
## Herramientas de R para la investigación de mercados

# 1. Segmentación de mercados: uso de herramientas cluster

José Ignacio Casas

# Segmentación de mercados

1. Objetivo, fases y elección de variables
2. Distancias entre objetos
3. Cluster jerárquico
4. Clusters no jerárquicos (datos numéricos): k-means, PAM, CLARA
5. Validación

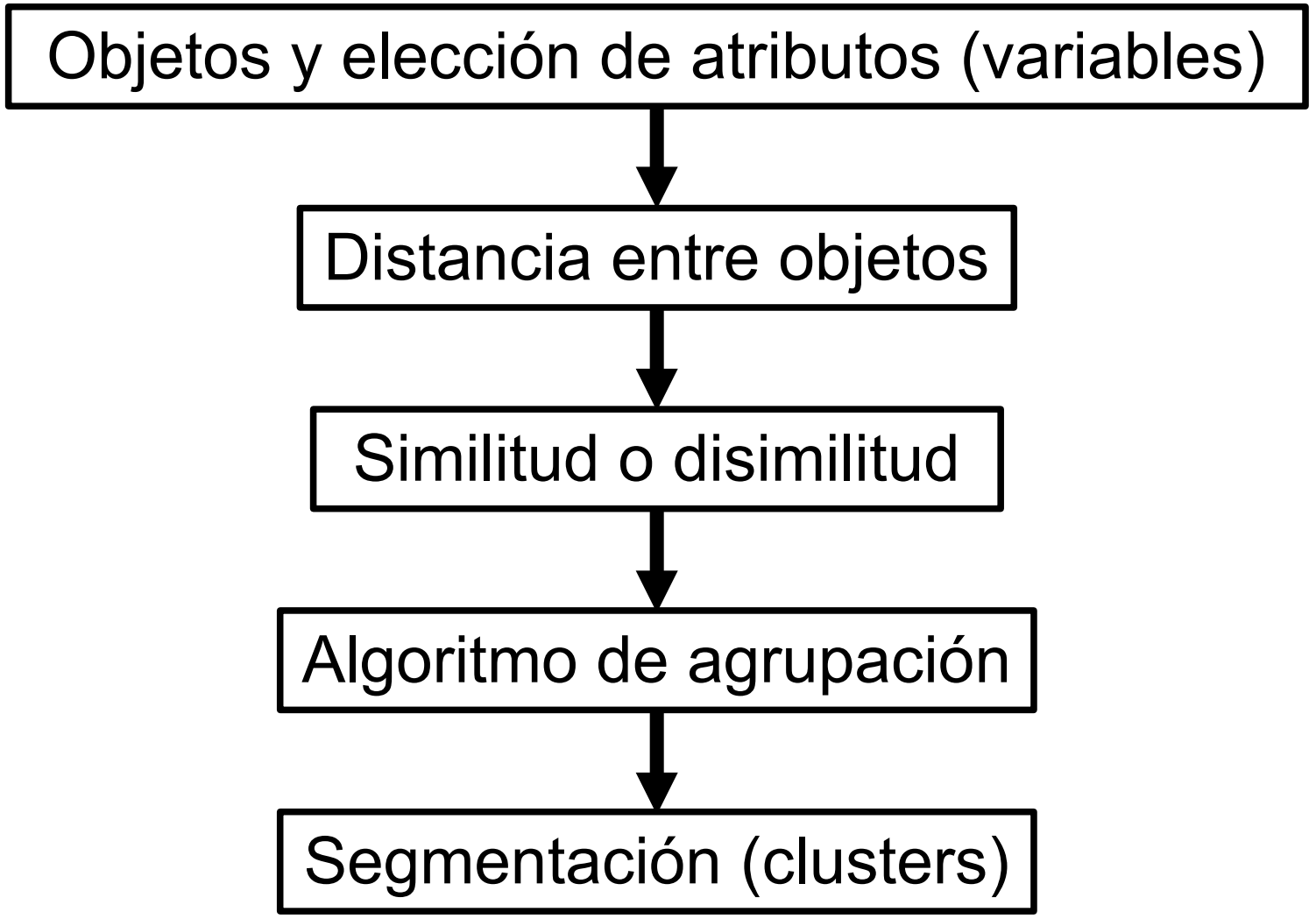
# 1.Objetivo de la segmentación

- Identificar grupos homogéneos de objetos: **clusters**

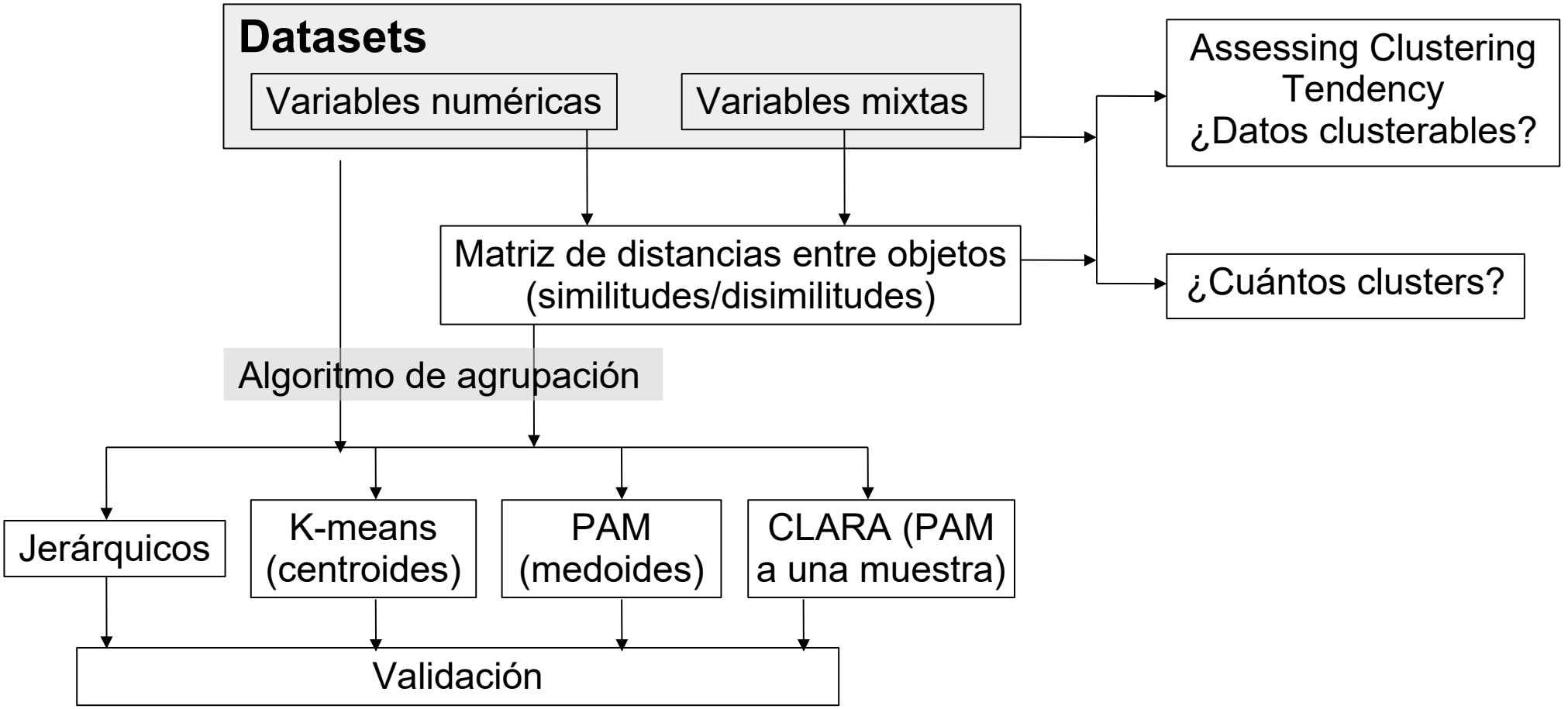
- Los objetos pueden ser individuos, grupos de clientes, empresas, etc

