

Datamining

Clustering Hierarchical

Andrés C. Medina

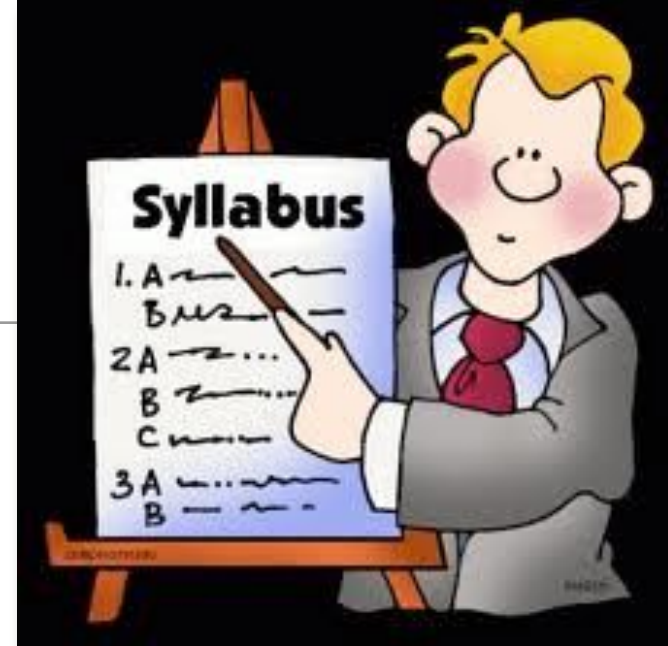
PhD(c) Engineering Systems

Lead Advanced Analytics Financial Retail Cencosud-Scotiabank

andres.medina.s@edu.uai.cl

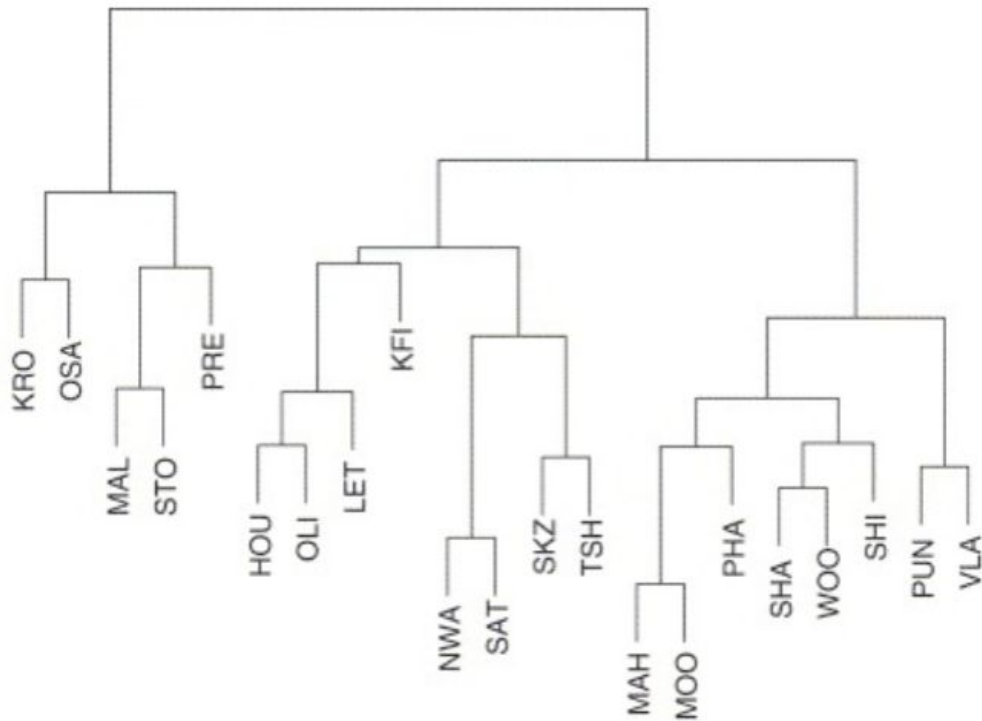
Contenidos

- Introducción
- **Clustering**
 - Introducción
 - Métodos de partición
 - Métodos de densidad
 - **Métodos jerárquicos**
 - Métodos difusos
 - Evaluación

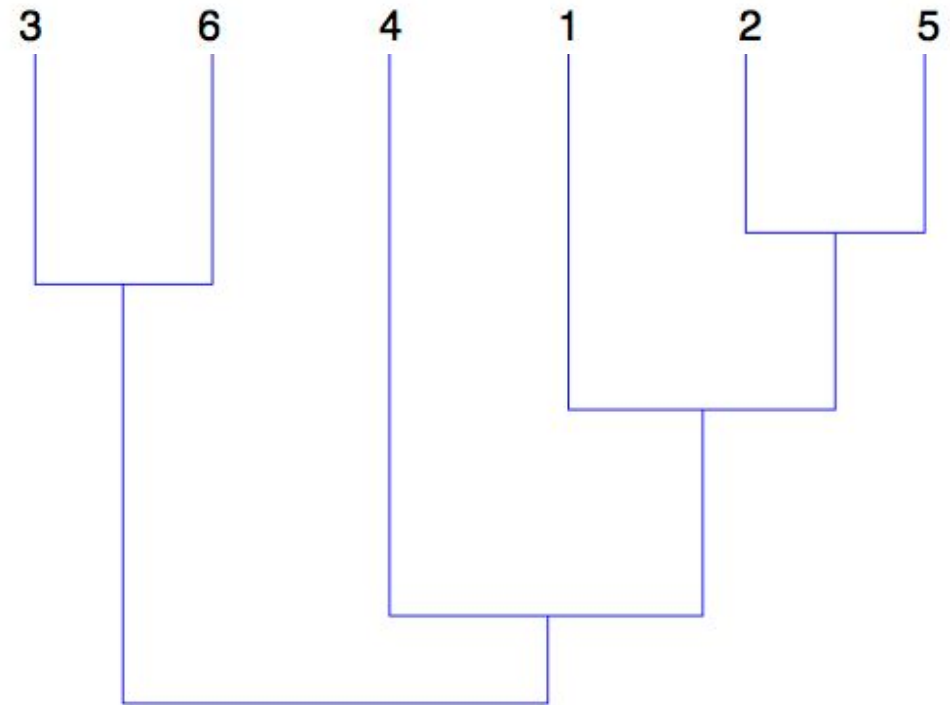


Métodos jerárquicos, introducción

- Los métodos jerárquicos pueden ser aglomerativos o divisivos, en ambos casos un dendrograma es generado mostrando la secuencia de uniones o divisiones de los clusters.



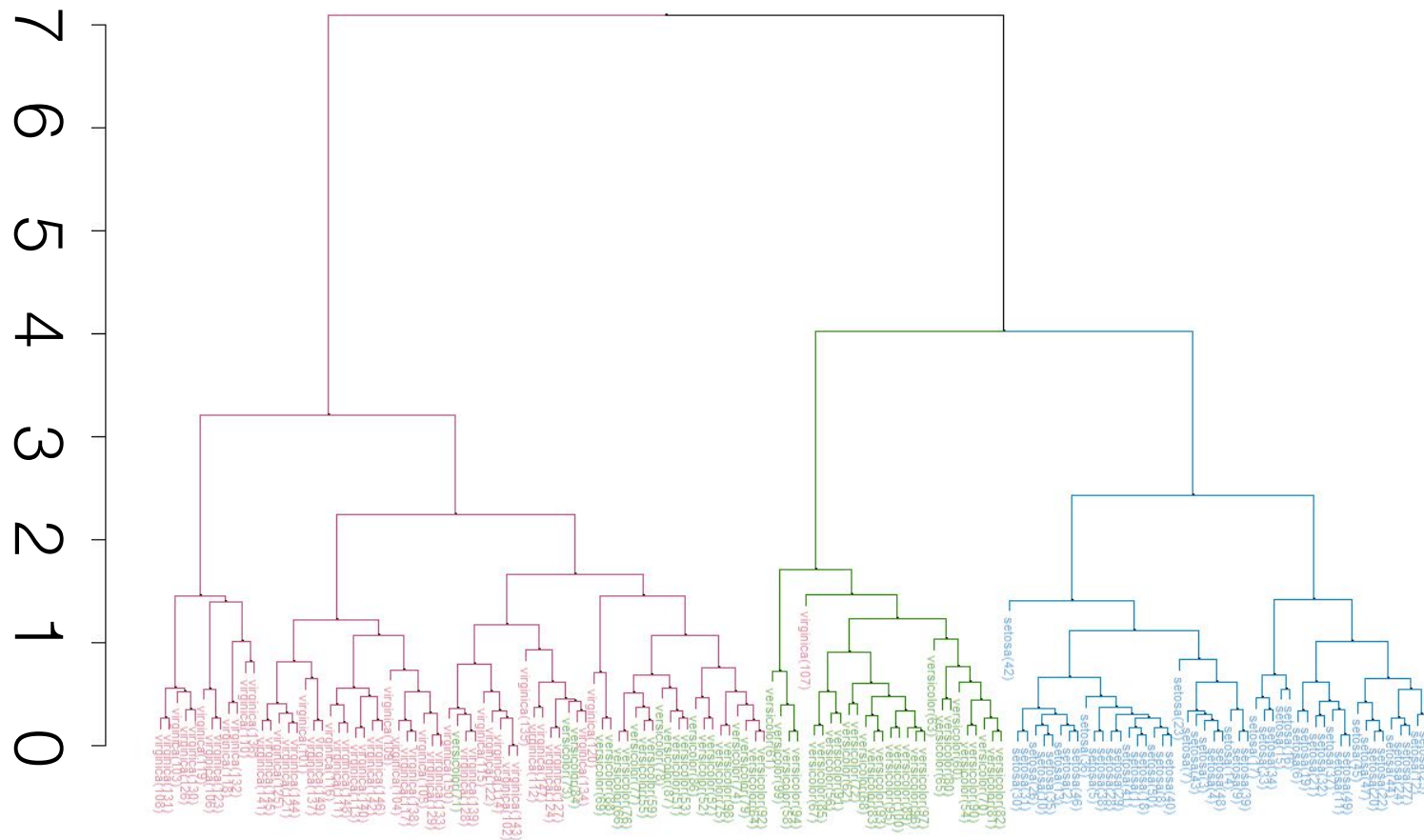
Aglomerativo



Divisivo

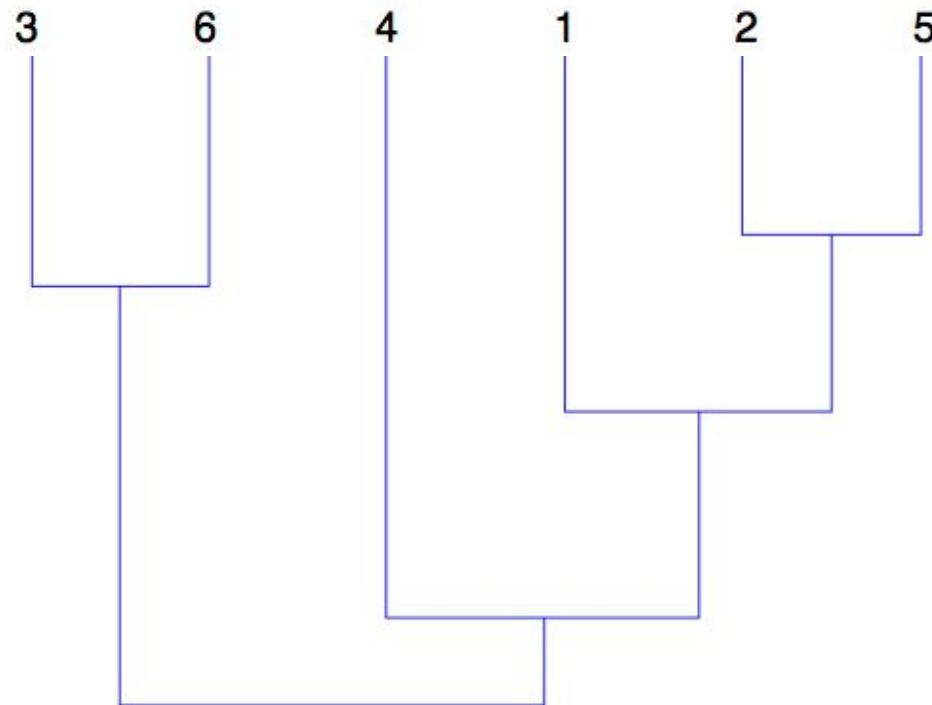
Métodos jerárquicos, dendrograma

- Un dendrograma es un diagrama del tipo árbol que ilustra el proceso de generación de clusters producido por los algoritmos jerárquicos. Mientras uno de los ejes muestra los puntos originales, el otro eje, puede mostrar la distancia existente entre los clusters.



