Clusters jerárquicos

Dr. Raimundo Sánchez

raimundo.sanchez@uai.cl

@raimun2

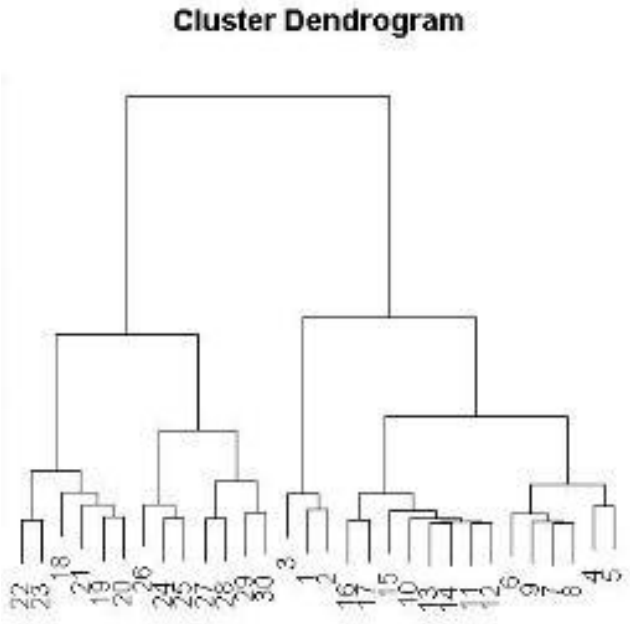
Idea general

El algoritmo básico para clustering aglomerativo es sencillo

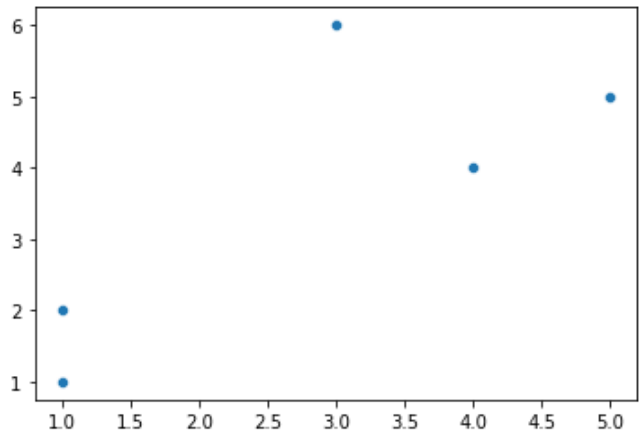
1. Deje que cada punto de datos sea un clúster
2. Calcular la matriz de proximidad (matriz de distancia entre cada clúster)
3. Repetir hasta que sólo quede un solo clúster
    1. Fusionar los dos clústeres más cercanos
    2. Actualizar la matriz de proximidad

El paso clave es el cálculo de la proximidad de dos clústeres

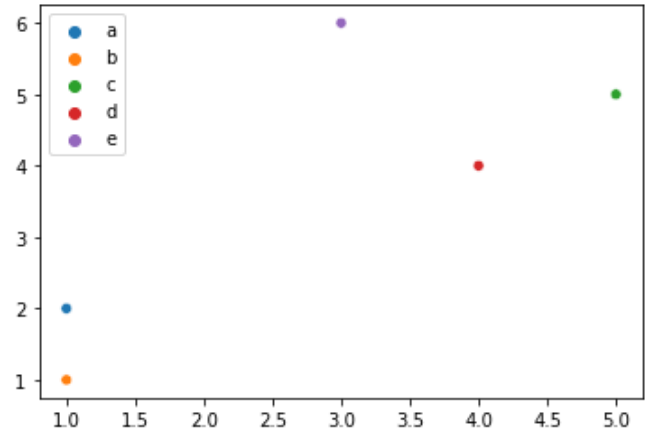
Diferentes enfoques para definir la distancia entre clústeres distinguen los diferentes algoritmos



Algoritmo



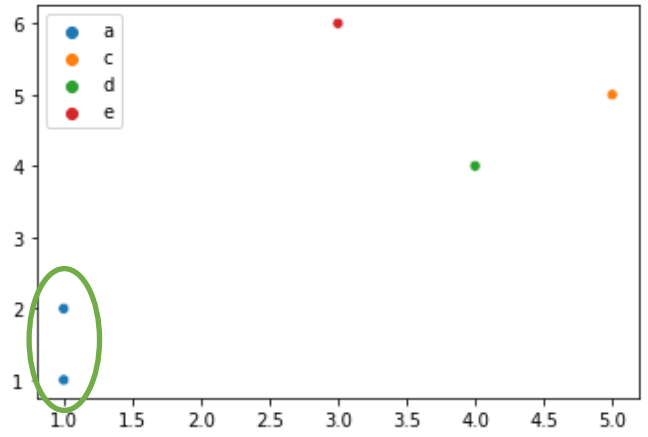
Deje que cada punto sea un clúster



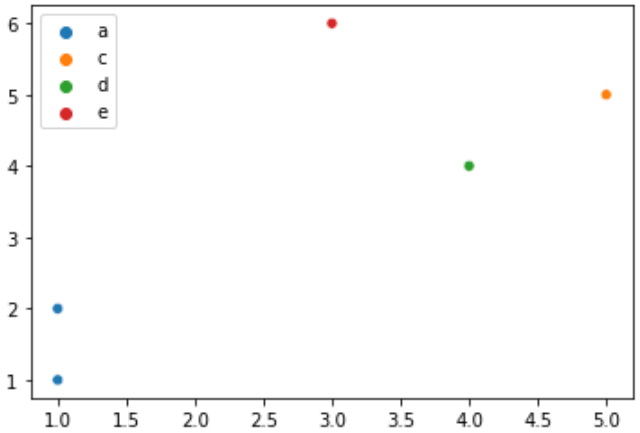
Calcular la matriz de proximidad (matriz de distancia entre cada clúster)

```
array([[0.        , 1.        , 5.        , 3.60555128, 4.47213595],
       [1.        , 0.        , 5.65685425, 4.24264069, 5.38516481],
       [5.        , 5.65685425, 0.        , 1.41421356, 2.23606798],
       [3.60555128, 4.24264069, 1.41421356, 0.        , 2.23606798],
       [4.47213595, 5.38516481, 2.23606798, 2.23606798, 0.        ]])
```

