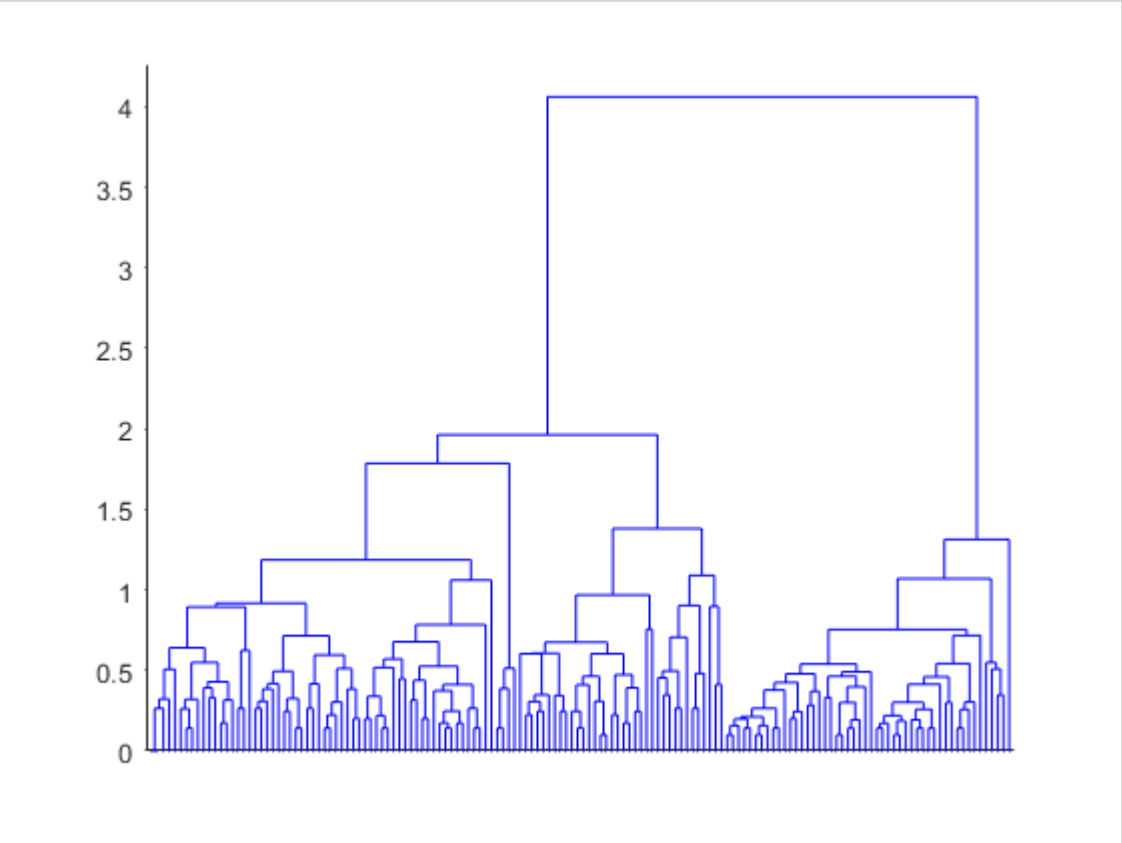


DigitalHouse > Coding School

DATA SCIENCE

Clase 48

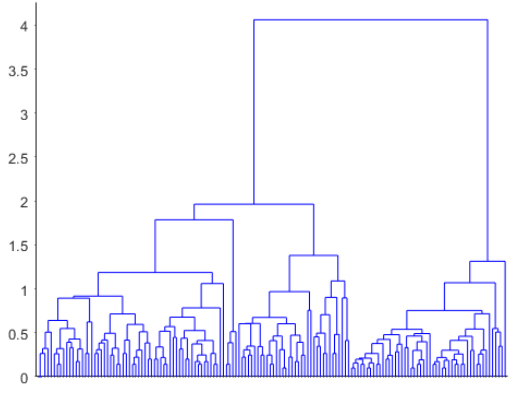
Introducción al clustering jerárquico

2017

Clustering Jerárquico

OBJETIVOS DE LA CLASE

1. **Introducir conceptos vinculados al clustering jerárquico**
2. **Explicar qué es un dendrograma**
3. **Explicar aspectos vinculados a las diferentes medidas de disimilitud**
4. **Identificar las diferencias y similitudes con el método de K-Means**