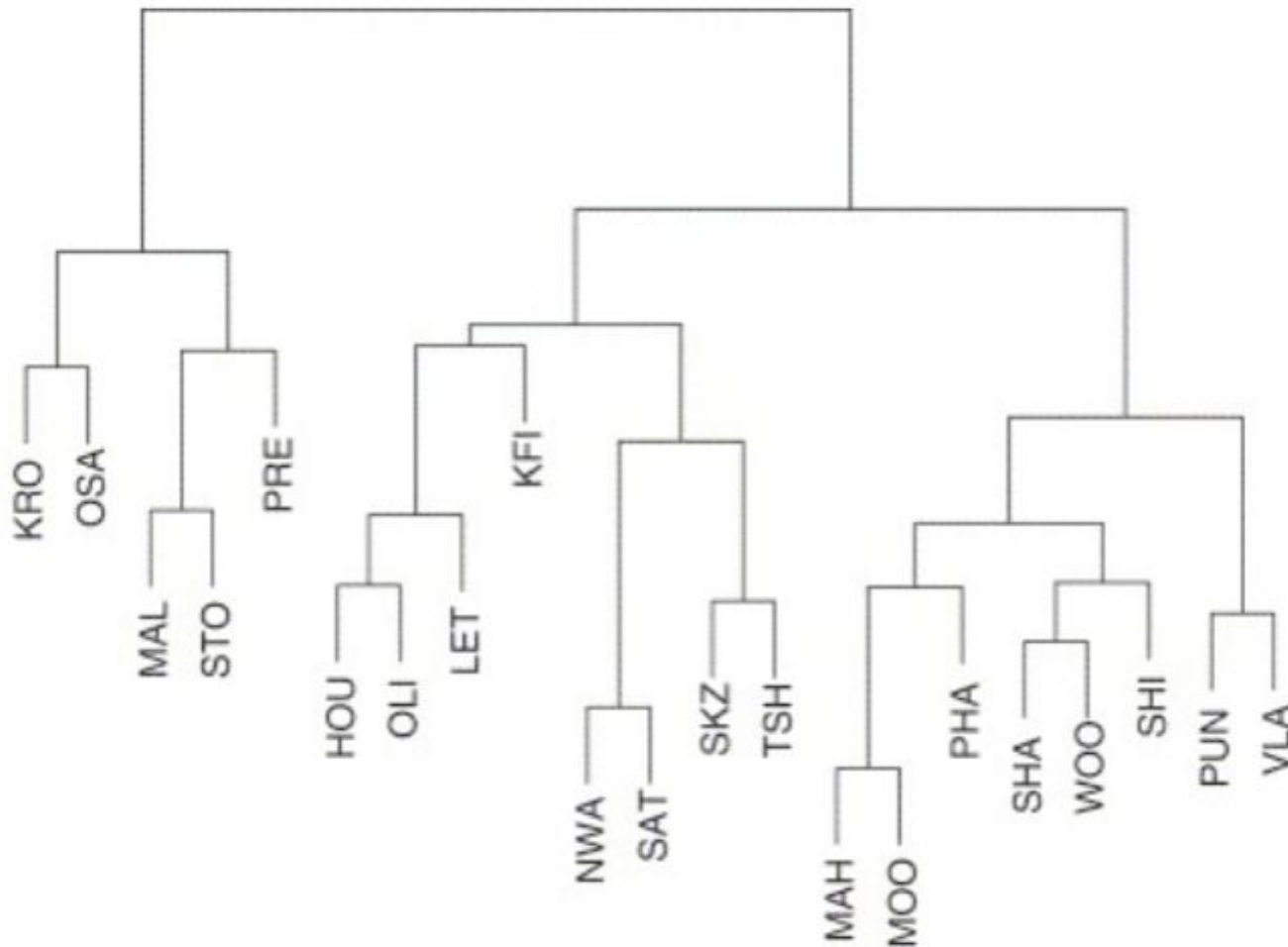
Métodos jerárquicos aglomerativos, introducción

- Método Jerárquico Divisivo: Comienza con todos los puntos como un solo cluster. Luego, en cada etapa, se selecciona un cluster y se separa en dos clusters. Este proceso se repite hasta que cada cluster tiene un punto.



Métodos jerárquicos aglomerativos, introducción

- Método Jerárquico Aglomerativo (MJA): Comienza con todos los puntos como clusters individuales. Luego, en cada etapa, se unen los cluster más cercanos entre si. Este proceso se repite hasta que se genera un solo gran cluster.



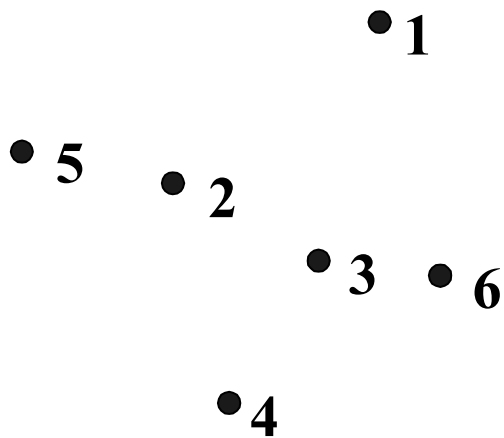
Métodos jerárquicos aglomerativos, algoritmo

- El algoritmo más básico del MJA es bastante sencillo
 - Sea cada punto un cluster
 - Calcula la matriz de proximidad/distancia entre cada cluster
 - Repetir**
 - Unir los cluster más cercanos
 - Recalcular/updatear la matriz de proximidad/distancia
 - Hasta** que exista un solo cluster
- Lo más importante de este proceso es el cálculo de la matriz de proximidad/distancia entre clusters
- Distintos enfoques de distancia entre clusters, segmentan los datos en forma distinta.

Métodos jerárquicos aglomerativos, ejemplo

- Comencemos con cada punto siendo un cluster y calculando la matriz de proximidad.
- Busquemos los cluster más cercanos.

	C1	C2	C3	C4	C5	C6
C1	0.00	0.54	0.51	0.95	0.84	0.59
C2		0.00	0.40	0.62	0.35	0.70
C3			0.00	0.45	0.74	0.30
C4				0.00	0.82	0.61
C5					0.00	1.04
C6						0.00



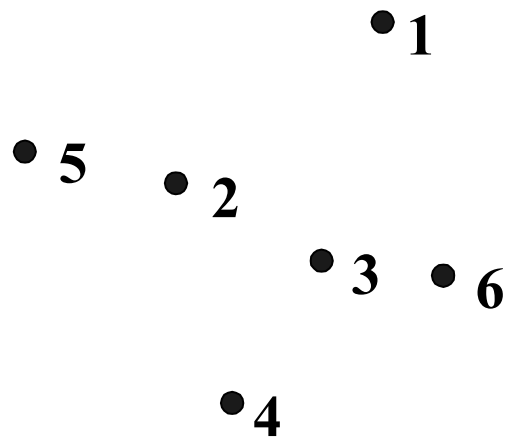
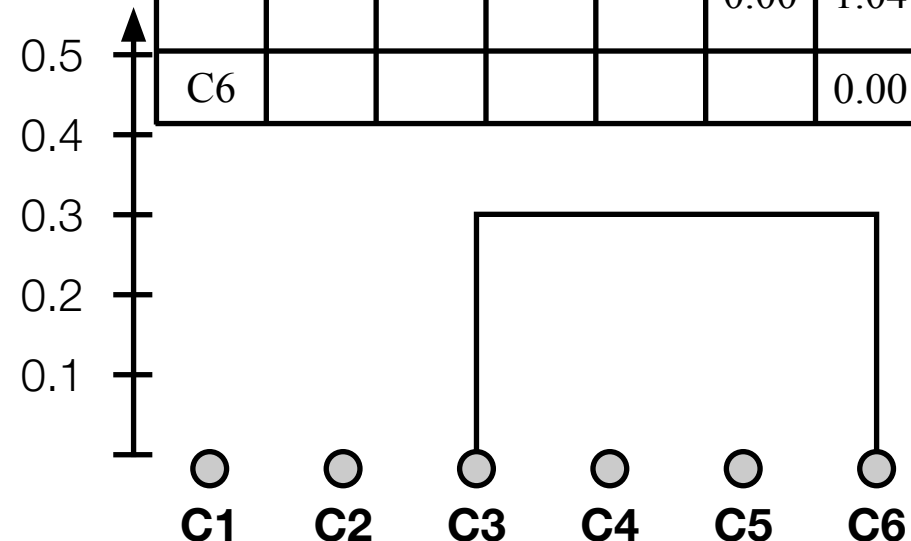
Datos originales

○ C1 ○ C2 ○ C3 ○ C4 ○ C5 ○ C6

Métodos jerárquicos aglomerativos, ejemplo

- Comencemos con cada punto siendo un cluster y calculando la matriz de proximidad.
- Busquemos los cluster más cercanos.
- Unamos los clusters.

	C1	C2	C3	C4	C5	C6
C1	0.00	0.54	0.51	0.95	0.84	0.59
C2		0.00	0.40	0.62	0.35	0.70
C3			0.00	0.45	0.74	0.30
C4				0.00	0.82	0.61
C5					0.00	1.04
C6						0.00

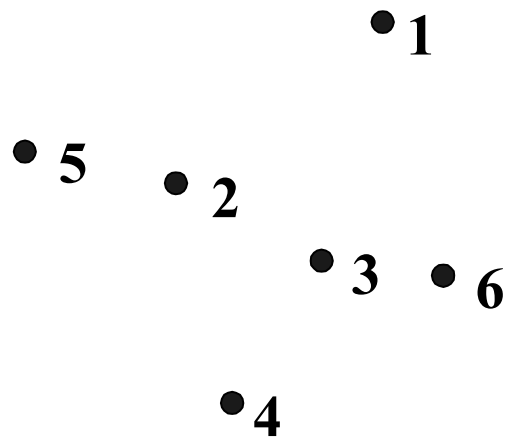


Datos originales

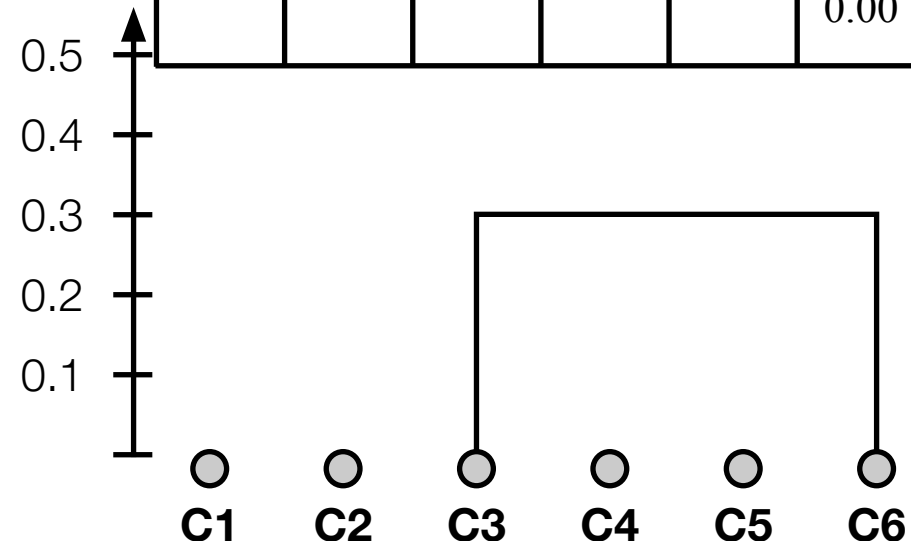
Métodos jerárquicos aglomerativos, ejemplo

- Comencemos con cada punto siendo un cluster y calculando la matriz de proximidad.
- Busquemos los cluster más cercanos.
- Unamos los clusters.
- Actualicemos la matriz de proximidad.

	C1	C2	C36	C4	C5
C1	0.00	0.54	???	0.95	0.84
C2		0.00	???	0.62	0.35
C36			0.00	???	???
C4				0.00	0.82
C5					0.00

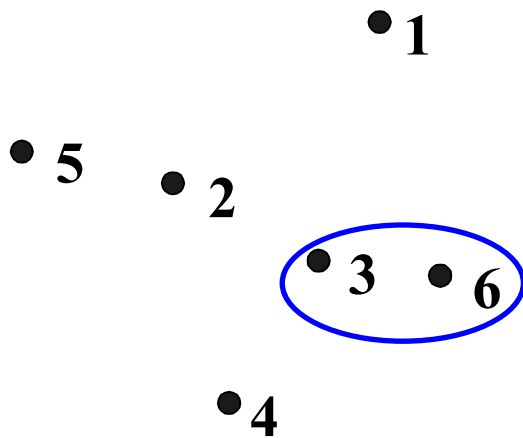


Datos originales



¿Cómo podemos actualizar la distancia entre clusters?