Merge the two closest clusters



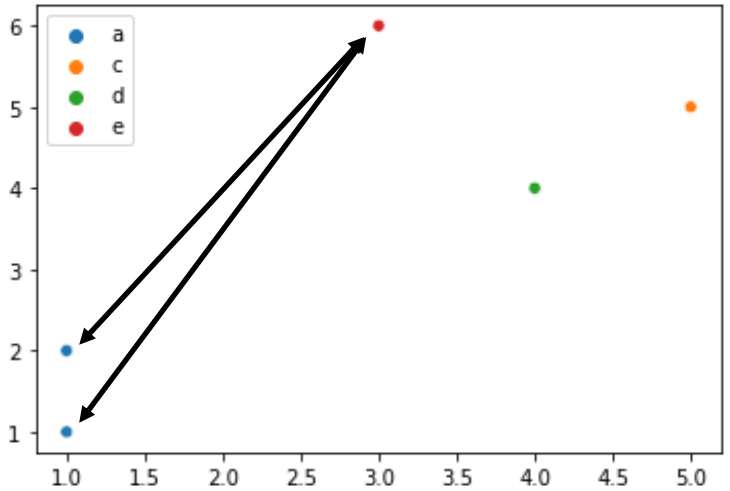
Algoritmo



Actualizar la matriz de proximidad

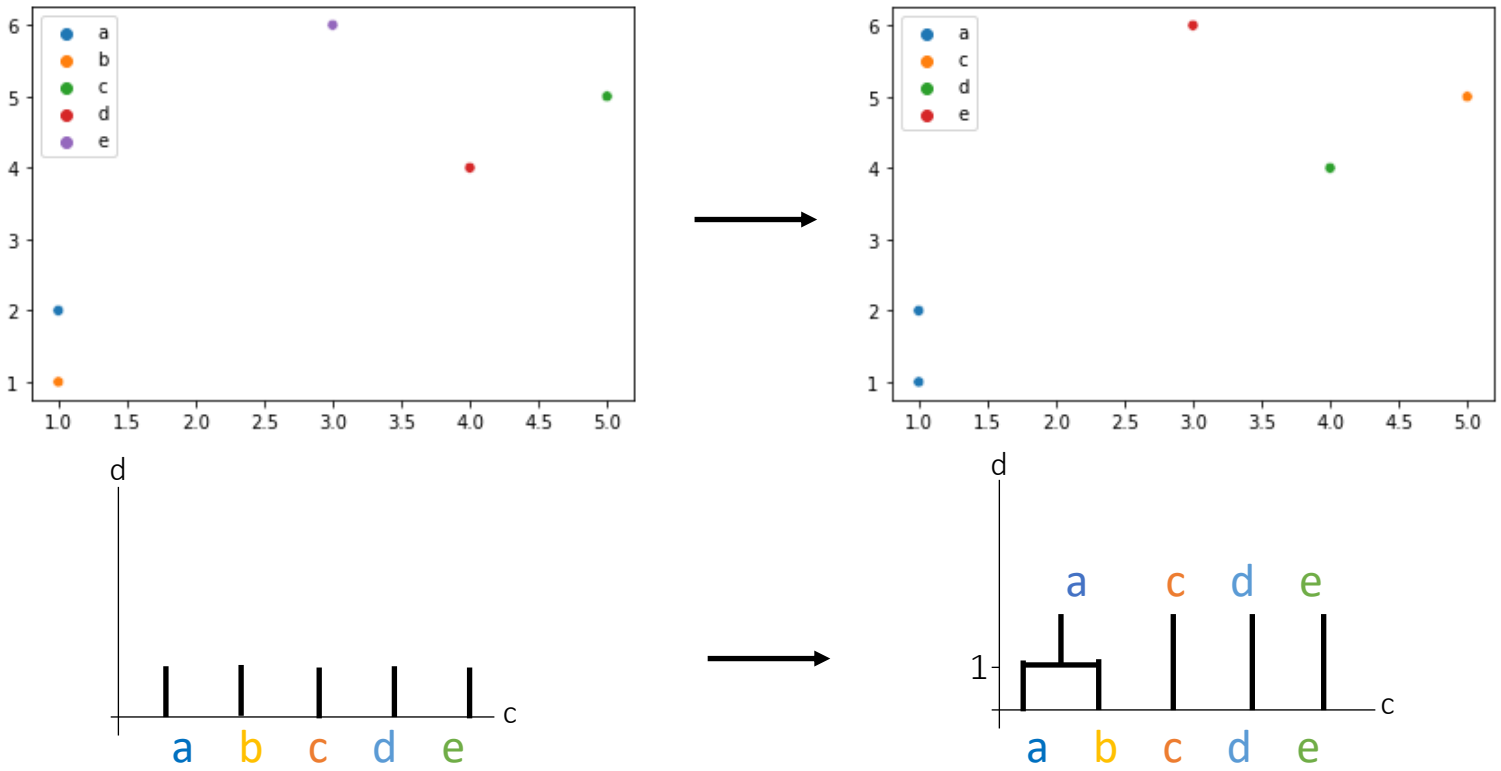
```
array([[0.        , 5.31507291, 3.90512484, 4.9244289 ],
       [5.31507291, 0.        , 1.41421356, 2.23606798],
       [3.90512484, 1.41421356, 0.        , 2.23606798],
       [4.9244289 , 2.23606798, 2.23606798, 0.        ]])
```

¿cómo?



$$D(C_i, C_j) = \mathbf{avg}\{d(x, y) | \mathbf{x} \in C_i, \mathbf{y} \in C_j\}$$

Algoritmo

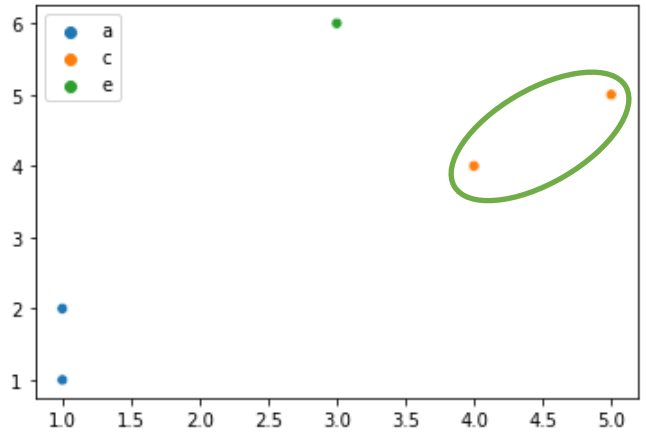


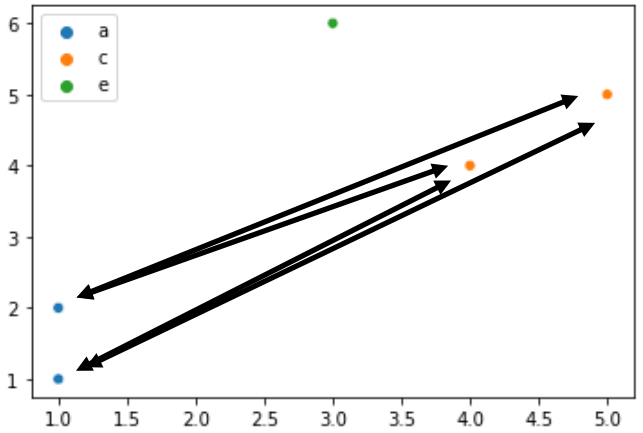
Algoritmo

repetir

```
array([[0.        , 5.31507291, 3.90512484, 4.9244289 ],
       [5.31507291, 0.        , 1.41421356, 2.23606798],
       [3.90512484, 1.41421356, 0.        , 2.23606798],
       [4.9244289 , 2.23606798, 2.23606798, 0.        ]])
```

Fusione los dos clústeres más cercanos



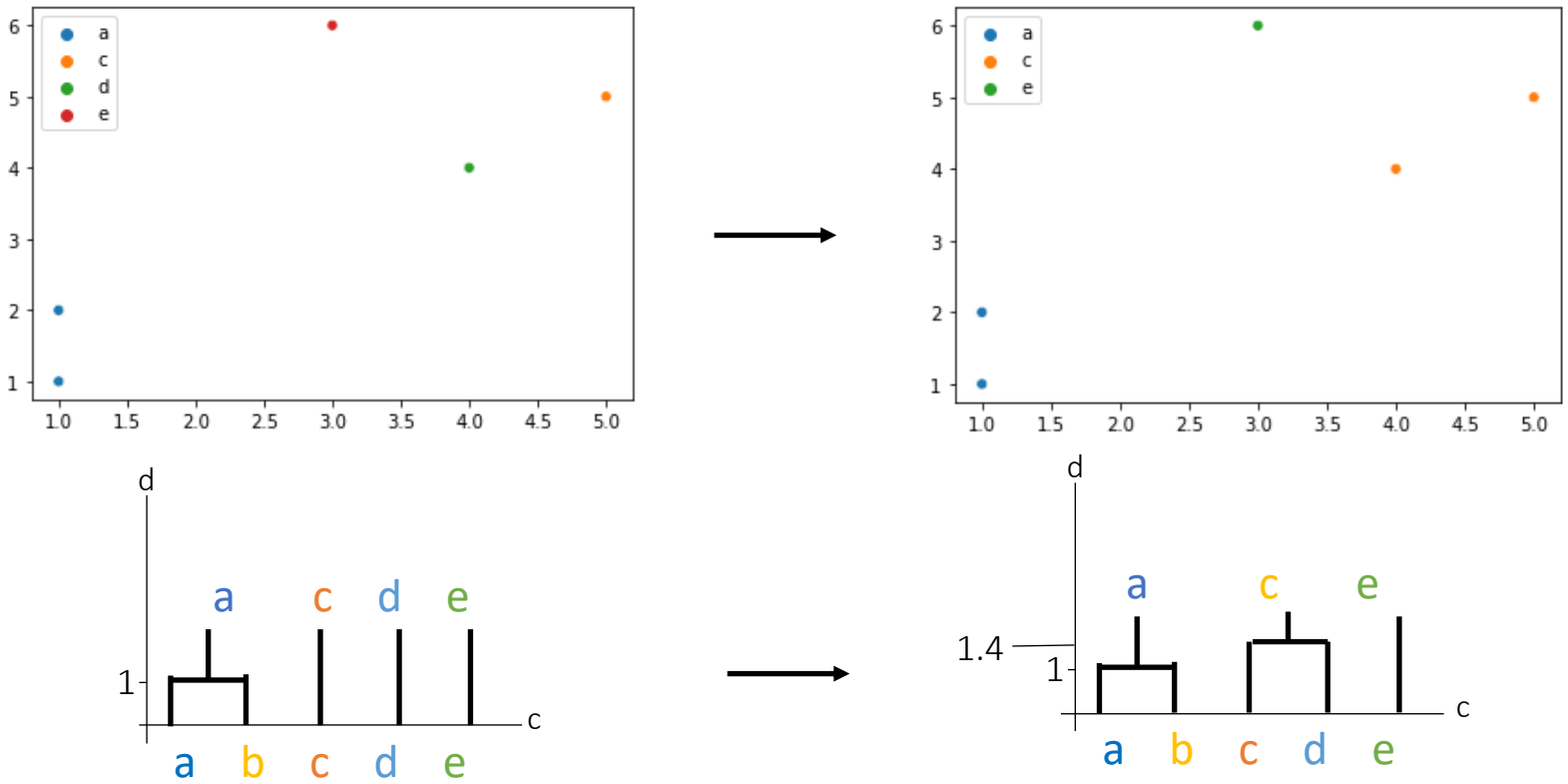


$$D(C_i, C_j) = \mathbf{avg}\{d(x, y) | \mathbf{x} \in C_i, \mathbf{y} \in C_j\}$$

