TÉCNICAS INTELIGENTES EN BIOINFORMÁTICA

CLUSTERING

Mario de J. Pérez Jiménez
Grupo de investigación en Computación Natural
Dpto. Ciencias de la Computación e Inteligencia Artificial
Universidad de Sevilla

Máster Universitario en Lógica, Computación e Inteligencia Artificial

Curso 2014-15

Introducción

Grandes avances en la Biología molecular/celular y en la Tecnología:

- Aumento exponencial de la información disponible para la investigación científica.
- Por ejemplo: el uso de la tecnica de **microarrays** permite almacenar millones de datos de expresión génica (regulación y patrones de comportamiento).
  - ⋆ Analizar grupos de genes con similar funcionalidad.
  - ⋆ Estudiar grupos de genes regulados de forma análoga.

Las investigaciones sobre grandes bases de datos (poblaciones de individuos) dependen de muchos factores.

- La variedad de parámetros hace que las investigaciones sean muy complejas.
- Las **técnicas de clasificación** facilitan la **organización** de las grandes bases de datos.

Para simplificar algunos problemas y hacerlos más tratables:

- Es conveniente agrupar objetos/individuos que tienen "**similares características**".

Minería de datos: conjunto de técnicas orientadas a la extracción de conocimiento útil de grandes bases de datos.

- Es una disciplina de la Inteligencia Artificial.

Tipos de datos

Los datos pueden ser de diversos tipos:

- Numéricos.
- Binarios.
- Ordinales.
- Nominales.
- De intervalos.
- Mixtos.
- Etc.

Clustering (I)

Una técnica de minería de datos: **Clustering**.

- Objetivo: a partir de una base de datos inicial (objetos/individuos), formar agrupaciones o clusters de acuerdo con una medidad de similitud (similaridad) entre ellos de tal manera que:
  - La similitud media entre los datos del mismo cluster sea "**alta**".
  - La similitud media entre los datos de distintos clusters sea "**baja**".

Etapas de un proceso de *clustering*:

- Representación de los objetos/individuos (vectores, coordenadas esféricas, grafos, etc.).
- Definición de una medida de similitud/similaridad (proximidad, distancia, etc.).
- Criterio de agrupación o clustering (distintas metodologías).
- Abstracción de los datos (transformación para otros procesos de análisis, etc.).
- Evaluación de los resultados (bondad del proceso, validación, etc.).

Clustering (II)

Existen diferentes métodos para clasificar datos (objetos/individuos) de acuerdo con sus similaridades:

- Jerárquico: la clasificación se estructura en niveles (niveles inferiores contenidos en niveles superiores).
  - ⋆ Aglomerativo (ascendente): se parte de tantos clusters como individuos y se irán formnado grupos según su similitud.
  - ⋆ Disociativo (descendente): Se parte de un único cluster y se va formando clusters según la disimilitud de sus componentes.
- No jerárquico: la clasificación se estructura en grupos sin que existan relaciones entre los diferentes grupos.

Población de individuos (I)

Objeto de estudio: una población de individuos:

- Los individuos tienen una serie de características o propiedades que interesan estudiar.

Por ejemplo:

- La información que proporciona los microarrays se expresa en un matrix $M = (a_{ij})_{1 \leq i \leq n, 1 \leq j \leq m}$: son los individuos de la población.
- Las filas representan los genes que se analizan y las columnas representan las condiciones experimentales.
- El valor a_{ij} representa la cantidad de RNAm expresado por el gen i en las condiciones dadas por j.

Formalización: Sea E un conjunto no vacío y $p \geq 1$ un número natural.

- Un p-individuo es un elemento del conjunto $E \times \overset{(p)}{\ldots} \times E = E^p$.
  - ⋆ Si $a = (a_1, a_2, \ldots, a_p) \in E^p$, entonces $a_r \in E$ codifica la propiedad r-ésima del individuo a.
- Una **población** de p-individuos es un subconjunto finito de E^p.
- Una métrica o distancia sobre E es una aplicación $d : E \times E \longrightarrow \mathbb{R}^+$ (números reales mayores o iguales que cero), tal que para cada $x, y \in E$:
  - ⋆ $d(x, y) = 0$ si y sólo si $x = y$.
  - ⋆ $d(x, y) = d(y, x)$.
  - ⋆ $d(x, y) \leq d(x, z) + d(z, y)$, para cada $z \in E$ (desigualdad triangular).