- Los objetos en cada **cluster** deben ser lo más parecidos entre sí y lo más diferentes de los objetos de los demás **clusters**

# 1.Fases de la segmentación



# 1.Análisis de Clusters



# 1.Elección de variables

- Qué variables incluir
  - Las que ofrecen una clara diferenciación entre objetos
  - Validez de criterio (*criterion validity*): en qué medida variables "independientes" para la agrupación de casos están asociadas con una o más variables "criterio" no incluidas en el análisis (compra, *willingness-to-pay*, etc.). Si existe esa relación, debería haber diferencias significativas de la variable "criterio" entre los distintos clusters
- Tamaño muestral vs. # variables a incluir vs. # clusters previsto. Reglas sugeridas:
  - Para clusters de tamaño similar: tamaño muestra = 10 x # Variables x # clusters
  - Tamaño muestral = 70 x # variables
- Variables con una alta correlación
  - El análisis de clusters no diferencia conceptualmente entre las variables incluidas. Si éstas muestran una alta correlación (>= 0,90) los aspectos específicos cubiertos por dichas variables estarán sobre-representados
  - En estos casos se sustituyen las variables por los factores derivados de un PCA o FA. **Inconvenientes**: puede reducir la identificación de clusters útiles para la segmentación. Es mejor reducir el número de variables, pero en caso de duda, la segmentación basada en factores subyacentes puede ser mejor

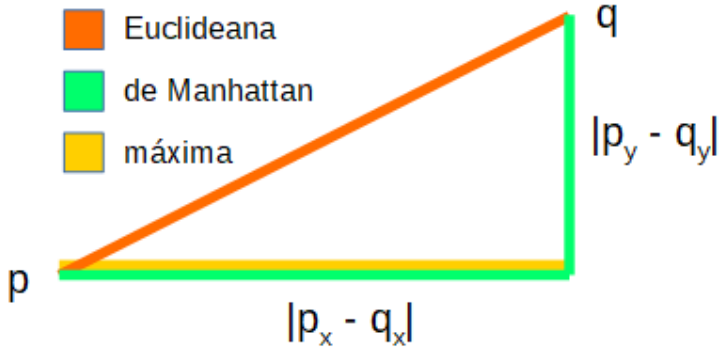
# 2.Distancias entre objetos

- Estandarización previa de las variables incluidas en el cálculo
- Distancias euclideanas (numéricas):
  - Euclideanas propiamente dichas
  - de Manhattan
  - máxima o de Chebyshev
  - de Minkoswki (L-Norm)
  - de Canberra
- Distancias no euclideanas:
  - Atributos binarios simétricos o asimétricos: coeficiente de **Jaccard**
  - Con datos mixtos: distancia de **Gower**
  - Similitud de coseno
- Distancias basadas en correlaciones
- Ejemplo comparativo de distancias entre 3 objetos

# 2.Distancias euclideanas

- Euclideana propiamente dicha $\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$
- de Manhattan $\sum_{i=1}^{n}|x_i-y_i|$
- máxima o de Chebyshev $max_i(|x_i-y_i|)$
- de Minkoswki (L-Norm)[1] $(\sum_{i=1}^{n}|x_i-y_i|^p)^{1/p}$
- de Canberra $d(p,q)=\sum_{i=1}^{n}\frac{|p_i-q_i|}{|p_i|+|q_i|}$
  - siendo $p$ y $q$ dos vectores n-dimensionales

**Gráficamente:**




(1) versión generalizada de las anteriores: p=2: Euclideana; p=1: de Manhattan; p=infinito: máxima

# 2.Distancias no euclideanas: atributos binarios

$p$ atributos binarios:

$p = q + r + s + t$

$q$ coincidencias positiva

$t$ coincidencias negativas

| | Objeto $j$ | | | |
|---|---|---|---|---|
| Objeto $i$ | | 1 | 0 | suma |
| | 1 | $q$ | $r$ | $q + r$ |
| | 0 | $s$ | $t$ | $s + t$ |
| | suma | $q + s$ | $r + t$ | $p$ |