Actualizar la matriz de proximidad

```
array([[0.        , 4.60977223, 4.9244289 ],
       [4.60977223, 0.        , 2.12132034],
       [4.9244289 , 2.12132034, 0.        ]])
```

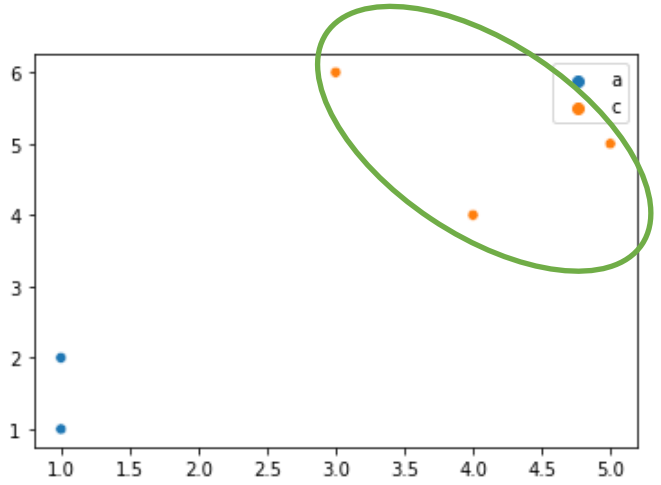
Algoritmo

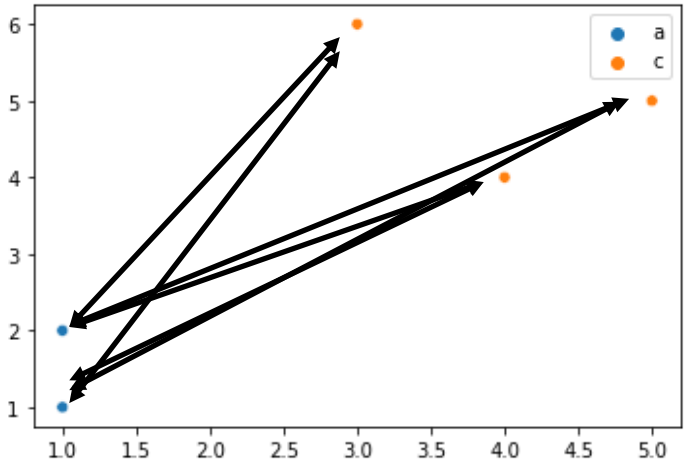


Algoritmo

```
array([[0.        , 4.60977223, 4.9244289 ],
       [4.60977223, 0.        , 2.12132034],
       [4.9244289 , 2.12132034, 0.        ]])
```

Fusione los dos clústeres más cercanos



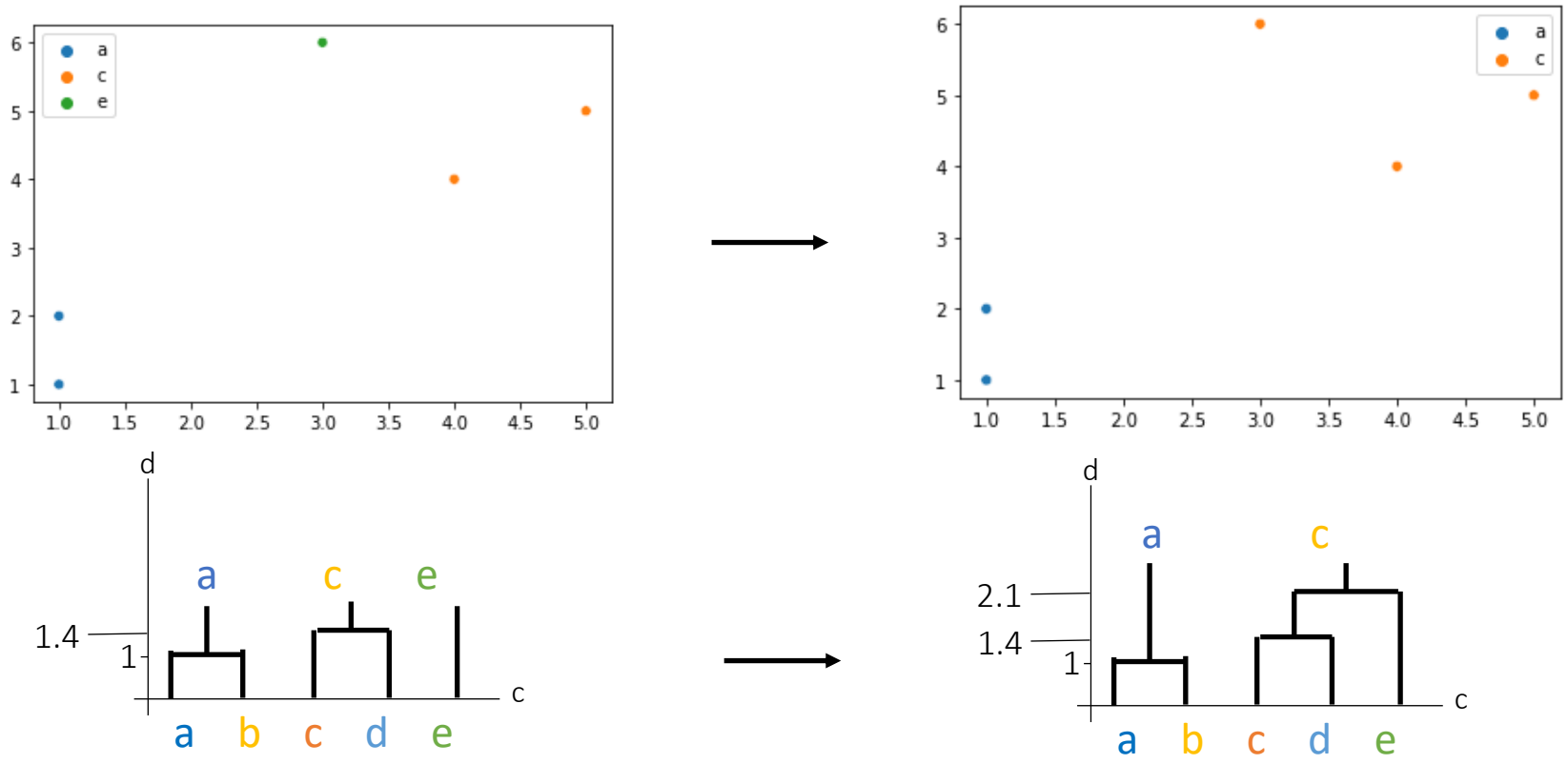


$$D(C_i, C_j) = \mathbf{avg}\{d(x, y) | \mathbf{x} \in C_i, \mathbf{y} \in C_j\}$$

Actualizar la matriz de proximidad

```
array([[0.        , 4.65026881],
       [4.65026881, 0.        ]])
```