Población de individuos (II)

Definición: Una similaridad sobre una población $\Omega = \{x_1, x_2, \ldots, x_n\}$ de p-individuos es una aplicación $s : \Omega \times \Omega \longrightarrow \mathbb{R}^+$ tal que

- Para cada par de individuos $x_i, x_j \in \Omega$ se tiene que $s(x_i, x_j) = s(x_j, x_i)$.
- Para cada par de individuos distintos $x_i, x_j \in \Omega$ se tiene que $s(x_i, x_i) = s(x_j, x_j) \geq s(x_i, x_j)$.

Una población $\Omega = \{x_1, x_2, \ldots, x_n\}$ de p-individuos se puede expresar así:

$$\begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \ldots & \ldots & \ldots & \ldots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}$$

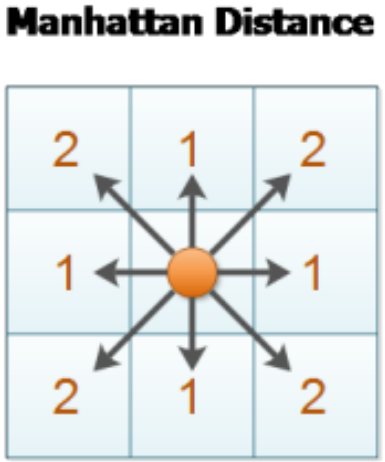
En donde el individuo i-ésimo x_i de la población es la tupla $(x_{i1}, x_{i2}, \ldots, x_{ip})$

Las componentes x_{ir} $(1 \leq r \leq p)$ del individuo son las propiedades o características que se analizan

Distancias (I)

Algunos ejemplos de distancia:

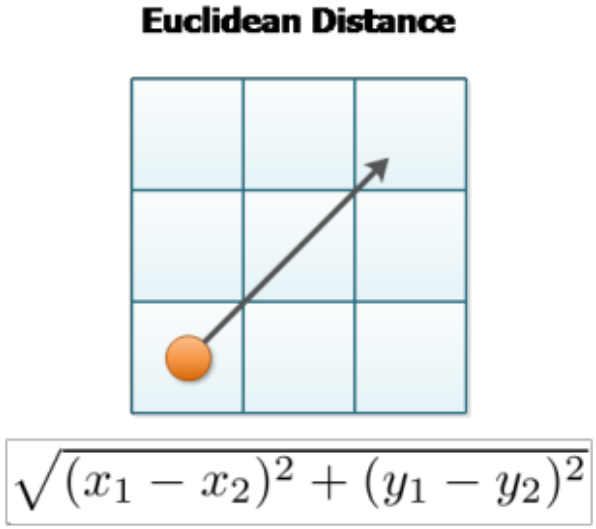
- Distancia de **Minkowski** (siendo $p, q \geq 1$): $d(x_i, x_j) = \sqrt[q]{\sum_{r=1}^{p}(|x_{i,r} - x_{j,r}|^q)}$.
- Distancia de **Manhattan** (caso $q = 1$): $d(x_i, x_j) = \sum_{r=1}^{p} |x_{i,r} - x_{j,r}|$.



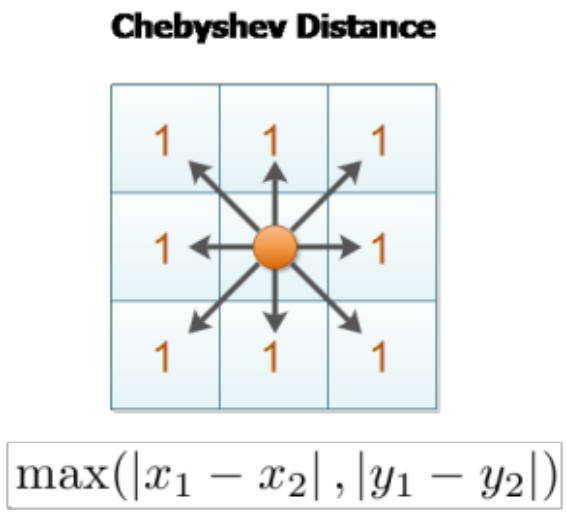
$$|x_1 - x_2| + |y_1 - y_2|$$

Distancias (II)

- Distancia **euclídea** (caso $q = 2$): $d(x_i, x_j) = \sqrt{\sum_{r=1}^{p}(|x_{i,r} - x_{j,r}|^2)}$.



