Ciencia de Datos

Victor Muñiz

victor_m@cimat.mx

Asistente:

Víctor Gómez

victor.gomez@cimat.mx

Maestría en Cómputo Estadístico.
Centro de Investigación en Matemáticas.
Unidad Monterrey.

Enero-Junio 2021

Métodos de aprendizaje no supervisado

Aprendizaje no supervisado

"Aprender sin maestro..."

Dados n observaciones $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, con $\mathbf{x} \in \mathbb{R}^d$ ¿Qué podemos decir acerca de su densidad conjunta $P(\mathbf{x})$?

- Dos grandes fuentes de información: $\mu_{\mathbf{x}}$ y $\mathbf{S}$
- Estimación (paramétrica o no paramétrica) de la densidad de $\mathbf{x}$. *solo para d pequeño...*
- Una opción: variables latentes a través de proyecciones en baja dimensión de características "interesantes" de $P(\mathbf{x})$ (PCA).
- Otra opción: encontrar regiones convexas que contienen modas de $P(\mathbf{x})$ (mezclas de densidades o **grupos**).
- ¿Cómo se agrupan los datos? El concepto de **similaridad** o **disimilaridad** es fundamental.

Aprendizaje no supervisado

"Aprender sin maestro..."

Dados n observaciones $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, con $\mathbf{x} \in \mathbb{R}^d$ ¿Qué podemos decir acerca de su densidad conjunta $P(\mathbf{x})$?

- Dos grandes fuentes de información: $\mu_{\mathbf{x}}$ y $\mathbf{S}$
- Estimación (paramétrica o no paramétrica) de la densidad de $\mathbf{x}$. *solo para d pequeño...*
- Una opción: variables latentes a través de proyecciones en baja dimensión de características "interesantes" de $P(\mathbf{x})$ (PCA).
- Otra opción: encontrar regiones convexas que contienen modas de $P(\mathbf{x})$ (mezclas de densidades o **grupos**).
- ¿Cómo se agrupan los datos? El concepto de **similaridad** o **disimilaridad** es fundamental.

Aprendizaje no supervisado

"Aprender sin maestro..."

Dados n observaciones $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, con $\mathbf{x} \in \mathbb{R}^d$ ¿Qué podemos decir acerca de su densidad conjunta $P(\mathbf{x})$?

- Dos grandes fuentes de información: $\mu_{\mathbf{x}}$ y $\mathbf{S}$
- Estimación (paramétrica o no paramétrica) de la densidad de $\mathbf{x}$. *solo para d pequeño...*
- Una opción: variables latentes a través de proyecciones en baja dimensión de características "interesantes" de $P(\mathbf{x})$ (PCA).
- Otra opción: encontrar regiones convexas que contienen modas de $P(\mathbf{x})$ (mezclas de densidades o **grupos**).
- ¿Cómo se agrupan los datos? El concepto de **similaridad** o **disimilaridad** es fundamental.

Aprendizaje no supervisado

"Aprender sin maestro..."

Dados n observaciones $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, con $\mathbf{x} \in \mathbb{R}^d$ ¿Qué podemos decir acerca de su densidad conjunta $P(\mathbf{x})$?

- Dos grandes fuentes de información: $\mu_{\mathbf{x}}$ y $\mathbf{S}$
- Estimación (paramétrica o no paramétrica) de la densidad de $\mathbf{x}$. *solo para d pequeño...*
- Una opción: variables latentes a través de proyecciones en baja dimensión de características "interesantes" de $P(\mathbf{x})$ (PCA).
- Otra opción: encontrar regiones convexas que contienen modas de $P(\mathbf{x})$ (mezclas de densidades o **grupos**).
- ¿Cómo se agrupan los datos? El concepto de **similaridad** o **disimilaridad** es fundamental.

Aprendizaje no supervisado

"Aprender sin maestro..."

Dados n observaciones $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, con $\mathbf{x} \in \mathbb{R}^d$ ¿Qué podemos decir acerca de su densidad conjunta $P(\mathbf{x})$?

- Dos grandes fuentes de información: $\mu_{\mathbf{x}}$ y $\mathbf{S}$
- Estimación (paramétrica o no paramétrica) de la densidad de $\mathbf{x}$. *solo para d pequeño...*
- Una opción: variables latentes a través de proyecciones en baja dimensión de características "interesantes" de $P(\mathbf{x})$ (PCA).
- Otra opción: encontrar regiones convexas que contienen modas de $P(\mathbf{x})$ (mezclas de densidades o **grupos**).
- ¿Cómo se agrupan los datos? El concepto de **similaridad** o **disimilaridad** es fundamental.

Aprendizaje no supervisado

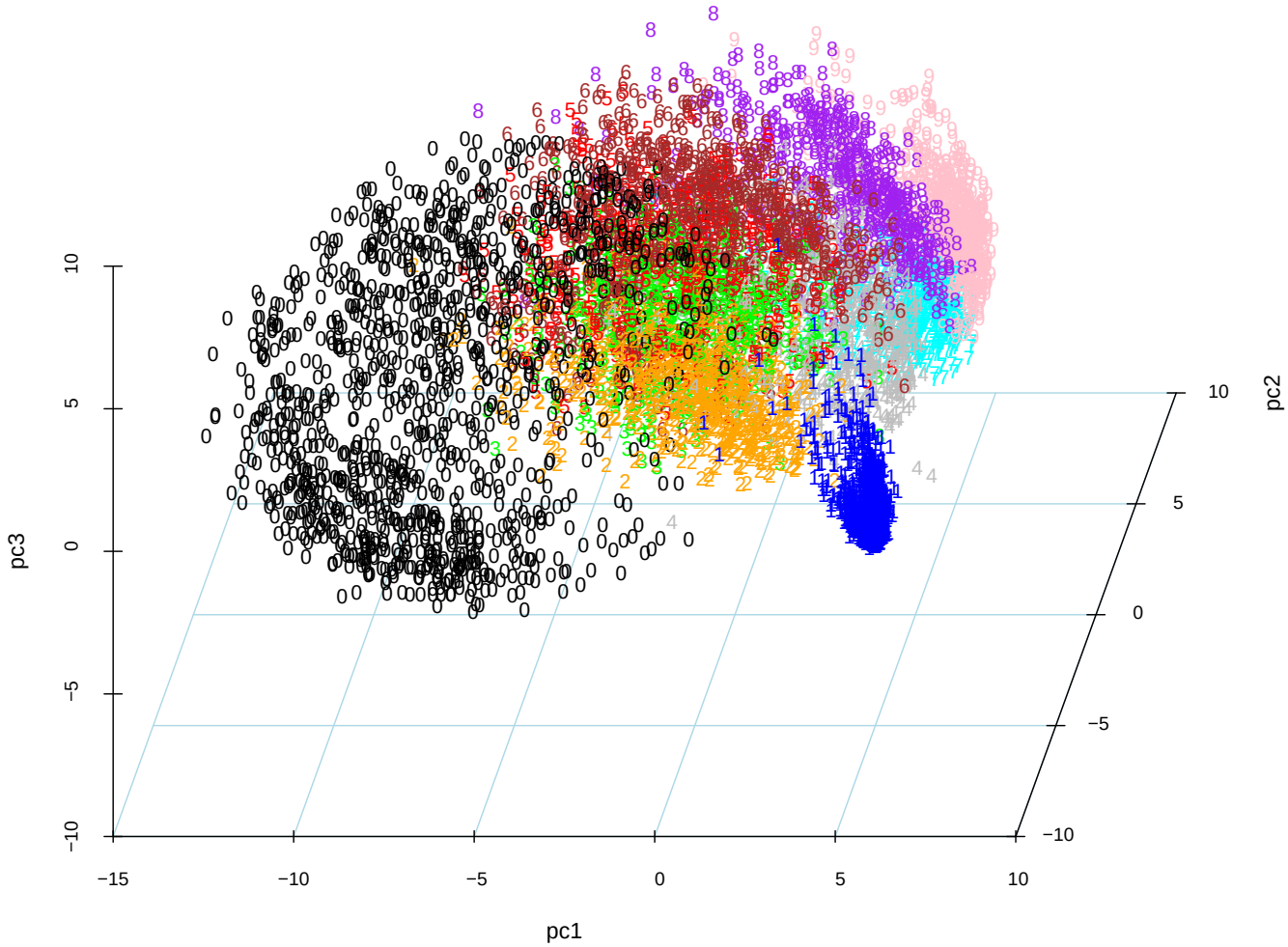
"Aprender sin maestro..."

