

# *Clustering*

## Estadística Multivariante

Sofía Almeida Bruno  
Daniel Bolaños Martínez  
José María Borrás Serrano  
Fernando de la Hoz Moreno  
Pedro Manuel Flores Crespo  
María Victoria Granados Pozo

18 de enero de 2020

# Clustering

- Objetivo: agrupar objetos similares.
- Dadas  $x_1, \dots, x_n$  medidas de  $p$  variables en  $n$  objetos considerados *heterogéneos*. El objetivo del análisis clúster es agrupar estos objetos en  $k$  clases *homogéneas*, donde  $k$  es también desconocido.

# Clustering

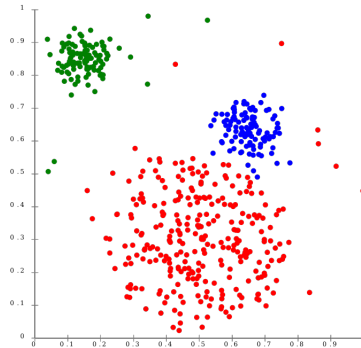


Figura: Ejemplo de *clustering*. [Chi11]

## Ejemplos de *Clustering*

- Biología: determinación de especies.
- *Marketing*: descubrimiento de grupos de clientes.

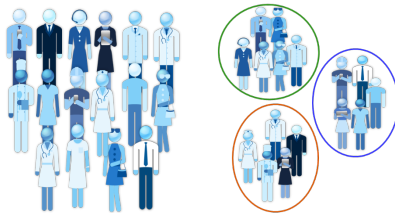


Figura: Ejemplo de *clustering*. [noa]

- Psicología: encontrar tipos de personalidad.
- Arqueología: datar objetos encontrados.
- Planificación urbana: identificar grupos de viviendas.

# *Clustering*

Para realizar un análisis clúster hay que:

- Elegir una medida de similitud.
- Elegir un algoritmo para construir los grupos.
  - ▶ Particionamiento.
  - ▶ Jerárquicos.

## *Medidas de similitud*

Realizar un agrupamiento simple a partir de un conjunto complejo de datos requiere una medida de “similitud”.

Para elegir esta medida es necesario tomar consideraciones iniciales como:

- Naturaleza de las variables (discreta, continua, binaria).
- Escalas de las medidas (nominal, ordinal, intervalo).
- Conocimiento sobre el problema.

Los valores de las variables consideradas deberán ser normalizados.

## *Distancias de similitud para pares de ítems*

La distancia estadística entre dos observaciones  $p$ -dimensionales  $x' = [x_1, \dots, x_p]$  e  $y' = [y_1, \dots, y_p]$  es:

$$d(x, y) = \sqrt{(x - y)^T A (x - y)}$$

Donde  $A = S^{-1}$  y  $S$  contiene las varianzas y covarianzas de la muestra.  $S$  no es conocida antes de aplicar clustering, por lo que se suele usar la distancia euclídea:

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + \dots + (x_p - y_p)^2} = \sqrt{(x - y)^T (x - y)}.$$

## *Otras medidas y coeficientes de similitud*

- Métrica de Minkowski:

$$d(x, y) = \left[ \sum_{i=1}^p |x_i - y_i|^m \right]^{1/m}$$

- Métrica de Canberra (variables no negativas):

$$d(x, y) = \sum_{i=1}^p \frac{|x_i - y_i|}{(x_i + y_i)}$$

- Coeficiente de Czekanowski (variables no negativas):

$$d(x, y) = 1 - \frac{2 \sum_{i=1}^p \min(x_i, y_i)}{\sum_{i=1}^p (x_i + y_i)}$$



## *Binarización de variables*

Si los ítems no pueden ser representados por medidas  $p$ -dimensionales significativas, las parejas de ítems se suelen comparar según la presencia o ausencia de ciertas características.