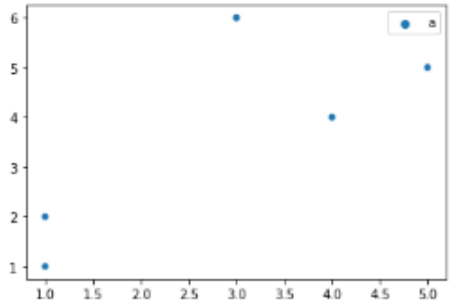
Algoritmo



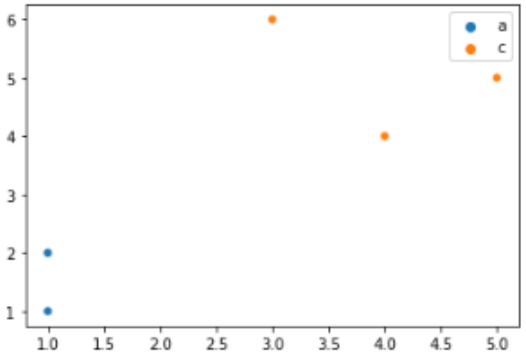
Algoritmo

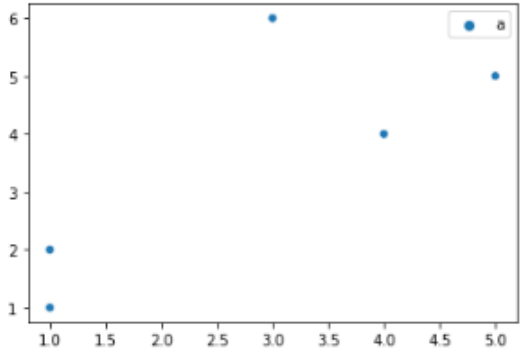
```
array([[0.        , 4.65026881],
       [4.65026881, 0.        ]])
```

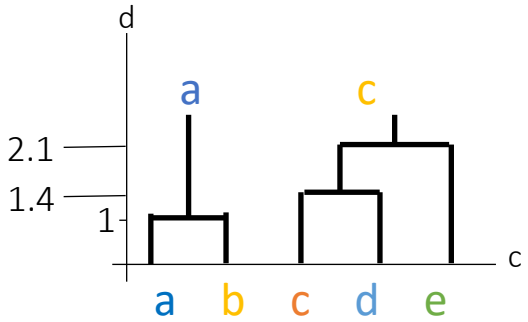
Fusione los dos clústeres más cercanos

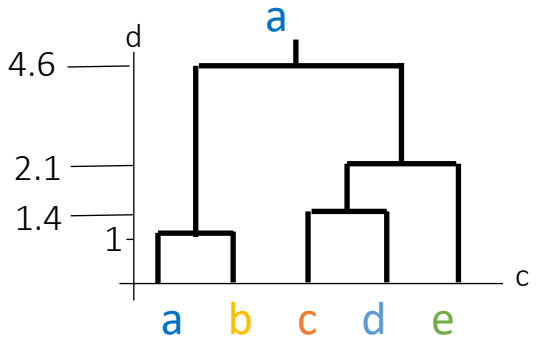


"Hasta que quede un solo clúster"



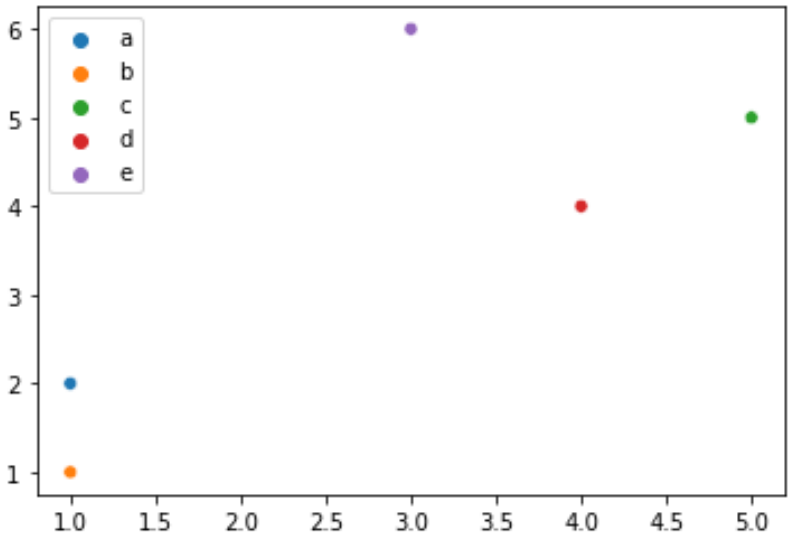






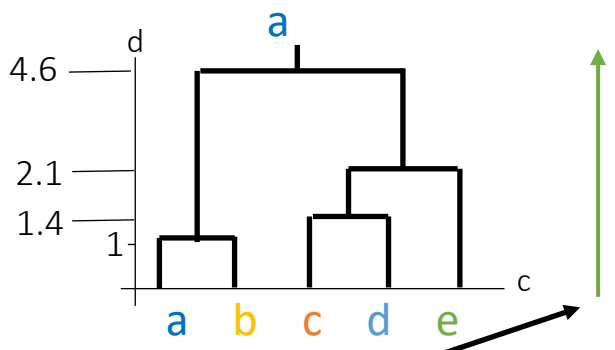
Resumen

Data



$$D(C_i, C_j) = \mathbf{avg}\{d(x, y) | \mathbf{x} \in C_i, \mathbf{y} \in C_j\}$$

Dendrograma

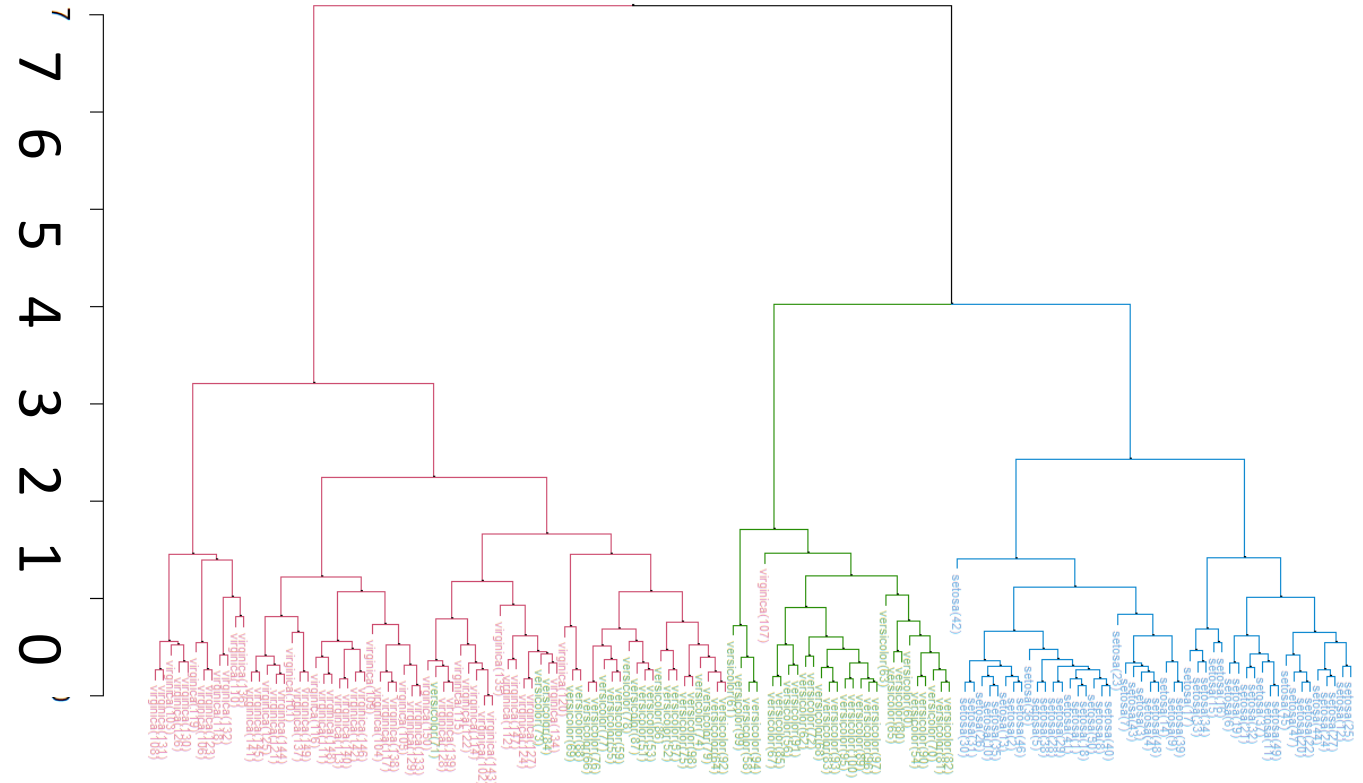


Jerárquico aglomerativo “average linkage”

Dendrogramas

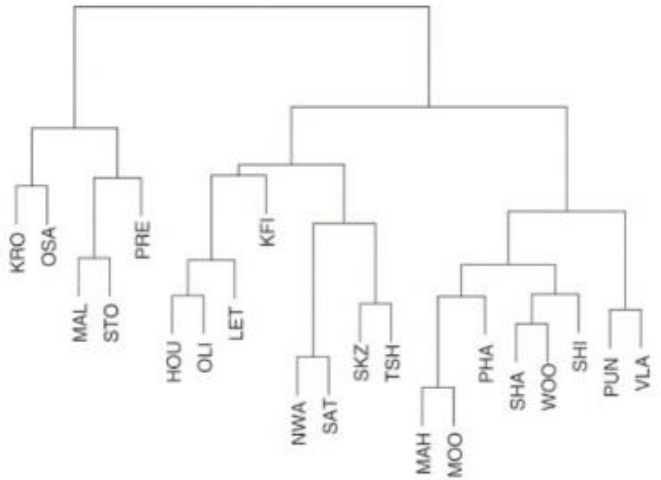
Un dendrograma es una estructura de tipo árbol que muestra el proceso generativo de agrupación en clústeres.