Matrices de Proximidad



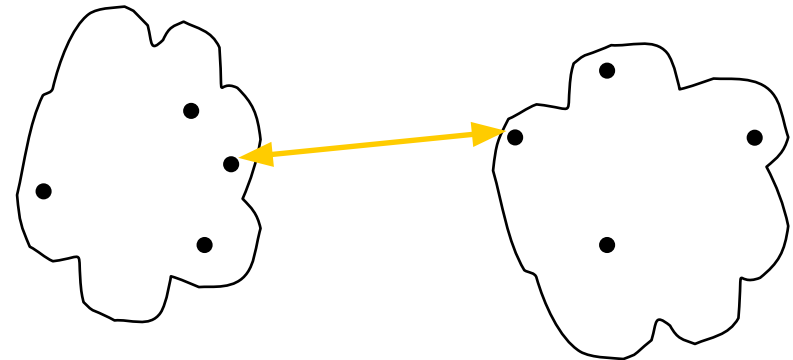
	C1	C2	C3	C4	C5	C6
C1	0.00	0.54	0.51	0.95	0.84	0.59
C2		0.00	0.40	0.62	0.35	0.70
C3			0.00	0.45	0.74	0.30
C4				0.00	0.82	0.61
C5					0.00	1.04
C6						0.00

	C1	C2	C36	C4	C5
C1	0.00	0.54	???	0.95	0.84
C2		0.00	???	0.62	0.35
C36			0.00	???	???
C4				0.00	0.82
C5					0.00

Datos originales

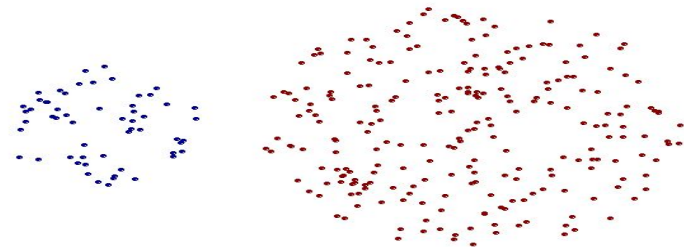
Métodos jerárquicos aglomerativos, single linkage

- **Single linkage:** La distancia entre clusters esta determinada por los puntos más similares (ceranos) entre los clusters.

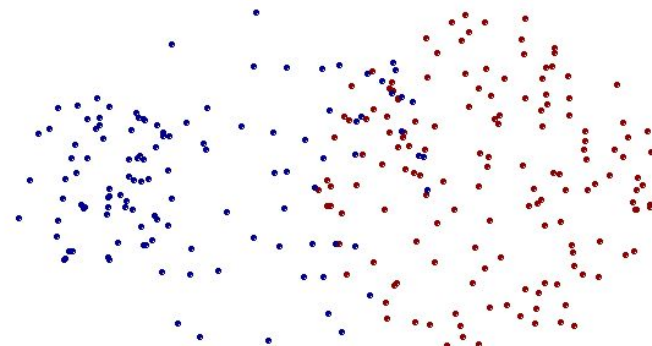


$$D(C_i, C_j) = \mathbf{min}\{d(x, y) | \mathbf{x} \in C_i, \mathbf{y} \in C_j\}$$

- **Ventaja:**
Genera cluster largos y delgados

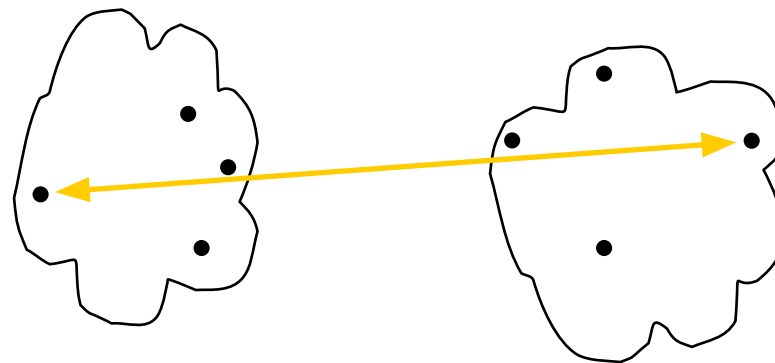


- **Limitación:**
Afectado por los datos atípicos outliers



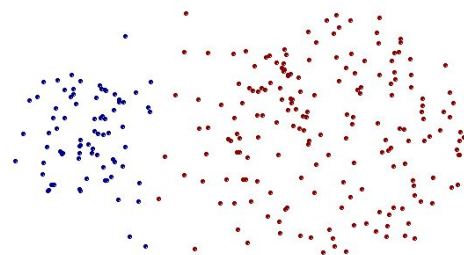
Métodos jerárquicos aglomerativos, complete linkage

- **Complete linkage:** La distancia entre clusters esta determinada por los puntos más disimiles (lejanos) entre los clusters.

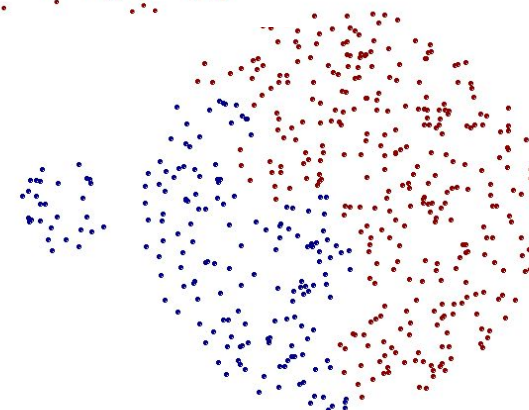


$$D(C_i, C_j) = \mathbf{max}\{d(x, y) | \mathbf{x} \in C_i, \mathbf{y} \in C_j\}$$

- **Ventaja:**
Menos susceptible a datos atípicos

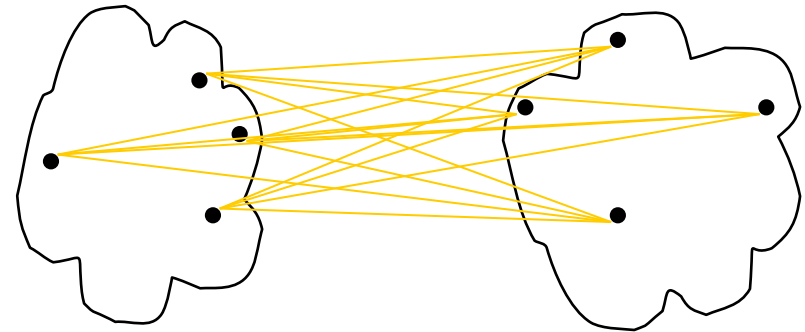


- **Limitación:**
 - Tiende a quebrar cluster grandes
 - Sesgado a generar clusters circulares



Métodos jerárquicos aglomerativos, average linkage

- **Average linkage:** La distancia entre clusters esta determinada por la distancia promedio entre todos los puntos de los clusters.

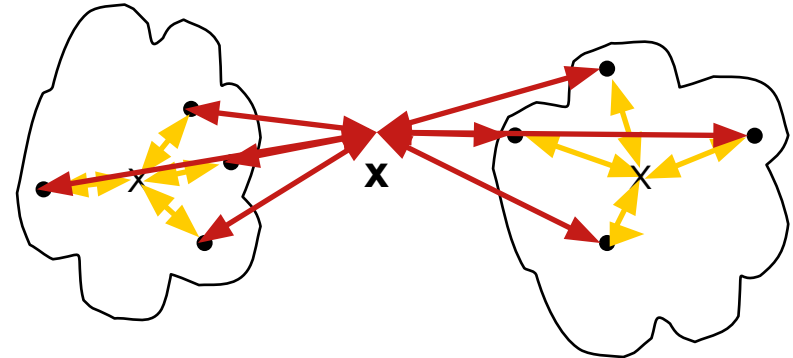


$$D(C_i, C_j) = \mathbf{avg}\{d(x, y) | \mathbf{x} \in C_i, \mathbf{y} \in C_j\}$$

- Acuerdo de compromiso entre Single y Complete Linkage
- **Ventaja:**
Menos susceptible a datos atípicos
- **Limitación:**
Sesgado a generar clusters circulares

Métodos jerárquicos aglomerativos, ward linkage

- **Ward linkage:** La distancia entre clusters esta determinada por el incremento del within cluster distance (usando squared euclidean distance) cuando se unen los clusters.

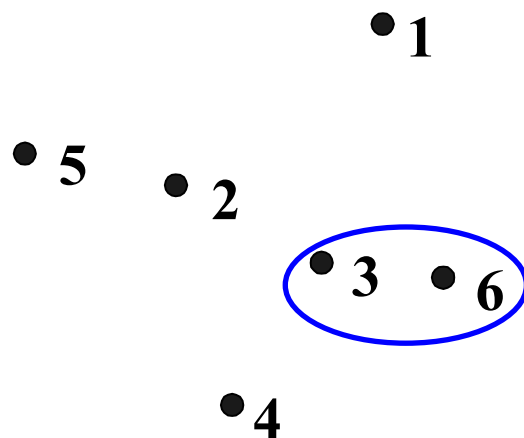


$$D(C_i, C_j) = wc(C_{ij}) - wc(C_i) - wc(C_j) = \frac{n_i n_j}{n_i + n_j} \|\bar{C}_i - \bar{C}_j\|^2$$

- Minimiza la distancia intra cluster y maximiza la distancia entre clusters.
- **Ventaja:**
Menos susceptible a datos atípicos
- **Limitación:**
Sesgado a generar clusters circulares

Métodos jerárquicos aglomerativos, ejemplo single linkage

- Actualicemos la matriz de proximidad usando single linkage.
- $d(C1, C3) = 0.51$ y $d(C1, C6) = 0.59$, como $d(C1, C3) < d(C1, C6)$ entonces $d(C1, C36) = d(C1, C3)$, es decir $d(C1, C36) = 0.51$.
- $d(C2, C3) = 0.40$ y $d(C2, C6) = 0.70 \Rightarrow d(C2, C36) = 0.40$.
 $d(C4, C3) = 0.45$ y $d(C4, C6) = 0.61 \Rightarrow d(C4, C36) = 0.45$.
 $d(C5, C3) = 0.74$ y $d(C5, C6) = 1.04 \Rightarrow d(C5, C36) = 0.74$.



Datos originales

	C1	C2	C3	C4	C5	C6
C1	0.00	0.54	0.51	0.95	0.84	0.59
C2		0.00	0.40	0.62	0.35	0.70
C3			0.00	0.45	0.74	0.30
C4				0.00	0.82	0.61
C5					0.00	1.04
C6						0.00

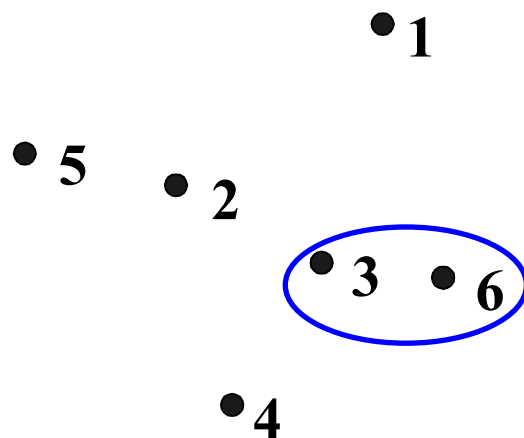
Matrices de Proximidad

	C1	C2	C36	C4	C5
C1	0.00	0.54	???	0.95	0.84
C2		0.00	???	0.62	0.35
C36			0.00	???	???
C4				0.00	0.82
C5					0.00

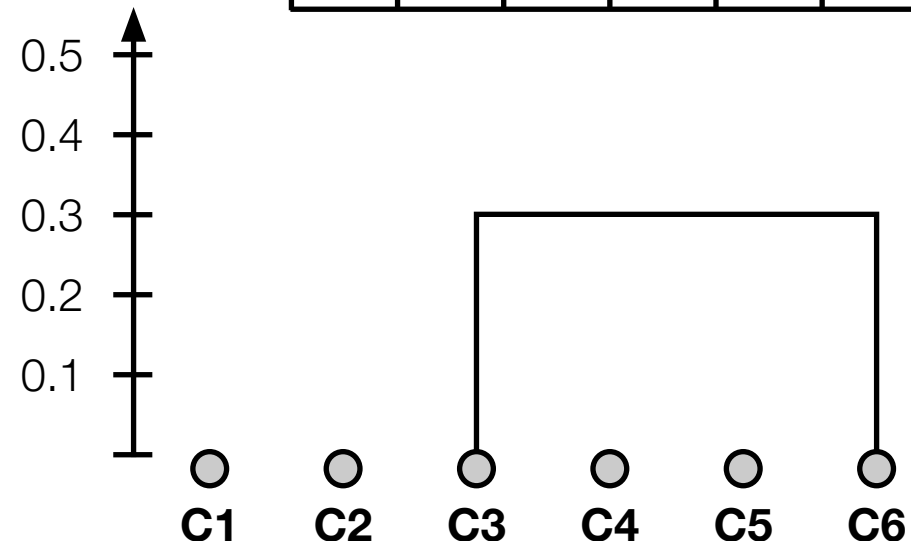
Métodos jerárquicos aglomerativos, ejemplo single linkage

- Comencemos con cada punto siendo un cluster y calculando la matriz de proximidad.
- **Busquemos los cluster más cercanos.**
- Unamos los clusters.
- Actualicemos la matriz de proximidad.