Dados n observaciones $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, con $\mathbf{x} \in \mathbb{R}^d$ ¿Qué podemos decir acerca de su densidad conjunta $P(\mathbf{x})$?

- Dos grandes fuentes de información: $\mu_{\mathbf{x}}$ y $\mathbf{S}$
- Estimación (paramétrica o no paramétrica) de la densidad de $\mathbf{x}$. *solo para d pequeño...*
- Una opción: variables latentes a través de proyecciones en baja dimensión de características "interesantes" de $P(\mathbf{x})$ (PCA).
- Otra opción: encontrar regiones convexas que contienen modas de $P(\mathbf{x})$ (mezclas de densidades o **grupos**).
- ¿Cómo se agrupan los datos? El concepto de **similaridad** o **disimilaridad** es fundamental.

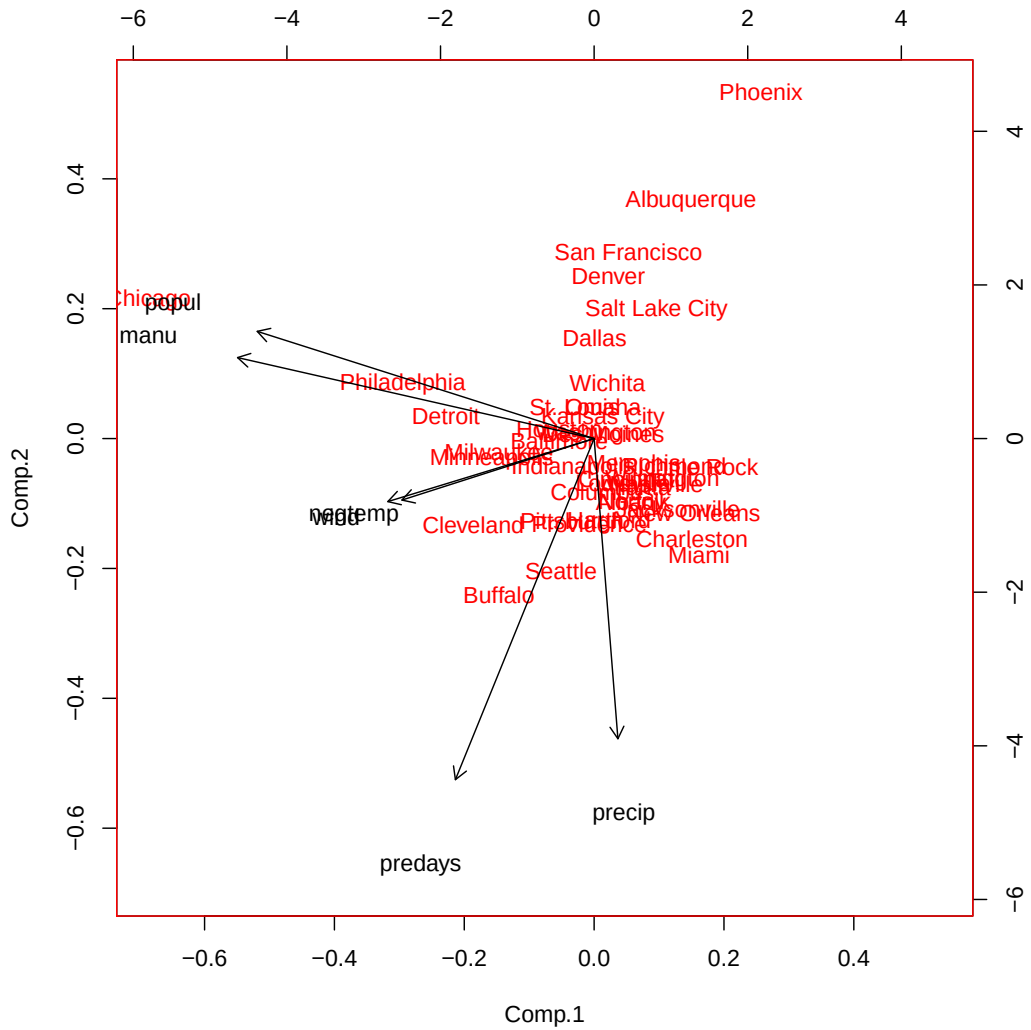
Aprendizaje no supervisado

Ejemplo: Dígitos escaneados. ¿Qué objetos son similares? (ya lo vimos antes...)



Aprendizaje no supervisado

- Podemos agrupar tanto observaciones como atributos. Recuerda el biplot de PCA.



Aprendizaje no supervisado

La disimilaridad o proximidad la representamos como una función de **distancia**

$$d_{ij} = d(\mathbf{x}_i, \mathbf{x}_j),$$

y para n objetos, esta disimilaridad la expresamos mediante

$$\mathbf{D}_{n \times n},$$

con $D_{ij} = d_{ij}$.

La función de distancia debe tener ciertas propiedades...

- $d(x, y) \geq 0$
- $d(x, y) = 0 \leftrightarrow x = y$
- $d(x, y) = d(y, x)$
- $d(x, y) \leq d(x, z) + d(z, y)$

En general, la disimilaridad es una pseudo-distancia.
Ejemplos de distancias...

Aprendizaje no supervisado

La disimilaridad o proximidad la representamos como una función de **distancia**

$$d_{ij} = d(\mathbf{x}_i, \mathbf{x}_j),$$

y para n objetos, esta disimilaridad la expresamos mediante

$$\mathbf{D}_{n \times n},$$

con $D_{ij} = d_{ij}$.

La función de distancia debe tener ciertas propiedades...

- $d(x, y) \geq 0$
- $d(x, y) = 0 \leftrightarrow x = y$
- $d(x, y) = d(y, x)$
- $d(x, y) \leq d(x, z) + d(z, y)$

En general, la disimilaridad es una pseudo-distancia.

Ejemplos de distancias...

Aprendizaje no supervisado

La disimilaridad o proximidad la representamos como una función de **distancia**

$$d_{ij} = d(\mathbf{x}_i, \mathbf{x}_j),$$

y para n objetos, esta disimilaridad la expresamos mediante

$$\mathbf{D}_{n \times n},$$

con $D_{ij} = d_{ij}$.

La función de distancia debe tener ciertas propiedades...