El eje X muestra los puntos de datos originales, mientras que el eje Y podría mostrar la distancia entre clústeres.

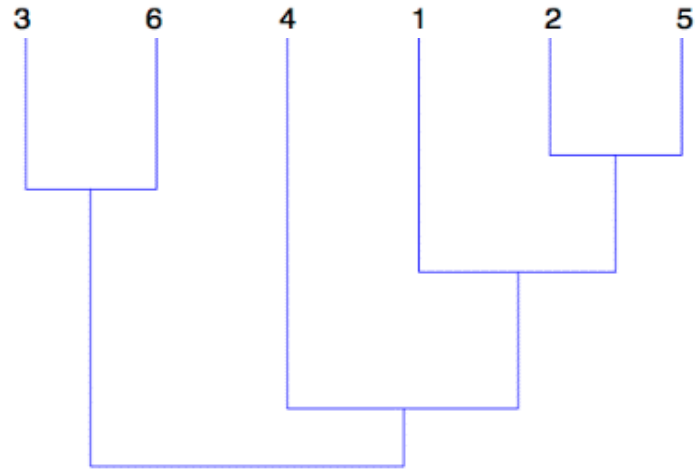


Dirección

Los métodos jerárquicos pueden ser aglomerantes o divisivos, en ambos casos se genera un dendrograma que muestra las secuencias de combinaciones o divisiones.



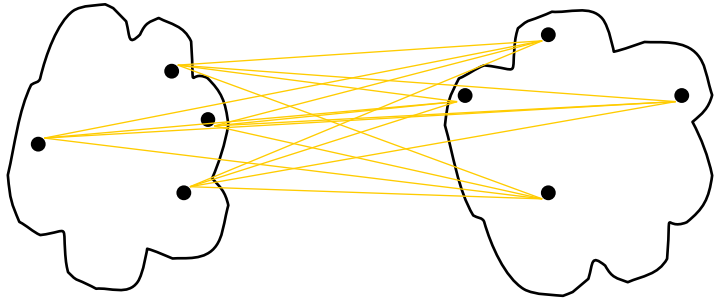
Aglomerativo



Divisivo

Average linkage

La distancia entre clústeres se basa en la distancia media entre todos los puntos de los diferentes clústeres

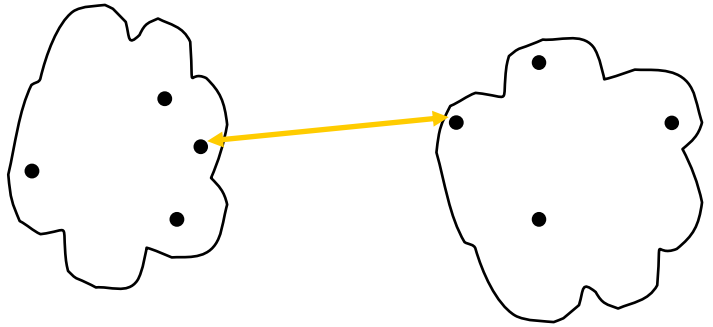


$$D(C_i, C_j) = \mathbf{avg}\{d(x, y) | \mathbf{x} \in C_i, \mathbf{y} \in C_j\}$$

- **Fortalezas: Menos susceptibles a los valores atípicos**
- **Limitaciones: Sesgado hacia los clusters globulares**

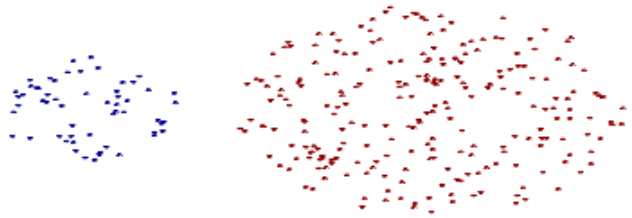
Single linkage

La distancia entre clústeres se basa en los dos puntos (más cercanos) más similares de los diferentes clústeres

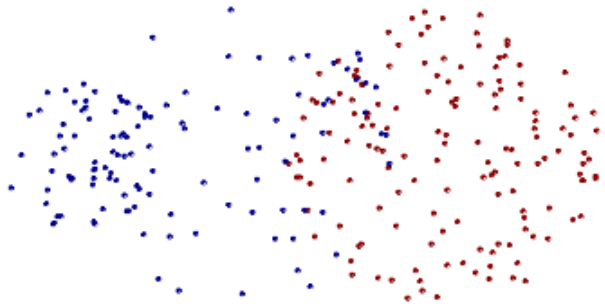


$$D(C_i, C_j) = \mathbf{min}\{d(x, y) | \mathbf{x} \in C_i, \mathbf{y} \in C_j\}$$

- **Fortalezas:**
  Produce clusters largos y delgados

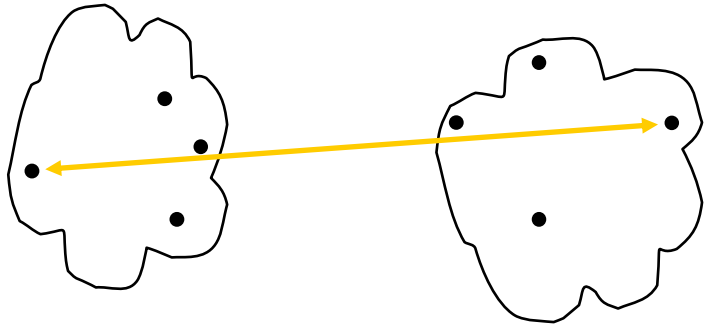


- **Limitaciones:**
  Sensible a los valores atípicos



Complete linkage

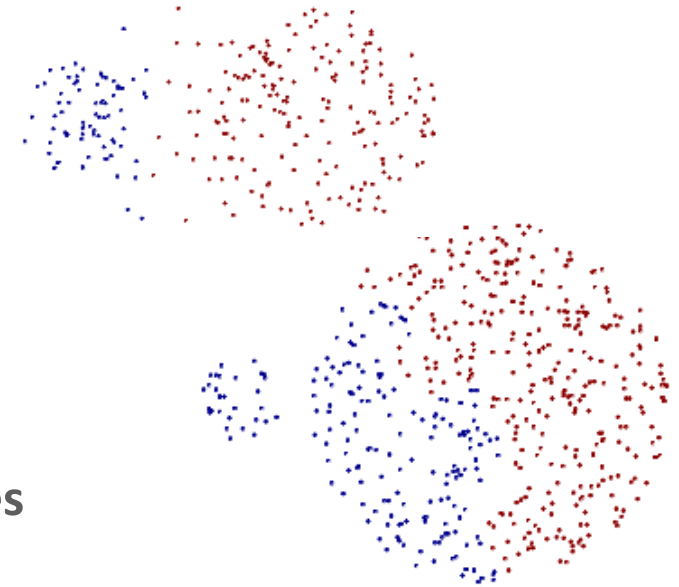
La distancia entre clústeres se basa en los dos puntos más diferentes (más distantes) de los diferentes clústeres



