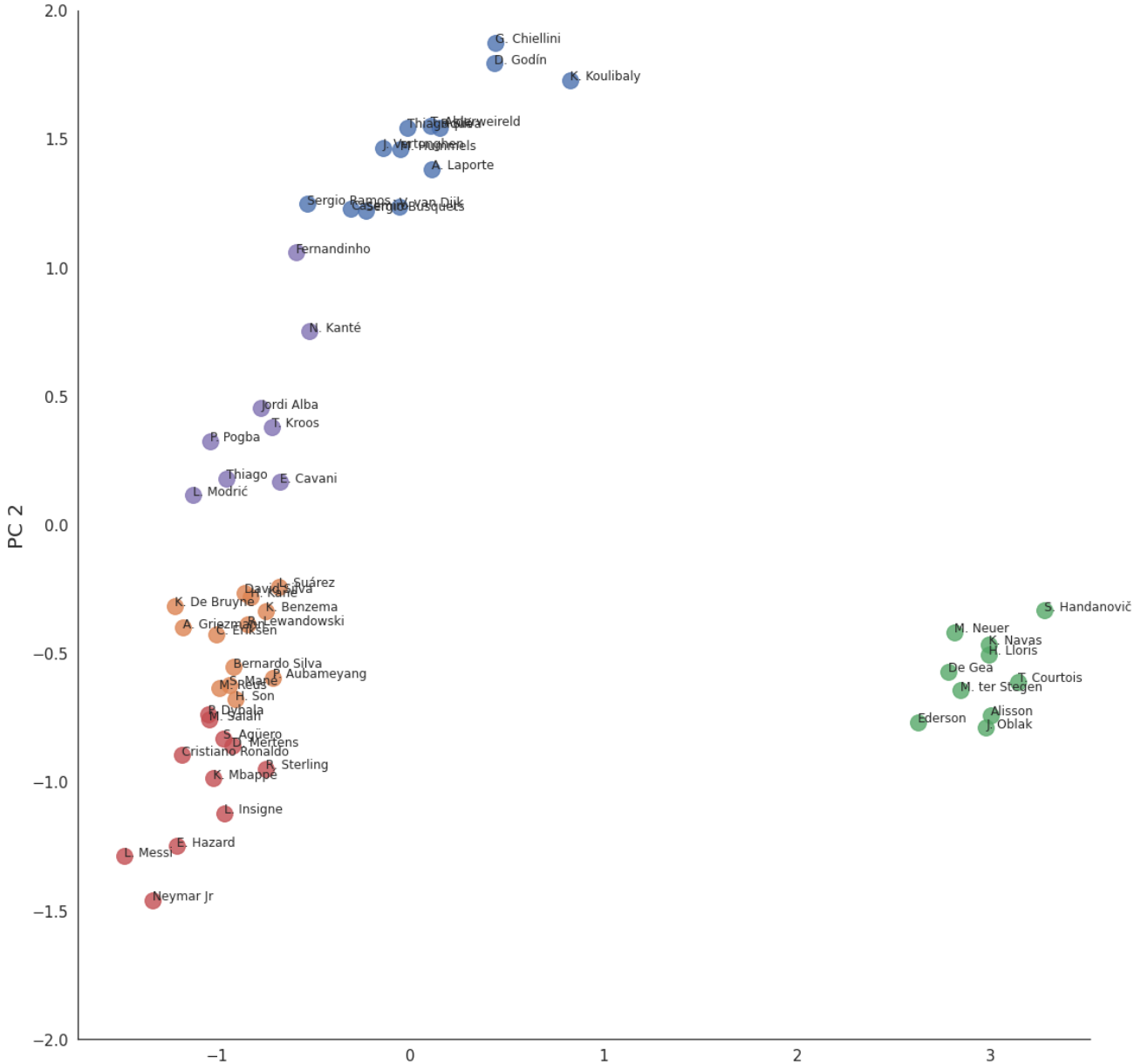
Puede ser utilizado como un paso previo a análisis predictivo. Genero clusters para darle una etiqueta a mis observaciones, luego hago predicción para nuevos datos

Kmeans Airbnb

https://shravan-kuchkula.github.io/nyc-airbnb-kmeans/#visualize-k-means-clusters-on-a-folium-map

https://towardsdatascience.com/use-data-science-to-find-your-next-airbnb-getaway-3cb9c8333ad1

Clusterización Jugadores de fútbol



https://towardsdatascience.com/grouping-soccer-players-with-similar-skillsets-in-fifa-20-part-1-k-means-clustering-c4a845db78bc

Aplicaciones

Investigación de mercados: segmentación de clientes, mercados, posicionamiento de productos

Análisis de redes sociales: reconocer comunidades

Ciencias sociales: Identificar estudiantes, empleados con características similares

Sistemas de recomendación: definir preferencias basado en características de los clusters

Análisis de crimen: identificar áreas con comportamientos de crimen similares

¿Para qué sirve?

•Resumir de alguna forma todas las variables utilizadas en una nueva variable de clases (clusters)

• Definición de una ‘variable latente’

•Conocer posibles perfiles de observaciones que existen en la base de datos

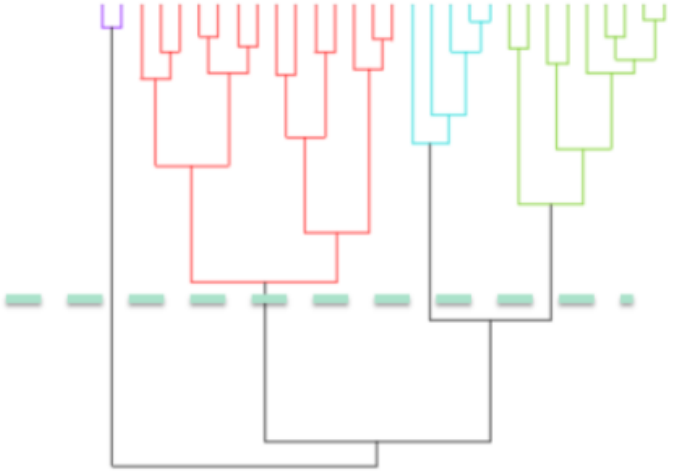
• Tomar decisiones de negocio basadas en la segmentación

Dos tipos de clusterización

- **Jerárquica**

  Crear una descomposición jerárquica de las observaciones basada en algún criterio (di)similitud

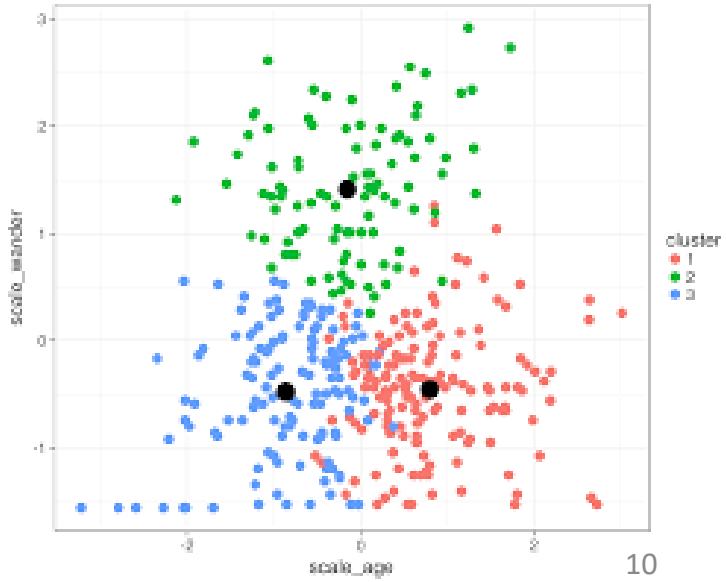
  - Aglomerativo
  - Divisivo



- **Particional**

  Construir particiones en los datos y evaluar, de acuerdo a un criterio, que tan buena es la partición