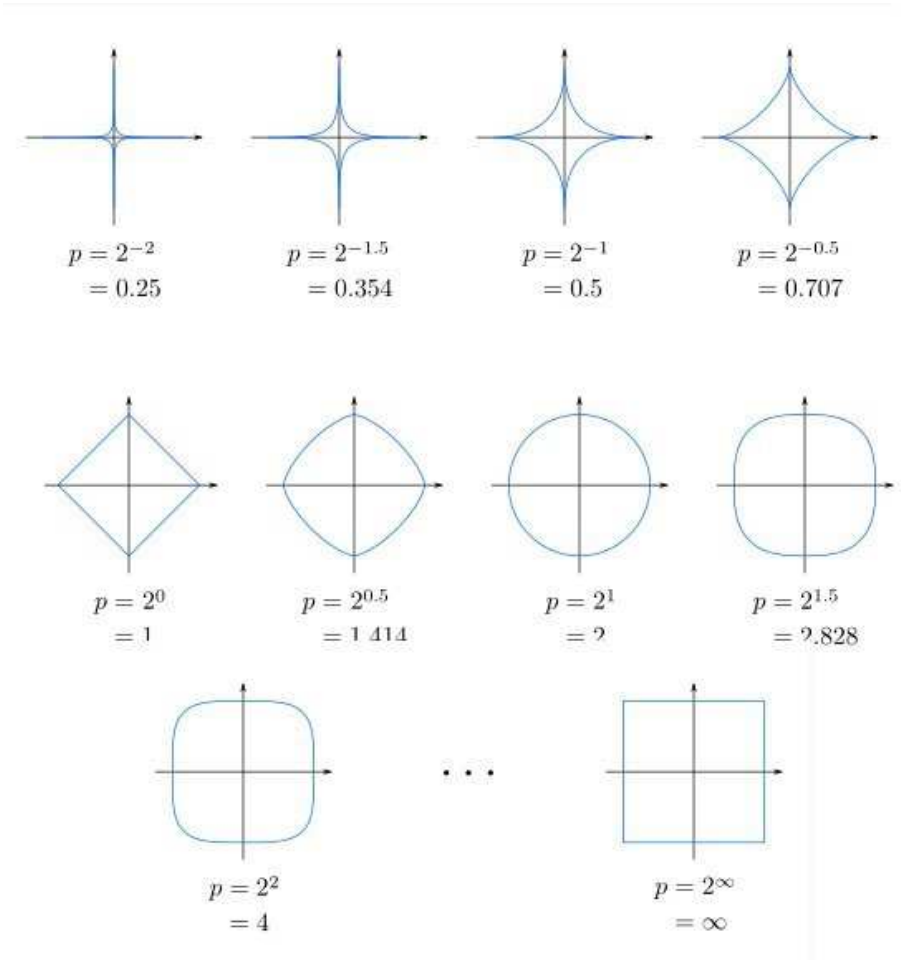
- $d(x, y) \geq 0$
- $d(x, y) = 0 \leftrightarrow x = y$
- $d(x, y) = d(y, x)$
- $d(x, y) \leq d(x, z) + d(z, y)$

En general, la disimilaridad es una pseudo-distancia.
Ejemplos de distancias...

Aprendizaje no supervisado

Distancias: considera el caso general

$$d_{ij}^p = \left(\sum_{k=1}^{d} |x_{ik} - x_{jk}|^p \right)^{1/p} \text{ (Minkowski)}$$



Aprendizaje no supervisado

Distancias y similaridades para diferentes tipos de variables

```
> head(data)
       Pob.1   Pob.2       GAct.1         GAct.2   GAct.3   GAct.4    GAct.5
703        0     114          0.0              0      0.0      0.7       0.0
994     1062     573          0.0              0      0.3      0.0       1.3
730      475     184          0.3              0      0.0      0.7       0.0
1113     441      80          0.0              0      0.3      0.0       0.7
810      325     241          0.9              0      1.0      0.0       0.0
714       86     202          0.0              0      0.0      1.4       0.0
       GAct.6    GAct.7    GAct.8  GAct.9 Comp.XXXX     Comp.2    Comp.3
703       0.0       0.0         0       0      0.00       0.50      0.00
994       0.0       0.0         0       0      0.35       0.00      1.55
730       0.0       0.0         0       0      0.00       0.00      0.35
1113      0.0       0.7         0       0      0.00       0.00      0.00
810       0.7       0.0         0       0      0.00       0.85      0.15
714       0.0       0.0         0       0      0.00       0.00      0.00
      Comp.4   Comp.5     Traf.1    Traf.2       V.1 Pob.NSE     Acc.1
703     0.00      0.0        317         3    116.16       4         0
994     4.70      0.0          0       138     94.00       4         0
730     0.35      0.0        162         5     70.00       4        10
1113    1.15      1.0        310        19    125.00       6         0
810     7.75      1.4        367        22    120.00       3         8
714     1.50      0.0        268         2     70.00       7         0
        Acc.2    Acc.3   Acc.4   Acc.5   Acc.6   Acc.7       G1       G2
703         5        0       7       0       2       8       10       10
994         5        6      10       0      10       0        0        0
730         9       10       3      10      10       8        4       10
1113        8       10      10       0       2       8        0        0
810         7       10      10       8      10       6        4       10
714         3       10      10       0       2       4       10        8
```

Aprendizaje no supervisado

En general, la similaridad s podemos considerarla como:

$$s_{ij} = 1 - d_{ij}$$

Distancias y similaridades para diferentes tipos de variables.

- Numéricas (contínuas o cuantitativas)
- Binarias
- Categóricas (nominales)
- Ordinales

Aprendizaje no supervisado

En general, la similaridad s podemos considerarla como:

$$s_{ij} = 1 - d_{ij}$$

Distancias y similaridades para diferentes tipos de variables.

- Numéricas (contínuas o cuantitativas)
- Binarias
- Categóricas (nominales)
- Ordinales

Aprendizaje no supervisado

En general, considerando diferentes tipos de variables, definimos

$$d_{ij} = \sum_{k=1}^{d} w_k d_k(x_{ik}, x_{jk}),$$

donde

$$\sum_{k=1}^{d} w_k = 1,$$

son una serie de pesos que pueden considerar, entre otras cosas, la naturaleza de cada variable.

Observa que, poniendo el mismo w_k no necesariamente tienen el mismo peso las variables. ¿Porqué?

Dispersión...

Aprendizaje no supervisado

En general, considerando diferentes tipos de variables, definimos

$$d_{ij} = \sum_{k=1}^{d} w_k d_k(x_{ik}, x_{jk}),$$

donde

$$\sum_{k=1}^{d} w_k = 1,$$

son una serie de pesos que pueden considerar, entre otras cosas, la naturaleza de cada variable.
Observa que, poniendo el mismo w_k no necesariamente tienen el mismo peso las variables. ¿Porqué?
Dispersión...

Aprendizaje no supervisado

En general, considerando diferentes tipos de variables, definimos

$$d_{ij} = \sum_{k=1}^{d} w_k d_k(x_{ik}, x_{jk}),$$

donde

$$\sum_{k=1}^{d} w_k = 1,$$

son una serie de pesos que pueden considerar, entre otras cosas, la naturaleza de cada variable.
Observa que, poniendo el mismo w_k no necesariamente tienen el mismo peso las variables. ¿Porqué?
Dispersión...

Aprendizaje no supervisado

Variables binarias

NUMBERS OF CHARACTERS OCCURRING IN, OR ABSENT FROM, TWO INDIVIDUALS: a (+, +) COMMON TO BOTH INDIVIDUALS; b (−, +) AND c (+, −) OCCURRING IN ONLY ONE INDIVIDUAL; AND d (−, −) ABSENT FROM BOTH

	Individual 1 +	Individual 1 −	Totals
Individual 2 +	a	b	$a+b$
Individual 2 −	c	d	$c+d$
Totals	$a+c$	$b+d$	v

$$s_{ij} = \frac{a+d}{a+b+c+d}, \qquad d_{ij} = \frac{b+c}{a+b+c+d}$$

Aprendizaje no supervisado

- Variables categóricas (nominales)

$$s_{ij} = \frac{u}{d}, \qquad d_{ij} = \frac{d-u}{d},$$

u es el número de coincidencias, es decir, el número de variables categóricas en los cuales los objetos i y j coinciden.

- Variables ordinales. Reemplazamos por

$$z_i = \frac{i-1}{M-1},$$

M es el número de categorias ordinales. Este valor lo consideramos como una variable cuantitativa. Generalmente, se estandarizan estas variables (cuantitativas y ordinales):

$$s_{ij}^k = \frac{|x_{ik} - x_{jk}|}{rango(k)}$$

Aprendizaje no supervisado

- Medida general de distancia o disimilaridad

$$d_{ij} = \frac{\sum_{k=1}^{d} \delta_{ij}^k d_{ij}^k}{\sum_{k=1}^{d} \delta_{ij}^k},$$

con δ_{ij}^k una variable que indica si el atributo k se observó en ambos objetos i, j. Esta variable puede tener asociado un peso w_k para cada variable asociada a los objetos i, j, en cuyo caso, $\delta_{ij}^k w_k = w_{i,j,k}$, como ya lo mencionamos en clase. Hay distintos esquemas para "pesar" las variables.
Ver el paper original: Gower (1971) A general coefficient of similarity and some of its properties. Biometrics 27 857-874.

