ntic master School

UNIVERSIDAD COMPLUTENSE DE MADRID

Técnicas no supervisadas: Clustering

Minería de Datos y Modelización Predictiva

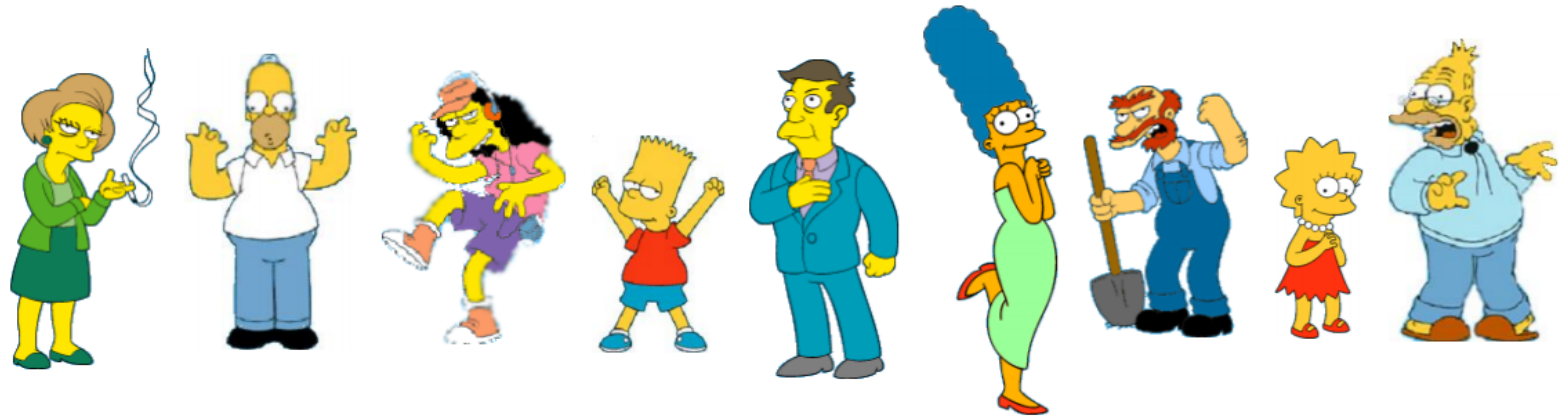
Máster Big Data, Data Science & Business Analytics

Universidad Complutense de Madrid

Curso 2024-2025

IDEA INTUITIVA

- El término clustering hace referencia a un amplio abanico de técnicas no supervisadas cuya finalidad es **encontrar patrones o grupos** (clusters) dentro de un conjunto de observaciones.
- Las particiones se establecen de forma que las observaciones que están dentro de un mismo grupo son similares entre ellas y distintas a las observaciones de otros grupos.
- Se trata de un método no supervisado ya que el proceso ignora la variable respuesta que indica a que grupo pertenece realmente cada observación (si es que existe tal variable).
- Hay que tener en cuenta que el clustering es eminentemente subjetivo!



FIJAR VARIOS CRITERIOS

- Una medida de **proximidad** entre elementos.
  - Similitud: $s(x_j, x_k)$ mayor cuanto más póximos
  - Disimilitud: $d(x_j, x_k)$ menor cuanto más póximos
- Un criterio para **evaluar** la "bondad" de la solución.
- Un algoritmo de **optimización** en base a las medidas anteriores.

BASADAS EN DISTANCIAS Y CORRELACIONES

- **Distancia Euclídea** –> $d_{euc} = \sqrt{\sum_i (x_{ij} - x_{ik})^2}$
- **Distancia de Manhattan** –> $d_{mht} = \sum_i |x_{ij} - x_{ik}|$
- **Distancia de Mahalanobis** (realmente es mixta ya que pondera la distancia por la correlación) –> $d_{mhala} = \sqrt{(x_j - x_k)^t S^{-1} (x_j - x_k)}$
- **Distancia de Pearson** –> $d_{pear} = 1 - r$

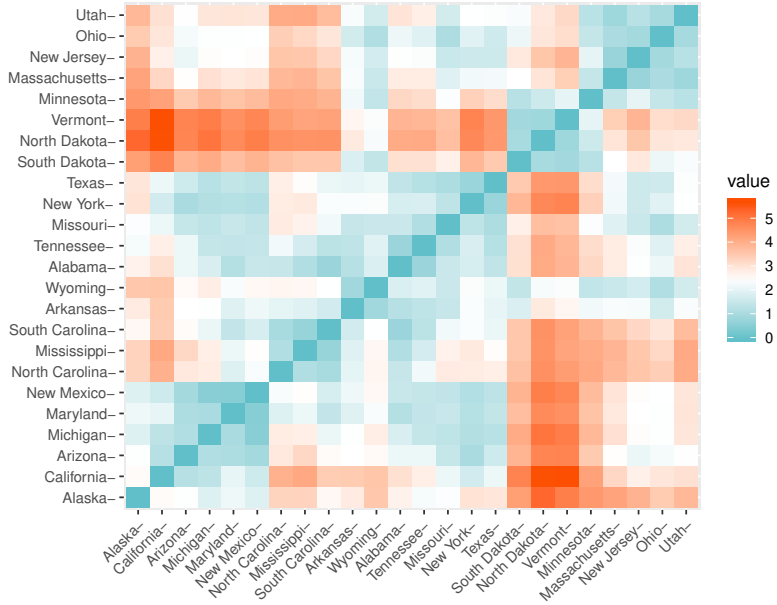
BASADAS EN ACUERDO

- **Distancia de Jaccard (dicotómicas)**
- **Distancia de Spearman**
- **Distancia de Kendall**

