Clustering: Distancias

Objetivo: Realizar *clusters* de objetos

Punto de partida: Un conjunto de objetos sobre los que se han observado diferentes variables.

Ejemplo:

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
1	5.1	3.5	1.4	0.2
2	4.9	3.0	1.4	0.2
3	4.7	3.2	1.3	0.2
4	4.6	3.1	1.5	0.2
5	5.0	3.6	1.4	0.2
6	5.4	3.9	1.7	0.4
7	4.6	3.4	1.4	0.3
8	5.0	3.4	1.5	0.2
9	4.4	2.9	1.4	0.2
...				

Clustering: Distancias

Tabla de datos

- $\Omega=\{\omega_1,\omega_2,\ldots,\omega_n\}$ es el conjunto de **objetos**
- $X_1,X_2,\ldots,X_p$ son las **variables**
- $X_j(\omega_i)=x_{ij}$ **valor** de la variable X_j sobre el objeto ω_i

Ω	X_1		X_j		X_p
ω_1	x_{11}		x_{1j}		x_{1p}
ω_i	x_{i1}		x_{ij}		x_{ip}
ω_n	x_{n1}		x_{nj}		x_{np}

Clustering: Distancias

- Las variables $X_1, X_2, \ldots, X_p$ pueden ser:
- **Cuantitativas**: $X_j(\omega_i) = x_{ij}$ es un número real
- **Cualitativas**: $X_j(\omega_i) = x_{ij}$ es un código

Dependiendo del contexto del problema:

- Las variables <u>binarias</u> ($X_j(\omega_i) = x_{ij}$ está en $\{0,1\}$) pueden ser tratadas como cuantitativas o como cualitativas.
- Las variables <u>ordinales</u> pueden ser tratadas como cuantitativas ($X_j(\omega_i) = x_{ij}$ está en $\{1,2,\ldots,r\}$) o como cualitativas.

Clustering: Distancias

» Una variable **cualitativa** X puede ser descompuesta en variables binarias; tantas como categorías o modalidades tenga:

- $X(\omega_i) = x_i$ está en $\{c_1, c_2, \ldots, c_r\}$.
- $X_k(\omega_i) = x_{ik}$ está en $\{0,1\}$:

$$\text{si } X(\omega_i) = c_k \text{ entonces } X_k(\omega_i) = 1, \text{ si no } X_k(\omega_i) = 0$$

» A esta forma de codificar una variable cualitativa se le denomima *disyuntiva completa*, por su correspondencia con la forma normal disyuntiva completa de las fórmulas de la lógica booleana.

Clustering: Distancias

» Variables cualitativas: *código disyuntivo completo*

Ω	X_1	X_2	X_3
ω_1	1	2	2
ω_2	1	4	3
ω_3	2	2	3
ω_4	1	3	1
ω_5	2	3	2
ω_6	2	4	3
ω_7	1	1	2
ω_8	1	1	1

Ω	X_{11}	X_{12}	X_{21}	X_{22}	X_{23}	X_{24}	X_{31}	X_{32}	X_{33}
ω_1	1	0	0	1	0	0	0	1	0
ω_2	1	0	0	0	0	1	0	0	1
ω_3	0	1	0	1	0	0	0	0	1
ω_4	1	0	0	0	1	0	1	0	0
ω_5	0	1	0	0	1	0	0	1	0
ω_6	0	1	0	0	0	1	0	0	1
ω_7	1	0	1	0	0	0	0	1	0
ω_8	1	0	1	0	0	0	1	0	0

Clustering: Distancias

›› Representación geométrica de los objetos:

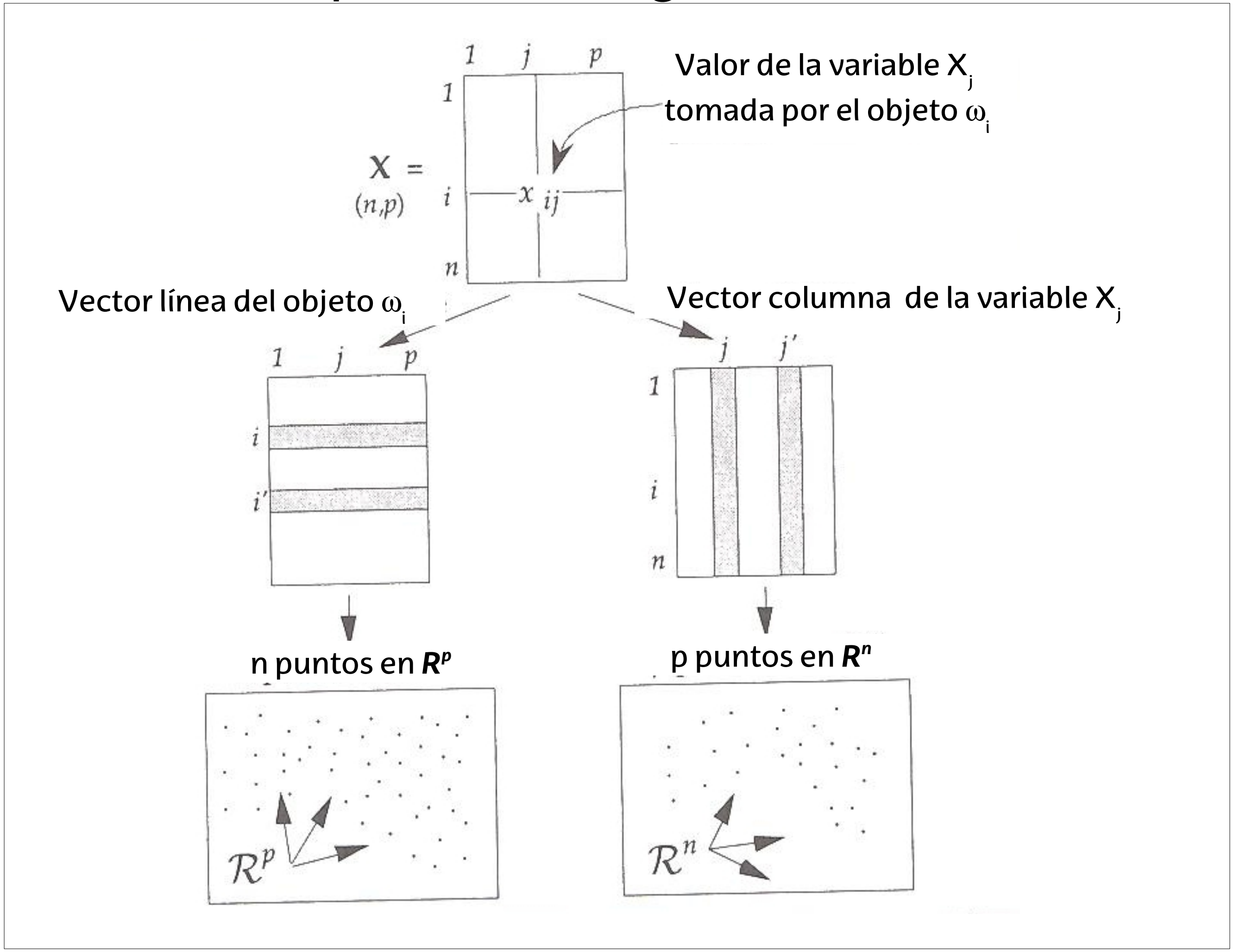
- Cada objeto ω_i de Ω es un punto del espacio $\boldsymbol{R}^p$.
- La desemejanza entre objetos se traduce en una distancia entre los puntos correspondientes.

›› Representación geométrica de las variables:

- Cada variable X_j es un punto del espacio $\boldsymbol{R}^n$.
- La semejanza entre variables se traduce en una *correlación* entre ellas. A partir la correlación se define una distancia entre variables.