  K-means, PAM, entre otros…



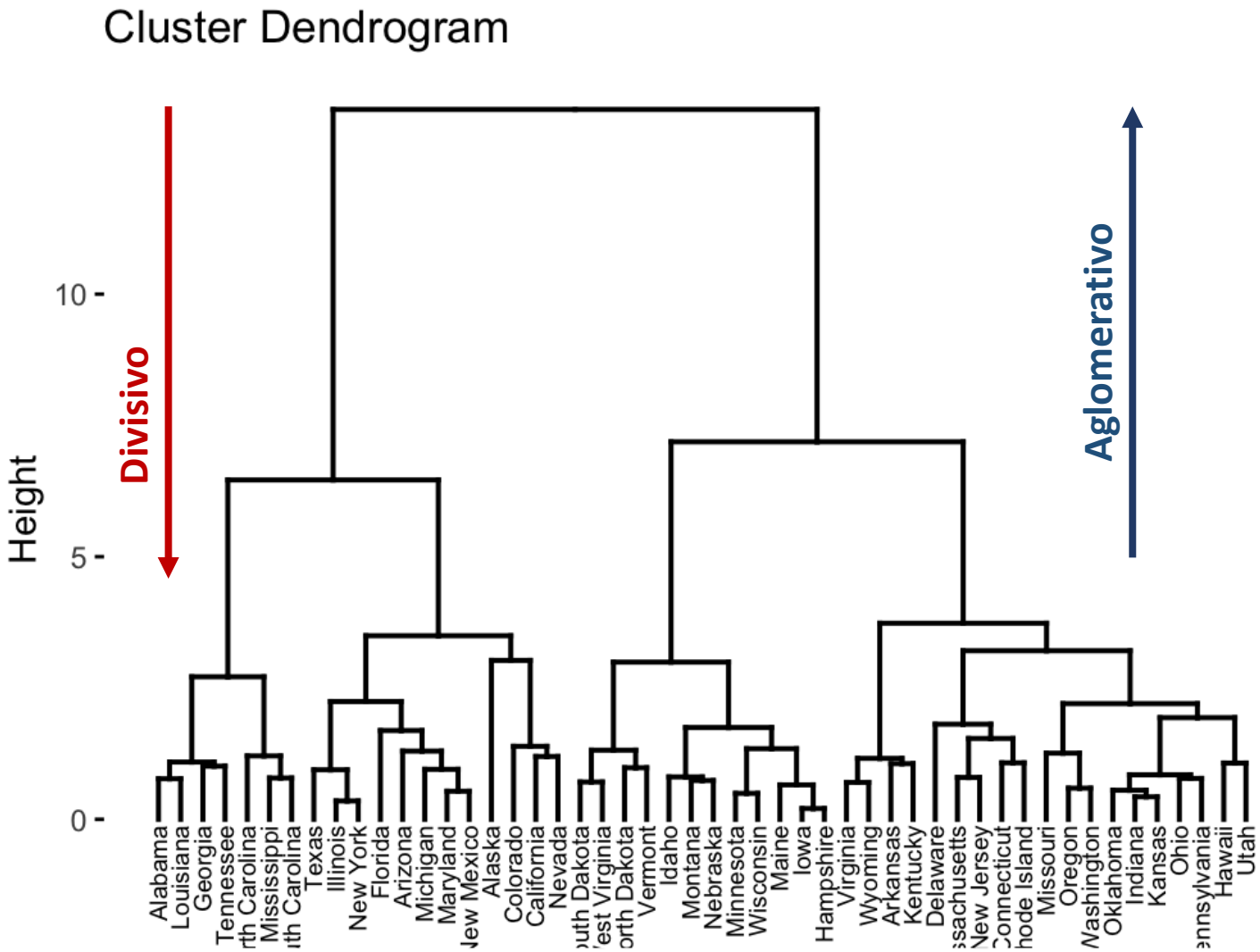
Clusterización Jerárquica - Aglomerativo

(De abajo hacia arriba)

Empieza con un cluster por cada observación

Va agrupando estos grupo individuales para formar nuevos clusters de dos o más observaciones

Continúa agrupando los clusters/observaciones hasta generar un único cluster con todas las observaciones



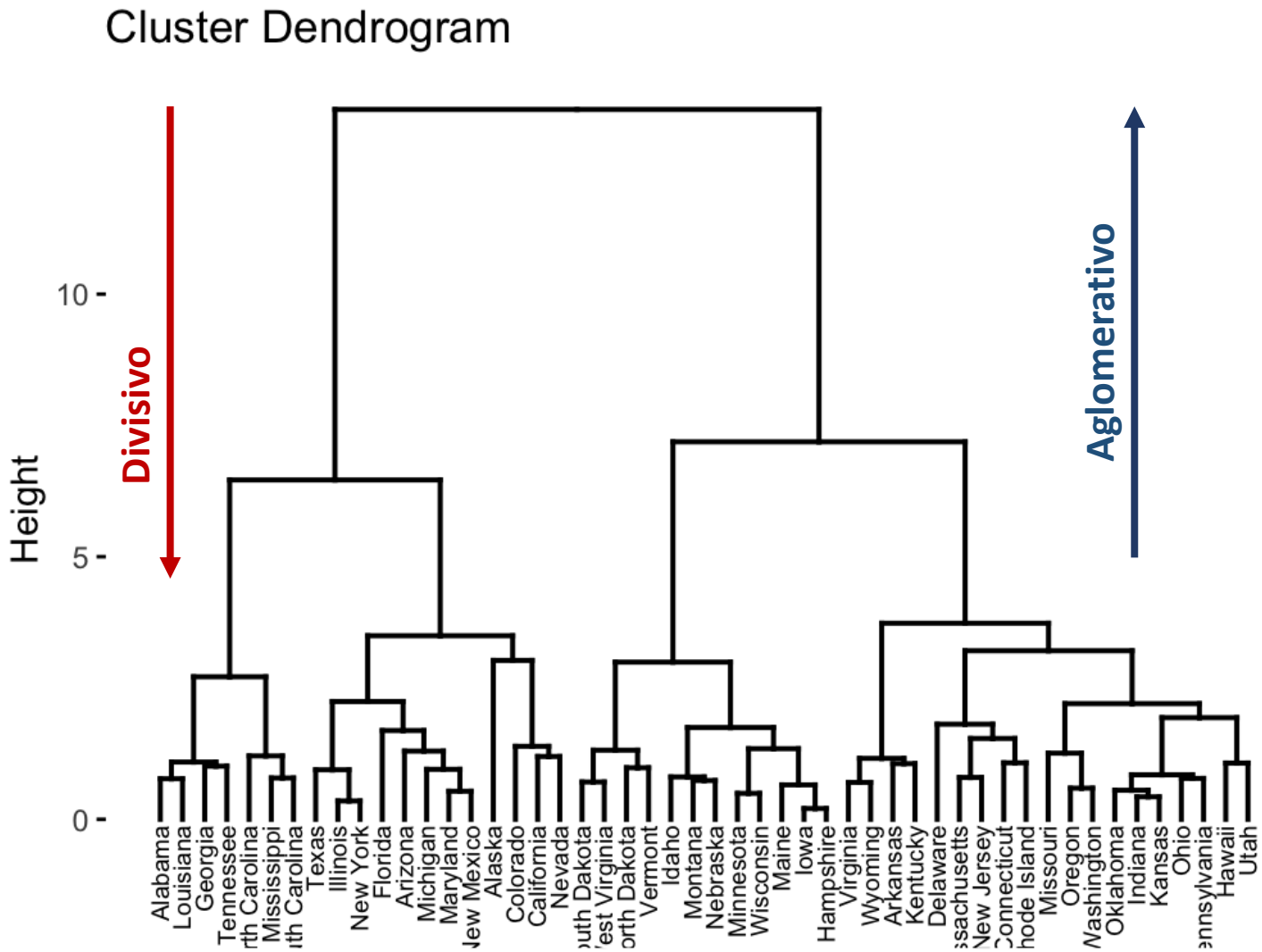
Clusterización Jerárquica - Divisivo

(De arriba hacia abajo)