Matemáticamente se consigue introduciendo una variable binaria, que toma el valor 1 si la característica está presente y el valor 0 si no.

Variables	1	2	3	4	5
Ítem $i$	1	0	0	1	1
Ítem $k$	1	1	0	1	0

Sea  $x_{ij}$  la puntuación de la  $j$ -ésima variable binaria en el  $i$ -ésimo ítem y  $x_{kj}$  la de la  $j$ -ésima en el  $k$ -ésimo ítem, con  $j = 1, \dots, p$ . Entonces:

$$(x_{ij} - x_{kj})^2 = \begin{cases} 0 & \text{si } x_{ij} = x_{kj} \\ 3n + 1 & \text{si } x_{ij} \neq x_{kj} \end{cases}$$

y  $\sum_{j=1}^p (x_{ij} - x_{kj})^2$  proporciona una forma de contar el número de disparidades. Una distancia grande corresponde a muchas disparidades, es decir, ítems desemejantes.

En el Ejemplo anterior:

$$\sum_{j=1}^5 (x_{ij} - x_{kj})^2 = (1 - 1)^2 + (0 - 1)^2 + (0 - 0)^2 + (1 - 1)^2 + (1 - 0)^2 = 2.$$

## Problemas de la distancia usada

La distancia usada valora por igual las parejas 1-1 y 0-0. En algunos casos no es cierto.

Existen diversos esquemas para definir los coeficientes de similitud. Organizando las frecuencias de las parejas que coinciden y las que no para los ítems  $i$  y  $k$  en forma de tabla de contingencia:

		Ítem $k$		
		1	0	Total
Ítem $i$	1	$a$	$b$	$a + b$
	0	$c$	$d$	$c + d$
Total		$a + c$	$b + d$	$p = a + b + c + d$

### Cuadro: Coeficientes de similitud para ítems *clustering*.

Coeficiente	Fundamento
1 $\frac{a+d}{p}$	Las parejas 1-1 y 0-0 ponderan lo mismo.
2 $\frac{2(a+d)}{2(a+d)+b+c}$	Las parejas 1-1 y 0-0 ponderan el doble.
3 $\frac{a+d}{a+d+2(b+c)}$	Las parejas que no coinciden ponderan el doble.
4 $\frac{a}{p}$	No hay parejas 0-0 en el numerador.
5 $\frac{a}{a+b+c}$	No hay parejas 0-0 en el numerador ni el denominador (Las parejas 0-0 son irrelevantes).
6 $\frac{2a}{2a+b+c}$	No hay parejas 0-0 en el numerador ni el denominador. Las parejas 1-1 ponderan el doble.
7 $\frac{a}{a+2(b+c)}$	No hay parejas 0-0 en el numerador ni el denominador. Las parejas que no coinciden ponderan el doble.
8 $\frac{a}{b+c}$	Proporción de parejas que coinciden (excluyendo las 0-0) en relación a las parejas que no coinciden.

## *Monotonicidad*

La monotonicidad es una característica que afecta a algunos procedimientos de *clustering* haciendo que sean invariantes al cambio de definición de similitud.

Un ejemplo de ello son los procedimientos jerárquicos de asociación simple y completa. Para estos métodos, cualquier elección entre los coeficientes de la Tabla anterior producirá las mismas agrupaciones.

## *Ejemplo de cálculo de coeficientes de similitud*

Veamos un ejemplo de cálculo de coeficientes de similitud.  
Definimos 5 individuos con las características mostradas:

	Altura (in)	Peso (lb)	Color de ojos	Color de pelo	Mano predominante	Género
Individuo 1	68	140	Verde	Rubio	Derecha	Femenino
Individuo 2	73	185	Marrón	Moreno	Derecha	Masculino
Individuo 3	67	165	Azul	Rubio	Derecha	Masculino
Individuo 4	64	120	Marrón	Moreno	Derecha	Femenino
Individuo 5	76	210	Marrón	Moreno	Izquierda	Masculino