- Disimilaridad binaria entre $i$ y $j$:

$$d(i,j) = \frac{r+s}{q+r+s+t}$$

- Disimilaridad asimétrica binaria:

$$d(i,j) = \frac{r+s}{q+r+s}$$

- Coeficiente de ***Jaccard***:

$$\text{sim}(i,j) = \frac{q}{q+r+s} = 1 - d(i,j)$$

# 2.Distancias no euclideanas: datos mixtos

Distancia de *Gower*: $d^2_{ij} = 1 - s_{ij}$

siendo:

$$s_{ij} = \frac{\sum_{h=1}^{p_1} (1 - |x_{ih} - x_{jh}|/G_h) + a + \alpha}{p_1 + (p_2 - d) + p_3}$$

$s_{ij}$: coeficiente de similitud de Gower,

$p_1$: número de variables cuantitativas continuas,

$p_2$: número de variables binarias,

$p_3$: número de variables cualitativas(no binarias),

$a$: número de coincidencias (1,1) en las variables binarias,

$d$: número de coincidencias (0,0) en las variables binarias,

$\alpha$: número de coincidencias en las variables cualitativas (no binarias) y

$G_h$: rango (o recorrido) de la h-ésima variable cuantitativa. (Rango = X máximo – X mínimo)

# 2.Distancias no euclideanas: similitud de coseno

- Coseno del ángulo que forman dos vectores: medida de similitud de sus orientaciones, independientemente de sus magnitudes
  - Con la misma orientación (ángulo de 0º): $\cos(\alpha) = 1$
  - Si son perpendiculares (ángulo de 90º): $\cos(\alpha) = 0$
  - Si tienen orientaciones opuestas (ángulo de 180º): $\cos(\alpha) = -1$

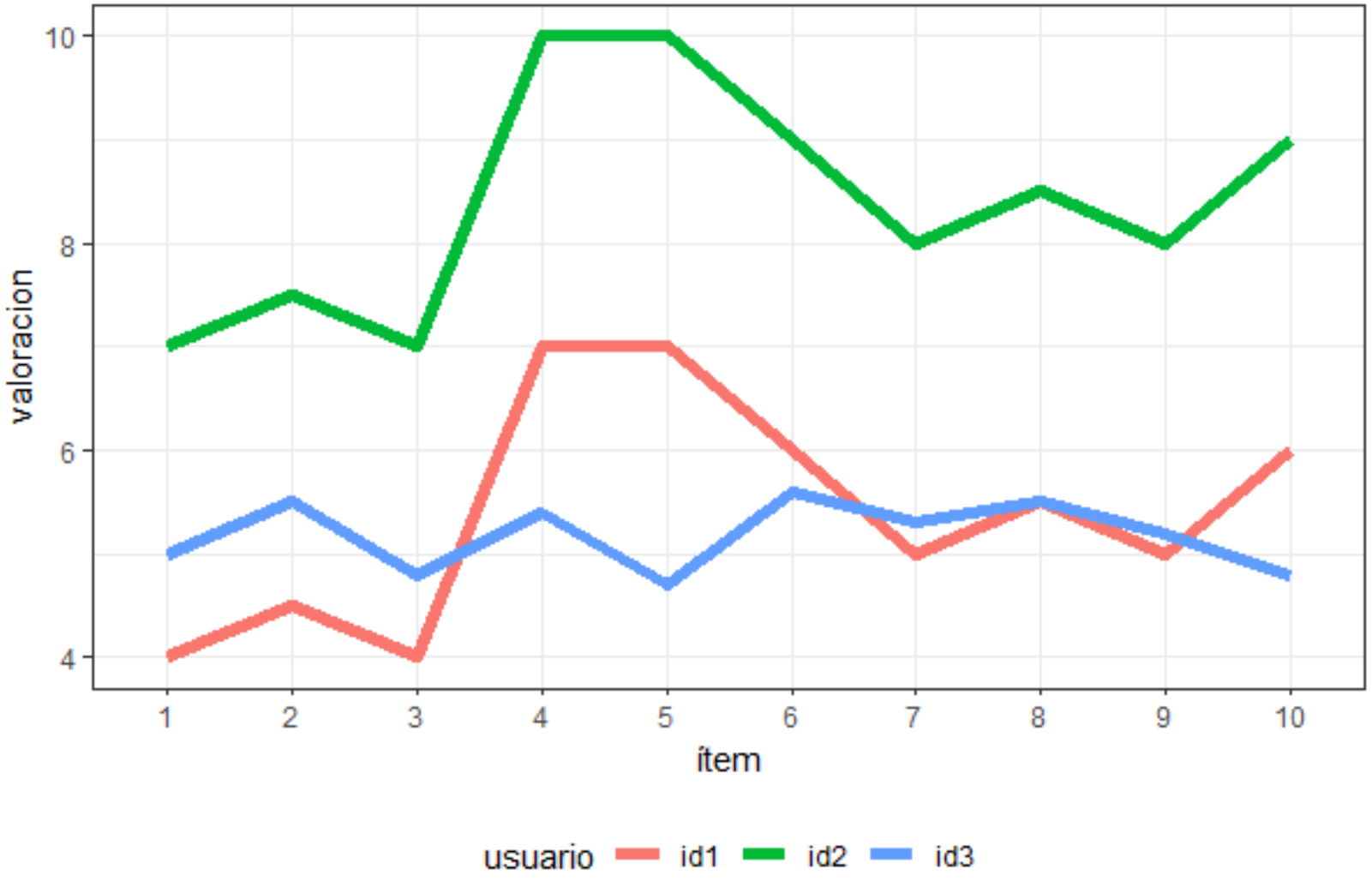
$$\cos(\alpha) = \frac{x \cdot y}{\|x\| \cdot \|y\|}$$