Clustering Jerárquico

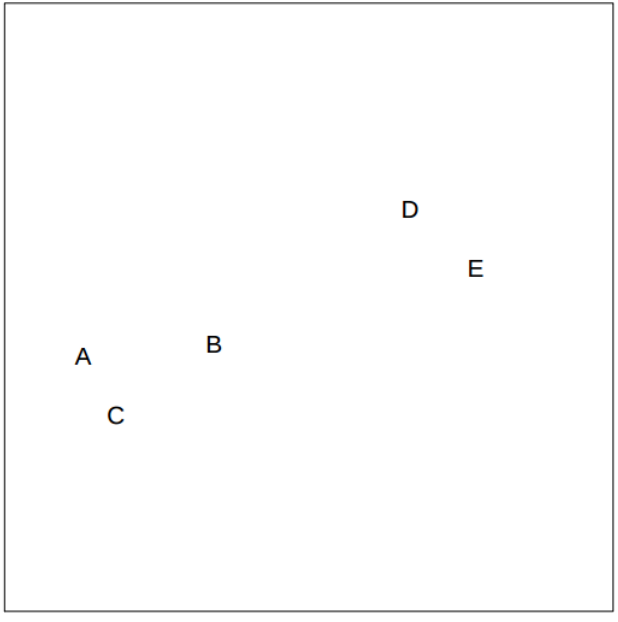
- Como ya vimos, K-Means requiere que especifiquemos de antemano cuál es el número de clusters que vamos a construir (k). Esto puede constituir una desventaja
- El clustering jerárquico es una técnica que no requiere esa definición de antemano: no es necesario que definamos previamente el k.
- Además ofrece una representación visual de las observaciones en forma de árbol que nos permite observar en una mirada los clusters obtenidos para cada uno de los posibles k (desde 1 hasta n).
- Vamos ver los métodos de clustering llamados “aglomerativos” o “bottom-up”, de los más comunes. La denominación se refiere al hecho de que el dendograma es construido comenzando desde las hojas (los “puntos”) hasta el tronco.