Clustering: Distancias

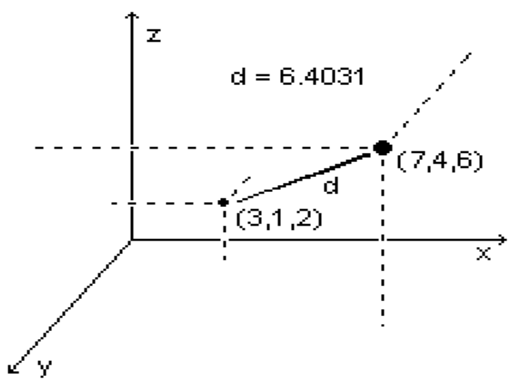
Representación geométrica



Clustering: Distancias

›› La *representación geométrica* de la tabla de datos induce a pensar las medidas de desemejanza como índices de *lejanía* o distancia, y las de semejanza como de *proximidad*.

- Distancia entre puntos
  - Distancia euclidiana



$$P = (p_1, p_2, \ldots, p_n) \qquad Q = (q_1, q_2, \ldots, q_n).$$

$$\sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

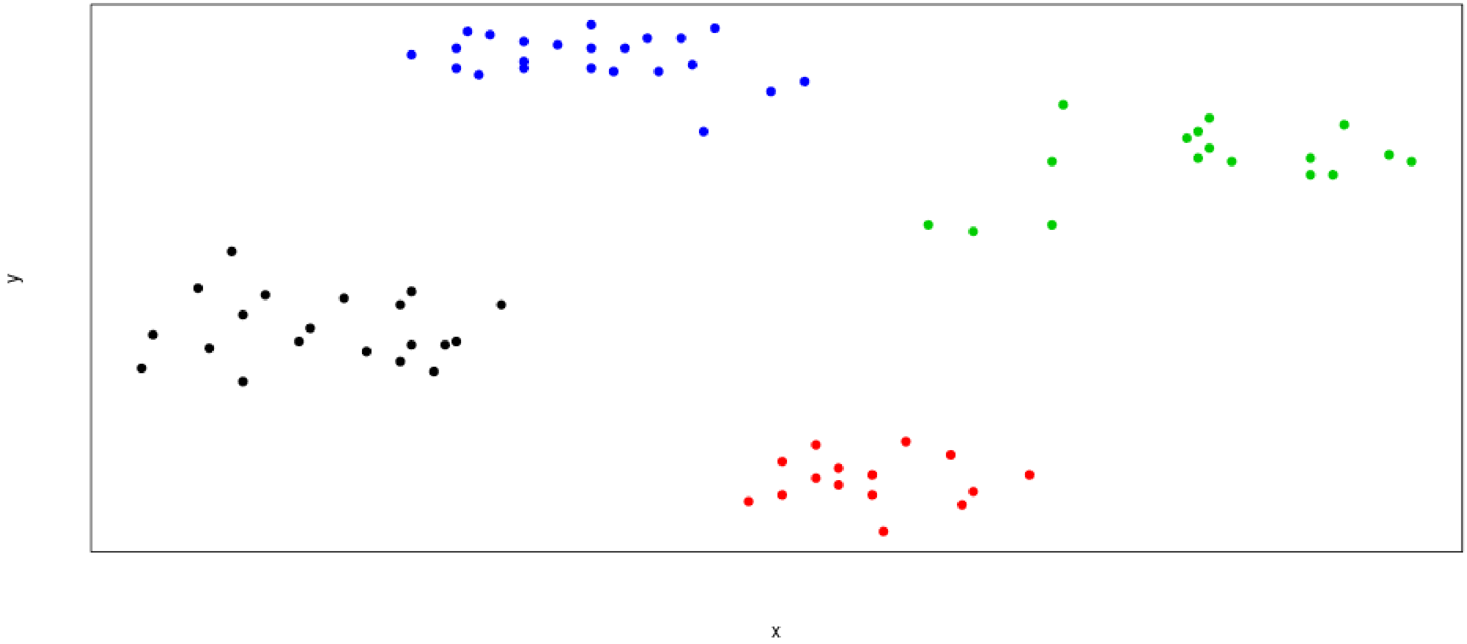
$$d = \sqrt{(x_2-x_1)^2+(y_2-y_1)^2+(z_2-z_1)^2} \quad = ((7-3)^2+(4-1)^2+(6-2)^2)^{1/2} = 41^{1/2} = 6.40$$