- Si los vectores se normalizan: **coseno centrado**, (equivalente a la correlación de Pearson)

# 2.Distancias basadas en correlaciones

- ***distancia = 1 - coeficiente de correlación***
- Casos más habituales:
  - Distancia de correlación de ***Pearson***: el grado de relación lineal entre dos objetos
  - Distancia de correlación de ***Spearman***: coeficiente de correlación de Pearson, pero aplicado después de transformar las puntuaciones originales en rangos
  - Distancia de ***Kendall***: comparación entre posiciones dentro del rango de variables ordinales

# 2.Ejemplo comparativo de distancias entre 3 objetos (1)



# 2.Ejemplo comparativo de distancias entre 3 objetos (y 2)

Matriz de distancias euclideanas:

| | id1 | id2 |
|---|---|---|
| id2 | 9,487 | |
| id3 | 3,496 | 10,743 |

Matriz de distancias basadas en la correlación de Pearson:

| | id1 | id2 |
|---|---|---|
| id2 | 0,000 | |
| id3 | 0,976 | 0,976 |

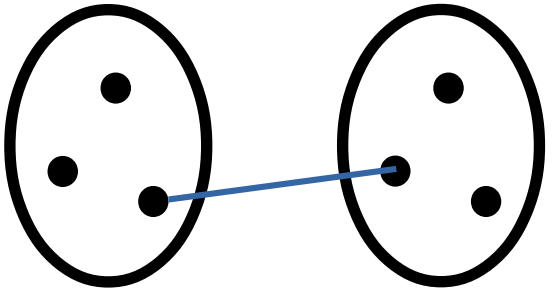
# 3.Clusters jerárquico

- **No** hay que fijar de antemano el número de clusters, pero consume muchos recursos: en cada paso se comparan todos los posibles pares
- 2 direcciones de agrupación: **Aglomerativo** (ascendente) y **Disociativo** (descendente). Resultado final: **dendrograma**
- Algoritmos de agrupación (*"linkage"*) de objetos y sucesivos clusters más habituales:
  - **single**: valor mínimo de las distancias entre todos los pares posibles de un objeto de cada cluster (tiende a crear clusters largos y poco compactos)
  - **complete**: valor máximo de las distancias entre todos los pares posibles de un objeto de cada cluster (método muy habitual)
  - **average**: promedio de las distancias entre todos los pares posibles de un objeto de cada cluster
  - **Ward**: encuentra el par de clústers que llevan al incremento mínimo del total de la varianza del clúster después de mezclados (método muy habitual)