	C1	C2	C36	C4	C5
C1	0.00	0.54	0.51	0.95	0.84
C2		0.00	0.40	0.62	0.35
C36			0.00	0.45	0.74
C4				0.00	0.82
C5					0.00



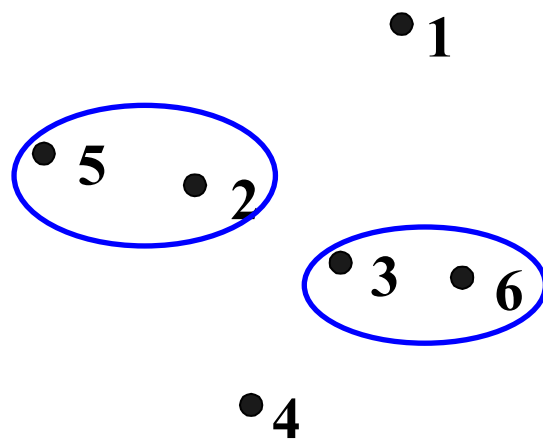
Datos originales



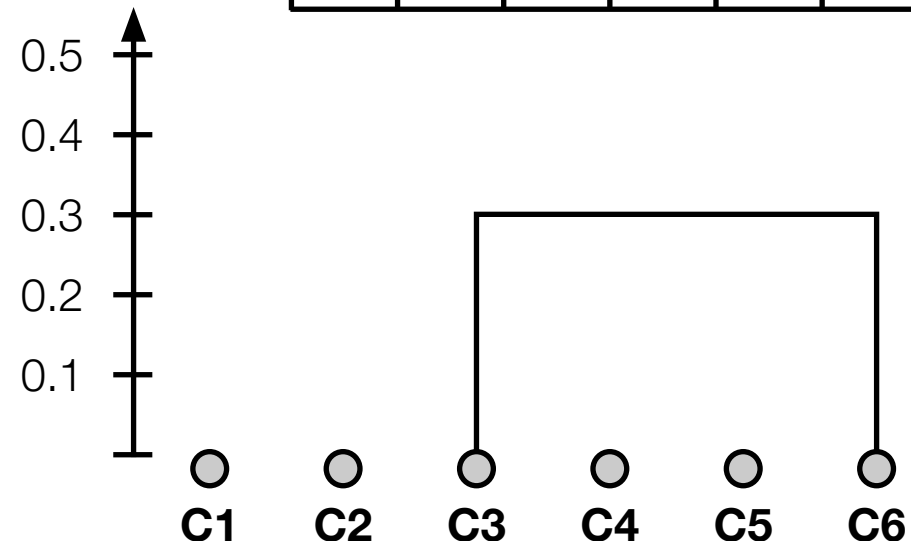
Métodos jerárquicos aglomerativos, ejemplo single linkage

- Comencemos con cada punto siendo un cluster y calculando la matriz de proximidad.
- Busquemos los cluster más cercanos.
- **Unamos los clusters.**
- Actualicemos la matriz de proximidad.

	C1	C2	C36	C4	C5
C1	0.00	0.54	0.51	0.95	0.84
C2		0.00	0.40	0.62	0.35
C36			0.00	0.45	0.74
C4				0.00	0.82
C5					0.00



Datos originales



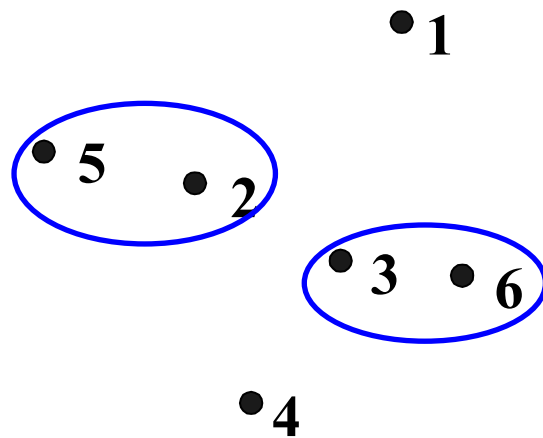
Métodos jerárquicos aglomerativos, ejemplo single linkage

- Comencemos con cada punto siendo un cluster y calculando la matriz de proximidad.
- Busquemos los cluster más cercanos.
- Unamos los clusters.

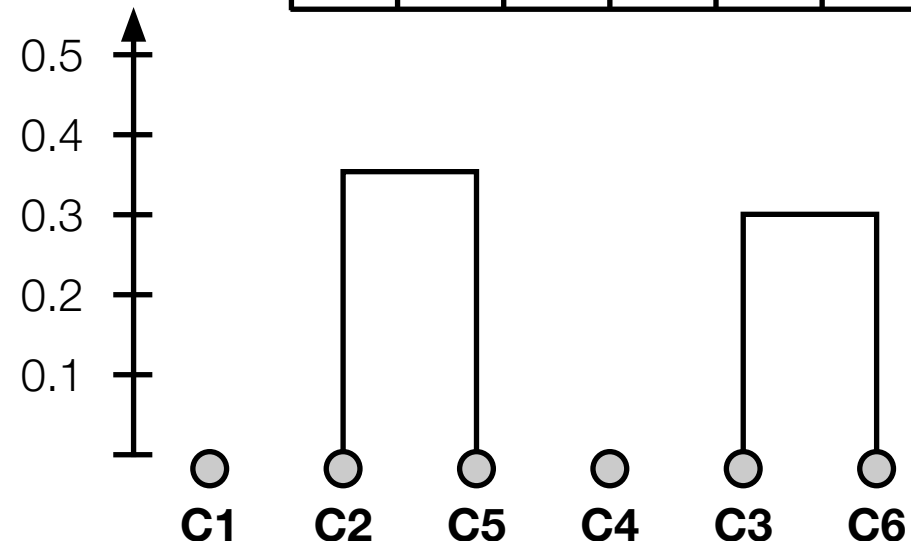
- **Actualicemos la matriz de proximidad.**

- $d(C1, C25) = 0.54$ $d(C25, C36) = 0.40$
 $d(C4, C25) = 0.62$

	C1	C2	C36	C4	C5
C1	0.00	0.54	0.51	0.95	0.84
C2		0.00	0.40	0.62	0.35
C36			0.00	0.45	0.74
C4				0.00	0.82
C5					0.00



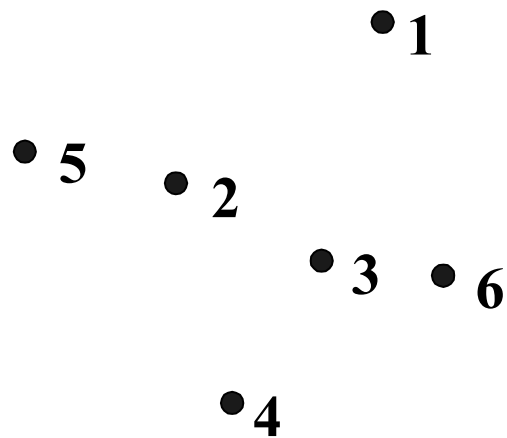
Datos originales



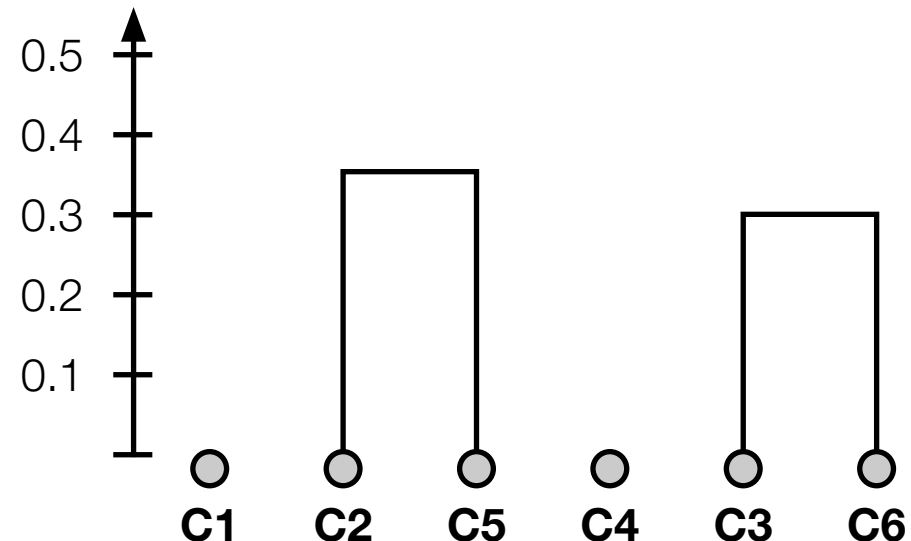
Métodos jerárquicos aglomerativos, ejemplo single linkage

- Comencemos con cada punto siendo un cluster y calculando la matriz de proximidad.
- **Busquemos los cluster más cercanos.**
- **Unamos los clusters.**
- Actualicemos la matriz de proximidad.

	C1	C25	C36	C4
C1	0.00	0.54	0.51	0.95
C25		0.00	0.40	0.62
C36			0.00	0.45
C4				0.00



Datos originales



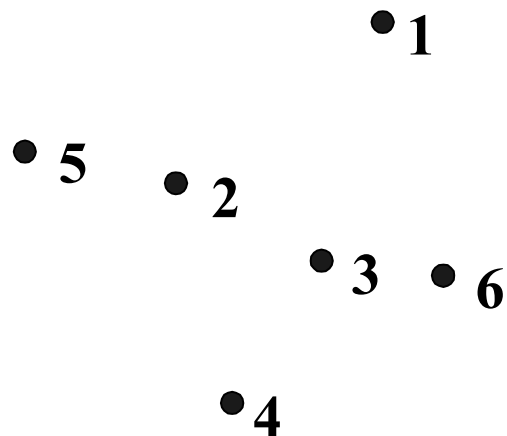
Métodos jerárquicos aglomerativos, ejemplo single linkage

- Comencemos con cada punto siendo un cluster y calculando la matriz de proximidad.
- Busquemos los cluster más cercanos.
- Unamos los clusters.

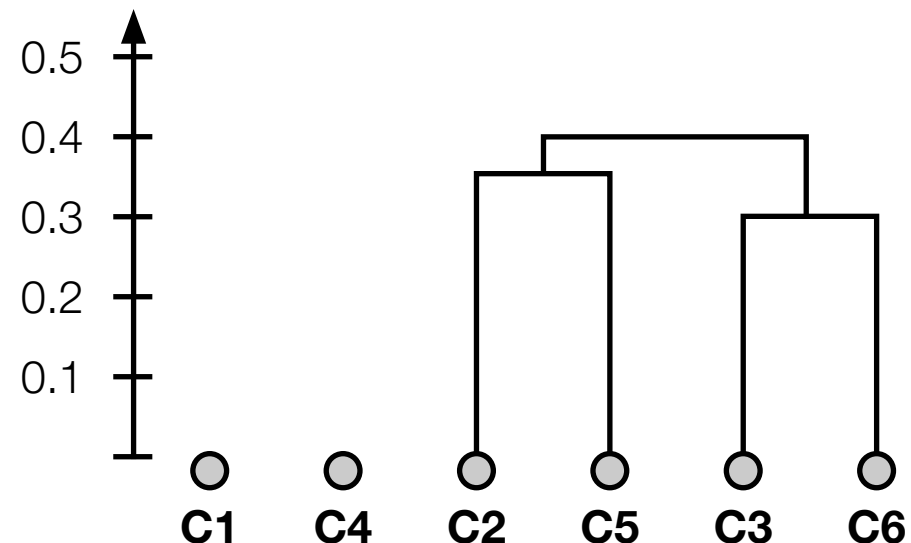
- **Actualicemos la matriz de proximidad.**

- $d(C1, C2536) = 0.51$
 $d(C4, C2536) = 0.45$

	C1	C25	C36	C4
C1	0.00	0.54	0.51	0.95
C25		0.00	0.40	0.62
C36			0.00	0.45
C4				0.00



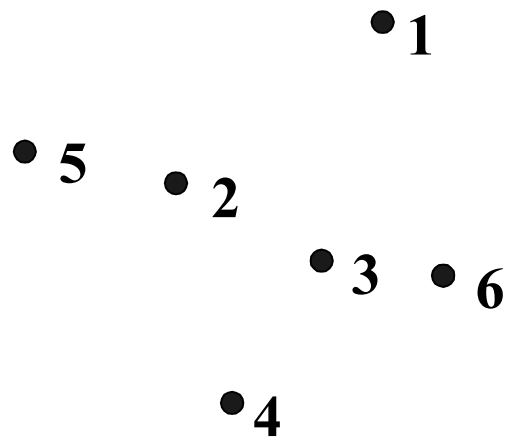
Datos originales



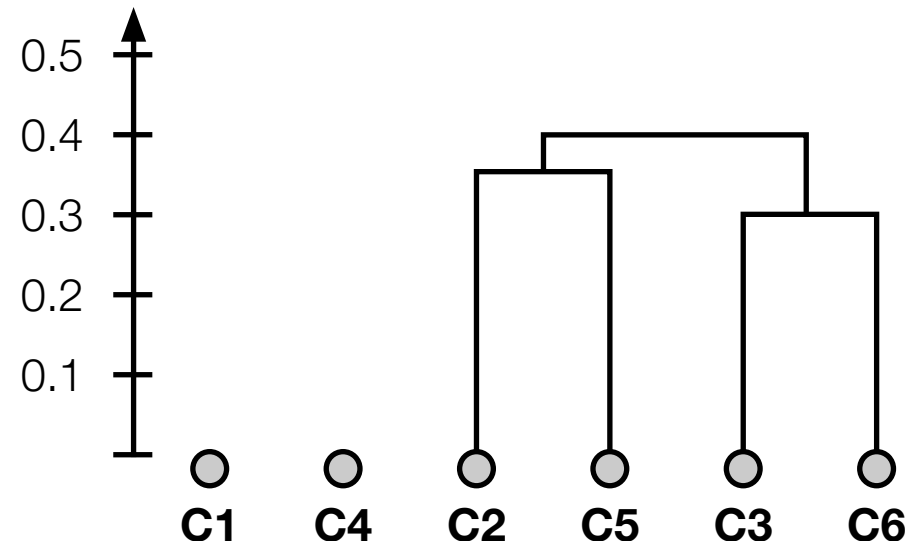
Métodos jerárquicos aglomerativos, ejemplo single linkage

- Comencemos con cada punto siendo un cluster y calculando la matriz de proximidad.
- **Busquemos los cluster más cercanos.**
- **Unamos los clusters.**
- Actualicemos la matriz de proximidad.