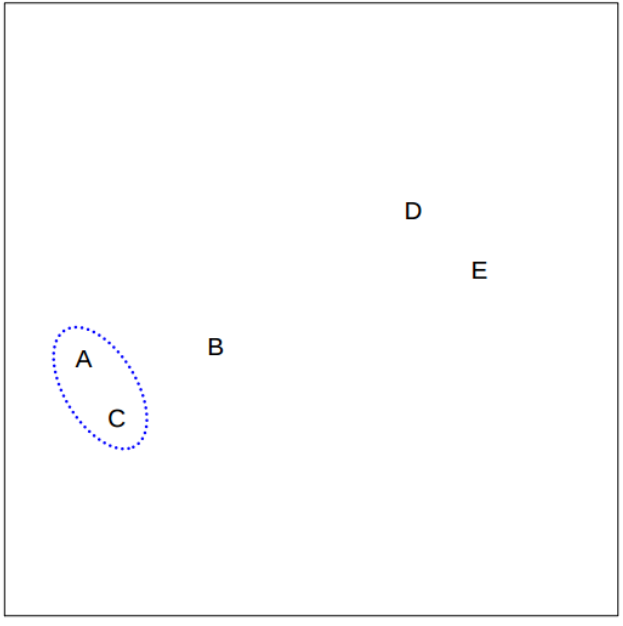
Clustering Jerárquico: Noción

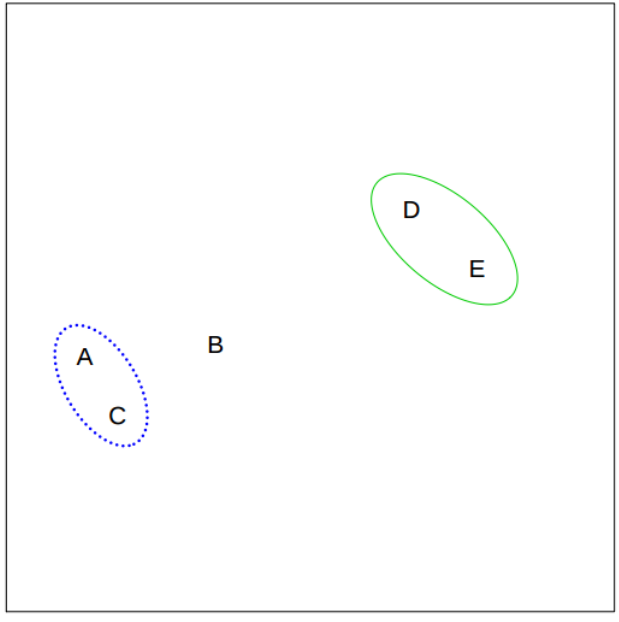
Visión general del algoritmo

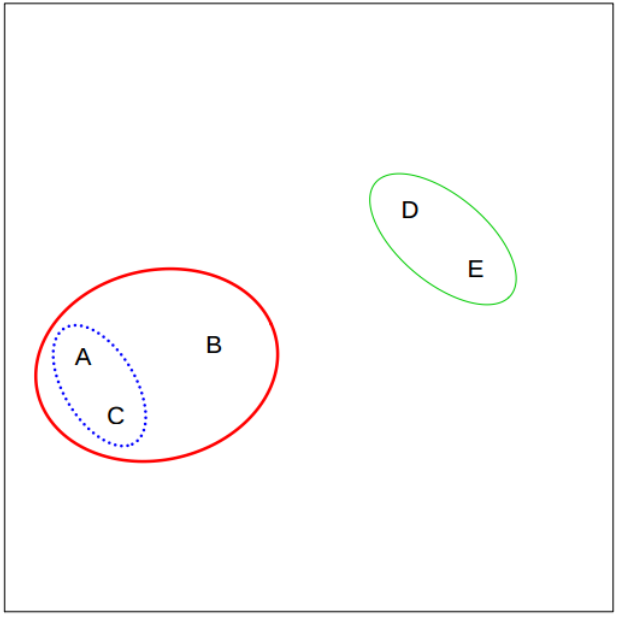
El enfoque de clusters jerárquico funciona de la siguiente forma

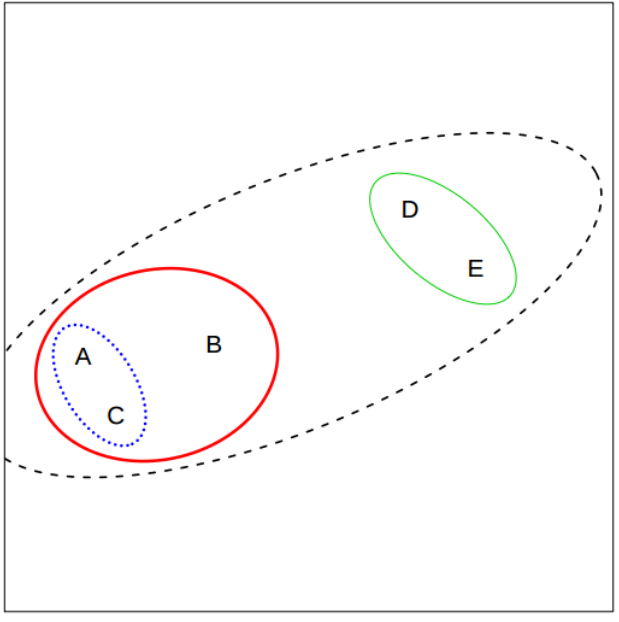
- Comienza con cada punto como un cluster
- Identifica los dos clusters más cercanos y conforma un cluster aglomerando esos dos
  - Para esto computa alguna medida de disimilaridad entre todos los clusters
- Se repite
- Hasta que todos los puntos conforman un único cluster