Clustering: Distancias

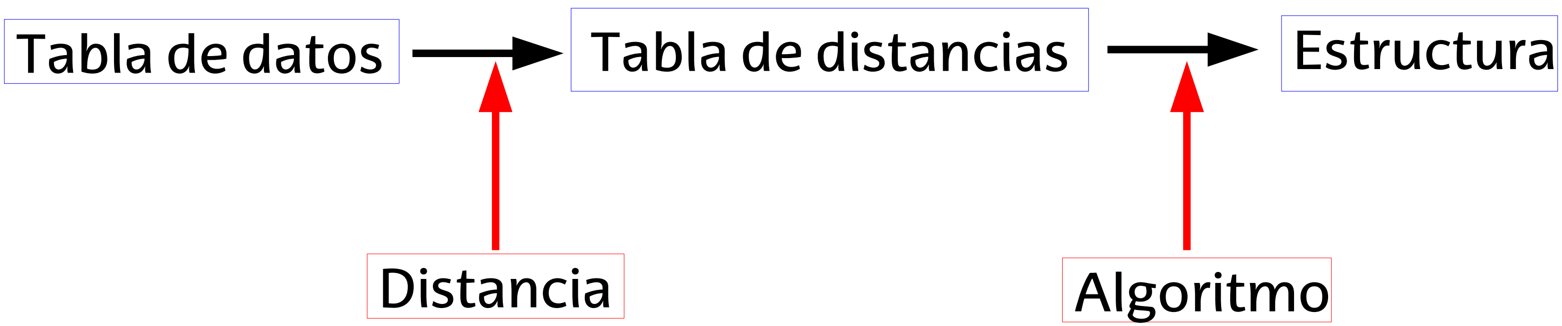
- El concepto de distancia tiene un papel fundamental en muchos métodos de análisis de datos, en particular en el *clustering*, (también en el análisis discriminante, el análisis de regresión, el análisis factorial, etc).
- El concepto de distancia euclidiana está muy relacionado con los conceptos básicos estadísticos de *varianza* (variación) y *correlación* (asociación).

Clustering: Distancias

- Formación de *clusters* según distancias

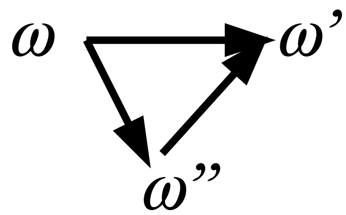


Clustering: Distancias



Clustering: Distancias

- Una **distancia**, por definición, cumple estas propiedades para todos los objetos *ω,ω',ω''* *de* Ω:
  - $d(\omega,\omega') \geq 0$ (no negatividad)
  - $d(\omega,\omega') = 0$ si y sólo si $\omega=\omega'$ (discernibilidad)
  - $d(\omega,\omega') = d(\omega',\omega)$ (simetría)
  - $d(\omega,\omega') \leq d(\omega,\omega'') + d(\omega'',\omega')$ (desigualdad triangular)



- No hay una guía para decidir cuál es la distancia conveniente para cada problema.

Clustering: Distancias

- Variables **cuantitativas.**

  Distancia de *Minkowsky* (norma $\mathrm{L_r}$)

  $d_r(\omega_i, \omega_{i'}) = [\ \Sigma_{j=1}{}^{p} |x_{ij} - x_{i'j}|^r\ ]^{1/r}$ **r**, factor de *Minkowsky*

- r=1 (rectangular) $d(\omega_i, \omega_{i'}) = \Sigma_{j=1}{}^{p} |x_{ij} - x_{i'j}|$

  (bloque de ciudad, *Manhattan*)

- r=2 (euclidiana) $d(\omega_i, \omega_{i'}) = [\ \Sigma_{j=1}{}^{p} |x_{ij} - x_{i'j}|^2\ ]^{1/2}$
- r=∞ (max, *Chebyshev*) $d(\omega_i, \omega_{i'}) = \max_{j=1}{}^{p} |x_{ij} - x_{i'j}|$