- Distancia de **Chebyshev** (chessboard distance): $d(x_i, x_j) = \text{máx}\{|x_{i,r} - x_{j,r}| \ : \ 1 \leq r \leq p\}$.



Clustering jerárquico aglomerativo

Estrategia de la distancia mínima o similitud máxima (amalgamiento simple: single linkage)

Un **algoritmo de construcción**:

Entrada: una familia básica de clusters de una población (cada cluster contiene un único individuo) y una matriz inicial de distancias entre pares de individuos.

1. Se eligen los dos clusters **más cercanos** y se considera una nueva familia en la que ese par es sustituido por un cluster que contiene a ambos.
2. Se actualiza la matriz de distancias calculando los valores para cada par de clusters en la nueva familia (en cada paso, el número total de clusters disminuye en una unidad).
3. Se vuelve al paso 1 hasta que exista un único cluster

Software: AGNES (Agglomerative Nesting)

Estrategia de la distancia mínima o similitud máxima (II)

Se parte de una población de 7 individuos. Clusters iniciales $\Delta_0 = \{\{a\}, \{b\}, \{c\}, \{d\}, \{e\}, \{f\}, \{g\}\}$
Matriz de distancias iniciales de la población es:

	a	b	c	d	e	f	g
a	0						
b	2.15	0					
c	0.7	1.53	0				
d	1.07	1.14	0.43	0			
e	0.85	1.38	**0.21**	0.29	0		
f	1.16	1.01	0.55	0.22	0.41	0	
g	1.56	2.83	1.86	2.04	2.02	2.05	0

- Nivel $K = 1$: clusters más próximos $\{c\}$ y $\{e\}$ (**0.21**)
  Entonces $\Delta_1 = \{\{a\}, \{b\}, \{\mathbf{c}, \mathbf{e}\}, \{d\}, \{f\}, \{g\}\}$
  Matriz de distancias actualizada:

	a	b	$\{c, e\}$	d	f	g
a	0					
b	2.15	0				
$\{c, e\}$	0.7	1.38	0			
d	1.07	1.14	0.29	0		
f	1.16	1.01	0.41	**0.22**	0.41	0
g	1.56	2.83	1.86	2.04	2.05	0

- Nivel $K = 2$: clusters más próximos $\{d\}$ y $\{f\}$ (**0.22**)
  Entonces $\Delta_2 = \{\{a\}, \{b\}, \{c, e\}, \{\mathbf{d}, \mathbf{f}\}, \{g\}\}$
  Matriz de distancias actualizada:

	a	b	$\{c, e\}$	$\{d, f\}$	g
a	0				
b	2.15	0			
$\{c, e\}$	0.7	1.38	0		
$\{d, f\}$	1.07	1.01	**0.29**	0	
g	1.56	2.83	1.86	2.04	0

- Nivel $K = 3$: clusters más próximos $\{c, e\}$ y $\{d, f\}$ (**0.29**)
  Entonces $\Delta_3 = \{\{a\}, \{b\}, \{\{\mathbf{c}, \mathbf{e}\}, \{\mathbf{d}, \mathbf{f}\}\}, \{g\}\}$
  Matriz de distancias actualizada:

	a	b	$\{\{c, e\},\{d, f\}\}$	g
a	0			
b	2.15	0		
$\{\{c, e\},\{d, f\}\}$	**0.7**	1.01	0	
g	1.56	2.83	1.86	0

- Nivel $K = 4$: clusters más próximos $\{a\}$ y $\{\{c, e\}, \{d, f\}\}$ (**0.7**)
  Entonces $\Delta_4 = \{\{\mathbf{a}\}, \{\{\mathbf{c}, \mathbf{e}\}, \{\mathbf{d}, \mathbf{f}\}\}\}, \{b\}, \{g\}\}$
  Matriz de distancias actualizada:

	$\{\{a\},\{\{c, e\}, \{d, f\}\}\}$	b	g
$\{\{a\},\{\{c, e\}, \{d, f\}\}$	0		
b	**1.01**	0	
g	1.56	2.83	0