	C1	C2536	C4
C1	0.00	0.51	0.95
C2536		0.00	0.45
C4			0.00



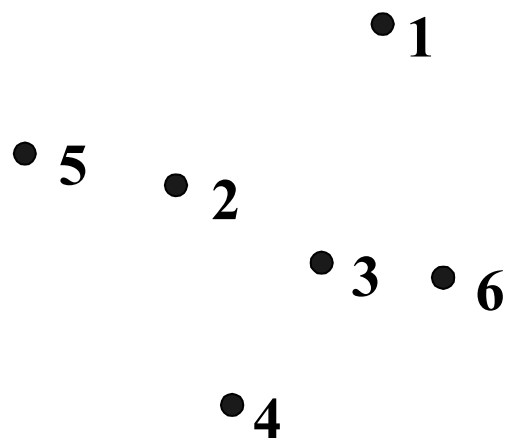
Datos originales



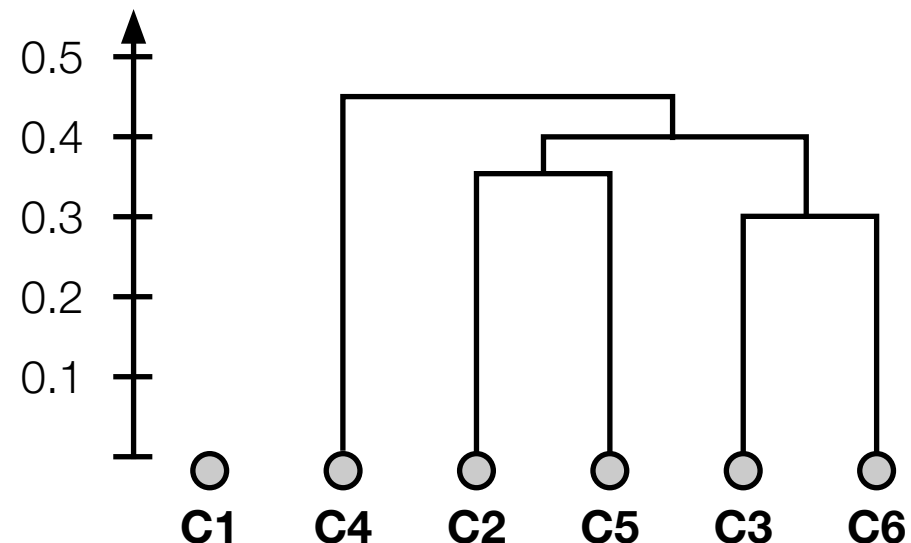
Métodos jerárquicos aglomerativos, ejemplo single linkage

- Comencemos con cada punto siendo un cluster y calculando la matriz de proximidad.
- Busquemos los cluster más cercanos.
- Unamos los clusters.
- **Actualicemos la matriz de proximidad.**
- $d(C1, C25364) = 0.51$

	C1	C2536	C4
C1	0.00	0.51	0.95
C2536		0.00	0.45
C4			0.00

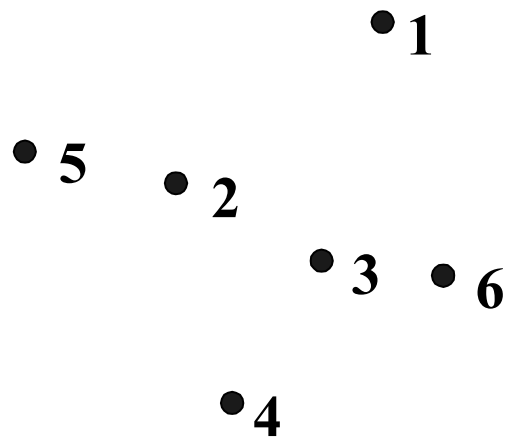


Datos originales

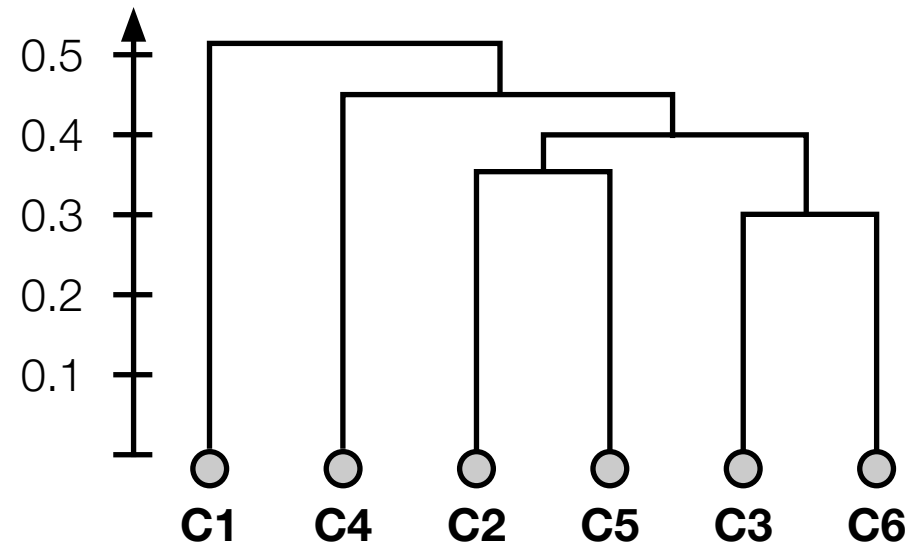


Métodos jerárquicos aglomerativos, ejemplo single linkage

- Comencemos con cada punto siendo un cluster y calculando la matriz de proximidad.
- Busquemos los cluster más cercanos.
- Unamos los clusters.
- Actualicemos la matriz de proximidad.



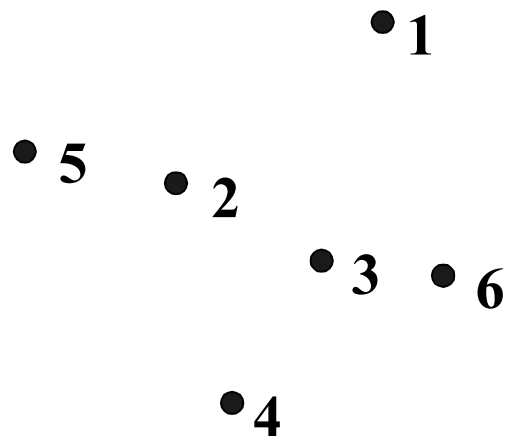
Datos originales



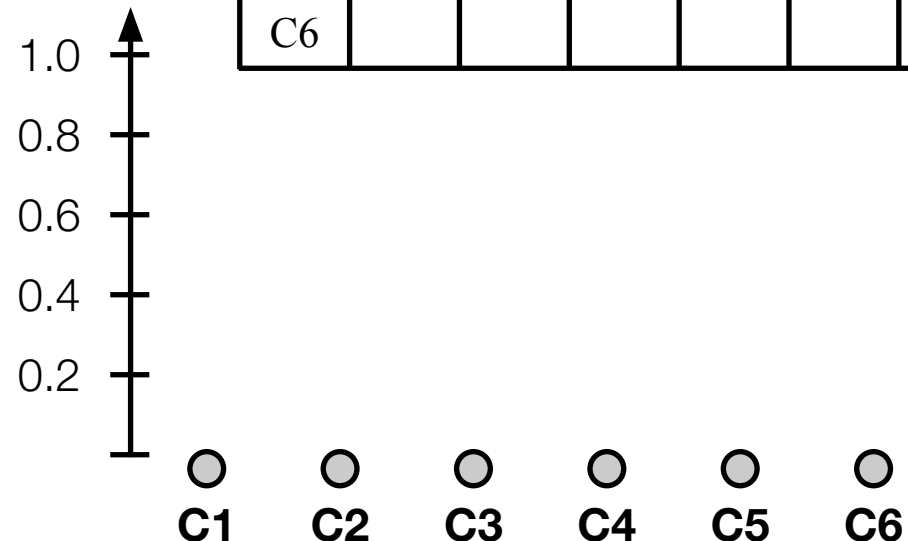
Métodos jerárquicos aglomerativos, ejemplo complete linkage

- Comencemos con cada punto siendo un cluster y calculando la matriz de proximidad.
- **Busquemos los cluster más cercanos.**
- **Unamos los clusters.**
- **Actualicemos la matriz de proximidad.**
- $d(C1, C6) = 0.59$ $d(C2, C6) = 0.70$
 $d(C4, C6) = 0.61$ $d(C5, C6) = 1.04$

	C1	C2	C3	C4	C5	C6
C1	0.00	0.54	0.51	0.95	0.84	0.59
C2		0.00	0.40	0.62	0.35	0.70
C3			0.00	0.45	0.74	0.30
C4				0.00	0.82	0.61
C5					0.00	1.04
C6						0.00



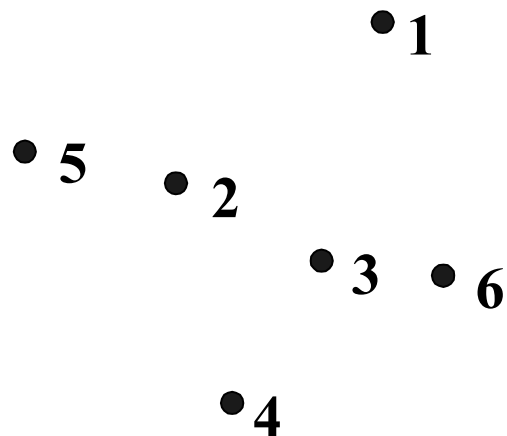
Datos originales



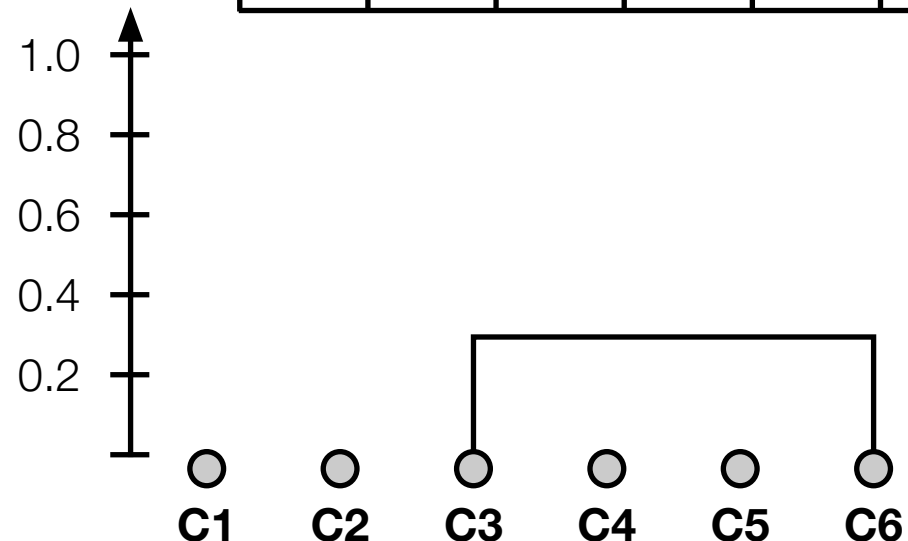
Métodos jerárquicos aglomerativos, ejemplo complete linkage

- Comencemos con cada punto siendo un cluster y calculando la matriz de proximidad.
- **Busquemos los cluster más cercanos.**
- **Unamos los clusters.**
- **Actualicemos la matriz de proximidad.**
- $d(C1, C25) = 0.84$ $d(C4, C25) = 0.82$
 $d(C36, C25) = 1.04$