Definimos seis variables binarias  $X_1, X_2, X_3, X_4, X_5, X_6$ :

$$X_1 = \begin{cases} 1 & \text{si Altura} \geq 72 \\ 0 & \text{si Altura} < 72 \end{cases}$$

$$X_2 = \begin{cases} 1 & \text{si Peso} \geq 150 \\ 0 & \text{si Peso} < 150 \end{cases}$$

$$X_3 = \begin{cases} 1 & \text{si Ojos marrones} \\ 0 & \text{si otro color de ojos} \end{cases}$$

$$X_4 = \begin{cases} 1 & \text{si Pelo rubio} \\ 0 & \text{si otro color de pelo} \end{cases}$$

$$X_5 = \begin{cases} 1 & \text{si Diestro} \\ 0 & \text{si Zurdo} \end{cases}$$

$$X_6 = \begin{cases} 1 & \text{si Masculino} \\ 0 & \text{si Femenino} \end{cases}.$$

Las puntuaciones de los individuos 1 y 2 en las  $p = 6$  variables binarias se muestran a continuación:

	$X_1$	$X_2$	$X_3$	$X_4$	$X_5$	$X_6$
Individuo 1	0	0	0	1	1	1
Individuo 2	1	1	1	0	1	0



El número de parejas que coinciden y las que no se indican a continuación en la tablas de contingencias:

		Individuo 2		Total
		1	0	
Individuo 1	1	1	2	3
	0	3	0	3
Total		4	2	6

Empleando el coeficiente de similitud 1, que pondera por igual las parejas que coinciden, calculamos:

$$\frac{a + d}{p} = \frac{1 + 0}{6} = \frac{1}{6}.$$

Continuando con el coeficiente de similitud 1, calculamos los demás números de similitud para cada pareja de individuos. Mostramos los resultados en la matriz simétrica 5 x 5.

		Individuo				
		1	2	3	4	5
Individuo	1	1				
	2	1/6	1			
	3	4/6	3/6	1		
	4	4/6	3/6	2/6	1	
	5	0	5/6	2/6	2/6	1

Basándonos en las magnitudes del coeficiente de similitud, debemos concluir que los individuos 2 y 5 son los más similares mientras que los individuos 1 y 5 son los menos similares.

## Construcción de similitudes a partir de distancias

Fijando:

$$s_{ik} = \frac{1}{1 + d_{ik}},$$

donde  $0 < s_{ik} \leq 1$  es la similitud entre los ítems  $i$  y  $k$ , entonces  $d_{ik}$  es la distancia correspondiente.

Pero no siempre podemos construir distancias “verdaderas”, a partir de similitudes. Sólo se pueden construir si la matriz de similitudes es definida no negativa y la máxima similitud cumple  $s_{ii} = 1$ .

Entonces  $d_{ik} = \sqrt{2(1 - s_{ik})}$ , cumple las propiedades de una distancia.

## *Medidas de similitud para pares de variables*

Para algunas aplicaciones, son las variables, en lugar de los ítems, las que deben ser agrupadas. Las medidas de similitud para variables suelen tomar la forma de coeficientes de correlaciones muestrales.

Cuando las variables son binarias, los datos se pueden organizar en una tabla de contingencia que tiene la siguiente forma:

		Variable $k$		Total
		1	0	
Variable $i$	1	$a$	$b$	$a + b$
	0	$c$	$d$	$c + d$
Total		$a + c$	$b + d$	$n = a + b + c + d$

La fórmula del coeficiente de correlación producto-momento aplicada a las variables binarias de la tabla de contingencia nos da:

$$r = \frac{ad - bc}{[(a + b)(c + d)(a + c)(b + d)]^{1/2}}.$$

Podemos tomar  $r$  como la medida de similitud entre las dos variables.

Se cumple la relación ( $r^2 = \chi^2/n$ ) para evaluar la independencia de dos variables categóricas. Para un  $n$  fijo, una similitud (o correlación) grande es consistente con la presencia de dependencia.