- Nivel $K = 5$: clusters más próximos $\{\{a\}, \{\{c, e\}, \{d, f\}\}\}$ y $\{b\}$ (**1.01**)
  Entonces $\Delta_5 = \{\{\{\mathbf{a}\}, \{\{\mathbf{c}, \mathbf{e}\}, \{\mathbf{d}, \mathbf{f}\}\}\}, \{\mathbf{b}\}\}, \{g\}\}$
  Matriz de distancias actualizada:

	$\{\{\{a\}, \{\{c, e\}, \{d, f\}\}\},\{b\}\}$	g
$\{\{\{a\}, \{\{c, e\}, \{d, f\}\}\},\{b\}\}$	0	
g	**1.56**	0

- Nivel $K = 6$: clusters más próximos $\{\{\{a\}, \{\{c, e\}, \{d, f\}\}\}\}, \{b\}\}$ y $\{g\}$ (**1.56**)
  Entonces $\Delta_6 = \{\{\{\{\mathbf{a}\}, \{\{\mathbf{c}, \mathbf{e}\}, \{\mathbf{d}, \mathbf{f}\}\}\}, \{\mathbf{b}\}, \{\mathbf{g}\}\}$

Representación gráfica mediante un árbol de clasificación: **dendograma**.

Otras estrategias

- Estrategia de la distancia máxima o similitud mínima (**amalgamiento completo: complete linkage**).

$$d(C_i, C_j) = \max\{d(x_l, x_m) \mid x_l \in C_i \wedge x_m \in C_j\}$$

Se unirán C_i y C_j para los valores min $\{d(C_i, C_j)\}$.

- Estrategia minimización de la distancia o similitud promedio no ponderada.
Sean dos clusters C_i y C_j. Supongamos que C_i está compuesto, a su vez, por dos clusters C_{i_1} y C_{i_2}. Entonces se considera la siguiente distancia:

$$d(C_i, C_j) = \frac{d(C_{i_1}, C_j) + d(C_{i_2}, C_j)}{2}$$

- Estrategia de minimización de la distancia o similitud promedio ponderada.
Sean dos clusters C_i y C_j. Supongamos que C_i está compuesto, a su vez, por dos clusters C_{i_1} y C_{i_2}. Supongamos que C_{i_1} tiene n_{i_1} elementos, C_{i_2} tiene n_{i_2} y C_j tiene n_j elementos. Entonces se considera la siguiente distancia:

$$d(C_i, C_j) = \frac{1}{(n_{i_1} + n_{i_2}) \cdot n_j} \sum_{i=1}^{n_{i_1}+n_{i_2}} \sum_{j=1}^{n_j} d(x_i, x_j)$$

En donde x_i denota un elemento arbitrario de C_i y x_j denota un elemento arbitrario de C_j.
Entonces se verifica que:

$$d(C_i, C_j) = \frac{n_{i_1} \cdot d(C_{i_1}, C_j) + n_{i_2} \cdot d(C_{i_2}, C_j)}{n_{i_1} + n_{i_2}}$$

Amalgamiento completo (complete linkage)

Se parte de una población de 7 individuos. Clusters iniciales $\Delta_0 = \{\{a\}, \{b\}, \{c\}, \{d\}, \{e\}, \{f\}, \{g\}\}$
Matriz de distancias iniciales de la población es:

	a	b	c	d	e	f	g
a	0						
b	2.15	0					
c	0.7	1.53	0				
d	1.07	1.14	0.43	0			
e	0.85	1.38	**0.21**	0.29	0		
f	1.16	1.01	0.55	0.22	0.41	0	
g	1.56	2.83	1.86	2.04	2.02	2.05	0

- Nivel $K = 1$: clusters más próximos $\{c\}$ y $\{e\}$ (**0.21**)
  Entonces $\Delta_1 = \{\{a\}, \{b\}, \{\mathbf{c}, \mathbf{e}\}, \{d\}, \{f\}, \{g\}\}$
  Matriz de distancias actualizada:

	a	b	$\{c, e\}$	d	f	g
a	0					
b	2.15	0				
$\{c, e\}$	0.85	1.53	0			
d	1.07	1.14	0.43	0		
f	1.16	1.01	0.55	**0.22**	0.41	0
g	1.56	2.83	2.02	2.04	2.05	0