	C1	C2	C36	C4	C5
C1	0.00	0.54	0.59	0.95	0.84
C2		0.00	0.70	0.62	0.35
C36			0.00	0.61	1.04
C4				0.00	0.82
C5					0.00



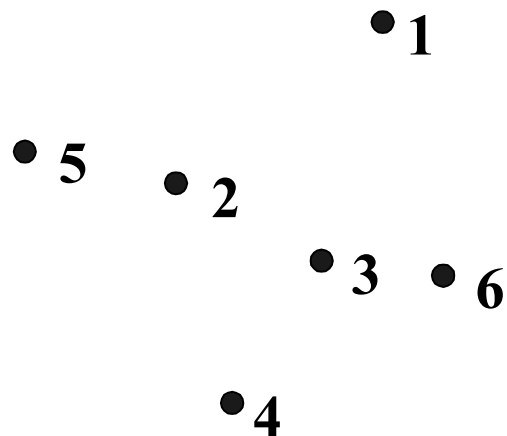
Datos originales



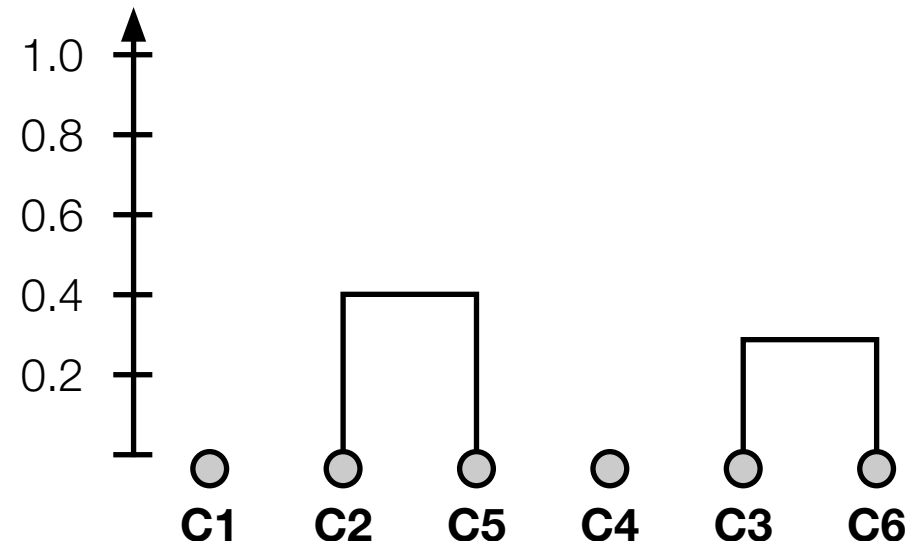
Métodos jerárquicos aglomerativos, ejemplo complete linkage

- Comencemos con cada punto siendo un cluster y calculando la matriz de proximidad.
- **Busquemos los cluster más cercanos.**
- **Unamos los clusters.**
- **Actualicemos la matriz de proximidad.**
- $d(C25, C136) = 1.04$
 $d(C4, C136) = 0.95$

	C1	C25	C36	C4
C1	0.00	0.84	0.59	0.95
C25		0.00	1.04	0.82
C36			0.00	0.61
C4				0.00



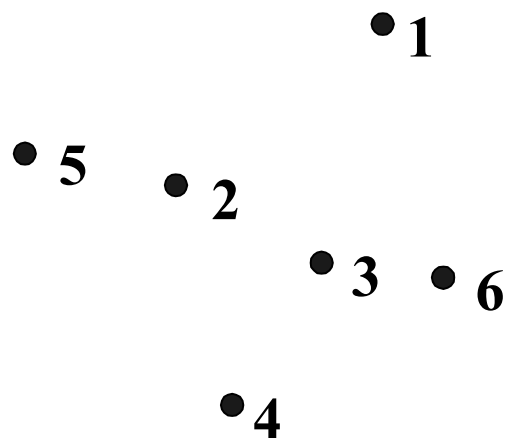
Datos originales



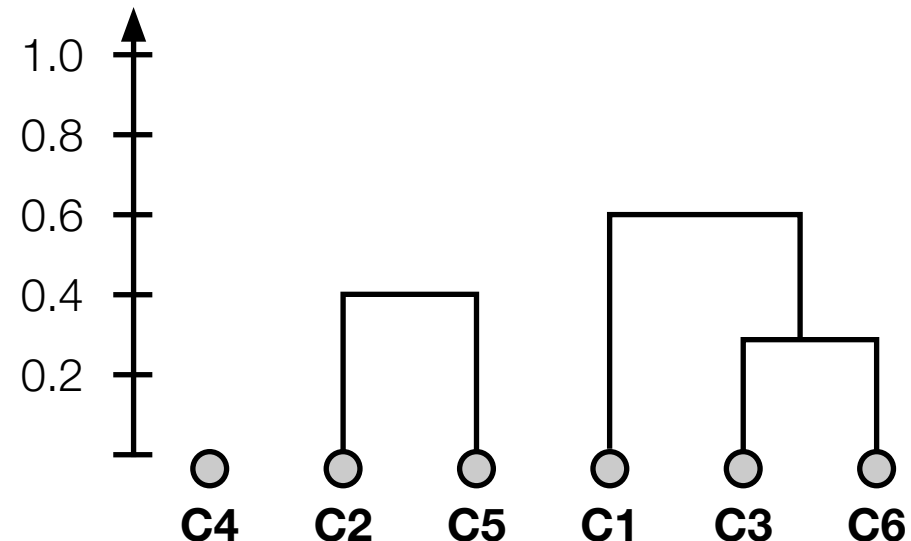
Métodos jerárquicos aglomerativos, ejemplo complete linkage

- Comencemos con cada punto siendo un cluster y calculando la matriz de proximidad.
- **Busquemos los cluster más cercanos.**
- **Unamos los clusters.**
- **Actualicemos la matriz de proximidad.**
- $d(C254, C136) = 1.04$

	C136	C25	C4
C136	0.00	1.04	0.95
C25		0.00	0.82
C4			0.00



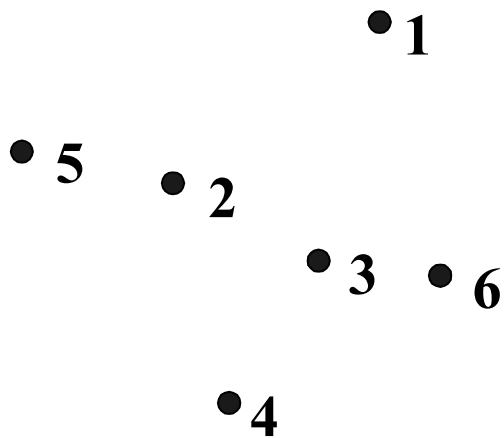
Datos originales



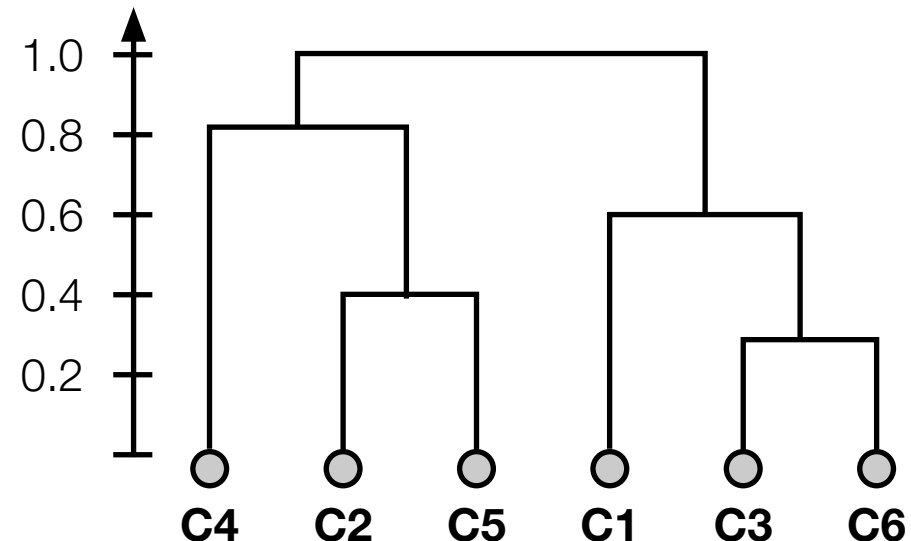
Métodos jerárquicos aglomerativos, ejemplo complete linkage

- Comencemos con cada punto siendo un cluster y calculando la matriz de proximidad.
- **Busquemos los cluster más cercanos.**
- **Unamos los clusters.**
- **Actualicemos la matriz de proximidad.**

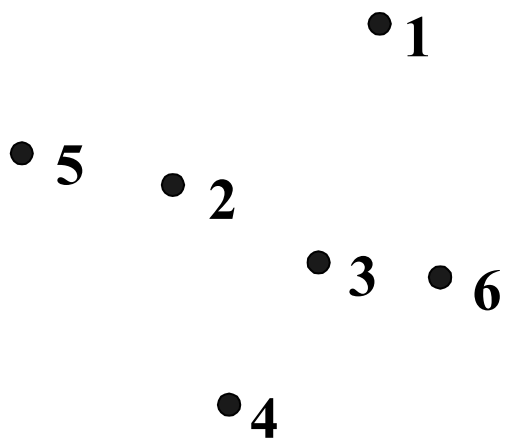
	C136	C254
C136	0.00	1.04
C254		0.00



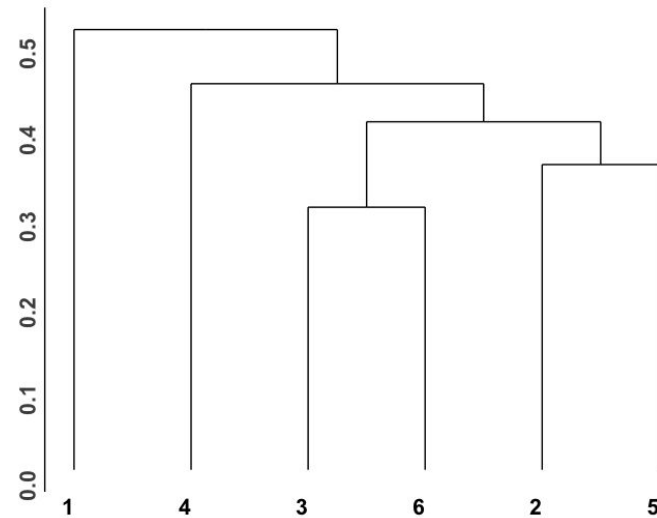
Datos originales



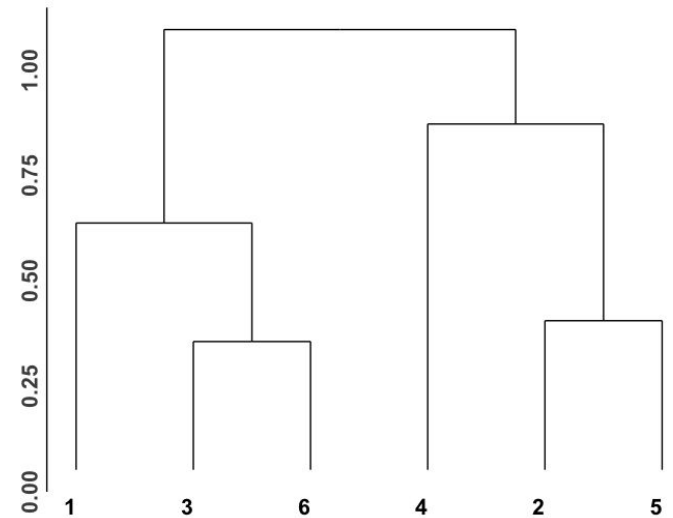
Métodos jerárquicos aglomerativos, ejemplos