Empieza con un cluster formado por todas las observaciones

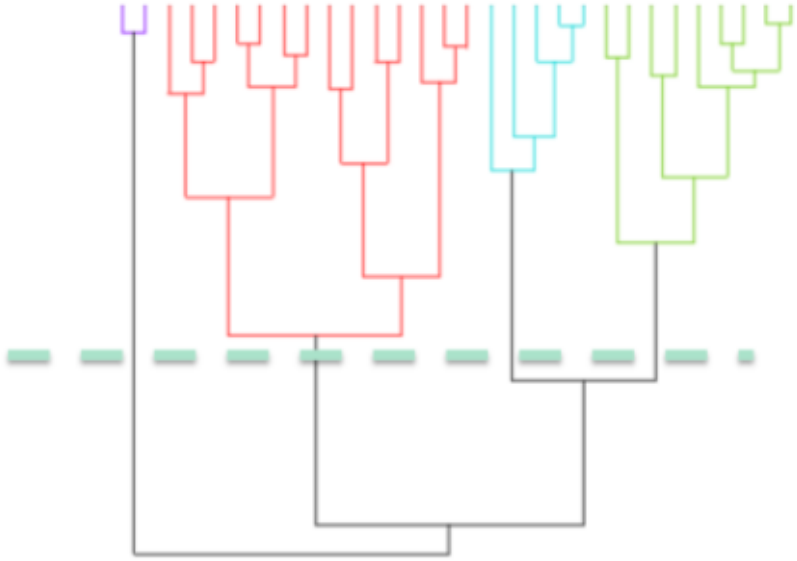
Va diviendo este gran cluster en clusters más pequeños

Continúa hasta generar tantos clusters como observaciones hay



Clusterización Jerárquica

Para decidir qué clusters deben combinarse (aglomerativo) o cuáles deben dividirse (divisivo), se necesita usar una medida de (di) similaridad entre las observaciones



La métrica utilizada generalmente es la distancia entre observaciones - clusters y un criterio de enlace (linkage) que cuantifica la (di)similitud de conjuntos como una función de la distancia entre pares de observaciones en los conjuntos

¿Cómo funciona?

Si es aglomerativo

1. Calcular la matriz de distancias/similitud

2. Cada observación define un cluster

3. Unir las dos observaciones/clusters más cercanas (de acuerdo a la matriz)
   Resultado 1: Un cluster de dos observaciones y (n-1) clusters de 1 observación

4. Actualizar la matriz de distancias/similitud

5. Repetir 3 y 4 hasta que quede un solo cluster

Distancia

La métrica de distancia puede ser

- Distancia euclideana
- Distancia euclideana^2
- Distancia Manhattan
- Distancia máxima
- Otras…

Debe cumplirse que

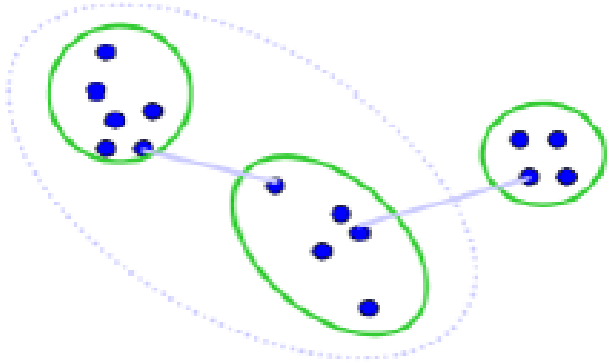
i. $d(x,y) \geq 0$	No negativa
ii. $d(x,y) = 0$	Si y solo si $x = y$
iii. $d(x,y) = d(y,x)$	Simetría
iv. $d(x,z) \leq d(x,y) + d(y,z)$	Desigualdad triangular

Tipos de Enlace

El criterio de enlace se define la distancia utilizada entre clusters

Enlace mínimo (simple): Evalúa la distancia entre clusters/observaciones como la **mínima**

$$D(A,B) = \min(d(a,b): a \in A, b \in B)$$

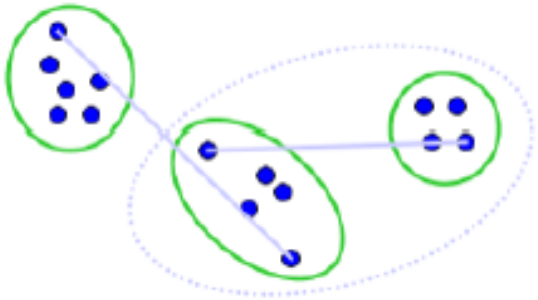


Lleva a clusters alargados

Tipos de Enlace

Enlace completo (máximo): Evalúa la distancia entre clusters/observaciones como la **máxima**

$$D(A, B) = \max(d(a, b): a \in A, b \in B)$$

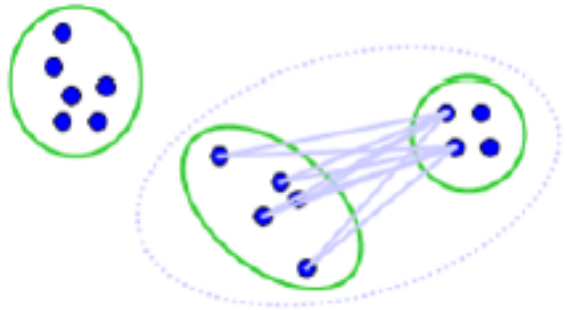


Lleva a clusters más compactos

Tipos de Enlace

Enlace promedio: Evalúa la distancia entre clusters/observaciones como el **promedio**

$$D(A,B) = \frac{1}{|A||B|}\sum_{A}\sum_{B} d(a,b)$$

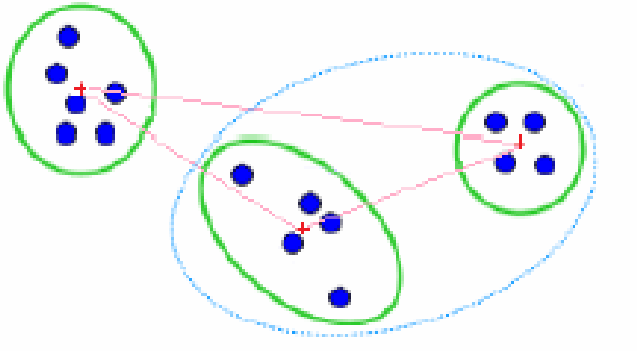


Es más robusto, especialmente en situaciones de ruido

Tipos de Enlace

Enlace de centroide: Evalúa la distancia entre clusters/observaciones como la distancia entre sus **centroides**