Clustering

Clustering

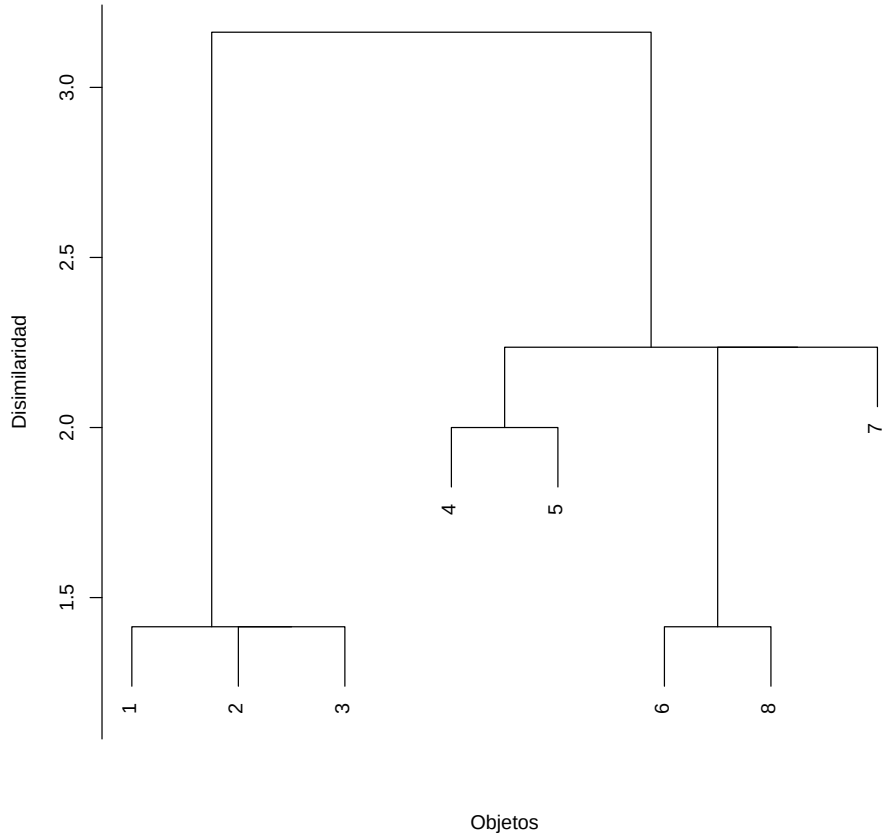
- También llamado segmentación.
- Consiste en agrupar o segmentar una colección de objetos en subconjuntos o clústers.
- Objetos dentro de cada cluster son mas **parecidos** que los objetos de otros clusters.

Clustering Jerárquico

Clustering Jerárquico

Produce representaciones jerárquicas donde los clústers en cada nivel de la jerarquía se crean uniendo clusters en el siguiente nivel.
En el nivel mas bajo, cada objeto representa un clúster.
En el nivel más alto, hay un solo cluster que representa todos los datos.

Clustering Jerárquico



Dos tipos (estrategias) para construirlos:

- Aglomerativo: bottom-up
- Divisivo: top-down

Clustering Jerárquico

Aglomerativo. Los objetos se "aglomeran" para formar clústers de 3 formas principalmente:

- **Single linkage** o vecino más cercano
- **Complete linkage** o vecino más lejano
- **Average linkage**

Clustering Jerárquico

Algoritmo de clustering aglomerativo:

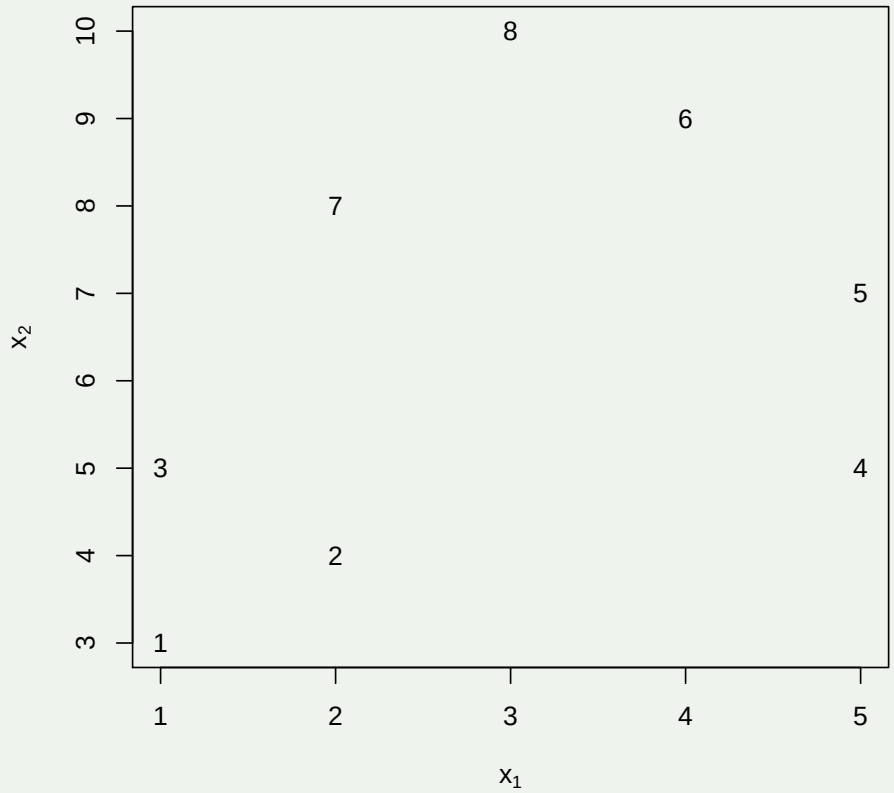
1: Input: Colección de objetos $\{\mathbf{x}_i\}_{i=1}^n$, donde n es el número inicial de clústers
2: Calcular la matriz de disimilaridades $\mathbf{D}_{n\times n}$, con $d_{ij} = d(\mathbf{x}_i, \mathbf{x}_j)$.
3: **for** $t = 1$ to $n - 1$ **do**
4: Calcular $d_{IJ} = \min(d_{ij})$ en $\mathbf{D}$. Unir los clústers I, J para formar un nuevo cluster IJ
5: Calcular disimilaridades $d_{IJ,K}$ entre el nuevo cluster IJ y los demás clusters $K \neq IJ$ según el método de conexión:

- Single: $d_{IJ,K} = \min(d_{I,K}, d_{J,K})$
- Complete: $d_{IJ,K} = \max(d_{I,K}, d_{J,K})$
- Average: $d_{IJ,K} = \sum_{i\in IJ}\sum_{k\in K} d_{ik}/(N_{IJ}N_K)$, N es el número de objetos en los clusters correspondientes

6: Obten $\mathbf{D}^{(t)}_{n-1\times n-1}$, reemplazando renglones y columnas I y J por una nueva columna IJ con las disimilaridades obtenidas en el paso anterior.
7: $\mathbf{D} = \mathbf{D}^{(t)}$
8: **end for**

Clustering

Ejemplo (Libro Izenman...)



Clustering

Disimilaridades.

	1	2	3	4	5	6	7	
1	0.000	1.414	2.000	4.472	5.657	6.708	5.099	7.2
2	1.414	0.000	1.414	3.162	4.243	5.385	4.000	6.0
3	2.000	1.414	0.000	4.000	4.472	5.000	3.162	5.3
4	4.472	3.162	4.000	0.000	2.000	4.123	4.243	5.3
5	5.657	4.243	4.472	2.000	0.000	2.236	3.162	3.6
6	6.708	5.385	5.000	4.123	2.236	0.000	2.236	1.4
7	5.099	4.000	3.162	4.243	3.162	2.236	0.000	2.2
8	7.280	6.083	5.385	5.385	3.606	1.414	2.236	0.0

Clustering

$k = 1$: Disimilaridades. Disimilaridad mínima.