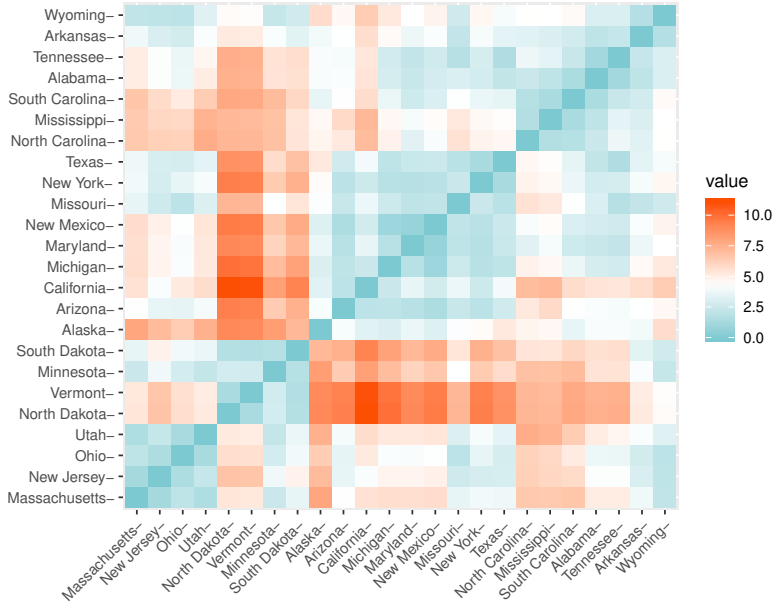
PARA DATOS DE NATURALEZA MIXTA

- **Coeficiente de similitud de Gower** Extensión muy útil que distingue los tipos de variables y calcula medida de acuerdo o distancia según el caso.