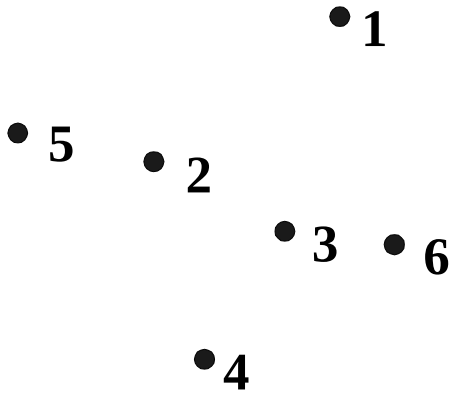
$$D(C_i, C_j) = \mathbf{max}\{d(x, y)|\mathbf{x} \in C_i, \mathbf{y} \in C_j\}$$

- **Fortalezas:**
  Menos susceptibles a los valores atípicos

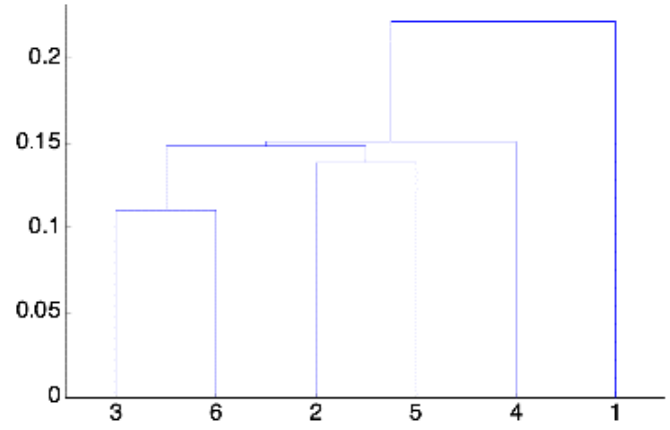
- **Limitaciones:**
  Tiende a romper grandes clusters
  Sesgado hacia los clusters globulares



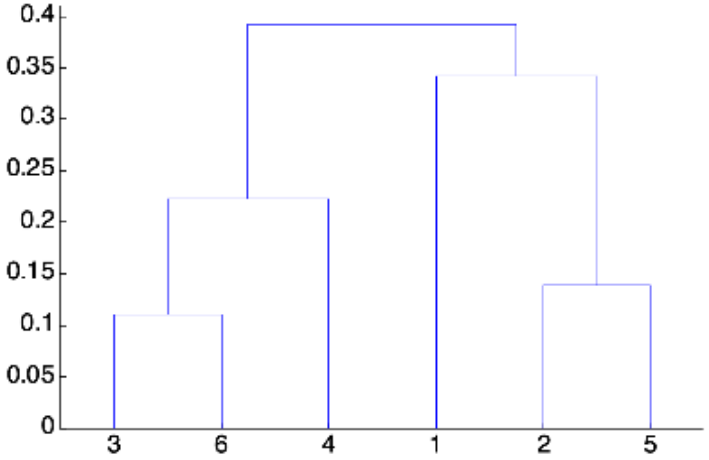
Métodos de aglomeración



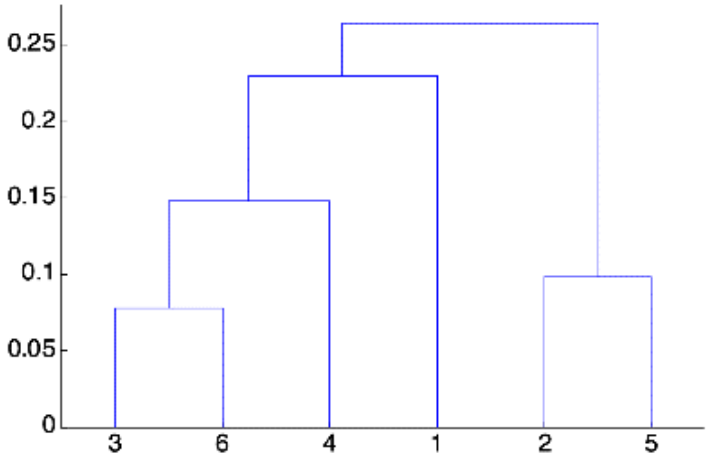
Original data



Single linkage



Complete linkage



Average linkage

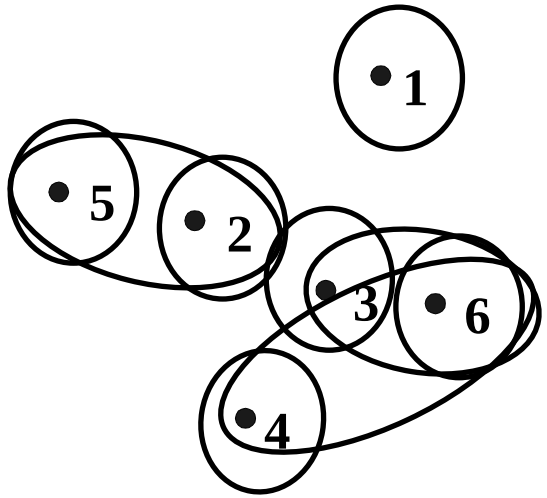
Nivel de análisis

Este proceso es subjetivo

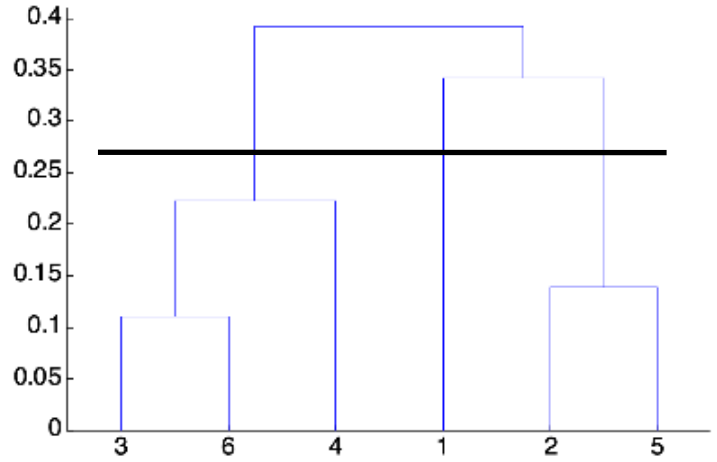
Aunque el dendrograma nos muestra toda la información, hay n clústeres posibles, donde n es el número de puntos de datos.

Se recomienda contar con un experto para analizar el dendrograma.

Hay métodos matemáticos, pero son heurísticos.

