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

16 de enero de 2020

Clustering

- Objetivo: agrupar objetos similares.
- Dadas $x_1, \cdots, x_n$ 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

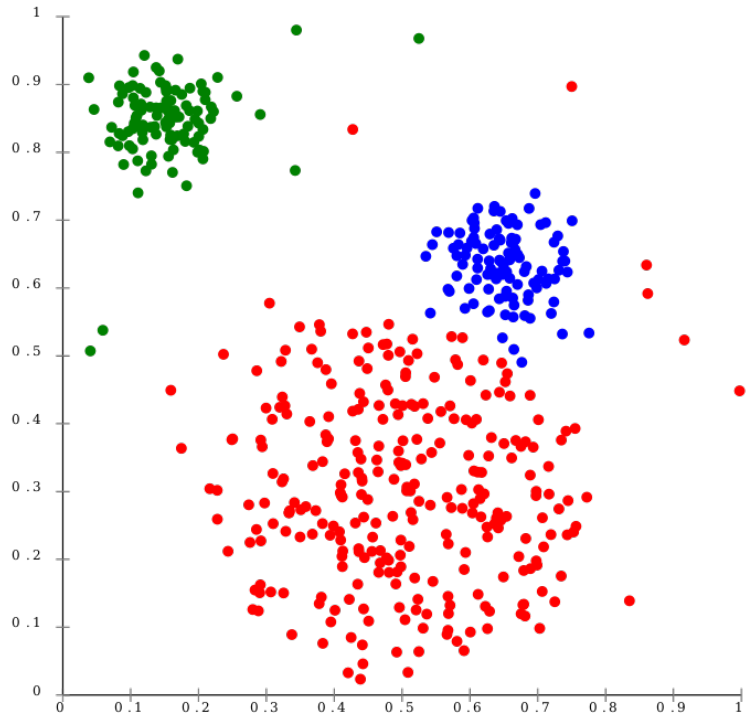

Figura: Ejemplo de *clustering*. [Chi11]

Ejemplos de *Clustering*

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

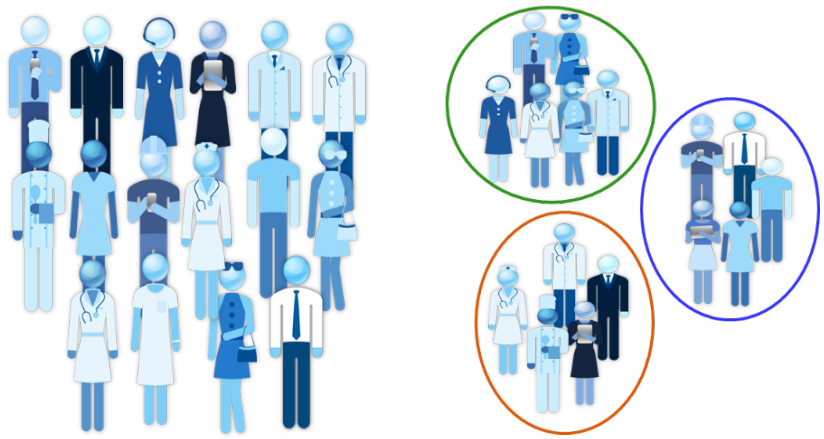

Figura: Ejemplo de *clustering*. [noa]

- 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.

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 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