$$D(A, B) = d(a_c, b_c)$$



También es más robusta, require de la definición del centroide

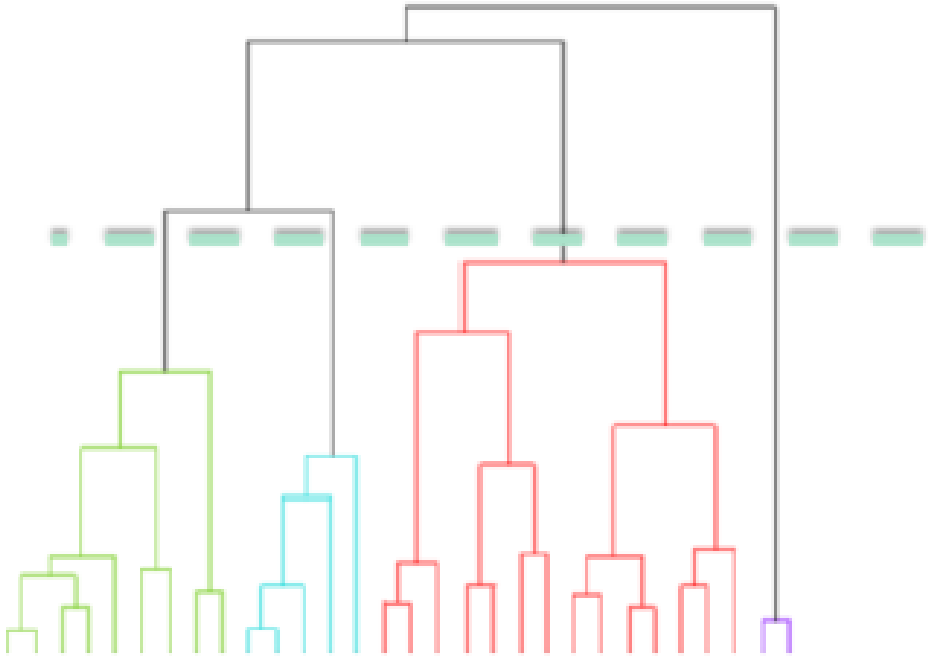
Tipos de Enlace

Ward: Usa el criterio de varianza minima. En cada paso del algoritmo se busca la (des)union que lleve al incremento mínimo de la varianza dentro de los clusters

$$\min SSE = \sum_{k=1}^{K} \sum_{i=1}^{n_k} \|x_{ki} - \mu_k\|^2$$

Dendograma

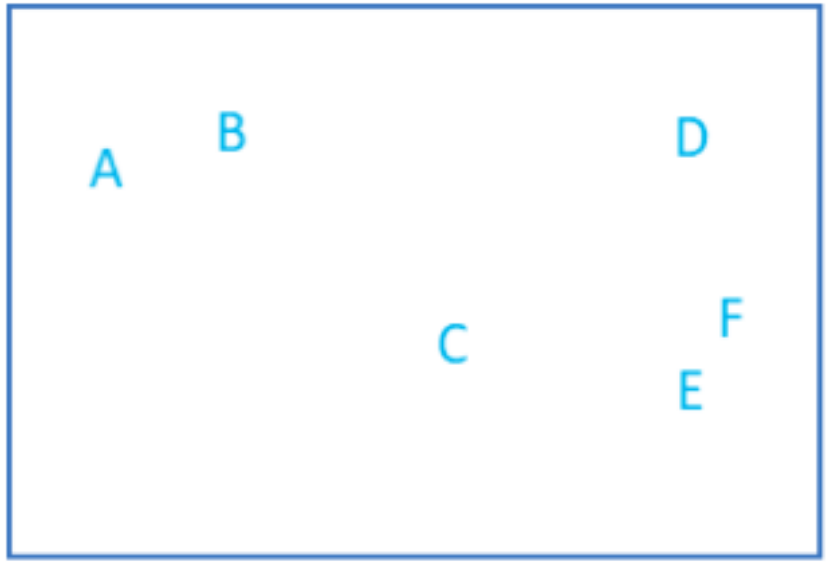
La representación de jerarquía se da en el dendograma

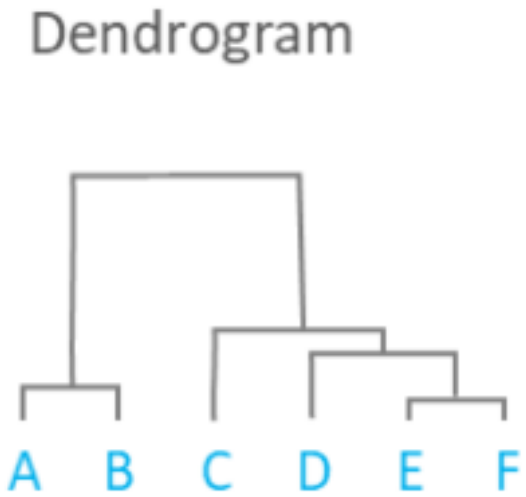


Puede ser utilizado para definir el número de clusters (al generar un corte en el dendograma)

Dendograma

La representación de jerarquía se da en el dendograma

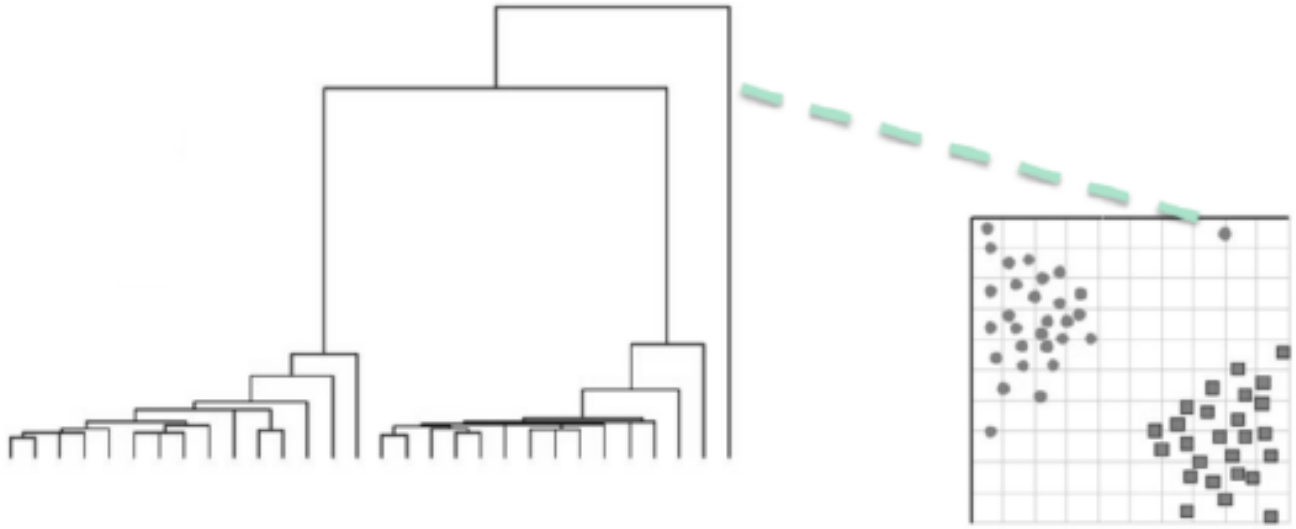




Puede ser utilizado para definir el número de clusters (al generar un corte en el dendograma)

Dendograma

Representación de la jerarquía



El dendograma también puede ser un herramienta para identificar datos atípicos o anomalías

Ventajas agrupamiento jerárquico

- Fácil de implementar (matemáticamente)
- Se obtiene una jerarquía sobre las observaciones
- No se necesita una definición de número de clusters apriori, el método mismo puede sugerir este número
- La estructura jerárquica muchas veces es intuitiva con el conocimiento experto

Desventajas agrupamiento jerárquico

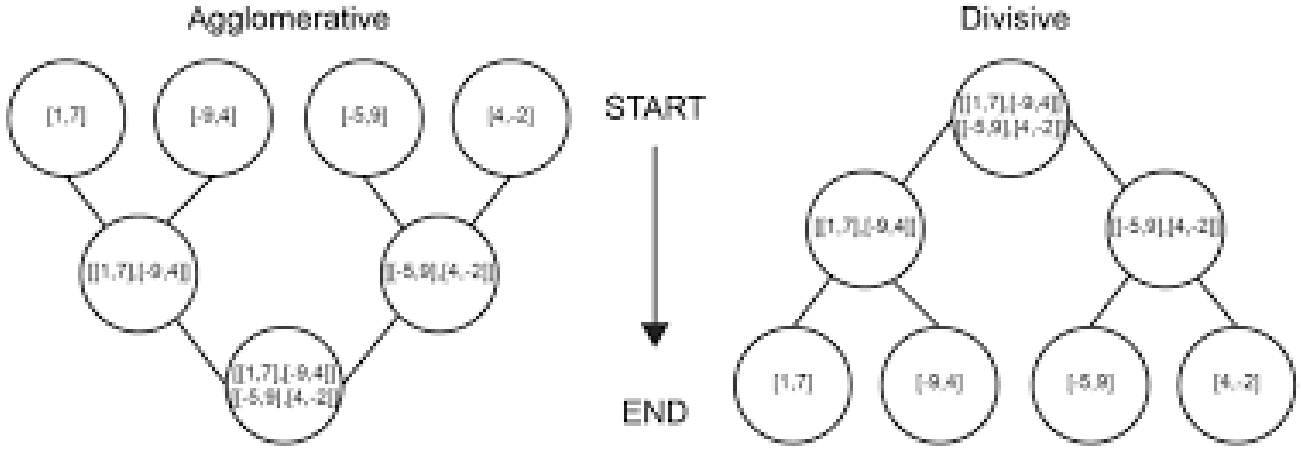
- Decisiones arbitrarias frente a la distancia y enlace utilizado

- No es fácil de escalar, aunque es rápido en muchos de los casos

¿Aglomerativo o divisivo?