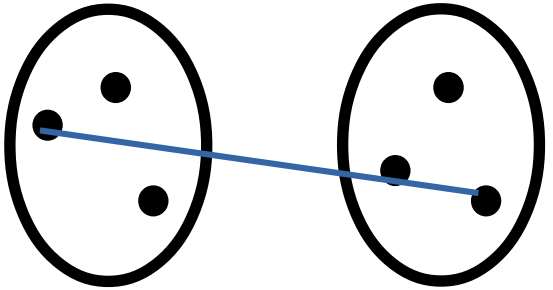
# 3.Clusters jerárquico

- Algoritmos de agrupación (*linkage*) de objetos y sucesivos clusters más habituales:

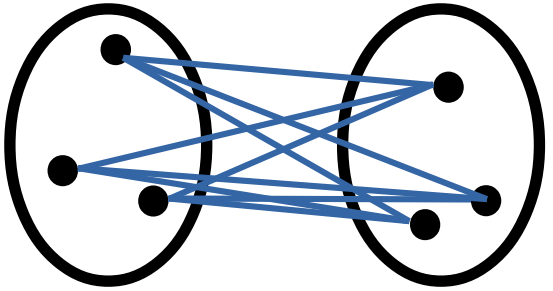
single



complete



average



# 4.k-means: cluster no jerárquico con datos numéricos

- Se fija de antemano el número ***k*** de **clusters**
- Cada **cluster** está representado por su centro (***centroide***), que se corresponde con la media de los objetos asignados al cluster
- Algoritmo de agrupación:
  a) elección aleatoria inicial de ***centroides***
  b) asignación de cada elemento al ***centroide*** más cercano (distancia euclideana)
  c) nuevo cálculo de ***centroides***
  d) repetición de los pasos *b* y *c* hasta que no cambian las asignaciones
- Ventajas: es muy eficiente
- Desventajas: a) la elección aleatoria inicial puede variar cada vez y ser errática; b) tiende a crear grupos esféricos de tamaño similar; c) sensible a los ***outliers***
- Soluciones posibles: a) probar diferente número de ***k* clusters** y comparar resultados; b) realizar varios cálculos (variará la elección aleatoria inicial) y comparar resultados

# 4.PAM: clusters no jerárquico con datos numéricos

- **P**artitioning **A**round **M**edoids: cada cluster representado por un **medoide** (el objeto ubicado más hacia el centro en todo el grupo)
- Se fija de antemano el número ***k*** de clusters
- Menos sensible que el método **k-means** al ruido y los outliers, ya que minimiza una suma de disimilaridades en vez de la suma de las distancias euclideanas al cuadrado
- Algoritmo de agrupación. Fase Build:
  a) Elección aleatoria de ***k* medoides** (objetos del dataset), a no ser que se escojan éstos previamente
  b) Calcular la matriz de disimilaridades
  c) Asignación de cada objeto do al **medoide** más cercano
- Algoritmo de agrupación. Fase swap:
  – Se escogen nuevos **medoides** si minimizan la suma de disimilaridades, y se vuelve al punto c)


José Ignacio Casas

# 4.CLARA: cluster no jerárquico con datos numéricos

- **CLARA** (**C**lustering **LAR**ge **A**pplications)
- Se apoya en **PAM** para analizar clusters en datasets de gran tamaño
- Hay que fijar de antemano el número ***k*** de clusters
- Aplica PAM **a una muestra** y encuentra los medoides de la muestra
- Para mejorar el resultado se toman varias muestras (5 muestras de tamaño 40+2k se consideran suficientes) y se escoge la mejor según el promedio de disparidad de todos los objetos