- Nivel $K = 2$: clusters más próximos $\{d\}$ y $\{f\}$ (**0.22**)
  Entonces $\Delta_2 = \{\{a\}, \{b\}, \{c, e\}, \{\mathbf{d}, \mathbf{f}\}, \{g\}\}$
  Matriz de distancias actualizada:

	a	b	$\{c, e\}$	$\{d, f\}$	g
a	0				
b	2.15	0			
$\{c, e\}$	0.85	1.53	0		
$\{d, f\}$	1.16	1.14	**0.55**	0	
g	1.56	2.83	2.02	2.05	0

- Nivel $K = 3$: clusters más próximos $\{c, e\}$ y $\{d, f\}$ (**0.55**)
  Entonces $\Delta_3 = \{\{a\}, \{b\}, \{\{\mathbf{c}, \mathbf{e}\}, \{\mathbf{d}, \mathbf{f}\}\}, \{g\}\}$
  Matriz de distancias actualizada:

	a	b	$\{\{c, e\},\{d, f\}\}$	g
a	0			
b	2.15	0		
$\{\{c, e\},\{d, f\}\}$	**1.16**	1.53	0	
g	1.56	2.83	2.05	0

- Nivel $K = 4$: clusters más próximos $\{a\}$ y $\{\{c, e\}, \{d, f\}\}$ (**1.16**)
  Entonces $\Delta_4 = \{\{\{\mathbf{a}\}, \{\{\mathbf{c}, \mathbf{e}\}, \{\mathbf{d}, \mathbf{f}\}\}\}, \{b\}, \{g\}\}$
  Matriz de distancias actualizada:

	$\{\{a\},\{\{c, e\}, \{d, f\}\}\}$	b	g
$\{\{a\},\{\{c, e\}, \{d, f\}\}$	0		
b	2.15	0	
g	**2.05**	2.83	0

- Nivel $K = 5$: clusters más próximos $\{\{a\}, \{\{c, e\}, \{d, f\}\}\}$ y $\{g\}$ (**2.05**)
  Entonces $\Delta_5 = \{\{\{\{\mathbf{a}\}, \{\{\mathbf{c}, \mathbf{e}\}, \{\mathbf{d}, \mathbf{f}\}\}\}, \{\mathbf{g}\}\}, \{b\}\}$
  Matriz de distancias actualizada:

	$\{\{\{a\}, \{\{c, e\}, \{d, f\}\}\},\{g\}\}$	b
$\{\{\{a\}, \{\{c, e\}, \{d, f\}\}\},\{g\}\}$	0	
b	**2.83**	0

- Nivel $K = 6$: clusters más próximos $\{\{\{a\}, \{\{c, e\}, \{d, f\}\}\}\}, \{g\}\}$ y $\{b\}$ (**2.83**)
  Entonces $\Delta_6 = \{\{\{\{\mathbf{a}\}, \{\{\mathbf{c}, \mathbf{e}\}, \{\mathbf{d}, \mathbf{f}\}\}\}, \{\mathbf{g}\}, \{\mathbf{b}\}\}$

Representación gráfica mediante un árbol de clasificación: **dendograma**.

Modelización de un clustering jerárquico aglomerativo

Se ha realizado una modelización computacional basada en sistemas P de un clustering jerárquico aglomerativo[1]

- Población de n individuos.
- Se estudian p características de cada individuo.
- Las características están codificadas a través de valores booleanos.
- La medida de similaridad usada es debida a Sokal y Michener (1963):

$$s(x_i, x_j) = \frac{1}{p} \cdot \sum_{r=1}^{p} (1 - |x_{i,r} - x_{j,r}|)$$

[1] M. Cardona, M.A. Colomer, M.J. Pérez-Jiménez. Hierarchical clustering with Membrane Computing. Computing and Informatics, **27**, 3+ (2008), 497-513

Clustering no jerárquico: agrupamiento por particiones

El **método de de las k-means**, MacQueen 1972.

- ⋆ Método basado en división o agrupación por particiones.
- ⋆ El método proporciona una clasificación de los datos en k clusters, siendo k un número prefijado.
- ⋆ El método trata de optimizar la función objetivo:

$$SSE = \sum_{i=1}^{k} \sum_{x \in C_i} d^2(x, \mu_i)$$

en donde C_i es un cluster y μ_i es el centroide correspondiente a dicho cluster.

El **centroide de un cluster** es un elemento que minimiza la suma de las distancias (similitudes) al resto de los elementos del cluster.

Clustering no jerárquico: agrupamiento por particiones

Un **algoritmo** que implementa el método de las k-means:

Entrada: Datos a clasificar (n), número de clusters k a formar, matriz de distancias (similitudes) y el número de iteraciones IT a realizar.