Casi todas las implementaciones utilizan el aglomerativo

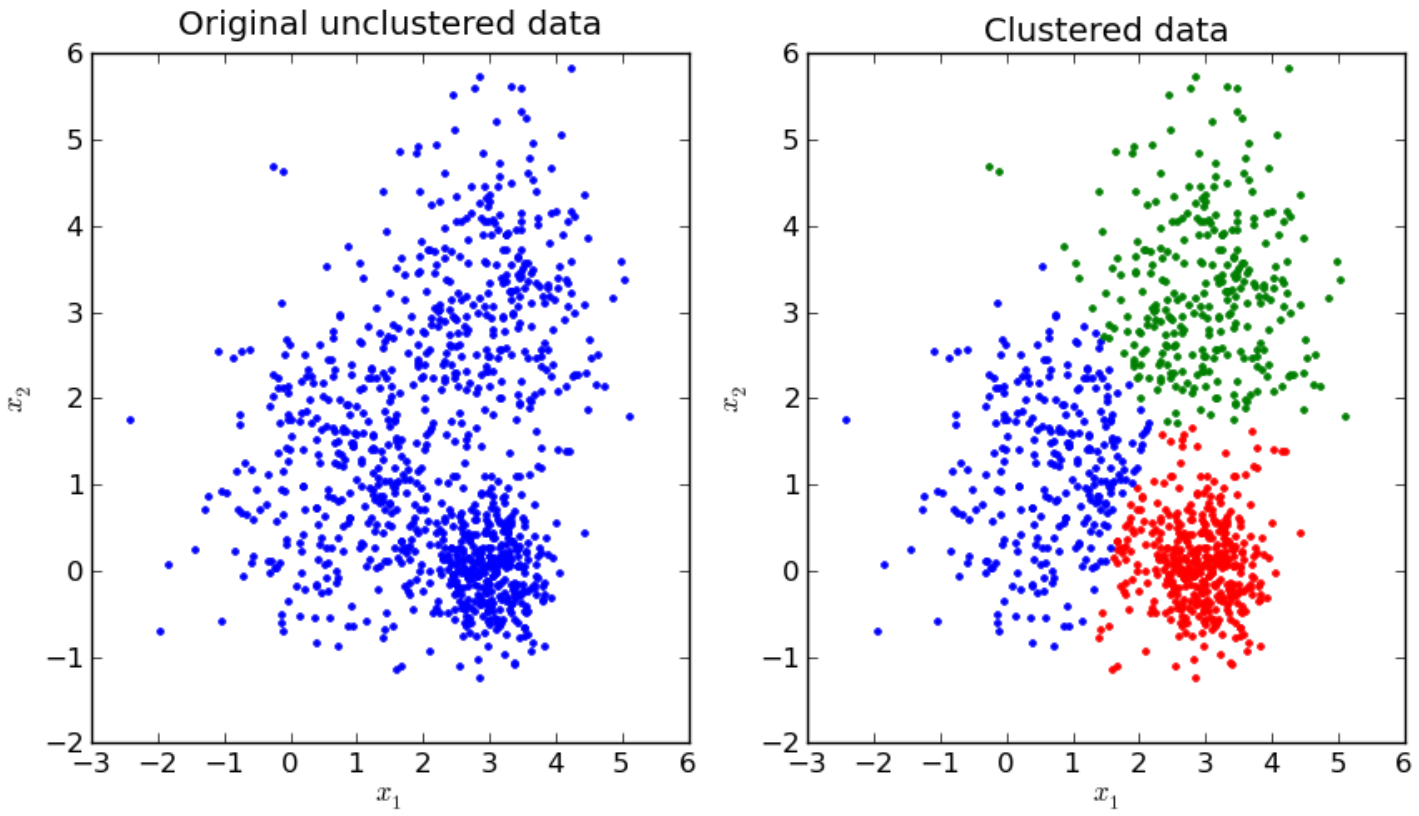
La clusterización divisiva resulta ser más costosa computacionalmente



Clusterización particional

No-jerárquica, cada observación es ubicada exactamente en uno de k-clusters

El analista usualmente decide el número de clusters



Uno de los algritmos mas conocidos(usados) es el k-means (k-medias)

Clusterización particional

Algoritmo K-means (k-medias)

1. Definir el valor de k: número de clusters

2. Inicializar los centros de los k clusters. Los centros pueden ser escogidos aleatoriamente o de acuerdo al conocimiento del analista

3. Decidir para cada observación a qué cluster pertenece, basado en una distancia**. La observación se asignará al cluster con el centro más cercano a ella.**

4. Recalcular los centros de los clusters (centroides)

5. Repetir 3 y 4 hasta que ninguna de las observaciones cambie de cluster

K-means

Elegir los centros al azar (puede ser en el espacio vectorial de las variables o unas observaciones específicas como centros) :

K-means

Calcular las distancias de cada observación a los centros y asignarlas al centro más cercano

K-means

Recalcular los centroides

K-means

Reasignar nuevamente las observaciones a los clusters de acuerdo a las distancias a los nuevos centros