## Ejemplo idiomas

Medimos las similitudes de 11 lenguajes en base a los primeros 10 números naturales en cada idioma.

Inglés (E)	Noruego (N)	Danés (Da)	Holandés (Du)	Alemán (G)	Francés (Fr)	Español (Sp)	Italiano (I)	Polaco (P)	Húngaro (H)	Finés (Fi)
one	en	en	een	eins	un	uno	uno	jeden	egy	yksi
two	to	to	tweet	zwei	deux	dos	due	dwa	ketto	kaksi
three	tre	tre	drie	drei	trois	tres	tre	trzy	három	kolme
four	fire	fire	vier	vier	quatre	cuatro	quattro	cztery	negy	nelja
five	fem	fem	vijf	funf	cinq	cinco	cinque	piec	öt	viisi
six	seks	seks	zes	sechs	six	seis	sei	szesc	hat	kuusi
seven	sju	syv	zeven	sieben	sept	siete	sette	siedem	het	seitsemán
eight	atte	otte	acht	acht	huit	ocho	otto	osiem	nyolc	kahdeksan
nine	ni	ni	negen	neun	neuf	nueve	nove	dziewiec	kilenc	yhdeksan
ten	ti	ti	tien	zehn	dix	diez	dieci	dziesięc	tíz	kymmenen

Los resultados confirman que el inglés, noruego, danés, holandés y alemán parecen formar un grupo. El francés, español, italiano y polaco forman otro, mientras que el húngaro y el finés no forman parte de ninguno.

[illegible]

## Número de clústeres

- En los algoritmos de *clustering*, uno de los problemas es determinar el número idóneo de clústeres  $k$ .
- Es un proceso ambiguo. Depende de las interpretaciones según la forma y la escala de la distribución de los datos y la solución deseada.
- Como  $k$  decrece de  $n$  a 1, el valor de la distancia debería aumentar ya que tendría que ser mayor cuando dos clústeres distintos se agrupan en uno solo.



## Número de clústeres-Método del codo

Consiste en dibujar la gráfica de las distancia a los centros de cada clúster en función del número de clústeres. Definimos:

$$SSE_k = \sum_{i=1}^{n_k} \| \mathbf{y}_i - \bar{\mathbf{y}}_k \|^2,$$

y para cada  $k$  dibujamos

$$D_k = \sum_{i=1}^k SSE_k.$$

## Número de clústeres-Método del codo

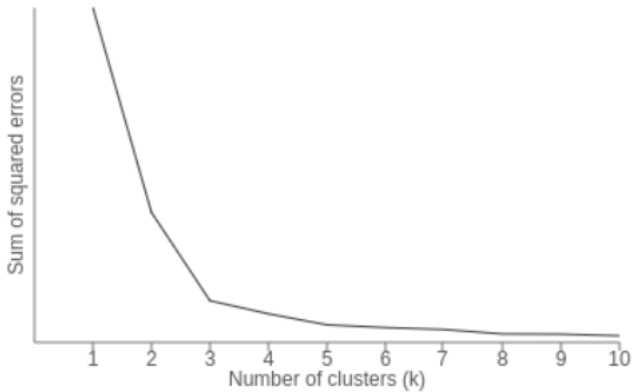


Figura: Ejemplo del método del codo.

## Número de clústeres-Estadístico $R^2$

Para  $n$  clústeres la suma total de las distancias al cuadrado es  $T = \sum_{i=1}^n \|\mathbf{y}_i - \bar{\mathbf{y}}\|^2$ . Así, para  $k$  clústeres definimos  $R^2$  como

$$R_k^2 = \frac{T - \sum_k SSE_k}{T}.$$

Para  $n$  clústeres  $SSE_k = 0$  por lo que  $R^2 = 1$ . Una gran disminución en  $R_k^2$  representaría un agrupamiento diferente.