1. Seleccionar una agrupación de los datos en k clusters.
2. Determinar los centroides de cada cluster.
3. $j \leftarrow 1$
4. Crear k nuevos clusters asignando cada dato al centroide más cercano.
5. Sustituir cada centroide de un nuevo cluster por el elemento que minimiza la suma de distancias al resto de datos del cluster.
6. Si $j \leq IT$, entonces volver al paso 2.
7. Si no, finalizar.

El proceso anterior puede ser convergente (en el sentido de que en un determinado paso, los nuevos clusters coinciden con los que ya se tenían) o no.

Complejidad: $O(n \cdot p \cdot k \cdot IT)$ (si no se fija el valor de k, se trata de un problema de la clase **NP**).

Agrupamiento por particiones

Ventajas:

- Es relativamente eficiente.
- Generalmente el proceso finaliza con un óptimo local y en un número reducido de iteraciones.

Desventajas:

- El resultado depende de la selección inicial; realizar ejecuciones correspondientes a diferentes selecciones.
- Hay que especificar de antemano el valor de k: usar un método jerárquico sobre una muestra de los datos para estimar k (por ejemplo, el número de clusters de una matriz de expresión génica no suele conocerse a priori).
- Es aplicable cuando está definida la media: se pueden usar otras medidas de centralización.
- Dificultad con datos no numéricos.
- No funciona bien cuando los clusters son de diferente tamaño, distinta densidad o no convexos (en su distribución espacial).
- Es muy sensible a los datos "anómalos" ya que distorsionan las medias.
- No es capaz de tratar con "ruido"

Las variantes se diferencian en:

- Selección de las k medias iniciales.
- Cálculo de las similitudes.
- Estrategias para calcular las medias.

En los datos categóricos se sustituirá la media por la moda.

Variantes de k-means

- **GRASP** (Greedy Randomized Adaptive Search Procedure) para evitar óptimos locales.
- **k-Modes**, Huang 1998: utiliza modas en vez de medias a fin de poder trabajar con atributos de tipo categórico.
- **k-Medoids**: utiliza medianas en vez de medias para limitar la influencia de los outliers.
  - ⋆ **PAM**: Partitioning Around Medoids, 1987.
  - ⋆ **CLARA**: Clustering LARge Applications, 1990.
  - ⋆ **CLARANS**: CLARA + Randomized Search, 1994.

Ejemplo (I)

Partimos de una población $\Omega = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7\}$ con 7 individuos y de cada uno de ellos analizamos dos propiedades o características (codificadas numericamente).
Usaremos como medidad de similitud la distancia euclídea.

	$x_{i,1}$	$x_{i,2}$
x_1	1	5
x_2	2	10
x_3	5	5
x_4	8	10
x_5	10	1
x_6	12	2
x_7	6	4

Apliquemos el método de las k-means para formar 3 clusters con un máximo de iteraciones $IT = 5$.
Se parte de la siguiente agrupación inicial: $\Delta_0 = \{\{\mathbf{x_1}, \mathbf{x_4}\}, \{\mathbf{x_2}, \mathbf{x_6}\}, \{\mathbf{x_3}, \mathbf{x_5}, \mathbf{x_7}\}\}$

Cluster 1: $\{\mathbf{x_1}, \mathbf{x_4}\}$; Cluster 2: $\{\mathbf{x_2}, \mathbf{x_6}\}$ y Cluster 3: $\{\mathbf{x_3}, \mathbf{x_5}, \mathbf{x_7}\}$

Hallemos los centroides iniciales:

- ⋆ Correspondiente al cluster 1: punto medio de x_1 y x_4; es decir, $\mathbf{C_1^0 = (4.5, 7.5)}$.
- ⋆ Correspondiente al cluster 2: punto medio de x_2 y x_6; es decir, $\mathbf{C_2^0 = (7, 6)}$.
- ⋆ Correspondiente al cluster 3: baricentro de x_3, x_5 y x_7; es decir, $\mathbf{C_3^0 = (7, 3.33)}$.