	1	2	3	4	5	6	7	
1	0.000	1.414	2.000	4.472	5.657	6.708	5.099	7.2
2	1.414	0.000	1.414	3.162	4.243	5.385	4.000	6.0
3	2.000	1.414	0.000	4.000	4.472	5.000	3.162	5.3
4	4.472	3.162	4.000	0.000	2.000	4.123	4.243	5.3
5	5.657	4.243	4.472	2.000	0.000	2.236	3.162	3.6
6	6.708	5.385	5.000	4.123	2.236	0.000	2.236	1.4
7	5.099	4.000	3.162	4.243	3.162	2.236	0.000	2.2
8	7.280	6.083	5.385	5.385	3.606	1.414	2.236	0.0

Clustering

Disimilaridades. Nuevo cluster.

	1	2	3	4	5	6	7	
1	0.000	1.414	2.000	4.472	5.657	6.708	5.099	7.2
2	1.414	0.000	1.414	3.162	4.243	5.385	4.000	6.0
3	2.000	1.414	0.000	4.000	4.472	5.000	3.162	5.3
4	4.472	3.162	4.000	0.000	2.000	4.123	4.243	5.3
5	5.657	4.243	4.472	2.000	0.000	2.236	3.162	3.6
6	6.708	5.385	5.000	4.123	2.236	0.000	2.236	1.4
7	5.099	4.000	3.162	4.243	3.162	2.236	0.000	2.2
8	7.280	6.083	5.385	5.385	3.606	1.414	2.236	0.0

Clustering

$k = 2$: Disimilaridades (minimas).

	1,2	3	4	5	6	7	8
1,2	0.000	1.414	3.162	4.243	5.385	4.000	6.083
3	1.414	0.000	4.000	4.472	5.000	3.162	5.385
4	3.162	4.000	0.000	2.000	4.123	4.243	5.385
5	4.243	4.472	2.000	0.000	2.236	3.162	3.606
6	5.385	5.000	4.123	2.236	0.000	2.236	1.414
7	4.000	3.162	4.243	3.162	2.236	0.000	2.236
8	6.083	5.385	5.385	3.606	1.414	2.236	0.000

Clustering

Disimilaridades (nuevo cluster).

	1,2	3	4	5	6	7	8
1,2	0.000	1.414	3.162	4.243	5.385	4.000	6.083
3	1.414	0.000	4.000	4.472	5.000	3.162	5.385
4	3.162	4.000	0.000	2.000	4.123	4.243	5.385
5	4.243	4.472	2.000	0.000	2.236	3.162	3.606
6	5.385	5.000	4.123	2.236	0.000	2.236	1.414
7	4.000	3.162	4.243	3.162	2.236	0.000	2.236
8	6.083	5.385	5.385	3.606	1.414	2.236	0.000

Clustering

$k = 3$: Disimilaridades (minimas).

	1,2,3	4	5	6	7	8
1,2,3	0.000	3.162	4.243	5.000	3.162	5.385
4	3.162	0.000	2.000	4.123	4.243	5.385
5	4.243	2.000	0.000	2.236	3.162	3.606
6	5.000	4.123	2.236	0.000	2.236	1.414
7	3.162	4.243	3.162	2.236	0.000	2.236
8	5.385	5.385	3.606	1.414	2.236	0.000

Clustering

Disimilaridades (nuevo cluster).

	1,2,3	4	5	6	7	8
1,2,3	0.000	3.162	4.243	5.000	3.162	5.385
4	3.162	0.000	2.000	4.123	4.243	5.385
5	4.243	2.000	0.000	2.236	3.162	3.606
6	5.000	4.123	2.236	0.000	2.236	1.414
7	3.162	4.243	3.162	2.236	0.000	2.236
8	5.385	5.385	3.606	1.414	2.236	0.000

Clustering

$k = 4$: Disimilaridades (minimas).

	1,2,3	4	5	6,8	7
1,2,3	0.000	3.162	4.243	5.000	3.162
4	3.162	0.000	2.000	4.123	4.243
5	4.243	2.000	0.000	2.236	3.162
6,8	5.000	4.123	2.236	0.000	2.236
7	3.162	4.243	3.162	2.236	0.000

Clustering

Disimilaridades (nuevo cluster).

	1,2,3	4	5	6,8	7
1,2,3	0.000	3.162	4.243	5.000	3.162
4	3.162	0.000	2.000	4.123	4.243
5	4.243	2.000	0.000	2.236	3.162
6,8	5.000	4.123	2.236	0.000	2.236
7	3.162	4.243	3.162	2.236	0.000

Clustering

$k = 5$: Disimilaridades (minimas).

	1,2,3	4,5	6,8	7
1,2,3	0.000	3.162	5.000	3.162
4,5	3.162	0.000	2.236	3.162
6,8	5.000	2.236	0.000	2.236
7	3.162	3.162	2.236	0.000

Clustering

Disimilaridades (nuevo cluster).

	1,2,3	4,5	6,8	7
1,2,3	0.000	3.162	5.000	3.162
4,5	3.162	0.000	2.236	3.162
6,8	5.000	2.236	0.000	2.236
7	3.162	3.162	2.236	0.000

Clustering

$k = 6$: Disimilaridades (minimas).

	1,2,3	4,5,6,8	7
1,2,3	0.000	3.162	3.162
4,5,6,8	3.162	0.000	2.236
7	3.162	2.236	0.000

Clustering

Disimilaridades (nuevo cluster).

	1,2,3	4,5,6,8	7
1,2,3	0.000	3.162	3.162
4,5,6,8	3.162	0.000	2.236
7	3.162	2.236	0.000