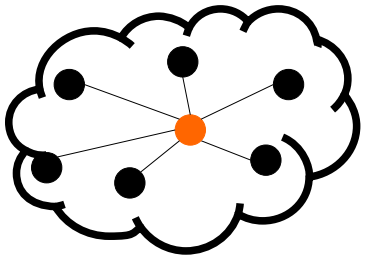
# 5.Validación

- Previa:
  - Assessing Clustering Tendency: ¿hay clusters?
  - Determinación del número óptimo de clusters
- Posterior:
  - Estadísticos de validación externa
  - Estadísticos de validación interna:
    - Coeficiente Silhouette
    - Índice Dunn

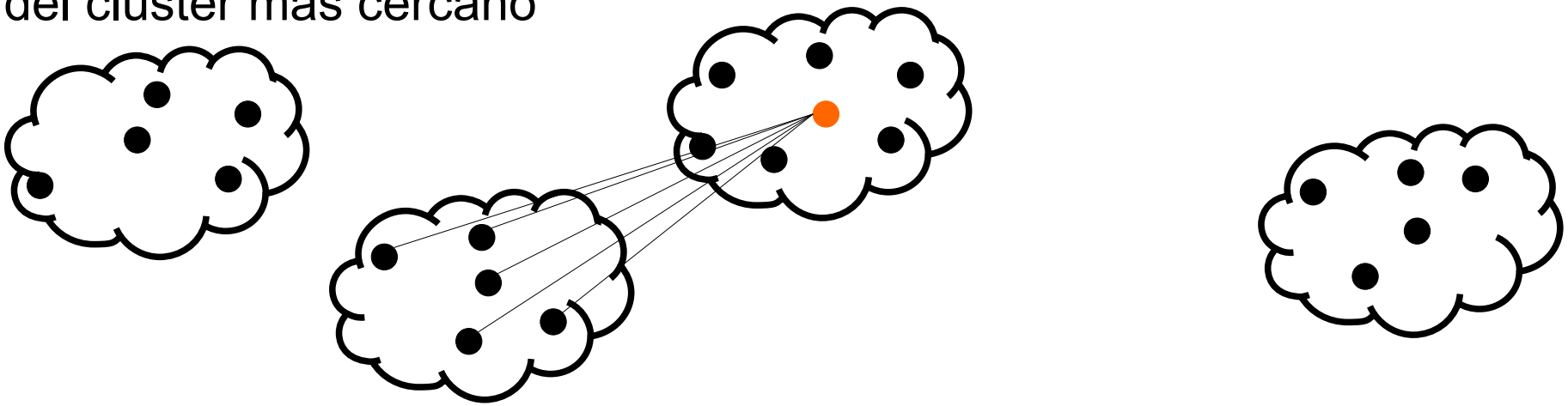
# 5.Coeficiente Silhouette (1)

Dado un objeto x de un cluster:

- Cohesión a(x) : distancia promedio de x a todos los demás objetos del mismo clúster



- Separación b(x) : distancia promedio de x a todos los demás objetos del clúster más cercano



# 5.Coeficiente Silhouette (2)

- El coeficiente Silhouette para el objeto x está definido como:

$$s(x)=\frac{b(x)-a(x)}{max\{a(x),b(x)\}}$$

- Donde el valor de s(x) puede variar entre -1 y 1:

  -1 = mal agrupamiento

  0 = indiferente

  1 = bueno

- El coeficiente Silhouette para un conjunto de objetos es:

$$SC=\frac{1}{N}\sum_{i=1}^{N} s(x)$$

# 5.Índice Dunn

- El Índice Dunn:

$$D = \frac{separación\ mínima\ intercluster}{máximo\ diámetro\ intracluster}$$

- Si existen clusters compactos y bien separados, el índice de Dunn será grande, dado que la distancia entre los clusters se supone grande y el diámetro de los mismos se espera sea pequeño
- Valores entre 0 e infinito
- Un solo cluster con un diámetro excesivamente puede distorsionar el valor del Índice de Dunn