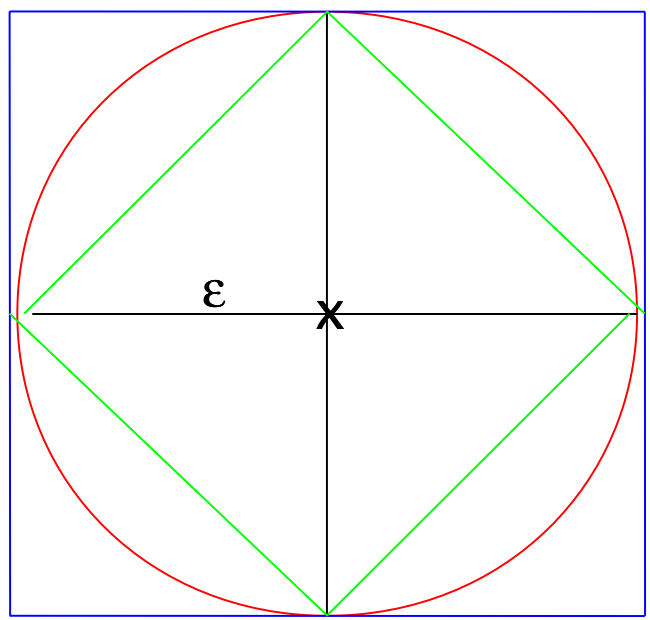
Clustering: Distancias

- Variables **cuantitativas.**

  Distancia de *Minkowsky* (norma L_r).

  Puntos y que están a la misma distancia de x:

$$B(x,\varepsilon) = \{y \mid d_r(x,y) = \varepsilon\}$$



r=1

r=2

r=∞

Clustering: Distancias

- Variables **cuantitativas.**

  Distancia euclidiana <u>cuadrática</u>

$$d(\omega_{i},\omega_{i'}) = \Sigma_{j=1}^{p}(x_{ij}-x_{i'j})^2$$

- La *distancia euclidiana cuadrática* está muy relacionada con el concepto estadístico básico de *varianza*.

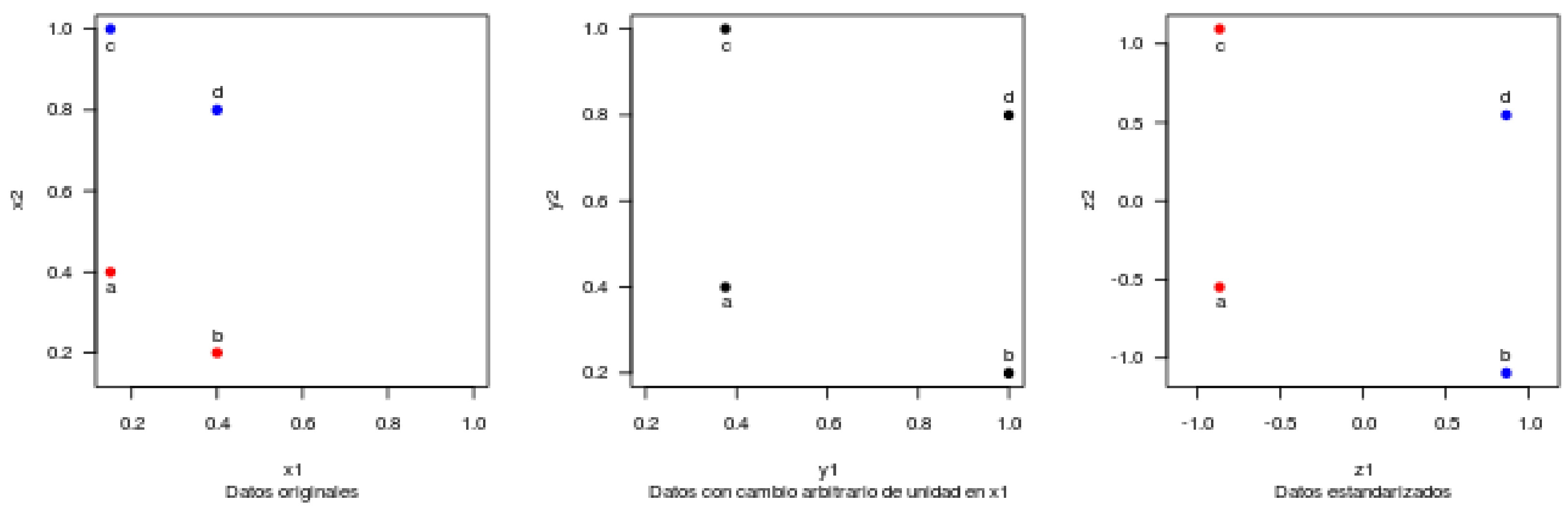
$$\mathbf{\Sigma_i\Sigma_{i'}d(\omega_{i},\omega_{i'})} = \Sigma_i\Sigma_{i'}\Sigma_{j=1}^{p}(x_{ij}-x_{i'j})^2 = 2n\cdot\Sigma_i\Sigma_{j=1}^{p}(x_{ij}-x_{0j})^2 =$$