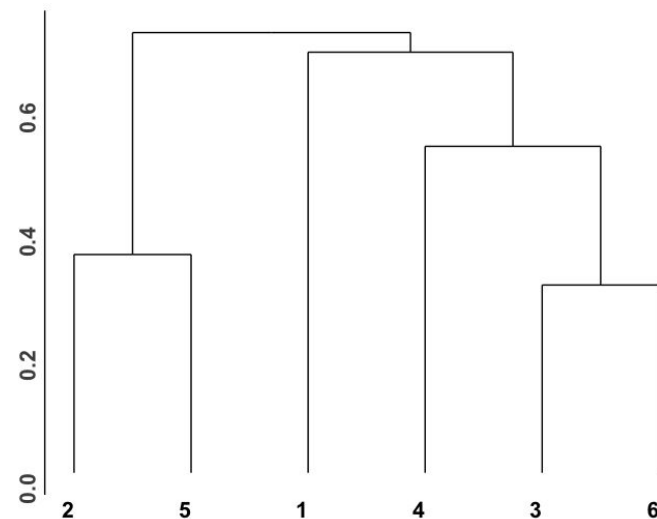
Datos originales



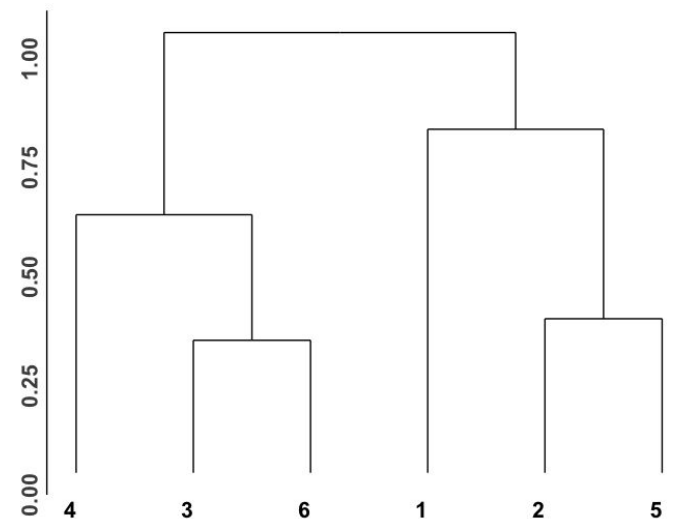
Single linkage



Complete linkage



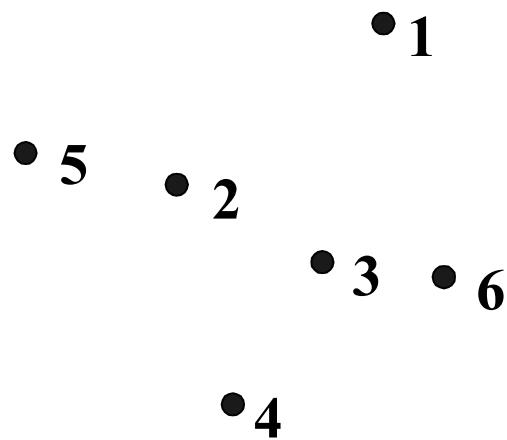
Average linkage



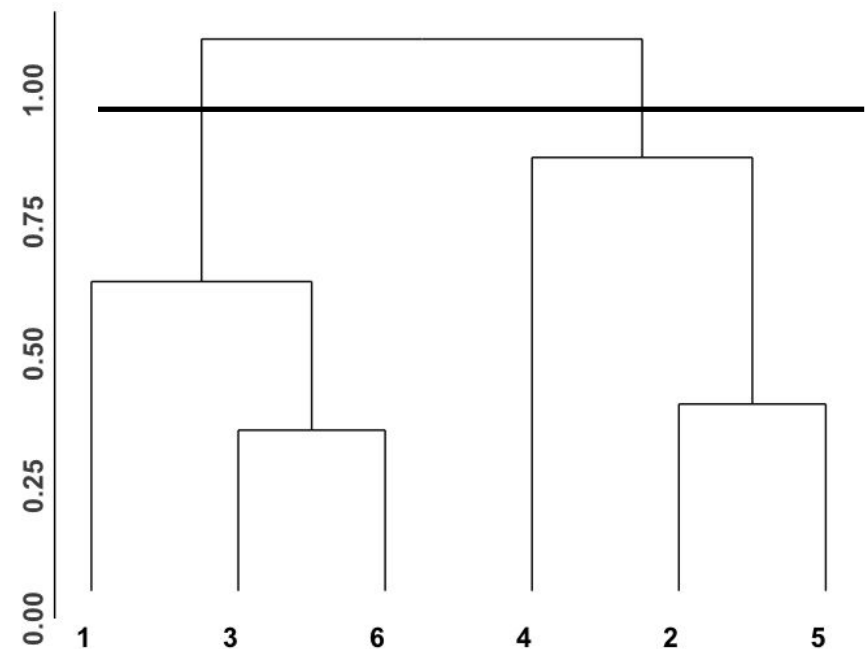
Ward linkage

Métodos jerárquicos aglomerativos, determinando el número de clusters

- A pesar de que el dendrograma muestra toda la información de los clusters generados, existen tantos posibles cluster como uniones posibles (el número de datos existentes)
- Como regla básica, los posibles clusters se determinan por uniones de clusters distantes entre ellas.



Original data



Complete linkage

Métodos jerárquicos aglomerativos, código en Python

- El cluster jerárquico utiliza la clase `sklearn.cluster.AgglomerativeClustering`
`AgglomerativeClustering(n_clusters=2, affinity='euclidean', linkage='ward', compute_full_tree = 'auto')`
- Parámetros
`n_clusters`: el número de clusters a buscar.
`affinity`: métrica de distancia, puede 'euclidean', 'manhattan', 'precomputed'.
`linkage`: 'ward', 'complete', 'average', 'single'.
`compute_full_tree`: 'auto', 'False', 'True', Si se quiere, se puede finalizar el algoritmo una vez alcanzado los `n_clusters`.
- Atributos
`n_clusters_`: número de clusters a buscar.
`labels_`: etiquetas de los puntos.
- Funciones
`fit(X)` y **`fit_predict(X)`**

Métodos jerárquicos aglomerativos, código en Python

- sklearn tiene varias deficiencias para este modelo. Siendo las más importantes, el definir un número de cluster y no generar un dendrograma en forma sencilla.
- Utilizaremos las funciones `linkage` y `dendrogram` de la biblioteca `scipy.cluster.hierarchy` en vez de sklearn.
- La función `linkage` aprende el modelo definido y retorna un dendrograma `linkage(y, method='single')`

Parámetros:

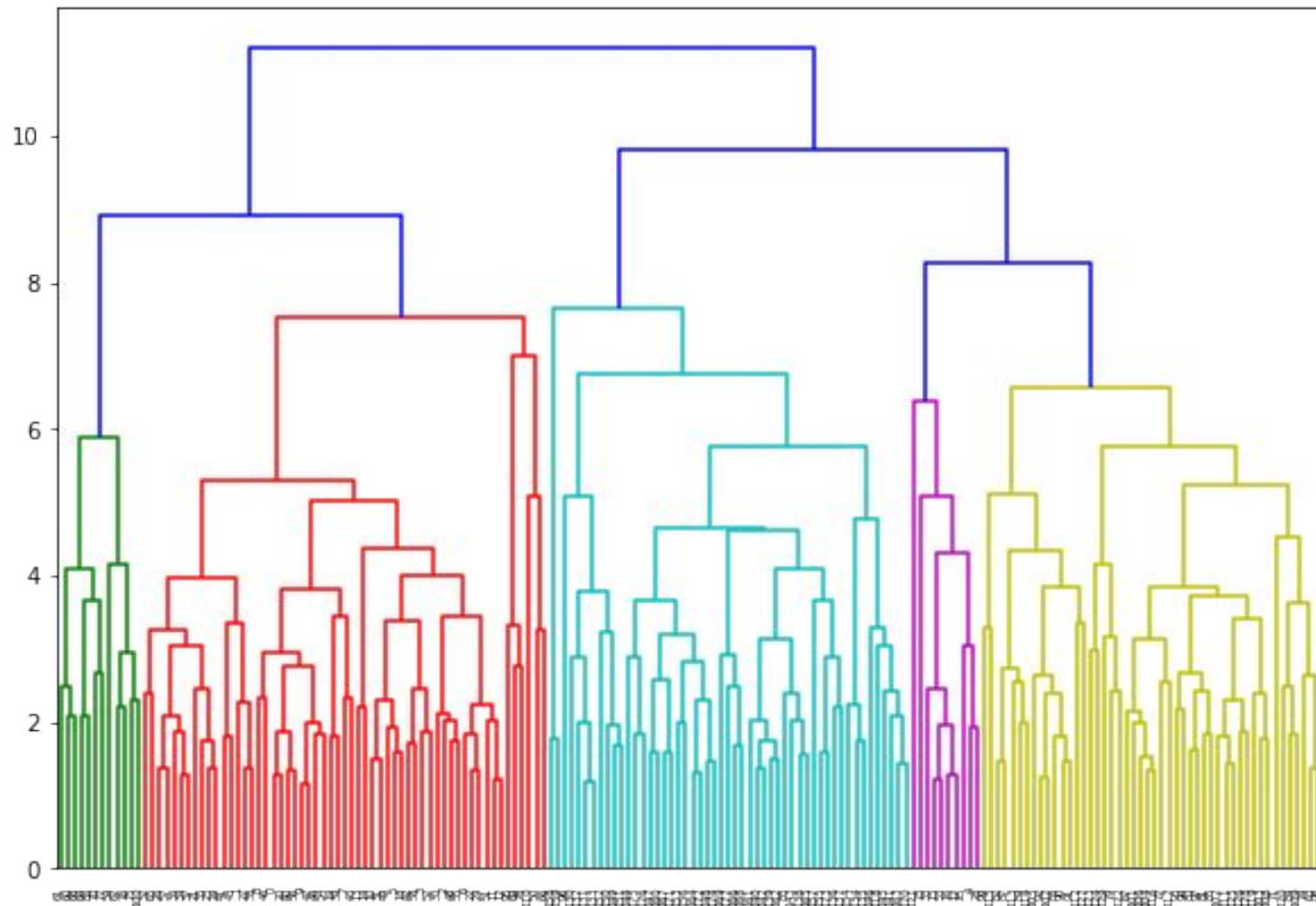
`y`: son los datos del modelo

`method`: 'ward', 'complete', 'average', 'single'.

- Para graficar basta con llamar a la función `dendrogram` con el modelo aprendido `dendrogram(modeloAprendido)`

Métodos jerárquicos aglomerativos, código en Python

- `modelo = shc.linkage(scaled_features, method='complete')`
`plt.figure(figsize=(10, 7))`
`objeto = shc.dendrogram(modelo)`



Métodos jerárquicos aglomerativos, código en Python

- Para extraer los clusters se utiliza la función **fcluster**, la cual recibe tres parámetros
 - **Z**: el modelo generado por la función linkage
 - **t**: el valor donde se quiere cortar el dendrograma
 - **criterion="distance"**. existen otros criterios de corte, pero distance es el que nos permite generar los clusters, basados en la altura que definimos.