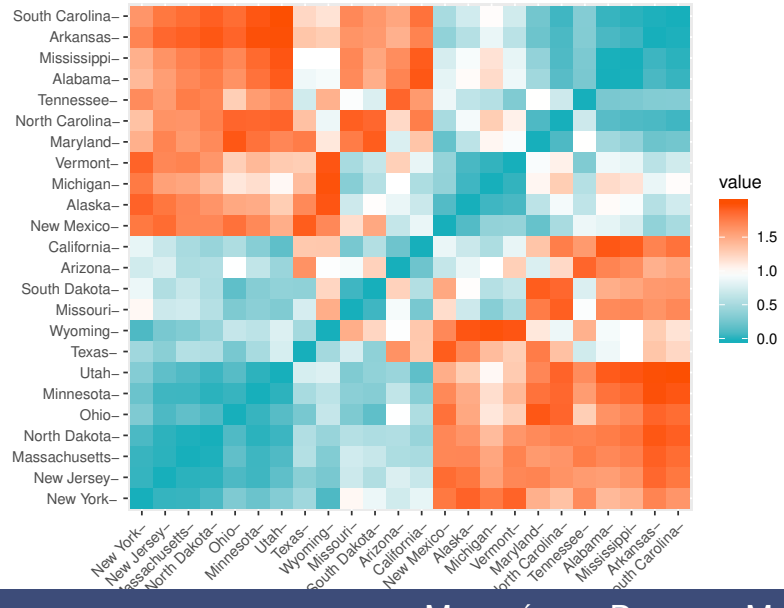
Distancia Euclídea



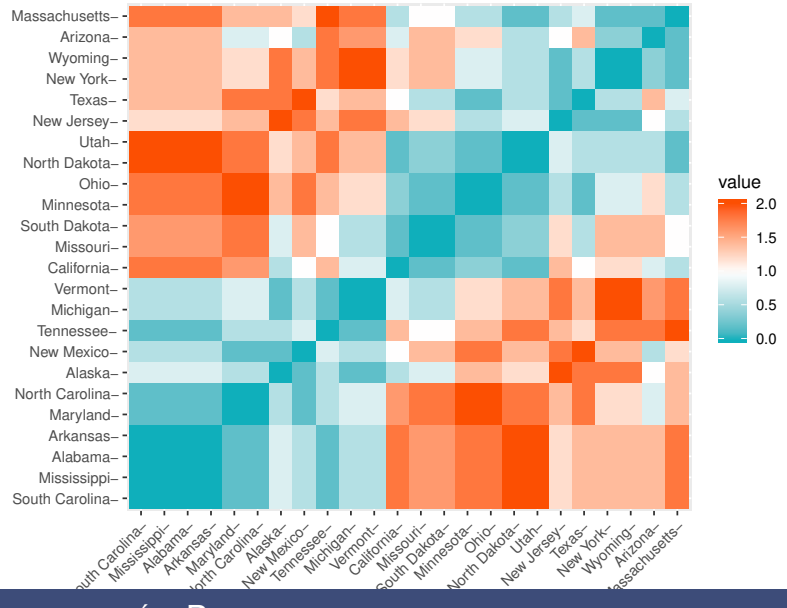
Distancia Manhattan



Distancia Pearson



Distancia Spearman



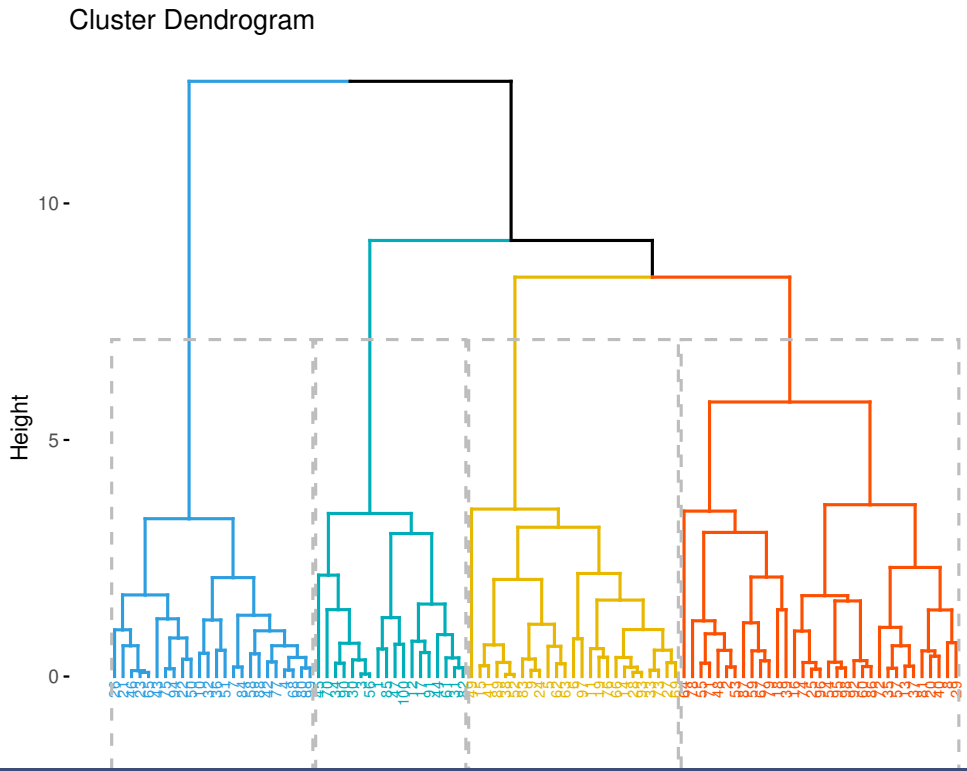
CLASIFICACIÓN

Dada la popularidad del clustering en disciplinas muy distintas (genómica, marketing, etc), se han desarrollado multitud de variantes y adaptaciones de sus métodos y algoritmos. Pueden diferenciarse tres grupos:

- **Partitioning Clustering**: Este tipo de algoritmos requieren que el usuario especifique de antemano el número de clusters que se van a crear (K-means, K-mediods, CLARA).
- **Hierarchical Clustering**: Este tipo de algoritmos no requieren que el usuario especifique de antemano el número de clusters. (agglomerative clustering, divisive clusterig).
- Métodos que combinan o modifican los anteriores (hierarchical K-means, fuzzy clustering, model based clustering y density based clustering).

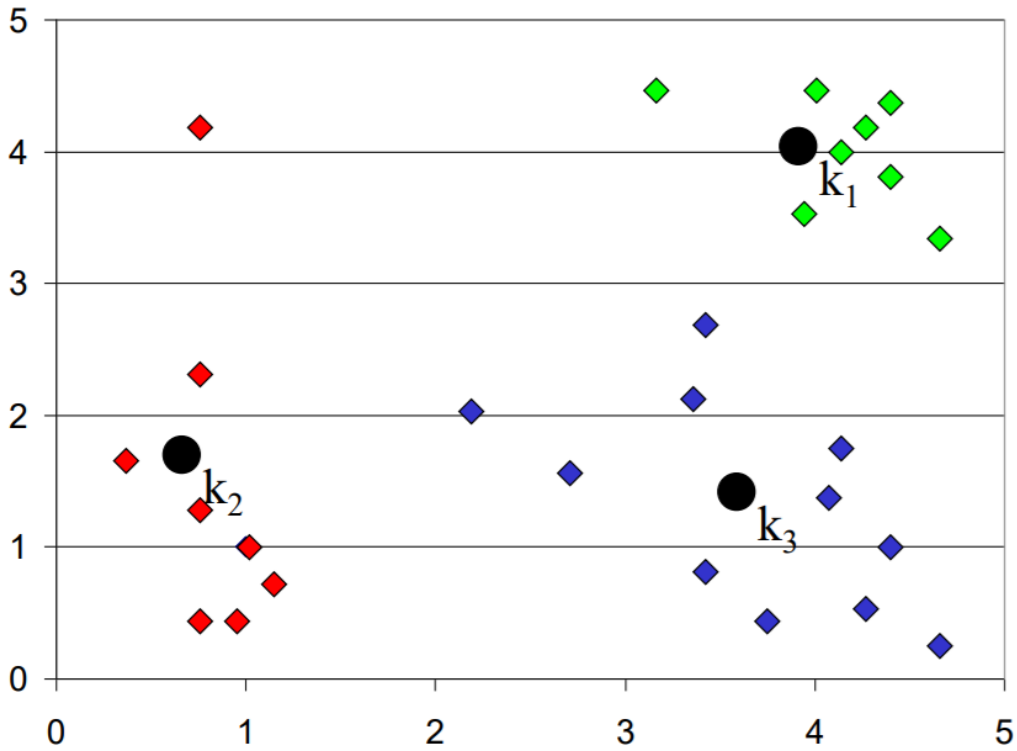
LOS DOS MÉTODOS PRINCIPALES

- **k-means clustering**, buscamos dividir las observaciones en un número de conglomerados previamente especificados.
- **Hierarchical clustering**, no sabemos de antemano cuántos clusters queremos; de hecho, terminamos con una representación visual en forma de árbol de las observaciones, llamada **dendrograma**, que nos permite ver de una vez las agrupaciones obtenidas para cada número posible de conglomerados, de 1 a n.



K-means. Idea intuitiva

- El método K-means clustering (MacQueen, 1967) agrupa las observaciones en K grupos distintos, donde el número K ha de ser especificado a priori.
- EL algoritmo encuentra los K mejores clusters, entendiendo como mejor cluster aquel cuya **varianza interna** (intra-cluster variation) sea lo más **pequeña** posible.
- Se trata, por lo tanto, de un problema de optimización, en el que se reparten las observaciones en k clusters de forma que la suma de las varianzas internas de todos ellos sea lo menor posible.



- La idea detrás de la agrupación K-means es que una buena agrupación es aquella para la cual la variación intra-cluster e más pequeña posible.
- La variación intra-cluster para el cluster Ck es una medida $W(C_k)$ de la cantidad por la cual las observaciones dentro de un cluster difieren entre sí.
- Por lo tanto, queremos resolver el problema

$$minimizar_{C_1,\ldots,C_k} = \sum_{k=1}^{K} W(C_k)$$

- En palabras, esta fórmula dice que queremos dividir las observaciones en k grupos de manera que la variación total dentro del grupo, sumada a todos los k grupos, sea lo más pequeña posible.