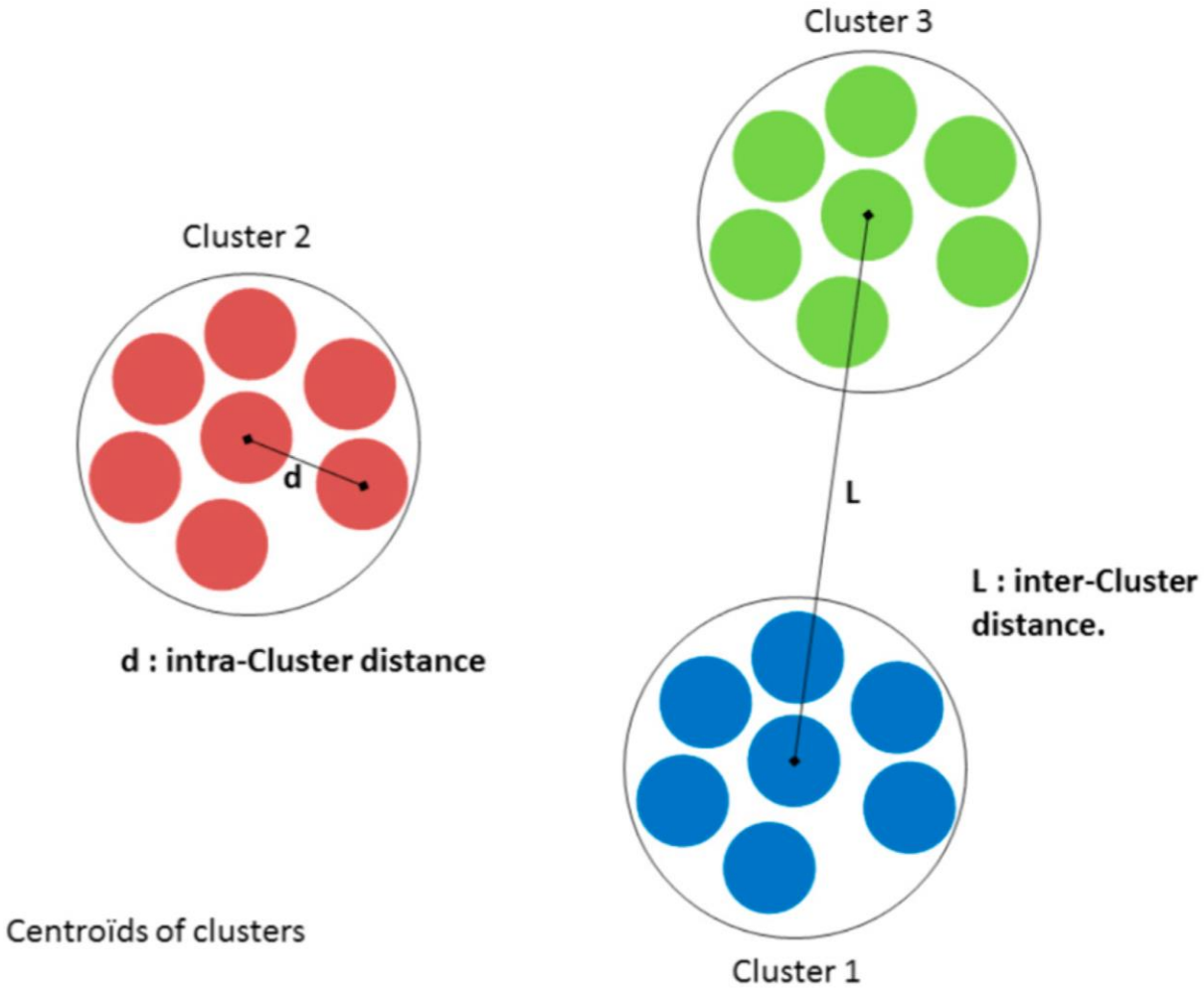
K-means

Recalculamos de nuevo los centroides y asignamos las observaciones. Si no hay cambios, termina el algoritmo

K-means

Para llegar a una partición K-means optimiza la homogeneidad intra-cluster.

$$\min SSE = \sum_{k=1}^{K} \sum_{i=1}^{n_k} \|x_{ki} - \mu_k\|^2$$

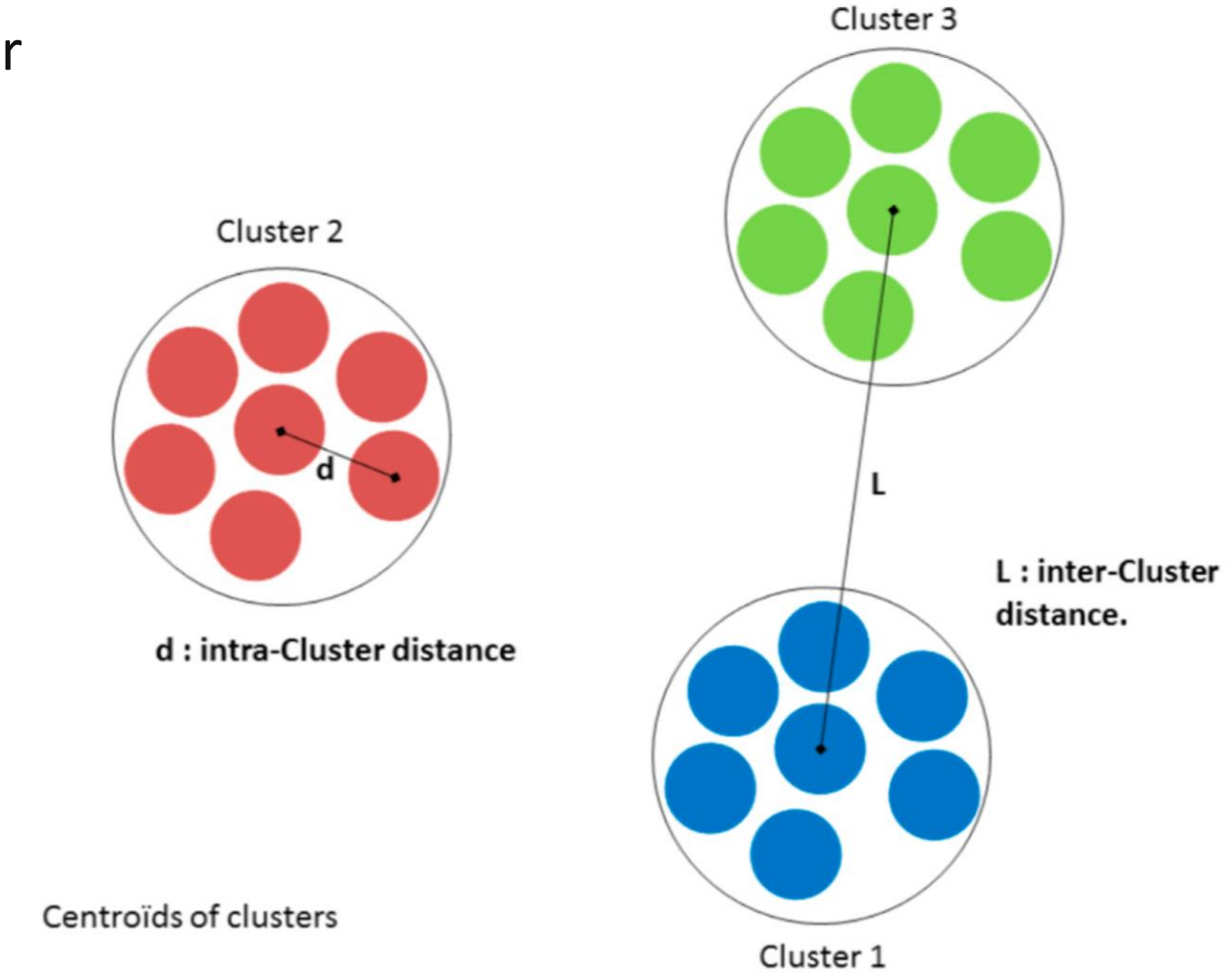


Minimiza la distorción total: suma total de las distancias de los puntos a su centroide