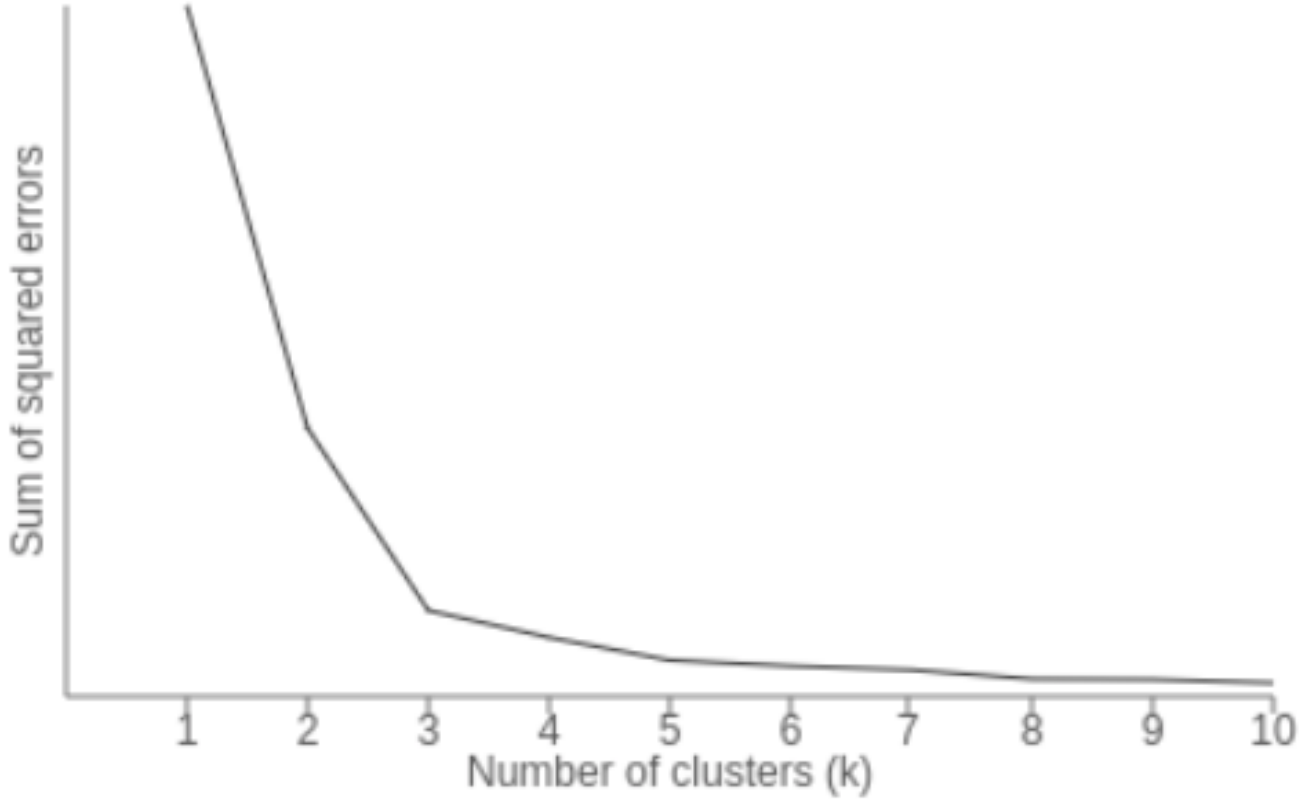

Figura: Ejemplo del método del codo.

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:

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

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 de 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*

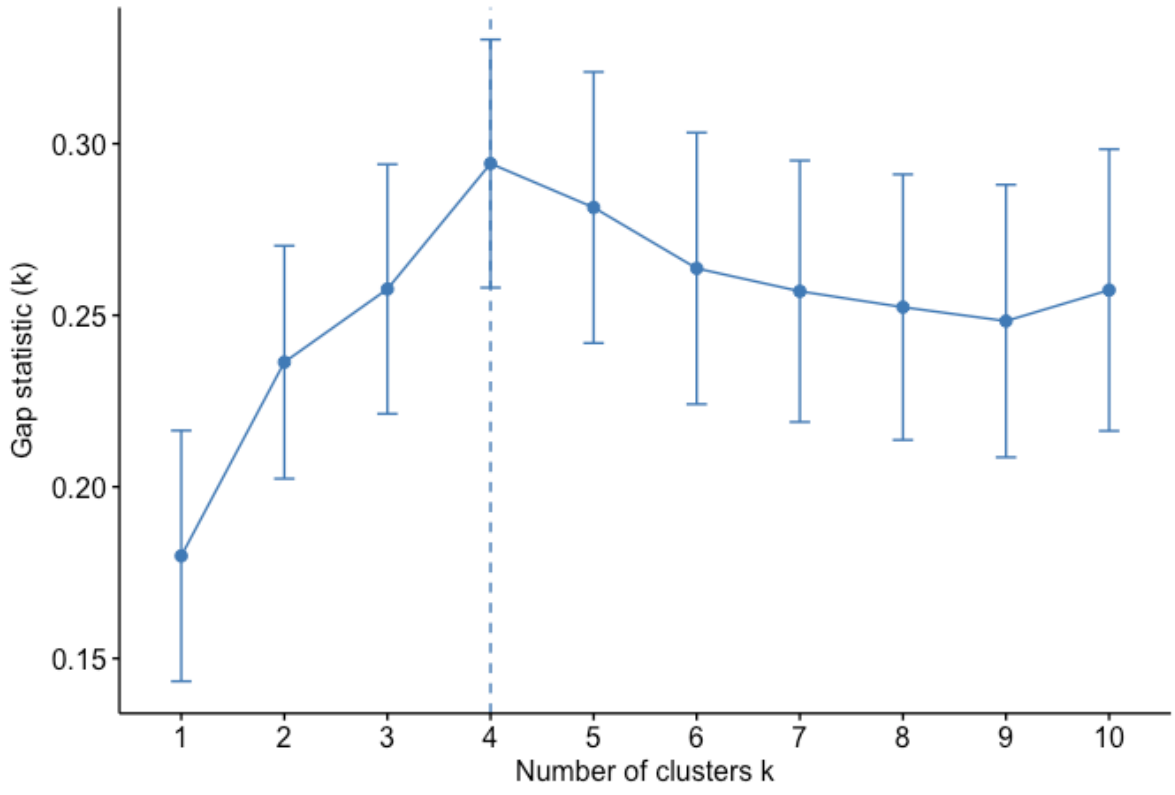

Figura: Ejemplo del método de la brecha.

Referencias I


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


Understanding data mining clustering methods.