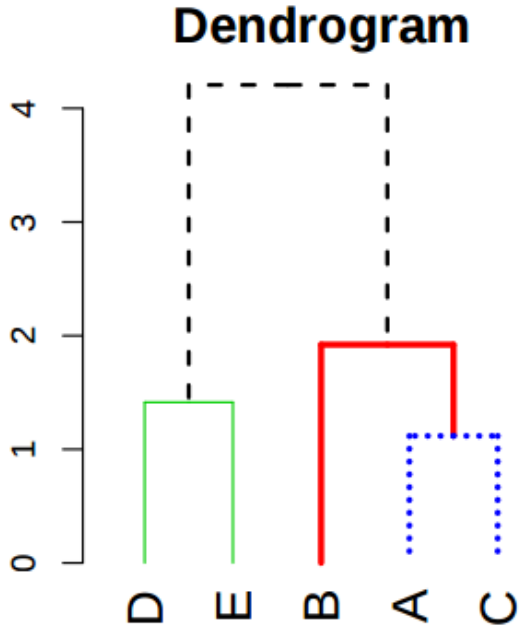

D
E
A
B
C


D
E
A
B
C


D
E
A
B
C


D
E
A
B
C

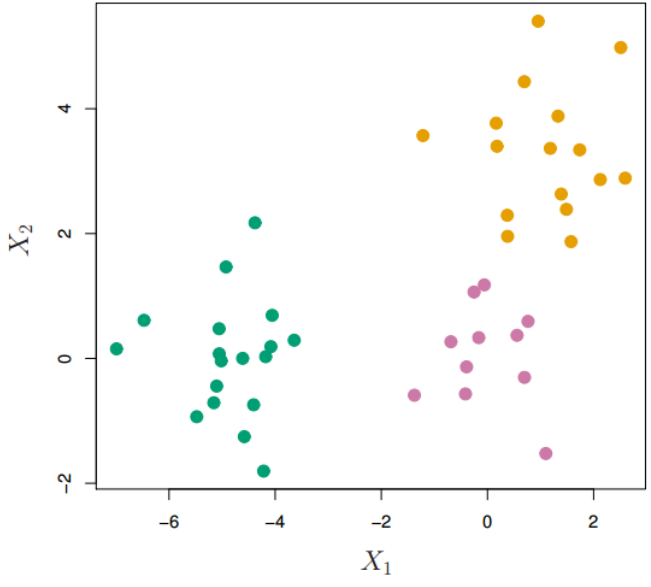

A
B
C
D
E

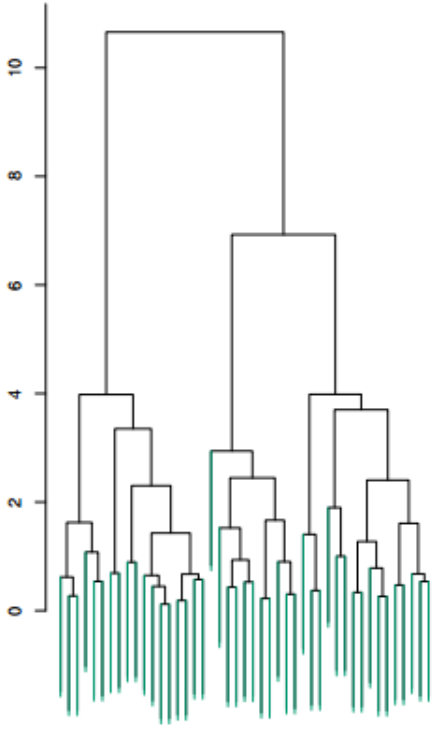

Dendrogram
0
1
2
3
4
D
E
B
A
C

¿Cómo interpretar el dendrograma?

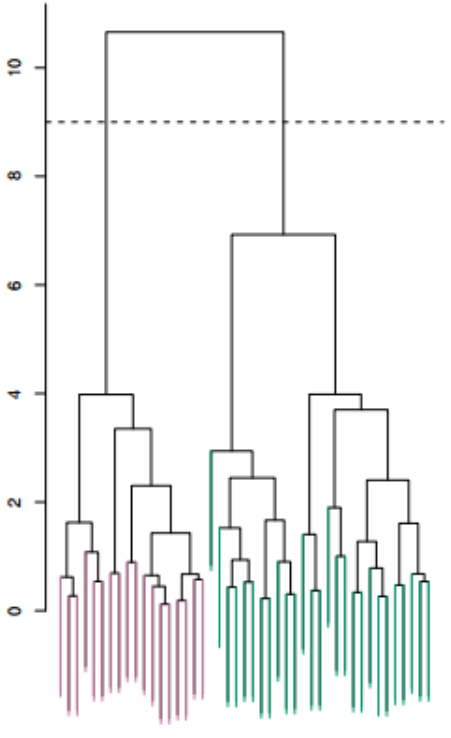
Para interpretar el dendrograma veamos un ejemplo un poco más complejo

- Comenzamos con 45 observaciones generadas en un espacio de dos dimensiones
- Al final del dendograma cada observación es un solo cluster.
- A medida que nos movemos hacia arriba del árbol algunas hojas se empiezan a fusionar: las que corresponden a observaciones muy similares.
- A medida que avanzamos más arriba del árbol, un número creciente de observaciones se han fusionado. Cuanto más temprano (más bajo en el árbol) dos observaciones se funden, más similares son entre sí.
- Las observaciones que se unen más arriba son las más diferentes