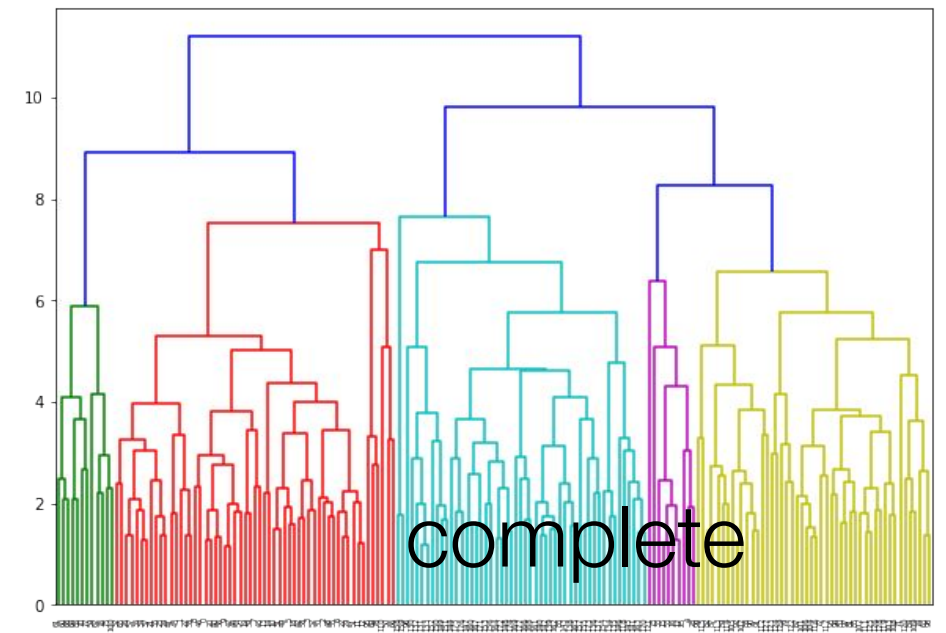
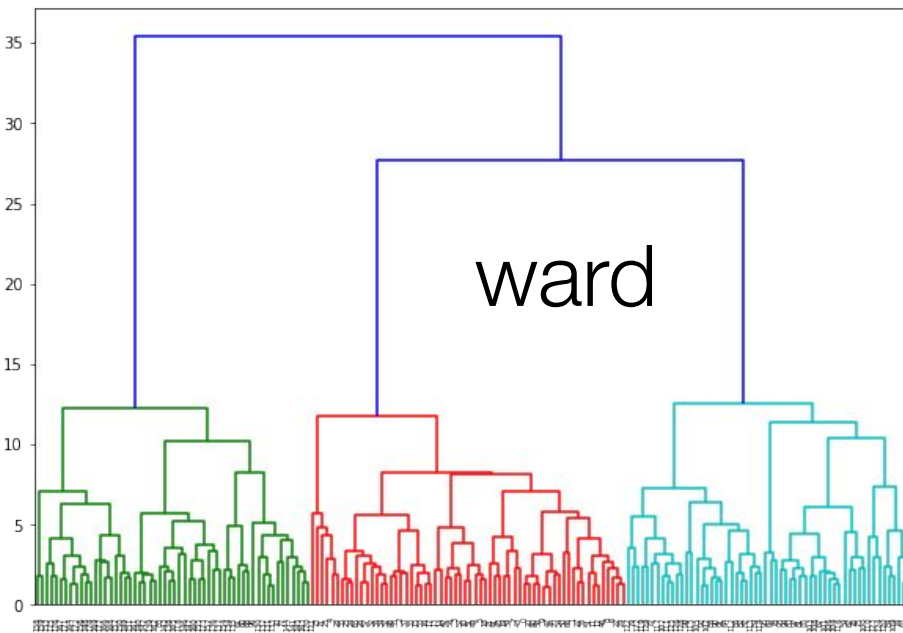
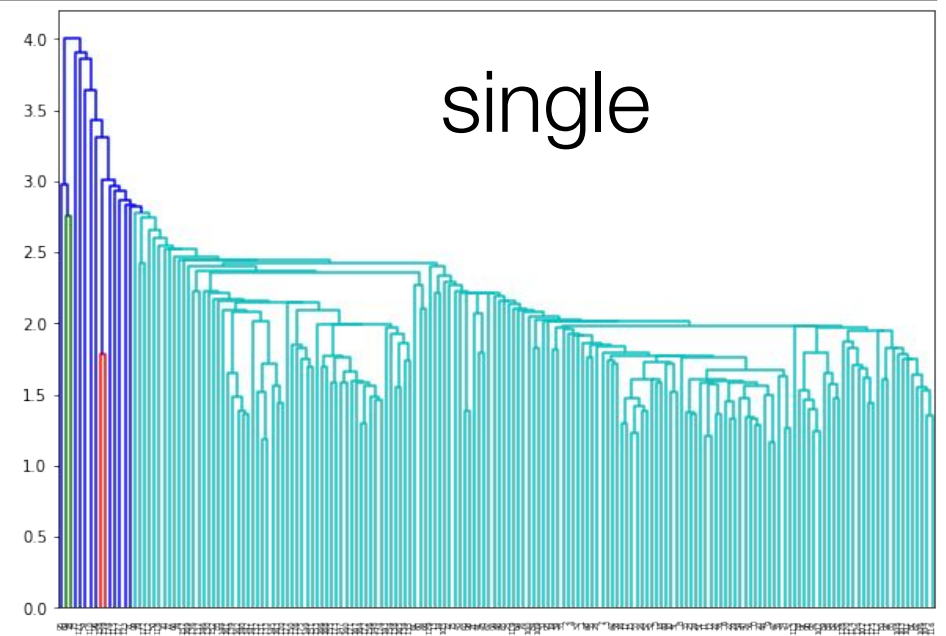
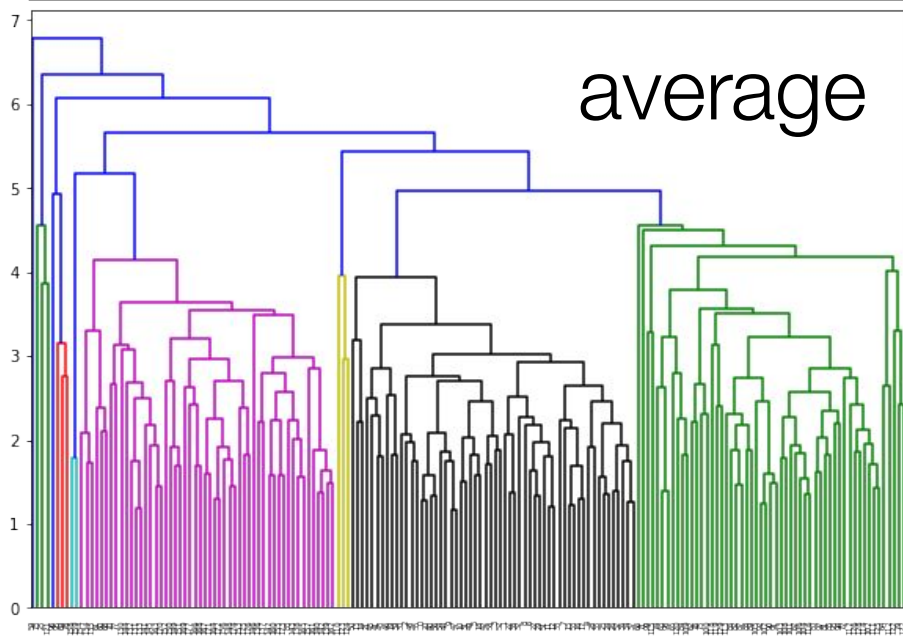
- Ejemplo:

```
clusters=shc.fcluster(modelo,t=10,criterion="distance")
```

clusters

[illegible]

Métodos jerárquicos aglomerativos, código en Python, ejemplos de dendrogramas



Métodos jerárquicos aglomerativos, problemas

- El algoritmo es muy lento computacionalmente
- Una vez que se ha definido combinar dos clusters, no existe la posibilidad de revertir esta decisión
- No existe una función objetivo para minimizar
- Diferentes enfoques tienen distintos problemas

Métodos jerárquicos aglomerativos, ejemplo

- A estudiantes le pedieron describir solo mencionando propiedades los siguientes conceptos:

-arm
-bacon
-chestnut
-eagle
-hair
-lips
-lobster
-ketchup
-sandwich
-scorpion
-thumb
-turkey

- 2625 propiedades fueron mencionadas.

317 propiedades distintas.

Mínimo y máximo número de propiedades distintas para un concepto 32 y 46 respectivamente.

Propiedad más mencionada para un concepto => 30

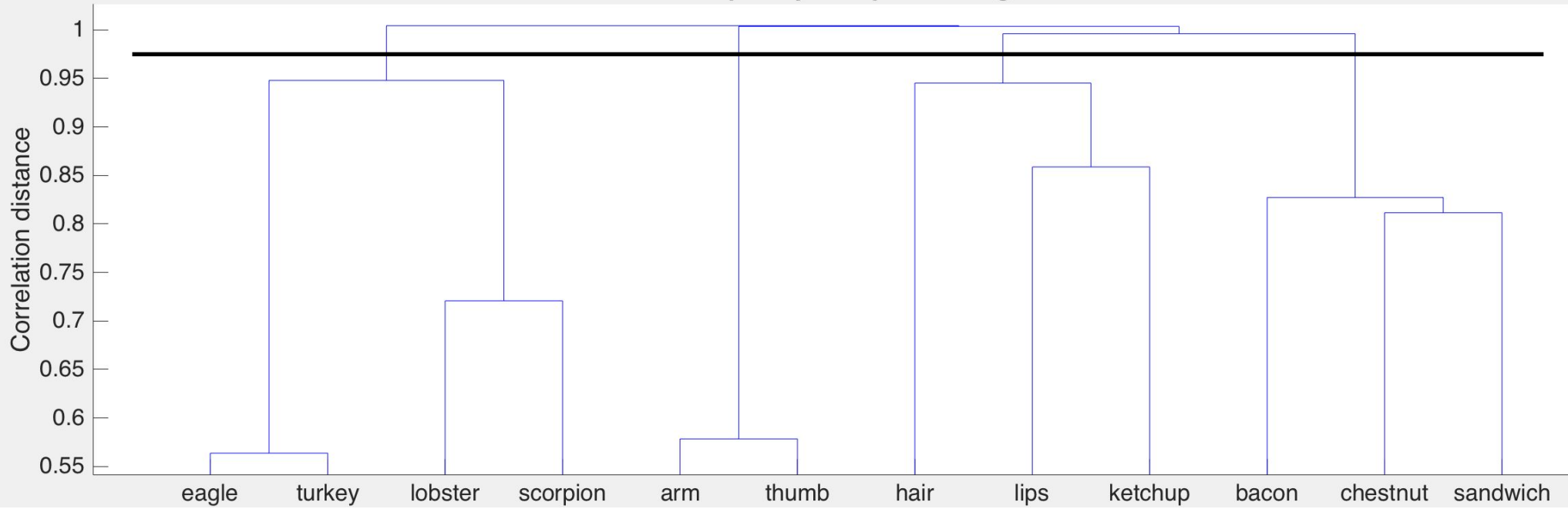
Ejemplo de los datos

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

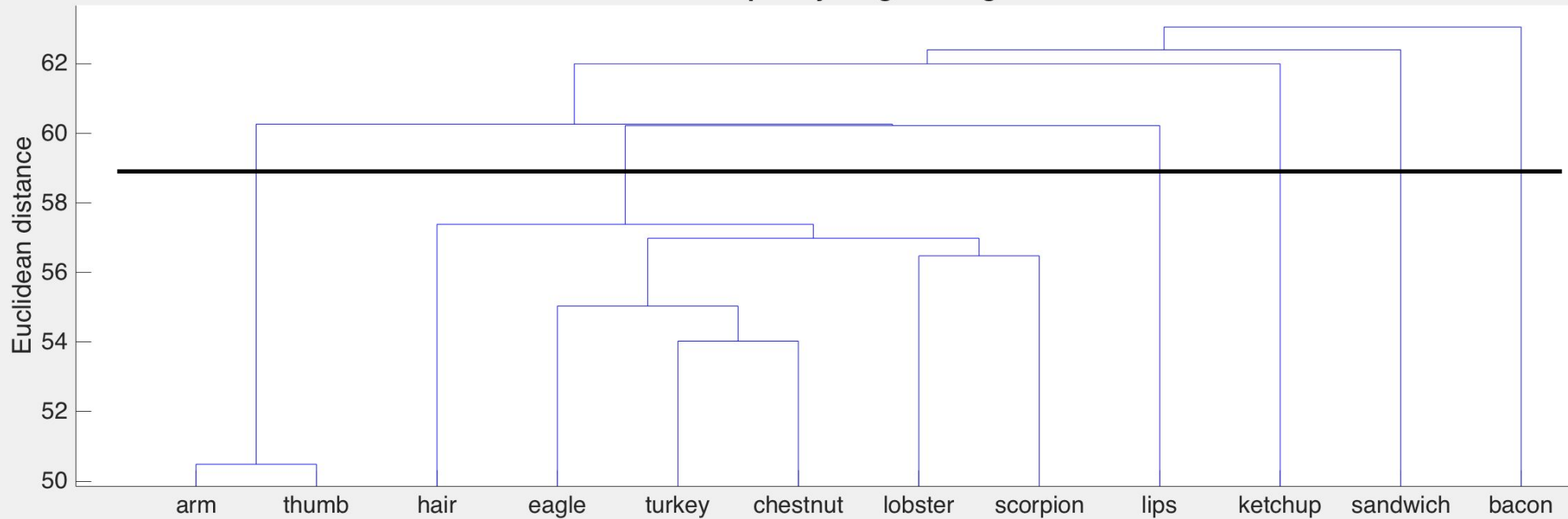
¿Qué clusters esperan?

¿Qué elección hay que realizar?

frequency Complete linkage



frequency Single linkage



Métodos jerárquicos aglomerativos, correlación, frecuencia.

- Scents—3 => bouquet, nail_polish, perfume.
- Big things—5 => castle, elephant, hippo, pyramid, rhino.
- Red things—7 => bacon, ham, heart, lips, lipstick, pig, tongue.
- Body parts—8 => arm, ear, finger, foot, leg, nose, thumb, toe.
- Long things—12 => bone, candle, crayon, eel, hair, hose, ..., whip, worm.
- Printed—12 => book, brochure, catalogue, ..., pamphlet, textbook.
- Dangerous things—14 => bomb, cigar, ..., poison, rattlesnake, revolver, ..., tobacco.
- Marine creatures—33 => carp, clam, cod, crab, ..., tortoise, trout, tuna, turtle, walrus.
- Food—46 => beer, biscuit, brandy, bread, butter, cake, ..., whisky, wine, yoghurt.
- Fruits & flowers—48 => apricot, aubergine, banana, birch, ..., watermelon, willow.
- Clothing—48 => apron, badge, bag, belt, bikini, blouse, ..., trousers, veil, wallet, wetsuit.
- Flying things—49 => aeroplane, bat_(animal), bee, ..., wasp, woodpecker, wren.
- Animals & insects—66 => alligator ant bear beaver toad tripod wolf zebra

Métodos jerárquicos aglomerativos, datos categóricos

- En el caso de datos categóricos hay que utilizar alguna función de distancia apropiada para este tipo de datos.
- La distancia de hamming corresponde al porcentaje de variables que difieren entre dos puntos.

Ejemplo

	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	V_9	V_{10}
$p_1 =$	1	4	3	3	2	3	1	0	4	0
$p_2 =$	0	4	3	2	2	3	1	4	4	1

$$\text{Hamming}(p_1, p_2) = 3/10 = 0.3$$

- Para ver otras posibles distancias

<https://docs.scipy.org/doc/scipy/reference/spatial.distance.html>

Métodos jerárquicos aglomerativos, datos categóricos

- En el caso de datos categóricos hay que utilizar la distancia de Hamming, que se encuentra en la librería `scipy.spatial.distance` función `pdist`.

- #Cargando librerías

```
from scipy.spatial.distance import pdist  
distCategorica = pdist(muestraDatos, 'hamming')
```

#Aplicando el modelo

```
modelo = shc.linkage(distCategorica, method='ward')
```

#Dendrograma final

```
plt.figure(figsize=(10, 7))
```

```
objeto = shc.dendrogram(modelo)
```