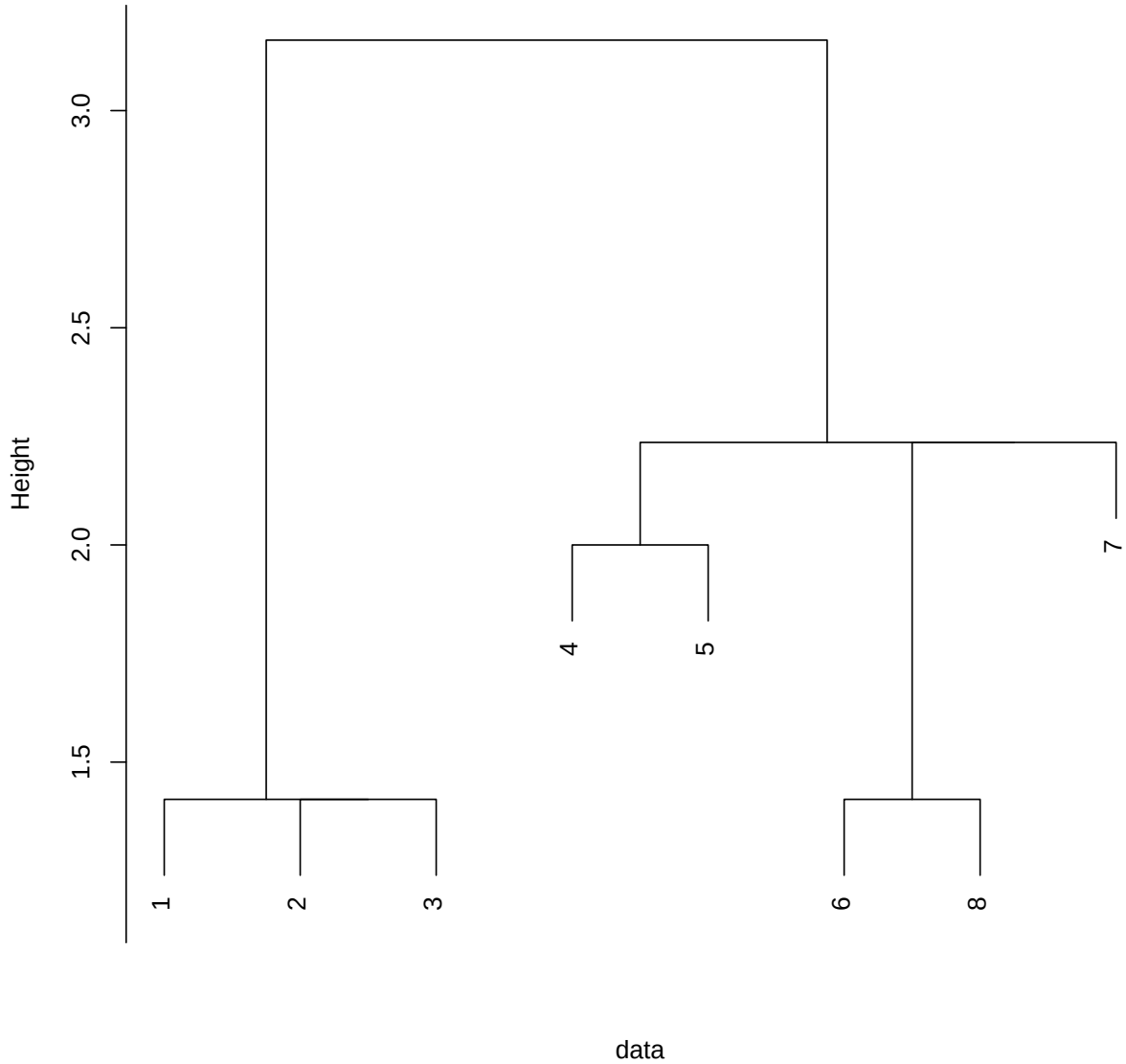
Clustering

$k = 7$: Disimilaridades (final).