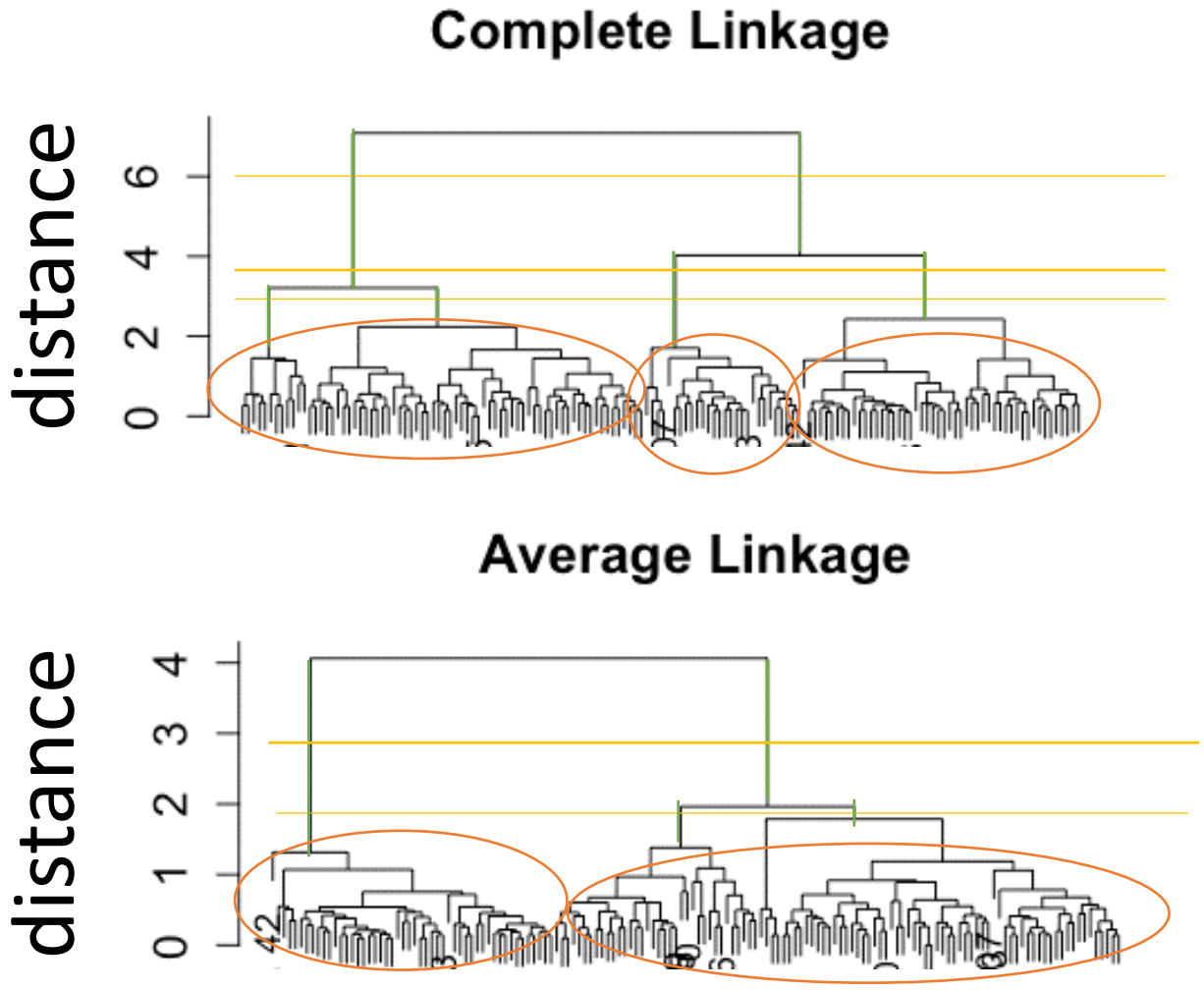


Original data



Complete linkage

Ejemplo


Complete Linkage
distance
0
2
4
6
Average Linkage
distance
0
1
2
3
4

Desventajas

- El algoritmo es demasiado caro O(n3): Hay n pasos, para unir clústeres, y en cada paso calculamos la matriz de proximidad O(n2).
- Una vez que se toma la decisión de combinar dos grupos, no se puede revertir
- Ninguna función objetivo se minimiza directamente

Documentos y datos categóricos

Los métodos jerárquicos se utilizan frecuentemente para describir vectores de documentos y datos categóricos

Para vectores de documentos:

- Distancia de correlación
- Distancia del coseno

Para datos categóricos:

- Disimilaridad de Gower
- Coeficiente de Jaccard
- IoF
- Goodall

Ejemplo vector de documentos

Se pidió a los alumnos que describieran los siguientes conceptos:

brazo, tocino, castaño, águila, pelo, labios, langosta, kétchup, sándwich, escorpión, pulgar, pavo

- Se mencionaron 2625 palabras.
- 317 atributos diferentes.
- Máximo de atributos diferentes para una palabra 46.
- Atributos mínimas diferentes para una palabra 32.
- Atributos más mencionada para una sola palabra => 30

Ejemplo de los datos

atributos

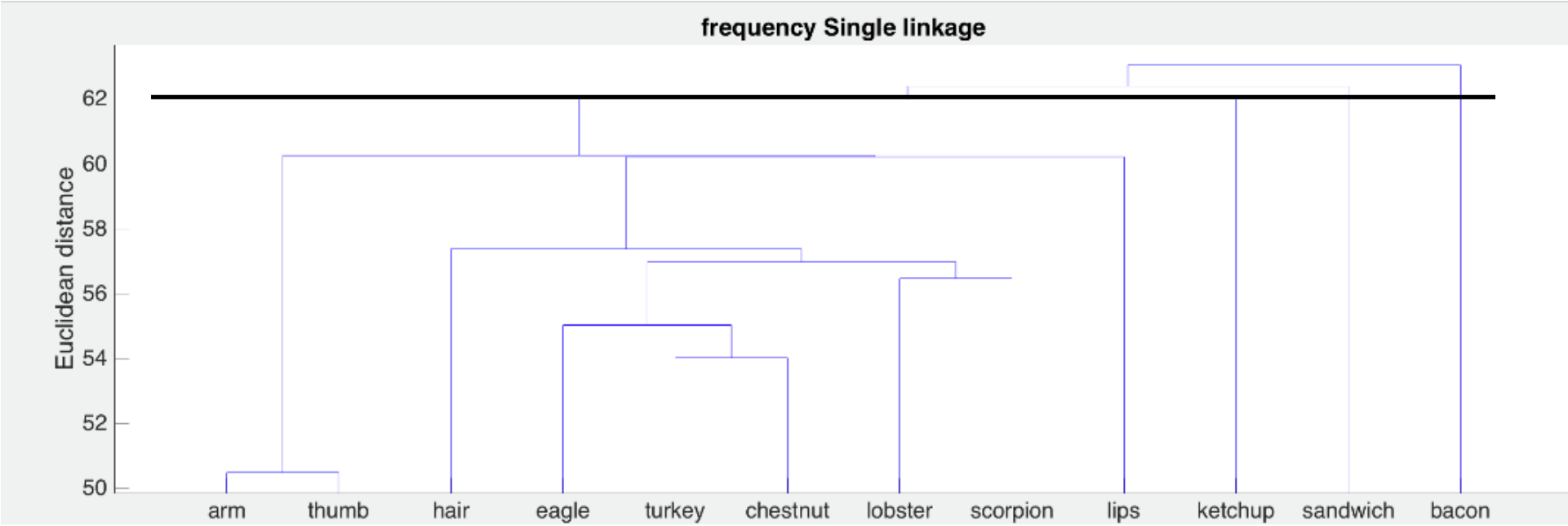
definiciones

4	0	0	4	0	0	0	2	0	0	0	0	5	0	0 …
0	0	0	2	0	13	0	0	0	2	0	0	0	0	0 …
0	0	0	0	0	0	0	0	0	0	0	0	4	0	0 …
0	0	0	0	3	0	0	0	0	0	2	9	9	0	2 …
0	0	0	13	0	2	0	0	0	0	0	0	0	0	0 …
0	0	0	2	0	3	0	0	0	0	0	0	3	0	0 …
0	0	0	0	0	3	0	0	0	0	0	0	0	0	0 …
0	0	17	0	0	0	0	0	0	0	0	0	0	0	0 …
0	0	0	10	0	3	0	0	0	0	0	0	6	0	0 …
…	…	…	…	…	…	…	…	…	…	…	…	…	…	… …

¿Qué clústeres esperas?

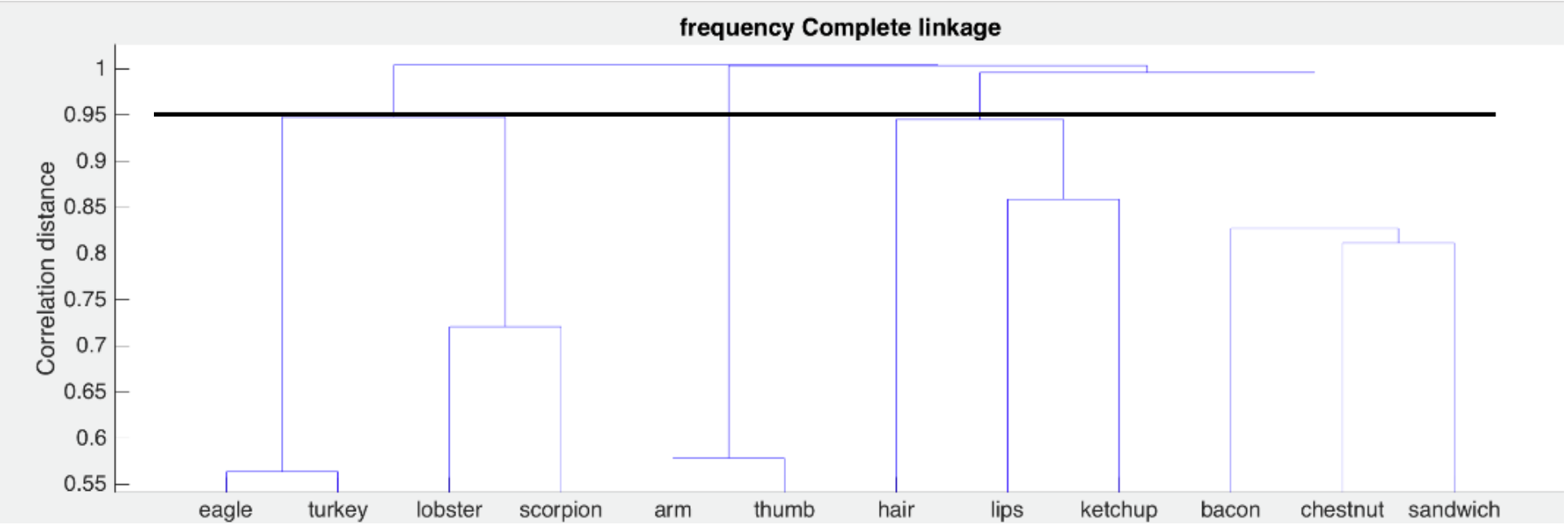
¿Qué decisiones debo tomar?