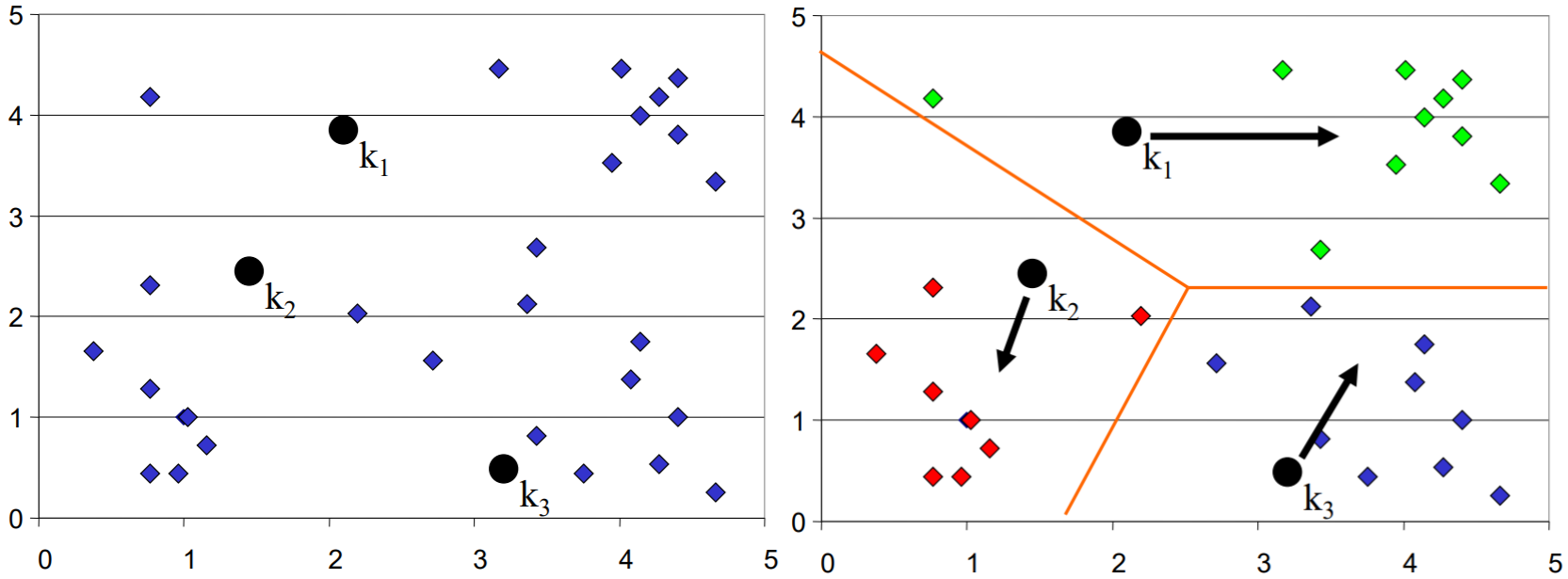
VARIACIÓN INTRA-CLUSTER

El algortimo estándar de Hartigan-Wong algorithm (Hartigan and Wong 1979), define la variación intra cluster como la suma de los cuadrados de las distancias Euclídeas entre cada item x_i y su correspondiente centroide μ_k.