	1,2,3	4,5,6,8,7
1,2,3	0.000	3.162
4,5,6,8,7	3.162	0.000

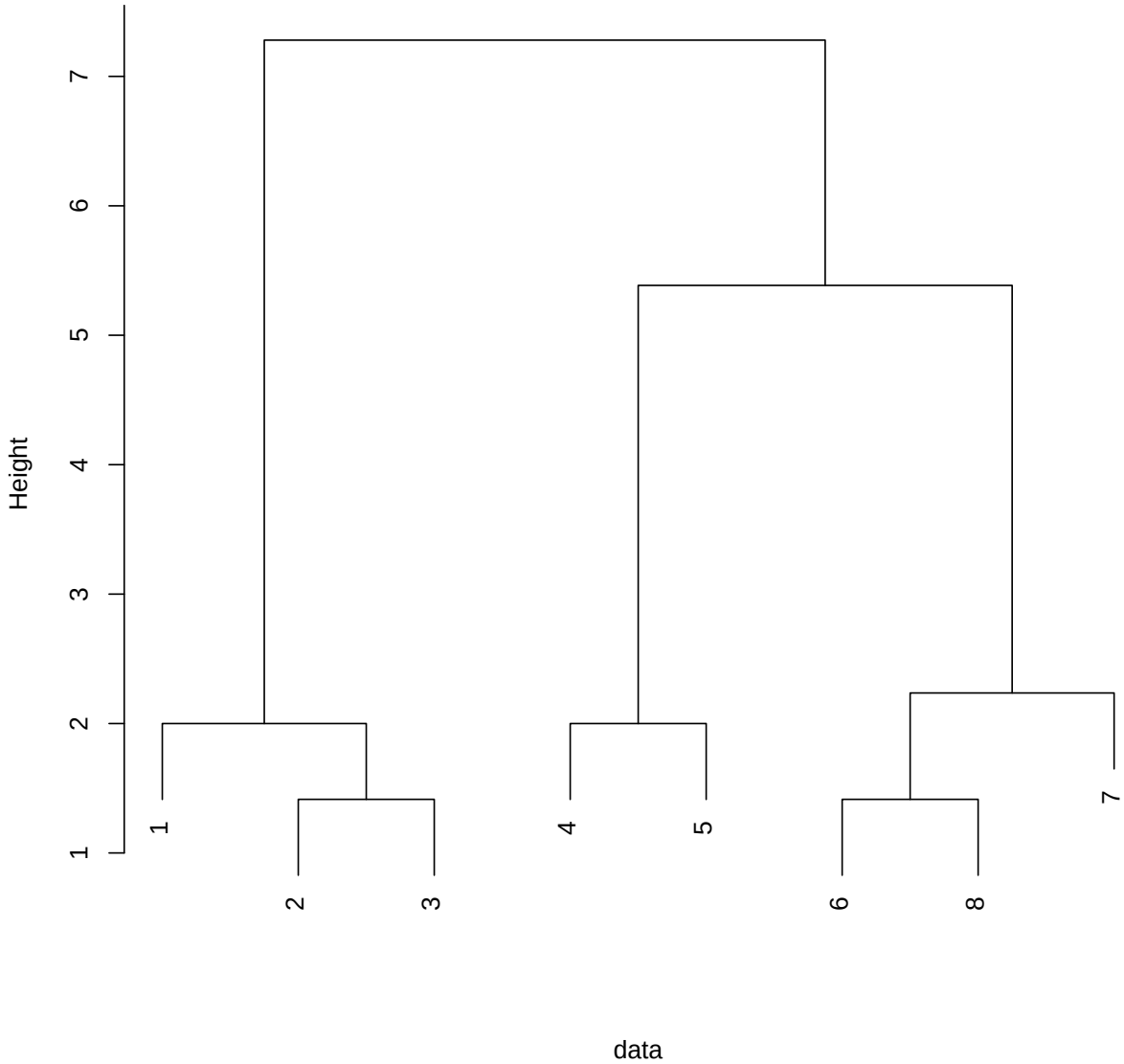
Clustering

Single linkage



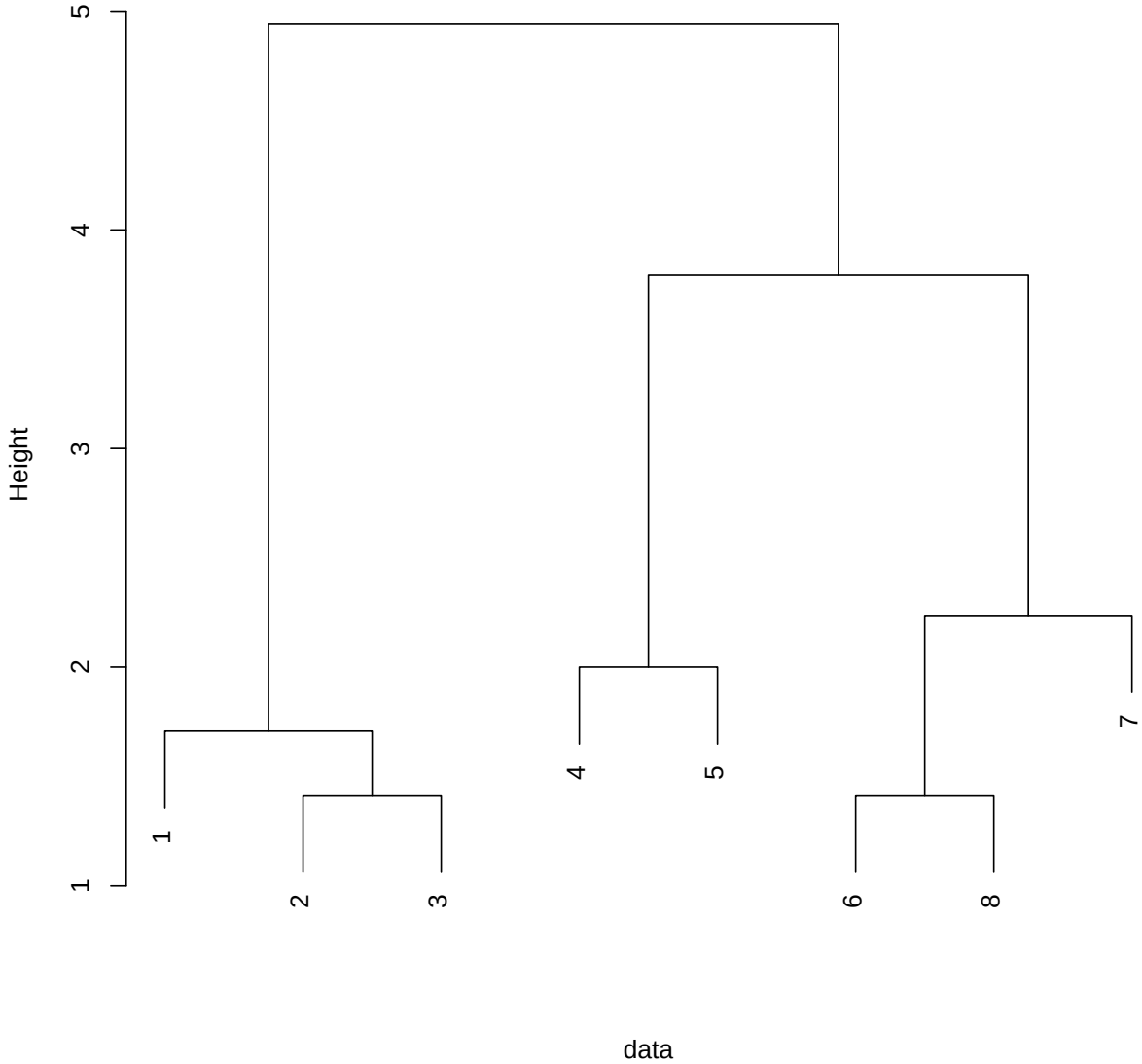
Clustering

Complete linkage



Clustering

Average linkage



Clustering

Clustering divisivo (top-down)

Partiendo de un clúster general (que contiene a todos los objetos), iterativamente se divide en un grupo "splinter" (como "sacar una astilla"), digamos, clúster A y el resto, digamos, clúster B

Clustering

Clustering divisivo (top-down)

1: Input: Colección de objetos $\mathbf{X}$
2: **for** $t = 1, 2, \ldots$ **do**
3: Calcula disimilaridades promedio d_{ij} de $\mathbf{X}$.
4: Construye

- A: objeto con máxima d_{ij}
- B: resto de los objetos

5: Para cada objeto en B, calcula

- $d^{(A)} = \bar{d}_{ij}$, las disimilaridades promedio entre objeto i y los restantes $j \neq i$ en B
- $d^{(B)} = \bar{d}_{ij}$, las disimilaridades promedio entre objeto i y los restantes $j \neq i$ en A

6: **if** $d^{(A)} - d^{(B)} < 0$ **then**
7: Stop
8: **end if**
9: Toma el objeto en B con máximo $(d^{(A)} - d^{(B)})$ y muévelo a A
10: Repite con $X = A$ y $X = B$
11: **end for**

Clustering

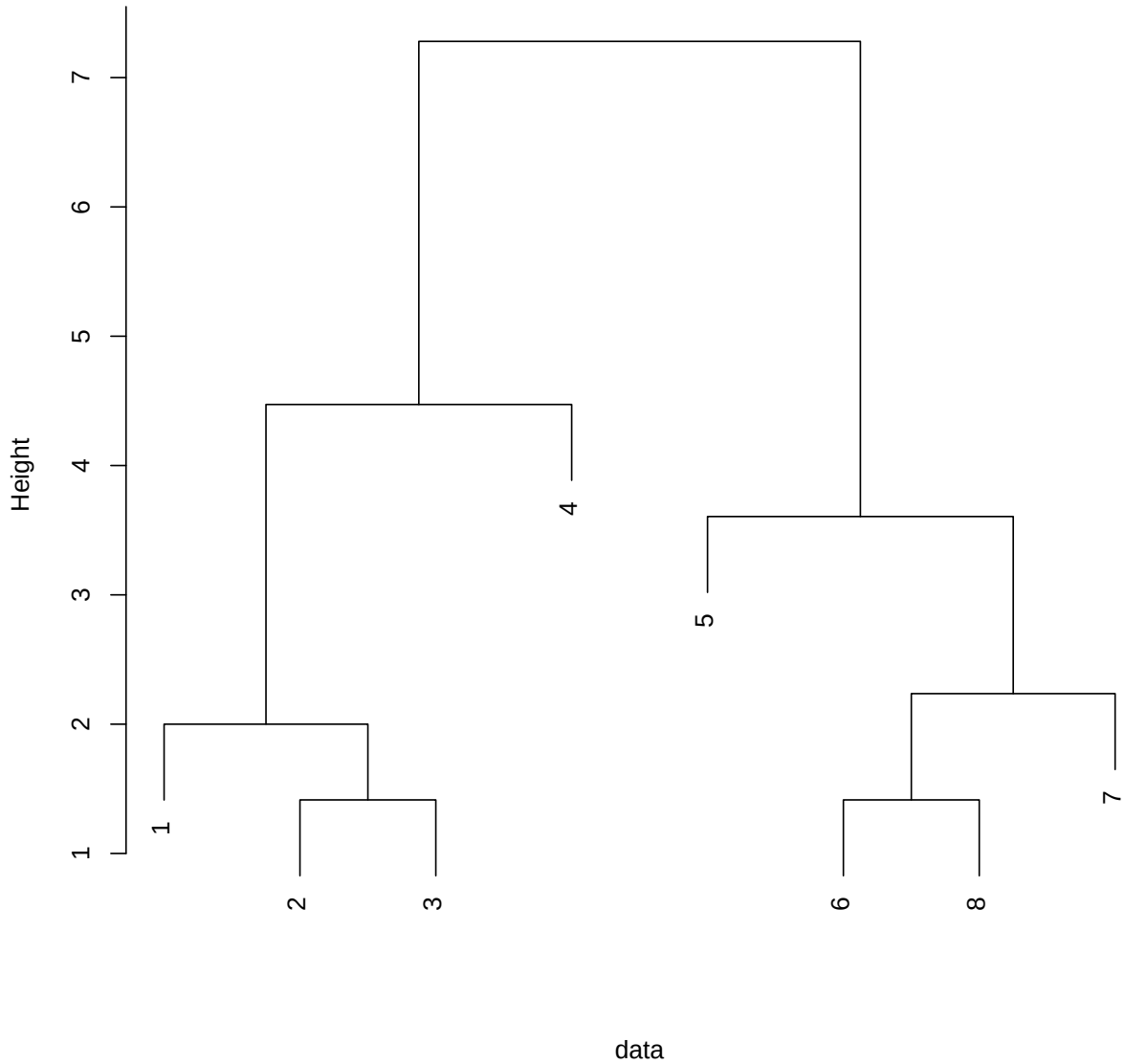
Clustering divisivo (top-down)

Observaciones:

- Generalmente empezamos con un clustering aglomerativo para llegar al clúster general, entonces son $n-1$ operaciones.
- Puede ser más costoso computacionalmente que el método aglomerativo, ya que, en el primer paso del clustering aglomerativo, consideramos **todas** las posibles fusiones de dos objetos, por lo que tenemos $n(n-1)/2$ combinaciones.
- En el algoritmo divisivo, tenemos $2^{n-1}-1$ posibilidades de dividir un objeto en dos clústers, que es mucho mayor que el algoritmo aglomerativo
- Hay implementaciones eficientes (`diana` en `R`).