Ejemplo vector de documentos



$$d(p,q) = \sqrt{\sum_{k=1}^{m}(p_k - q_k)^2}$$

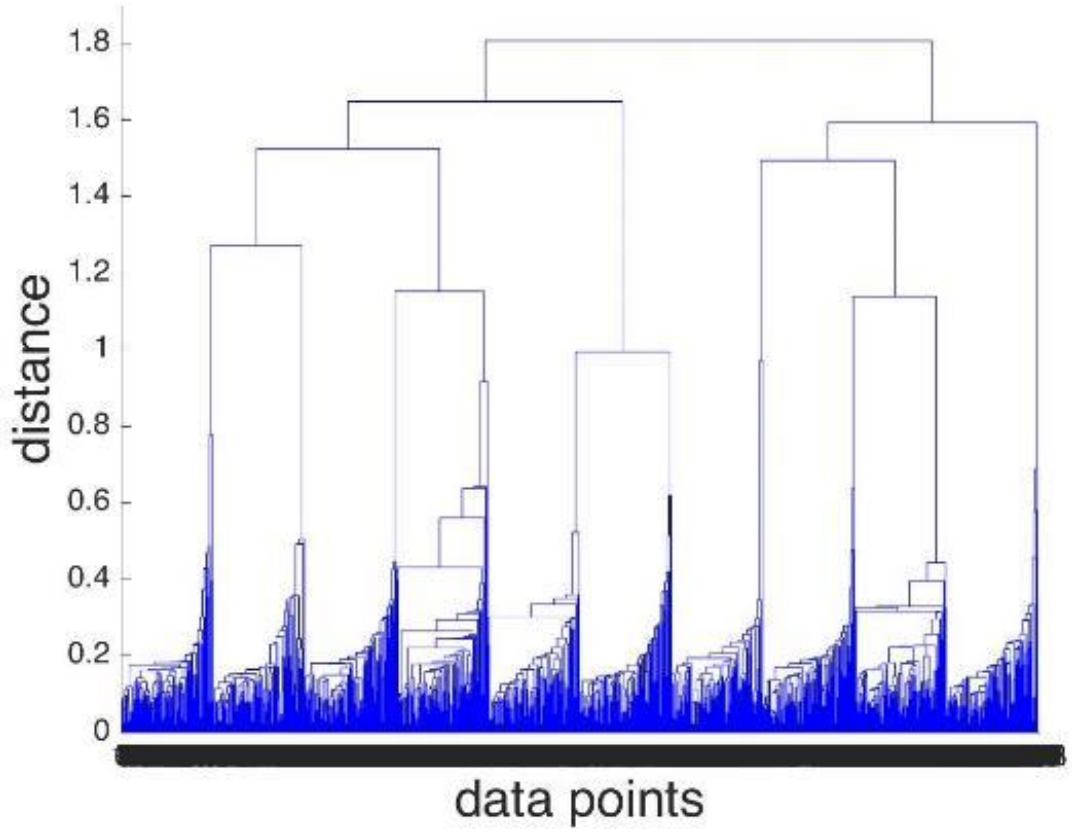
Ejemplo vector de documentos



$$Cor_D(x,y) = \frac{Cov_D(x,y)}{\sigma_D(x)\sigma_d(y)}$$

$$Cov_D^2(x,y) = \frac{1}{N^2}\sum_i^N \sum_j^N \hat{X}_{ij}\hat{Y}_{ij} \qquad Var_D^2(x) = Cov_D^2(x,x)$$

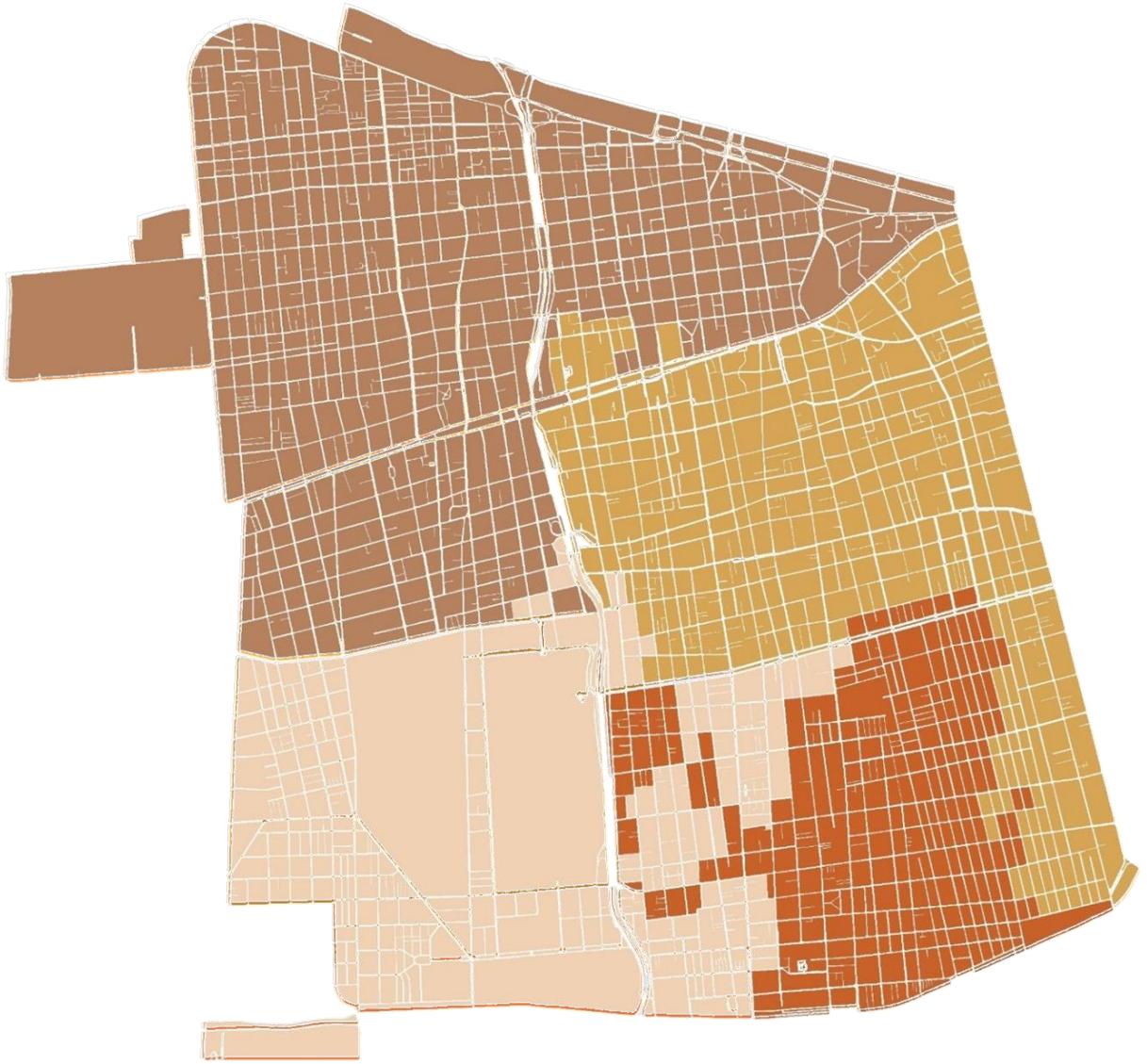
Aplicación centroides iniciales Kmedias



Regionalización jerárquica

- Algoritmo jerárquico con restricción de vecindad
- Utilizado para combinar zonas geográficas
- En cada iteración, cada entidad se auto-organiza espacialmente según las reglas:
  - Identifica y evalúa vecinos.
  - Selecciona el vecino más parecido
  - Se fusiona

	AREA	POPULATION
Mean	[illegible]	[illegible]
Mín	[illegible]	[illegible]
Max	[illegible]	[illegible]
Var Coeff.	[illegible]	[illegible]



Clusters jerárquicos

Dr. Raimundo Sánchez

raimundo.sanchez@uai.cl

@raimun2