$$W(C_k) = \sum_{x_i \in C_k} (x_i - \mu_k)^2$$

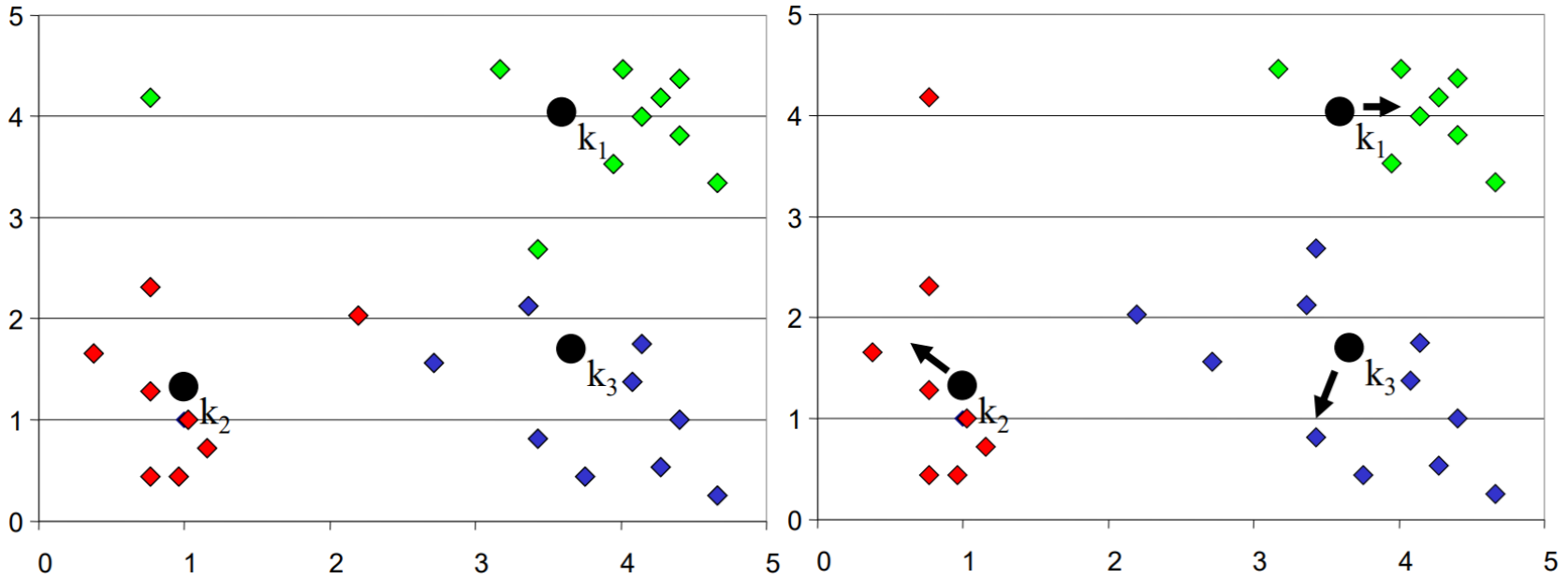
PASOS DEL ALGORITMO K-MEANS

1. Especificar el número K de clusters que se quieren crear.
2. Seleccionar de forma aleatoria k observaciones del conjunto de datos como centroides iniciales.
3. Asignar cada una de las observaciones al centroide más cercano.
4. Para cada uno de los K clusters recalcular su centroide.
5. Repetir los pasos 3 y 4 hasta que las asignaciones no cambien o se alcance el número máximo de iteraciones establecido.



PASOS DEL ALGORITMO K-MEANS

1. Especificar el número K de clusters que se quieren crear.
2. Seleccionar de forma aleatoria k observaciones del conjunto de datos como centroides iniciales.
3. Asignar cada una de las observaciones al centroide más cercano.
4. Para cada uno de los K clusters recalcular su centroide.
5. Repetir los pasos 3 y 4 hasta que las asignaciones no cambien o se alcance el número máximo de iteraciones establecido.



- K-mans es uno de los métodos de clustering más utilizados por la sencillez y velocidad de su algoritmo, pero presenta una serie de limitaciones.
- Requiere que se indique de antemano el número de clusters que se van a crear.
  - Esto puede ser complicado si no se dispone de información adicional sobre los datos con los que se trabaja.
  - Una posible solución es aplicar el algoritmo para un rango de valores k y evaluar con cual se consiguen mejores resultados, por ejemplo, menor suma total de varianza interna.
- Las agrupaciones resultantes pueden ser áltamente sensibles a la asignación aleatoria de los centroides iniciales.
  - Para minimizar este problema se recomienda repetir el proceso de clustering entre 20-50 veces y seleccionar como resultado definitivo el que tenga menor suma total de varianza interna. Aun así, no se garantiza que para un mismo conjunto de datos los resultados sean exactamente iguales.
- Presenta problemas de robustez frente a outliers. La única solución es excluirlos o recurrir a otros métodos de clustering más robustos como K-medoids (PAM).

CARACTERÍSTICAS Y TIPOS

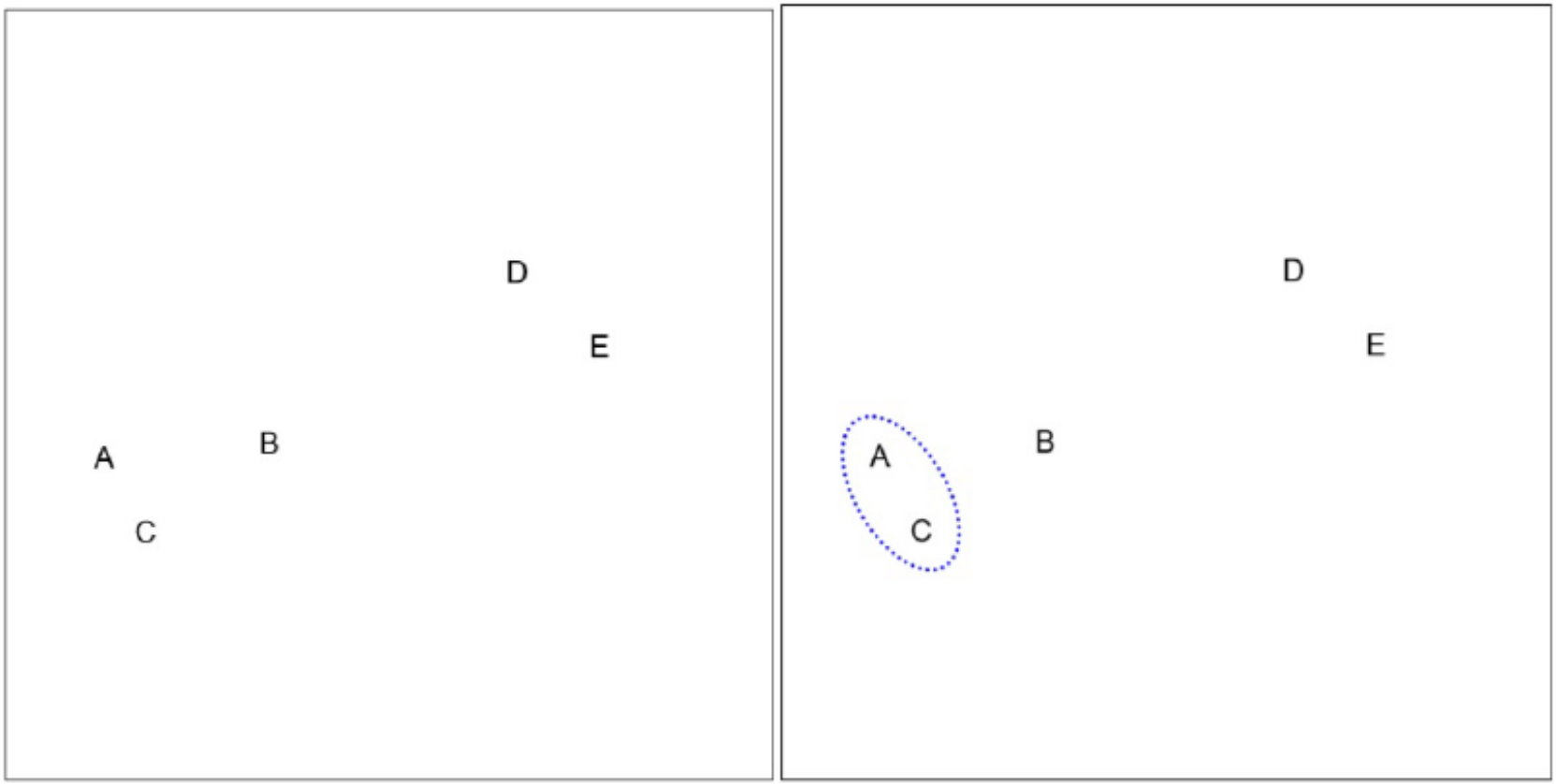
Cluster jerárquico (Hierarchical clustering) es una alternativa a los métodos de particionamiento clustering que no requiere que se pre-especifique el número de clusters. Los métodos que engloba el cluster jerárquico se subdividen en dos tipos dependiendo de la estrategia seguida para crear los grupos:

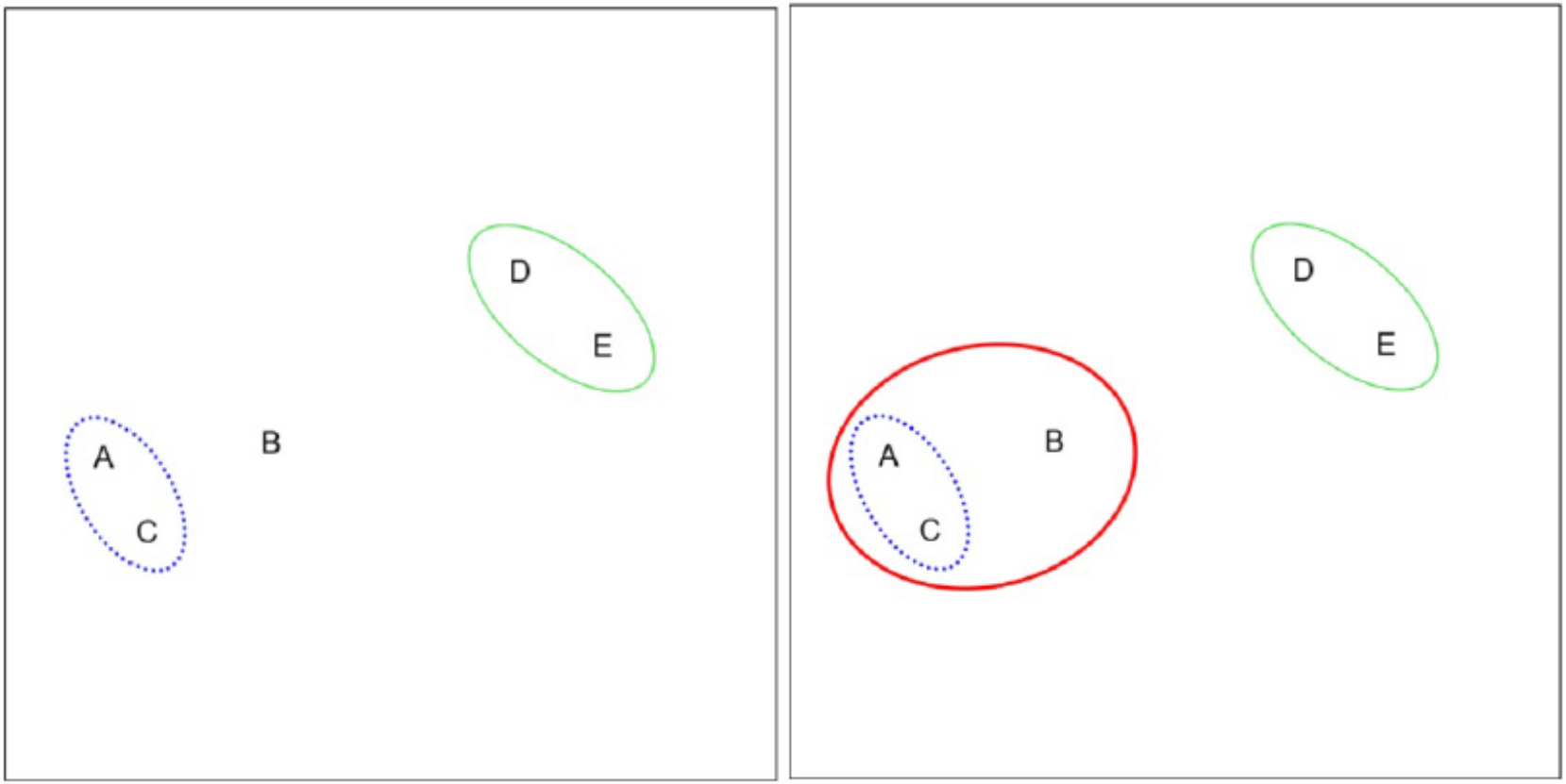
- Cluster jerárquico **aglomerativo (bottom-up)**: el agrupamiento se inicia en la base del árbol, donde cada observación forma un cluster individual. Los clusters se van combinado a medida que la estructura crece hasta converger en una única rama central.
- Cluster jerárquico **divisivo (top-down)**: es la estrategia opuesta al cluster aglomerativo, se inicia con todas las observaciones contenidas en un mismo cluster y se suceden divisiones hasta que cada observación forma un cluster individual.