Clustering

Clustering divisivo (top-down)



Clustering

Ejemplo (Primate Scapular (Izenman))

Mediciones de huesos de los hombros de primates.

```
A data frame with 105 observations on the following 11 va
‘genus’ a numeric vector
‘AD.BD’ a numeric vector
‘AD.CD’ a numeric vector
‘EA.CD’ a numeric vector
‘Dx.CD’ a numeric vector
‘SH.ACR’ a numeric vector
‘EAD’ a numeric vector
‘beta’ a numeric vector
‘gamma’ a numeric vector
‘class’ a factor with levels ‘Gorilla’ ‘Homo’ ‘Hylobates’
    ‘Pongo’
‘classdigit’ a factor with levels ‘1’ ‘2’ ‘3’ ‘4’ ‘5’
```

notebooks/5-clustering.ipynb

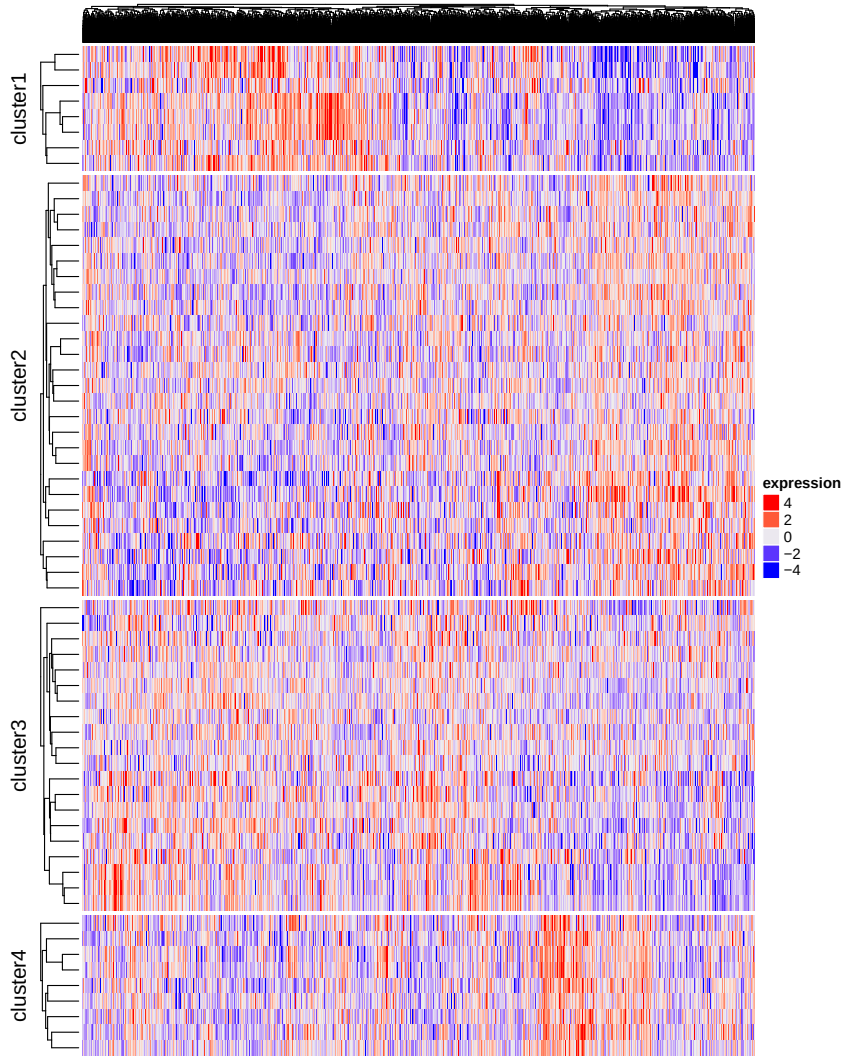
Clustering

Ejemplo (NCI60 Microarray data (Hastie et al.))

```
Description:

     NCI microarray data. The data contains expression levels
     genes from 64 cancer cell lines. Cancer type is also reco

Format:
     The format is a list containing two elements: 'data' and

     'data' is a 64 by 6830 matrix of the expression values wh
     'labs' is a vector listing the cancer types for the 64 ce

Source:
     The data come from Ross et al. (Nat Genet., 2000). More
     information can be obtained at
     http://genome-www.stanford.edu/nci60/
```

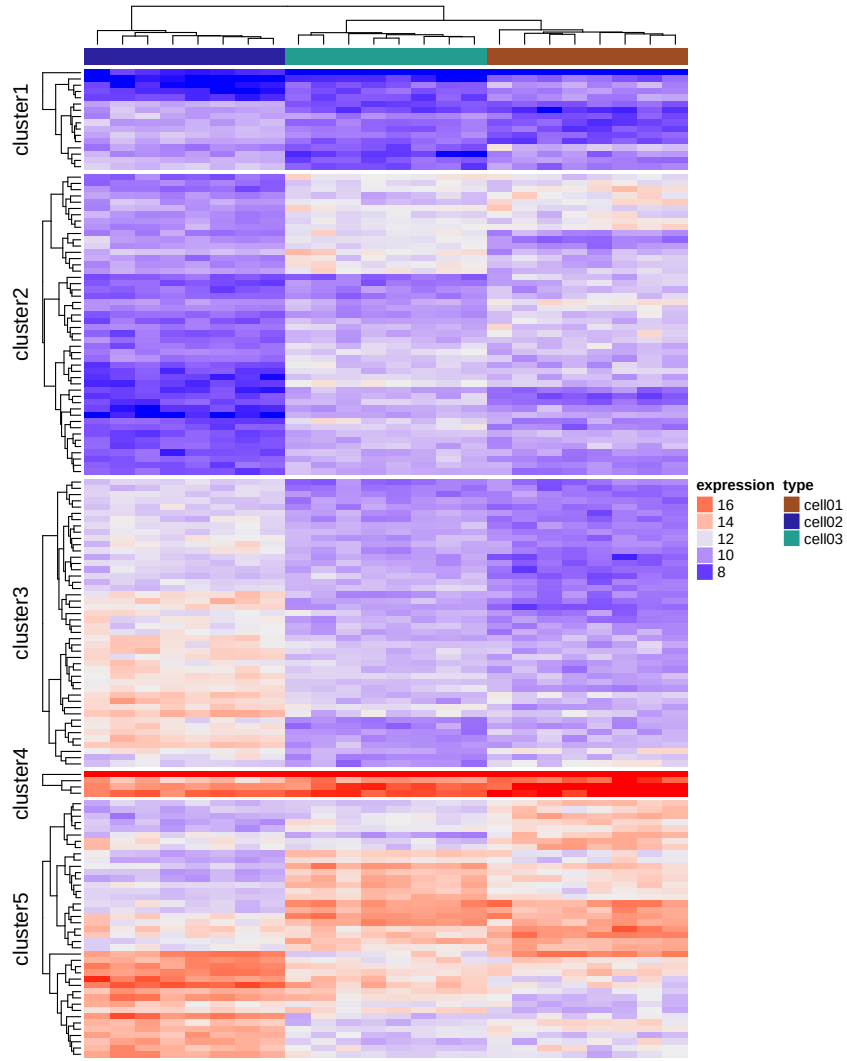
Clustering

Ejemplo: NCI60 Microarray data (Hastie et al.) $k = 4$ clusters.



Clustering

Otro ejemplo de microarreglos. Encontrando estructura.



Clustering jerárquico

Sobre la implementación.

- `Python`: Están los módulos `scipy.cluster.hierarchy` y `sklearn.cluster.AgglomerativeClustering`. Un poco limitados en cuanto a la visualización
- `R`. La función estándar es `hclust` (librería `stats`), pero no tiene implementado el clustering divisivo.
- Otra opción son las funciones `agnes` (aglomerative nesting) y `diana` (divisive analysis clustering), de la librería `cluster`.
- Para el heatmap, la función estándar es `heatmap`, en la librería `stats`.
- Hay varias librerías para resaltar y graficar los resultados. Como referencia, puedes recurrir a
  - `factoextra`
  - `dendextend`
  - `ComplexHeatmap`, del proyecto BioConductor.