También podríamos tener en cuenta el cambio en  $R^2$  al unir los clústeres  $R$  y  $S$  como  $SR^2 = R_k^2 - R_{k-1}^2$ . El estadístico  $SR^2$  representa, en función de  $T$ , la proporción de  $SSE_t - (SSE_r + SSE_s)$  donde los clústeres  $C_R$  y  $C_S$  se han unido para formar el clúster  $C_T$ . Cuanto mayor sea el índice mayor será la pérdida de homogeneidad.

## Número de clústeres-Varianza agrupada

Para un solo clúster

$$s^2 = \sum_{i=1}^n \|\mathbf{y}_i - \bar{\mathbf{y}}\|^2 / p(n - 1).$$

Para el clúster  $C_k$

$$s^2 = \sum_{i=1}^{n_k} \|\mathbf{y}_i - \bar{\mathbf{y}}_k\|^2 / p(n_k - 1).$$

Valores grandes de la varianza agrupada indica que los clústeres no son homogéneos. Por lo tanto, si tiende a cero para algún  $k < n$  indica la formación de un clúster homogéneo.

## Número de clústeres-Pseudo estadísticos

El pseudo estadístico  $F$  se define como

$$F_k^* = \frac{(T - \sum_k SSE_k)/(k - 1)}{\sum_k SSE_k/(n - k)}.$$

El pseudo estadístico  $t^2$  se define como

$$\text{pseudo } t^2 = \frac{[SSE_t - (SSE_r + SSE_s)](n_R + n_S - 2)}{SSE_r + SSE_s}.$$

## Número de clústeres-*Silhouette method*

Definimos el índice:

$$s(i) = \frac{b(i) - a(i)}{\max\{b(i), a(i)\}}, \quad \forall i = 1, \dots, n$$

donde

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, j \neq i} d(i, j)$$

y

$$b(i) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i, j).$$

Se escoge el  $k$  que maximice el valor medio de  $s(i)$ .

## Número de clústeres-*Silhouette method*

k	Silhouette coeff.
2	0.7049787496083262
3	0.5882004012129721
4	0.6505186632729437
5	0.5745566973301872
6	0.43902711183132426

Cuadro: Ejemplo *silhouette method*.

Vemos que se obtienen los mejores resultados con 2 o 4 clústeres.

## Número de clústeres-*Gap method*

El  $k$  elegido será aquel que maximice el valor de:

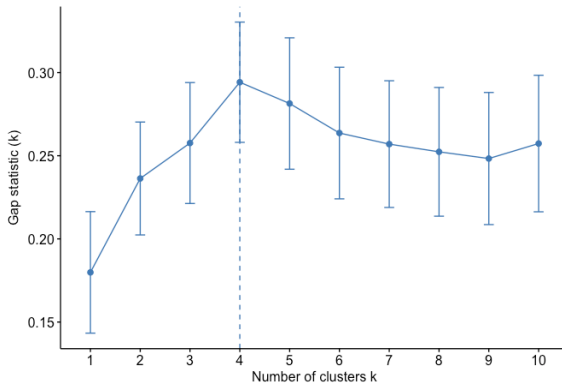
$$Gap(k) = E_n^* \{\log(W_k)\} - \log(W_k).$$

En la fórmula anterior  $E_n^*$  denota la media de una muestra de tamaño  $n$  y

$$W_k = \sum_{R=1}^k \frac{1}{2n_R} \sum_{ij \in C_R} d(i, j).$$



## Número de clústeres-*Gap method*



**Figura:** Ejemplo del método de la brecha.

## Referencias I



Chire, *Cluster analysis with optics on a density-based data set.*,  
[https://commons.wikimedia.org/wiki/File:  
OPTICS-Gaussian-data.svg](https://commons.wikimedia.org/wiki/File:OPTICS-Gaussian-data.svg), October 2011.



*Understanding data mining clustering methods.*