K-means

Se busca tener alta homogenidad intra-cluster y heterogeneidad entre clusters

$$\min SSE = \sum_{k=1}^{K} \sum_{i=1}^{n_k} \|x_{ki} - \mu_k\|^2$$

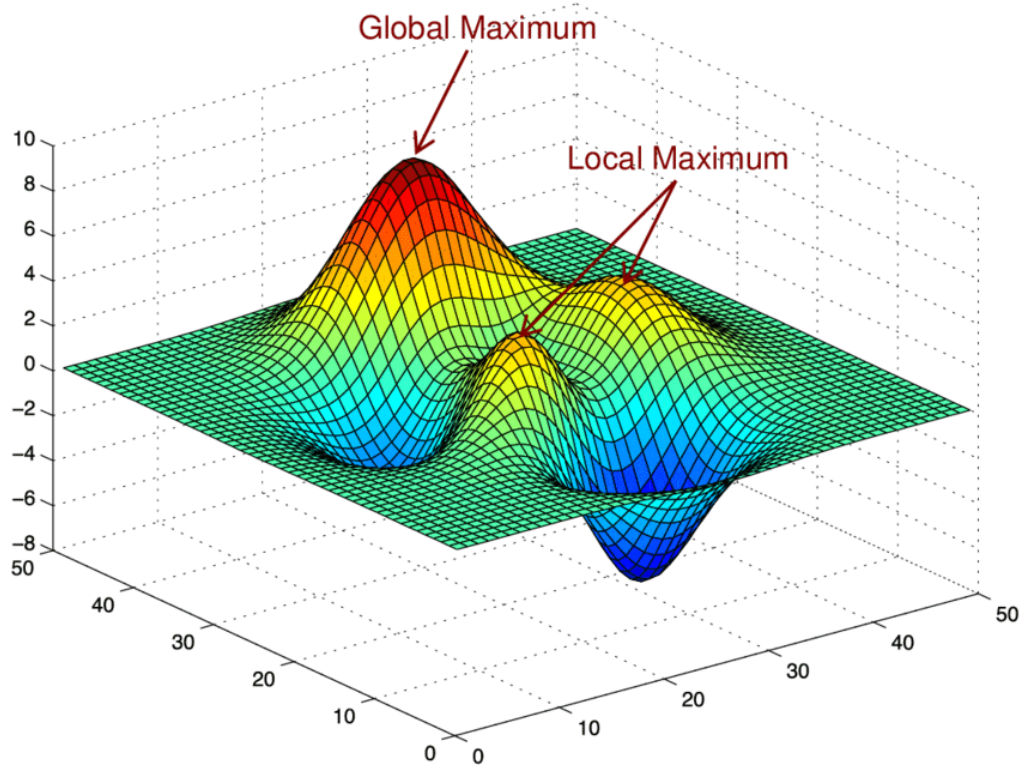


Observaciones dentro de un mismo cluster deben ser muy similares, y muy diferentes a observaciones en otros clusters

K-means

Lograr una única solución en este proceso de optimización es complejo, no necesariamente se llega al óptimo global.

- Usar siempre una semilla al ejecutarlo
- Validar y hacer multiples corridas



Sólo en 1 dimension se llega a una única solución

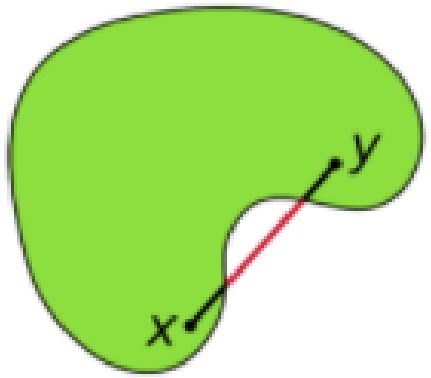
Fortalezas K-means

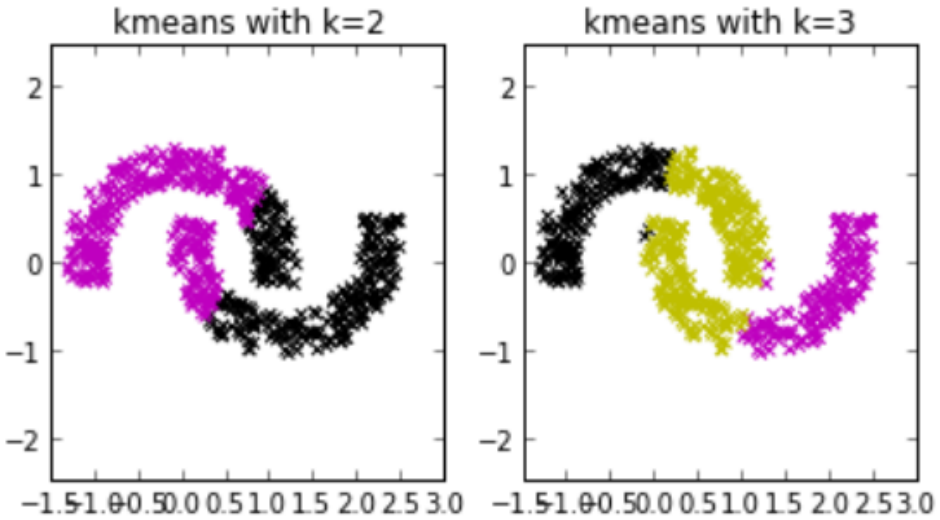
- Muy sencillo de implementar
- Intuitivo: la función objetivo que se optimiza (homogeneidad dentro)
- Relativamente eficiente

Debilidades K-means

- Sólo se puede aplicar cuando está definida la media para calcular los centroides. Sólo variables numéricas
- Qué hacer con datos categóricos?
- Usualmente termina en óptimos locales, correr varias veces con inicios diferentes
- Requiere una especificación de k apriori
- Sensible a datos atípicos o ruido
- No es lo ideal para generar clusters de estructura no convexas

K means y estructuras no convexas





https://pafnuty.wordpress.com/2013/08/14/non-convex-sets-with-k-means-and-hierarchical-clustering/#:~:text=%E2%80%9CK%2Dmeans%20can't%20handle%20non%2Dconvex%20sets,is%20also%20within%20the%20object.

Otros métodos

- K-medoids – PAM (Partitioning around medoids)
  Igual que kmeans, no se usa el centroide sino el punto más central

- CLARA

- K-modas

K-means++

Es un algoritmo para escoger los centros

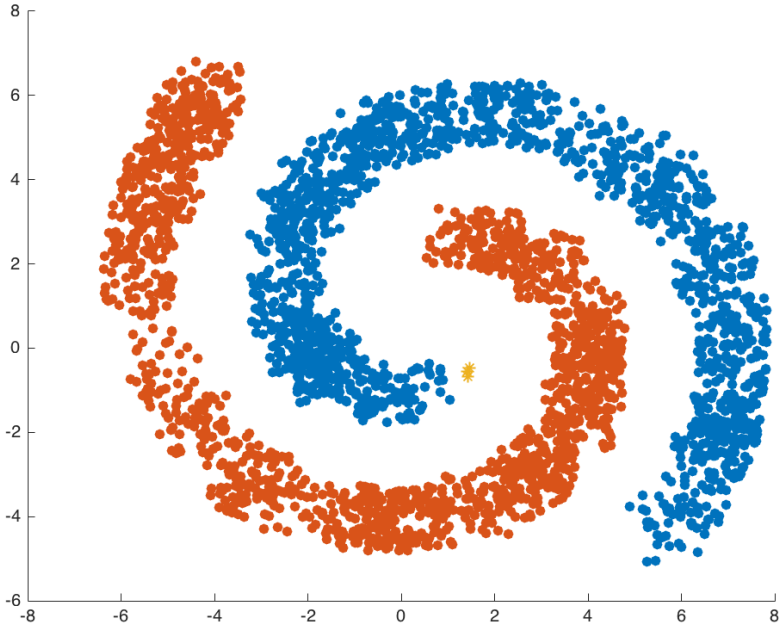
1. Escoger un centro, en el espacio de variables o entre las observaciones, utilizando una variable aleatoria uniforme (un punto al azar)
2. Calcular distancias de cada observación al centro definido, d(x,c)
3. Escoger otro centro, usando una probabilidad de selección proporcional a d(x,c)
4. Repetir 2 y 3 hasta que escojan los k centros
5. Ajustar k-means con los centros definidos en 4

El objetivo es definir los centros de los clusters iniciales lo más lejos possible uno de los otros

DBSCAN
“Density-based spatial clustering of applications with noise”

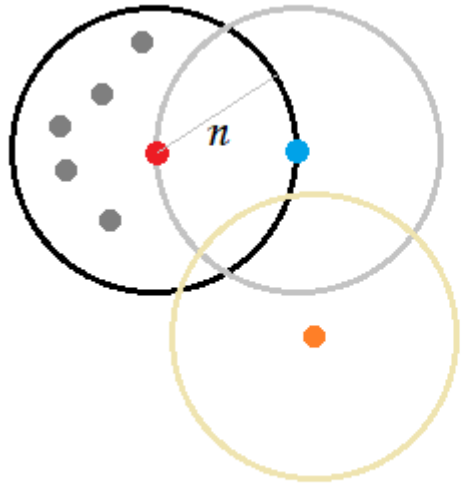
Agrupamiento especial basado en densidad de aplicaciones con ruido