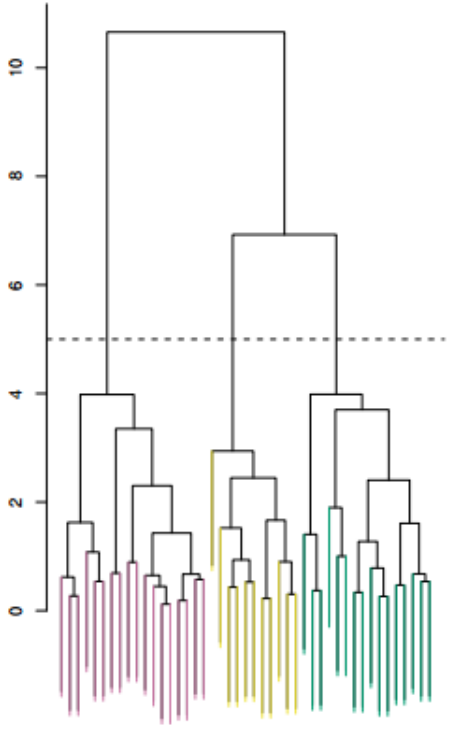




Un cluster



Dos clusters



Tres clusters

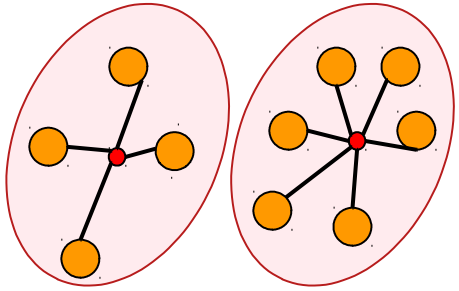
¿Cómo formar los clusters?

Problema a resolver

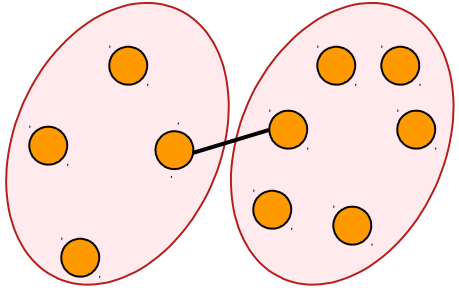
¿Cómo definimos la forma en que dos clusters se unen? ¿Qué medida usar? Algunas de las más utilizadas son:

TIpo de linkage	Descripción
Single	Calcula todas los pares de distancia entre los miembros del cluster A y el cluster B y utiliza la **mínima**.
Completo	Calcula todas los pares de distancia entre los miembros del cluster A y el cluster B y utiliza la **máxima**.
Average	Calcula todas los pares de distancia entre los miembros del cluster A y el cluster B y utiliza el **promedio de todas.**
Ward	Calcula la diferencia en la varianza total generada al aglomerar los diferentes clusters t busca la mínima

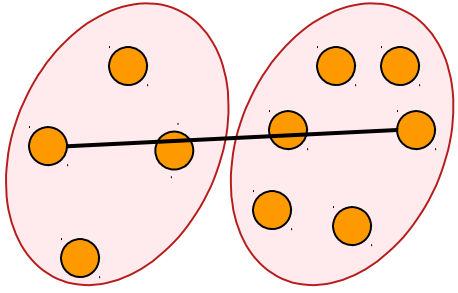
Ward Linkage



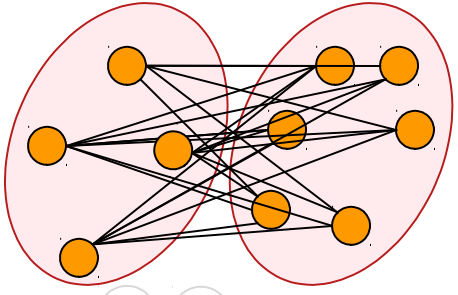
Single Linkage



Complete Linkage



Average Linkage



El tipo de linkage puede generar resultados diferentes

- Complete y average linkage tienden a generar clusters con tamaños similares.
- En cambio, single linkage tiende a generar clusters extendidos en los que las hojas se van fusionando una por una.

¿Cómo podemos evaluar un clustering jerárquico?

- Una forma es a través del coeficiente Copehenético (Copehenetic Coefficient)
  - Compara las distancias originales entre los puntos con las distancias que surgen del agrupamiento generado por el proceso de clustering
    - La idea es que valores cercanos a uno del coeficiente indica que las dos distancias están muy correlacionadas.

$$c = \frac{\sum_{i<j}(x(i,j) - \bar{x})(t(i,j) - \bar{t})}{\sqrt{[\sum_{i<j}(x(i,j) - \bar{x})^2][\sum_{i<j}(t(i,j) - \bar{t})^2]}}.$$

    - x(i,j) son las distancias entre los puntos
    - t(i,j) son las distancias entre los clusters (el punto en el que se unen en el dendograma)

CONCLUSIÓN

- Ofrecen una ventaja respecto a K-Means dado que no requieren definir a priori la cantidad de clusters a crear.
- Ofrecen una representación gráfica (dendograma) del proceso de generación de clusters
- Hay varios métodos para definir la forma de unión (linkage) entre los diferentes clusters