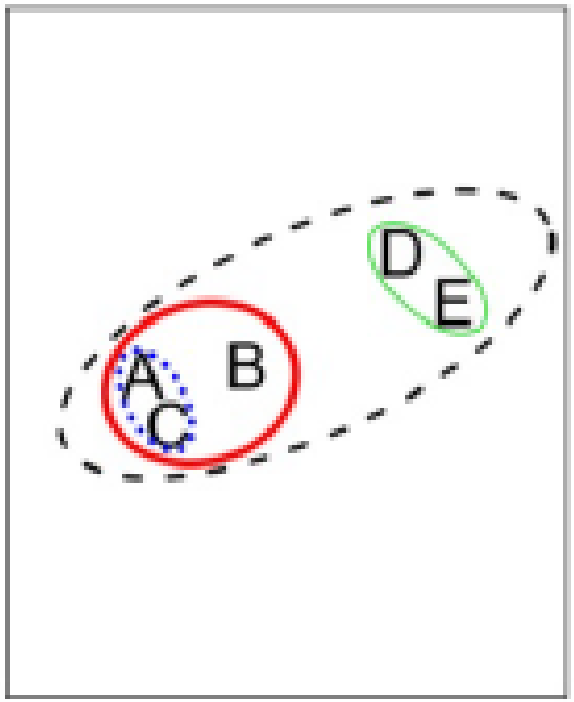
Los resultados pueden representarse de forma **muy intuitiva** en una estructura de árbol llamada **dendrograma**.

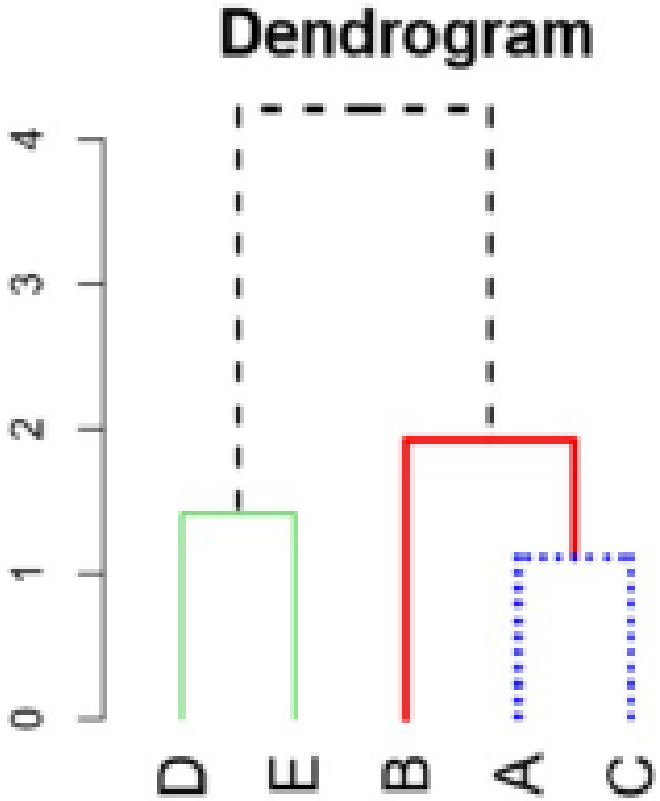
PRINCIPALES INCONVENIENTES

- Alto conste computacional
- Sensible a las primeras agrupaciones
- Compelejo de interpretar en alta dimensionalidad


D
E
A
B
C
D
E
A
B
C


D
E
A
B
C
D
E
A
B
C


A
B
C
D
E


Dendrogram
0
1
2
3
4
D
E
B
A
C

El proceso se inicia considerando cada una de las observaciones como un cluster individual, formando así la base del dendrograma.

1. Se inicia un proceso iterativo hasta que todas las observaciones pertenecen a un único cluster:
   1. Se calcula la distancia entre cada posible par de los n clusters.El investigador debe **determinar el tipo de medida** a emplear para cuantificar la **similitud** entre observaciones o grupos (distancia y **linkage**).
   2. Los dos clusters más similares se fusionan, de forma que quedan n-1 clusters.
2. Determinar dónde cortar la estructura de árbol generada (dendrograma).

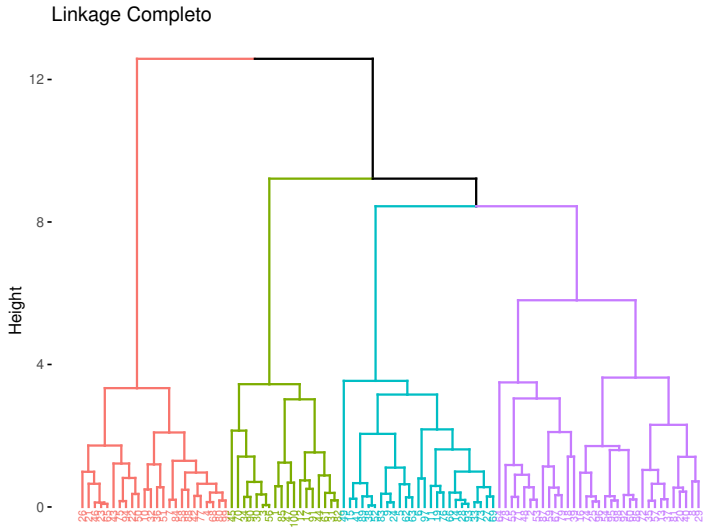
CONCEPTO DE LINKAGE

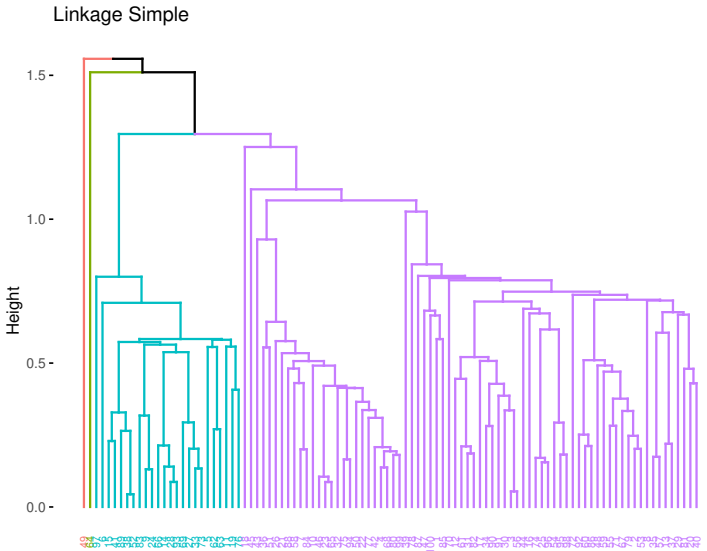
Para que el proceso de agrupamiento pueda llevarse a cabo, es necesario definir cómo se cuantifica la similitud entre dos clusters. Es decir, se tiene que **extender el concepto de distancia** entre pares de observaciones para que sea aplicable a pares de grupos, cada uno formado por varias observaciones. A este proceso se le conoce como **linkage**.