Agrupa puntos que estan muy juntos (puntos con muchos vecinos cercanos)



Resalta como outliers aquellos puntos en regiones de baja densidad (puntos cuyos vecinos estan muy lejos)

DBSCAN

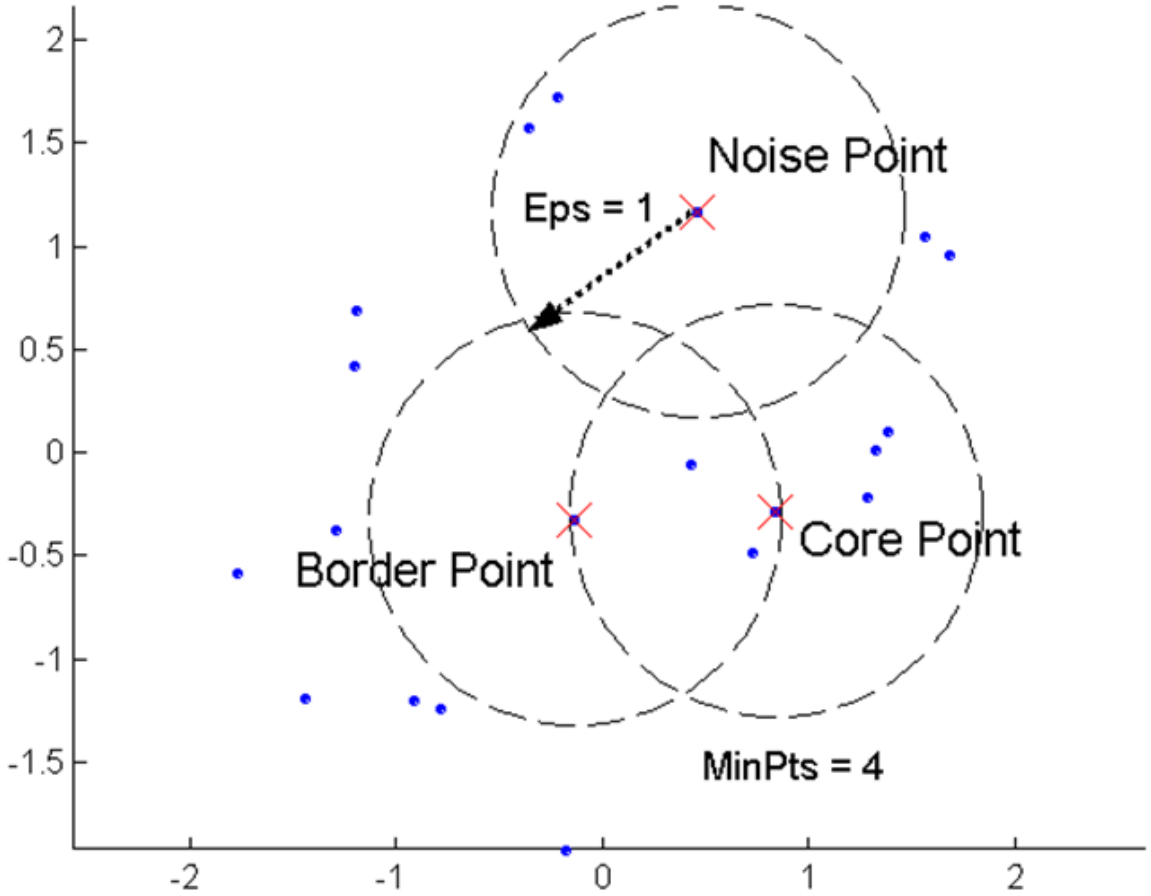


- Core Point
- Border Point
- Noise Point

n = Neighbourhood
m = 4

DBSCAN CLUSTERING

Abhijit Annaldas



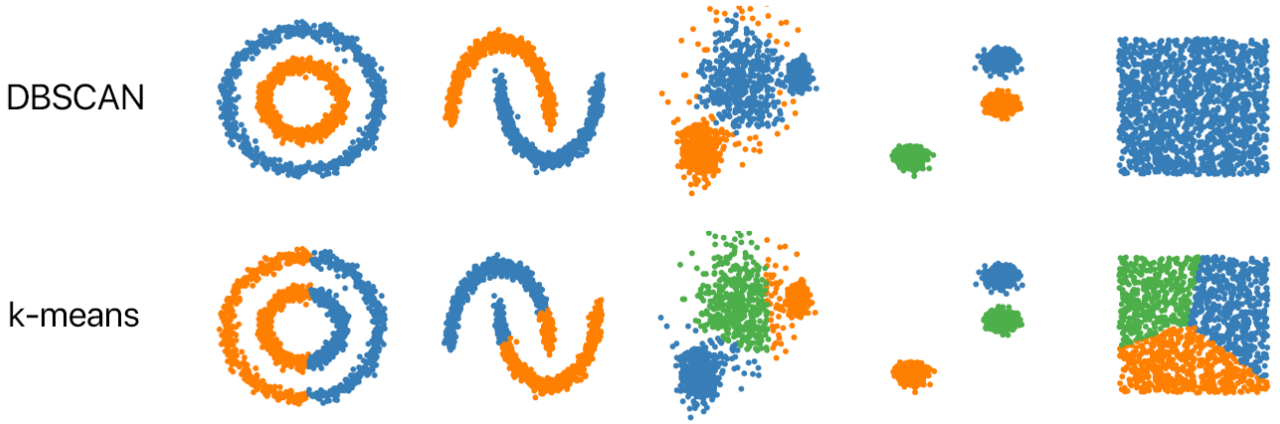
Épsilon es el radio que define el área para definir los vecinos de un punto

minPts es el mínimo de puntos por vecindad que determina si es un core point (muchos puntos vecinos), border point (pocos vecinos pero es vecino de un core point) o un noise point (sin vecinos cercanos)

DBSCAN

Ventajas

- Robusto frente a ruido y outliers
- Los clusters pueden tener distintas formas (convexas)
- No requiere definir el número de clusters



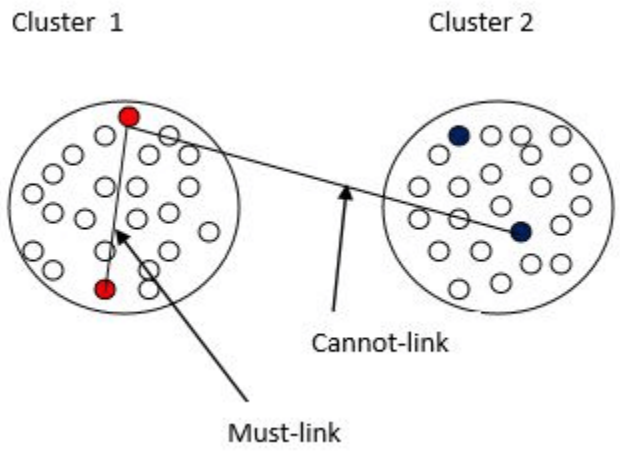
Desventajas

- No es tan útil en altas dimensionalidades
- Difícil con densidades variables
- Pueden quedar observaciones sin cluster

Conocimiento Experto

¿Cómo podríamos incluir el conocimiento del negocio en los algoritmos?

Se pueden utilizar restricciones de link obligatorio (must-link) y links no posibles (can't-link), quienes deberían o no estar en el mismo cluster?



Si se ponen muchas restricciones puede llevar a soluciones inviables (problemas de convergencia)
También podemos hacer una sugerencia de inicio de algoritmos

Conclusiones

- Una gran herramienta de análisis
- Permite encontrar patrones y hallazgos, que a “simple vista” no son fáciles de determinar
- Puede ser un paso previo a modelos predictivos y aprendizaje supervisado
- Fácil de explicar