Ejemplo (II)

A continuación hallamos la nueva matriz de distancia de los individuos a sus centroides respectivos:

	C_1^0	C_2^0	C_3^0
x_1	18.5	37	38.89
x_2	12.5	41	69.89
x_3	6.5	5	6.89
x_4	18.5	25	45.89
x_5	72.5	34	14.29
x_6	86.5	41	26.69
x_7	14.5	5	1.49

La nueva agrupación obtenida es la siguiente: $\Delta_1 = \{\{\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_4}\}, \{\mathbf{x_3}\}, \{\mathbf{x_5}, \mathbf{x_6}, \mathbf{x_7}\}\}$

Cluster 1: $\{\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_4}\}$; Cluster 2: $\{\mathbf{x_3}\}$ y Cluster 3: $\{\mathbf{x_5}, \mathbf{x_6}, \mathbf{x_7}\}$

Hallemos los **nuevos centroides**:

- ⋆ Correspondiente al cluster 1: baricentro de x_1, x_2 y x_4; es decir, $\mathbf{C_1^1 = (3.66, 8.33)}$.
- ⋆ Correspondiente al cluster 2: el único dato que existe x_3; es decir, $\mathbf{C_2^1 = (5, 5)}$.
- ⋆ Correspondiente al cluster 3: baricentro de x_3, x_5 y x_7; es decir, $\mathbf{C_3^1 = (9.33, 2.33)}$.

Ejemplo (III)

A continuación hallamos la nueva matriz de distancia de los individuos a sus centroides respectivos:

	C_1^1	C_2^1	C_3^1
x_1	18.16	16	76.51
x_2	5.54	34	112.55
x_3	12.88	0	25.87
x_4	21.62	34	60.59
x_5	93.22	41	2.21
x_6	109.62	58	7.23
x_7	24.22	2	13.87

La nueva agrupación obtenida es la siguiente: $\Delta_2 = \{\{\mathbf{x_2}, \mathbf{x_4}\}, \{\mathbf{x_1}, \mathbf{x_3}, \mathbf{x_7}\}, \{\mathbf{x_5}, \mathbf{x_6}\}\}$

Cluster 1: $\{\mathbf{x_2}, \mathbf{x_4}\}$; Cluster 2: $\{\mathbf{x_1}, \mathbf{x_3}, \mathbf{x_7}\}$ y Cluster 3: $\{\mathbf{x_5}, \mathbf{x_6}\}$

Hallemos los **nuevos centroides**:

- Correspondiente al cluster 1: punto medio de x_2 y x_4; es decir, $\mathbf{C_1^2 = (5, 10)}$.
- Correspondiente al cluster 2: baricentro de x_1, x_3 y x_7; es decir, $\mathbf{C_2^2 = (4, 4.66)}$.
- Correspondiente al cluster 3: punto medio de x_5 y x_6; es decir, $\mathbf{C_3^2 = (11, 1.5)}$.

Ejemplo (IV)

A continuación hallamos la nueva matriz de distancia de los individuos a sus centroides respectivos:

	C_1^2	C_2^2	C_3^2
x_1	41	9.11	112.25
x_2	9	32.51	153.25
x_3	25	1.11	48.25
x_4	9	44.51	81.25
x_5	106	49.39	1.25
x_6	113	71.07	1.25
x_7	37	4.44	31.25

La nueva agrupación obtenida es la siguiente: $\Delta_3 = \{\{x_2, x_4\}, \{x_1, x_3, x_7\}, \{x_5, x_6\}\}$

Cluster 1: $\{x_2, x_4\}$; Cluster 2: $\{x_1, x_3, x_7\}$ y Cluster 3: $\{x_5, x_6\}$

Por tanto, en este caso el **método** es **convergente**.

En consecuencia, no hace falta llegar al número de iteraciones prefijadas ($IT = 5$).

Medida de calidad de un proceso de clustering: Evaluación

Puntos claves del proceso de clustering:

- Elección de la medida de similitud o distancia.
- Elección del algoritmo de clustering.
- Elección del número de clustering.

Fijar un criterio para medir la calidad del procesos.

- Objetivo: a partir de una base de datos inicial (objetos/individuos), formar agrupaciones o clusters de acuerdo con una medidad de similitud entre ellos de tal manera que:
  - ⋆ La similitud media entre los datos del mismo cluster sea "alta" (*similitud intra-cluster*).
  - ⋆ La similitud media entre los datos de distintos clusters sea "baja" (*similitud inter-clusters*).

Así pues, hay que:

- Minimizar la distancia intra-cluster (**cohesión**).
- Maximizar la distancia inter-cluster (**separación**).