$$= 2n\cdot\Sigma_i d(\omega_{i},\omega_{0}) = \mathbf{2n^2\cdot\Sigma_{j=1}^{p}Var(X_j)}$$

siendo ω_0 el objeto cuyas coordenadas son las medias de las variables: $x_{0j}= \Sigma_i x_{ij}/n$

Clustering: Distancias

›› Variables **cuantitativas**. En el caso de variables heterogéneas las <u>unidades de medición</u> pueden tener gran influencia en la realización de *clusters*.



›› Solución: <u>Estandarización</u> de las variables

Clustering: Distancias

- Variables **cuantitativas**.

  Distancia de Mahalanobis. Generalización de la *euclidiana cuadrática*.

  Toma en cuenta las relaciones entre las variables.

$$d(\omega_i, \omega_{i'}) = \sum_{j=1}^{p} \sum_{j'=1}^{p} w_{jj'} \cdot (x_{ij} - x_{i'j}) \cdot (x_{ij'} - x_{i'j'})$$

matricialmente

$$d_{\Sigma}(\omega_i, \omega_{i'}) = (x_i - x_{i'}) \cdot \Sigma^{-1} \cdot (x_i - x_{i'})^t$$

Σ^{-1} es la inversa de la matriz de *covarianzas*

- Si $w_{jj'}=0$ cuando $j \neq j'$, la distancia es la *euclidiana cuadrática* con las variables *estandarizadas*.

Clustering: Distancias

» Variables **cuantitativas**. *Distancia de Mahalanobis*.

- Toma en cuenta las relaciones entre las variables. En las demás distancias anteriores se considera que las variables son mutuamente *independientes*. En la práctica, esto NO suele ser lo habitual.
- Caso extremo: *redundancia* de dos variables, relación lineal. Cada una de las dos aporta un sumando a la distancia: hay una variable que se cuenta dos veces (!).
- *Mahalanobis*: transforma las variables de forma que la matriz de correlaciones sea la identidad. La distancia es *invariante* respecto a cualquier combinación lineal de variables.

Clustering: Distancias

- Variables **binarias**.

Considerando todas las variables binarias para dos objetos ω_i y $\omega_{i'}$ se cuentan las veces en que ambos tienen *0-0*s, *1-1*s, *0-1*s y *1-0*s. *Tabla de coocurrencias*:

$$
\begin{array}{cc|cc|c}
 & \omega_{i'} & 1 & 0 & \\
\hline
\omega_i & 1 & a & b & a+b \\
 & 0 & c & d & c+d \\
\hline
 & & a+c & b+d & a+b+c+d
\end{array}
$$

$$d(\omega_i, \omega_{i'}) = (b+c)/(a+b+c+d)$$

$$d(\omega_i, \omega_{i'}) = (b+c)/(a+b+c) \quad \text{(coeficiente de } \textit{Jaccard}\text{)}$$

Clustering: Distancias

- Variables mixtas: **cuantitativas** y **binarias**.
  Situación más habitual.

$$d(\omega_i, \omega_{i'}) = \sum_{j=1}^{p} w_j \cdot d_j(\omega_i, \omega_{i'})$$

(*Gower*)

w_j peso de la variable X_j (arbitrario; usuario)

$d_j(\omega_i, \omega_{i'})$ distancia en la variable X_j (normalizada a [0,1])