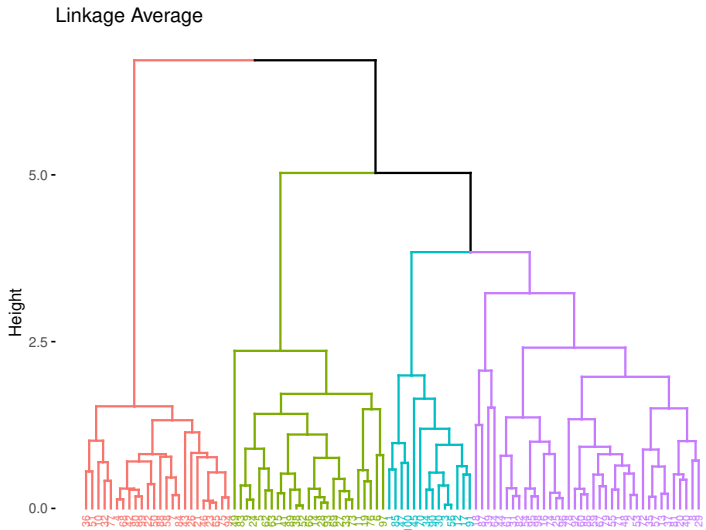
Distinta elección, distinto resultado!

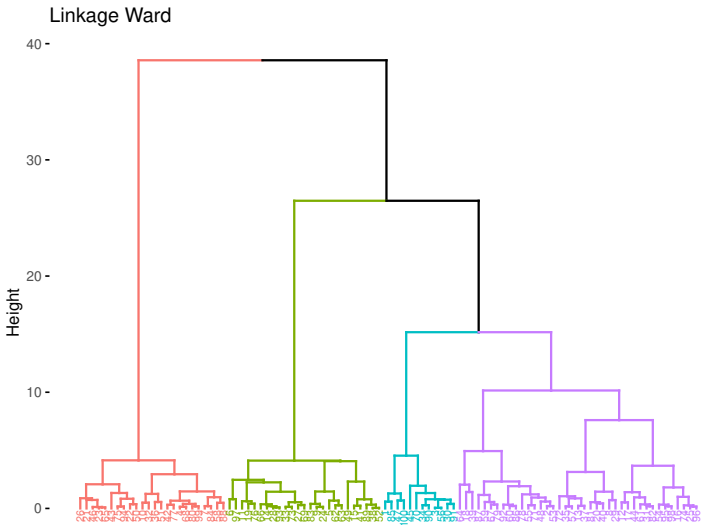
- **Completo**: Se calcula la distancia entre todos los posibles pares formados por una observación del cluster A y una del cluster B. La mayor de todas ellas se selecciona como la distancia entre los dos clusters. Se trata de la medida más conservadora (maximal intercluster dissimilarity).
- **Simple**: Se calcula la distancia entre todos los posibles pares formados por una observación del cluster A y una del cluster B. La menor de todas ellas se selecciona como la distancia entre los dos clusters. Se trata de la medida menos conservadora (minimal intercluster dissimilarity).
- **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 (mean intercluster dissimilarity).
- **Centroide**:Se calcula el centroide de cada uno de los clusters y se selecciona la distancia entre ellos como la distancia entre los dos clusters.
- **Ward**: El conocido método de mínima varianza, cuyo objetivo es minimizar la suma total de varianza intra-cluster. En cada paso se identifican aquellos 2 clusters cuya fusión conlleva menor incremento de la varianza total intra-cluster.

Tipos de Linkage









IDEA INTUITIVA

- Es posible que los resultados del algoritmo k-means no sean muy intuitivos cuando:
  - Los datos no son separables en conjuntos esféricos
  - Incorrecta selección del valor k
  - La inicialización de los centroides es mala
- Una posible estrategia es considerar los centroides estimados por el método de clustering jerárquico para k grupos e inicializar racionalmente los centroides del k-means.

Como se avisaba desde un principio, el análisis cluster es una técnica no supervisada cargada de subjetividad. ¿Cómo medir la similitud? ¿Cómo decidir el número adecuado de clusters? ¿Qué variables considerar para mejorar el análisis?
Por ello, es buena práctica considerar varios métodos y compararlos en cuanto a diversos criterios de "bondad" de ajuste en realción a la interpretabilidad de los grupos creados.

Algunas posibilidades en sklearn

- **DBSCAN** - Density-Based Spatial Clustering of Applications with Noise. Busca observaciones con alta densidad (de alguna forma que se encuentren rodeadas de otras que potencialmente podrían pertenencer a su grupo) y realiza la expansión del grupo desde ellas. Apropiado para conjuntos de datos que tengan potenciales grupos de densidad similar.
  `https://scikit-learn.org/stable/modules/clustering.html#dbscan`
- **AffinityPropagation** - Algoritmo basado en el envío de mensajes entre distintas observaciones relativas a la similitud entre ellas evaluando la representatividad de cada observación por medio de otras y buscando finalmente las observaciones más representativas como centros de distintos grupos. `https://scikit-learn.org/stable/modules/clustering.html#affinity-propagation`
- **Spectral clustering** - Calcula una matriz de adyacencia entre observaciones en dimensión reducida, en base a la cual genera una agrupación por algún algoritmo conocido (kmeans). `https://scikit-learn.org/stable/modules/clustering.html#spectral-clustering`