Universidad del País Vasco Euskal Herriko Unibertsitatea

KISA

Clustering: Distancias

Distancia entre variables

A partir de las *correlaciones*

Variables heterogéneas: Estandarización.

$$z_{ij} = (x_{ij}-x_{0j})/s_j \,,$$

$$x_{0j} = \Sigma_{i=1}{}^{n} x_{ij}/n \,, \quad s_j{}^2 = \Sigma_{i=1}{}^{n} \ (x_{ij}-x_{0j})^2/n$$

Coeficiente de correlación lineal:

$$\boldsymbol{r_{jj'}} = \Sigma_{i=1}{}^{n} \ z_{ij}\cdot z_{ij'}/n$$

Distancia

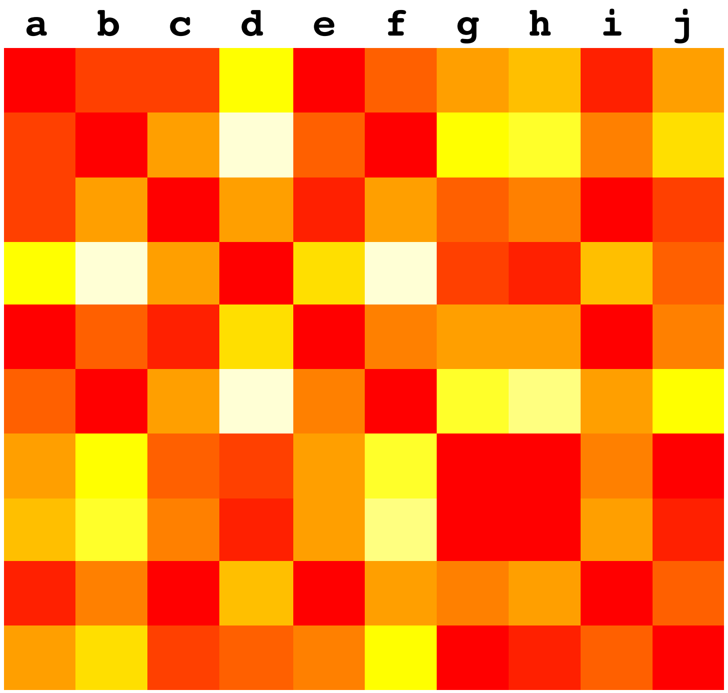
$$\boldsymbol{d(X_{j'},X_{j'})} = \Sigma_{i=1}{}^{n}(z_{ij}-z_{ij'})^2 = 2\cdot\boldsymbol{(1-r_{jj'})}$$

Clustering: Distancias

›› <u>Visualización</u> de la *tabla de distancias*:

Cuanto más cerca están los objetos entre sí, mayor es la intensidad del color (*mapa de calor*).

	a	b	c	d	e	f	g	h	i	j
a	0	3	3	10	1	4	7	8	2	6
b	3	0	6	13	4	1	10	11	5	9
c	3	6	0	7	2	7	4	5	1	3
d	10	13	7	0	9	14	3	2	8	4
e	1	4	2	9	0	5	6	7	1	5
f	4	1	7	14	5	0	11	12	6	10
g	7	10	4	3	6	11	0	1	5	1
h	8	11	5	2	7	12	1	0	6	2
i	2	5	1	8	1	6	5	6	0	4
j	6	9	3	4	5	10	1	2	4	0



Clustering: Distancias

›› <u>Visualización</u> de la *tabla de distancias <u>reordenada</u>*:

Cuanto más cerca están los objetos entre sí, mayor es la intensidad del color (*mapa de calor*).

	f	b	a	e	i	c	j	g	h	d
f	0	1	4	5	6	7	10	11	12	14
b	1	0	3	4	5	6	9	10	11	13
a	4	3	0	1	2	3	6	7	8	10
e	5	4	1	0	1	2	5	6	7	9
i	6	5	2	1	0	1	4	5	6	8
c	7	6	3	2	1	0	3	4	5	7
j	10	9	6	5	4	3	0	1	2	4
g	11	10	7	6	5	4	1	0	1	3
h	12	11	8	7	6	5	2	1	0	2
d	14	13	10	9	8	7	4	3	2	0