Métodos de clusters basados en particiones mediante reasignación.

La idea central de la mayoría de estos procedimientos es elegir alguna partición inicial de individuos y después intercambiar los miembros de estos clusters para obtener una partición mejor. Los diversos algoritmos existentes se diferencian sobre todo en lo que se entiende por una partición mejor y en los métodos que deben usarse para conseguir mejoras. Así, los métodos estudiados ahora comienzan con una partición inicial de los individuos en grupos o bien con un conjunto de puntos iniciales (denominados puntos semillas) sobre los cuales pueden formarse los clusters.

Sin embargo, también veremos algoritmos que no tienen este tipo de restricciones.

El problema de la partición inicial

Suponiendo que queremos establecer una división de las observaciones en K clusters, la primera partición puede determinarse eligiendo los denominados puntos semillas. Estos puntos semillas pueden determinarse según diferentes criterios. Entre ellos:

- Los K primeros puntos u observaciones
- K individuos elegidos de manera aleatoria
- K individuos elegidos de forma regularmente esparcidos, siguiendo algún tipo de progresión.
- Elegir los k-individuos mediante densidad de observaciones (Algoritmo de Ashtran; Algoritmo de Ball & Hall)

A continuación, el resto de los elementos (n-K) se asignan a su punto semilla más cercano, formando así los K primeros clusters. A partir de aquí, los algoritmos irán reasignando los puntos a los diferentes clusters atendiendo a algún criterio hasta que llegue a un punto donde la reasignación posible no mejore al estado anterior (convergencia).

Veremos los más usados: Método de Forgy y método de k-medias. También se verán otros métodos donde no es necesario establecer una partición inicial y determinar el número de clusters de antemano.

Algoritmos de reasignación

Admiten la "reasignación" de un individuo y, por tanto, requieren una partición inicial. Esto es, una vez considerado un individuo como miembro de un cluster, en un siguiente paso del análisis, puede salirse de él e integrarse en otro si de esta forma se mejora (optimiza) la partición. Esta posibilidad permite la sucesiva mejora de la partición inicial. Por lo general, estos métodos asumen a priori un número de clusters a formar. Son llamados así porque pretenden obtener la partición que optimice una cierta medida numérica definida. Los distintos métodos de optimización se diferencian entre sí en la manera de obtener la partición inicial y en la medida a optimizar en el proceso.

Método de Forgy

Forgy, sugiere un algoritmo simple consistente en la siguiente secuencia de pasos:

1. Comenzar con una partición inicial. Ir al paso segundo si se comienza con un conjunto de puntos semilla. Ir al paso tercero si se comienza con una partición de los casos.

2. Colocar cada individuo en el cluster con la semilla más próxima. Las semillas permanecen fijas para cada ciclo completo que recorra el conjunto de datos.

3. Calcular los nuevos puntos semilla como los centroides de los clusters.

4. Alternar los pasos segundo y tercero hasta que el proceso converja, o sea, continuar hasta que ningún individuo cambie de cluster en el paso segundo.

Es importante que, en este método, los centroides resultantes que se usan para calcular la distancia se establecen tras cada ciclo completo de reasignaciones.

Método de K-medias

Es quizás el más popular dentro de los algoritmos de reasignación.

1. Comenzar con una partición inicial. Ir al paso segundo si se comienza con un conjunto de puntos semilla. Ir al paso tercero si se comienza con una partición de los casos.

2. Colocar cada individuo en el cluster con la semilla más próxima de forma sucesiva y recalculando los centroides de cada cluster tras cada asignación individual.

3. Repetir el paso segundo hasta que el proceso converja, o sea, continuar hasta que ningún individuo cambie de cluster en el paso segundo.

Es importante que, en este método, los centroides resultantes que se usan para calcular la distancia se establecen tras cada reasignación.

Anexo: Algoritmo de Ashtran

Astrahan propuso el siguiente algoritmo para elegir puntos semilla:

Para cada individuo se calcula la *densidad*, entendiendo por tal el número de casos que distan de él una cierta distancia, digamos d_1.

Ordenar los casos por densidades y elegir aquel que tenga la mayor densidad como primer punto semilla.

Elegir de forma sucesiva los puntos semilla en orden de densidad decreciente sujeto a que cada nueva semilla tenga al menos una distancia mínima, d_2, con los otros puntos elegidos anteriormente.

Continuar eligiendo semillas hasta que todos los casos que faltan tengan densidad cero, o sea, hay al menos una distancia d_1 de cada punto a otro.

En el caso de que, por este procedimiento, se produjera un exceso de puntos generados, se agruparán de forma jerárquica hasta que haya exactamente *K*. Por ejemplo, el método del centroide puede ser elegido para tal cuestión.

Se consideran 7 individuos que toman valores respecto de 2 variables. Para el cálculo de la matriz de distancias utilizamos el cuadrado de la distancia euclídea.

	X_1	X_2
I_1	1	5
I_2	2	10
I_3	5	5
I_4	8	10
I_5	10	1
I_6	12	2
I_7	6	4

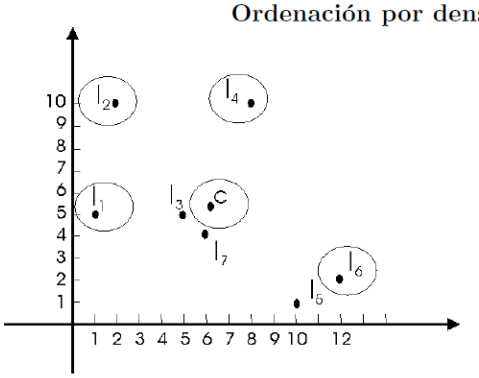
Distancia euclídea $\Longrightarrow$

	I_1	I_2	I_3	I_4	I_5	I_6	I_7
I_1	0						
I_2	26	0					
I_3	16	34	0				
I_4	74	36	34	0			
I_5	97	145	41	85	0		
I_6	130	164	58	80	5	0	
I_7	26	50	2	40	25	40	0

Densidades

- $d_1 = 60$ $\quad d(I_1) = 3 \quad d(I_2) = 2 \quad d(I_3) = 0 \quad d(I_4) = 3 \quad d(I_5) = 3 \quad d(I_6) = 3 \quad d(I_7) = 0$

Ordenación por densidades $(I_1 I_4 I_5 I_6)(I_2)(I_3 I_7)$



$1^{\underline{o}}$ punto semilla $\quad I_1 \quad d_2 = 40$

$2^{\underline{o}}$ punto semilla $\quad I_4 \quad d_3 = 30$

$3^{\underline{o}}$ punto semilla $\quad I_5 \quad d_4 = 15$

$4^{\underline{o}}$ punto semilla $\quad I_2 \quad d_5 = 10$

$5^{\underline{o}}$ punto semilla $\quad I_3 \quad d_6 = 5$

Anexo: Algoritmo de Ball & Hall

Tomar el vector de medias de los datos como el primer punto semilla; posteriormente se seleccionan los puntos semilla examinando los individuos sucesivamente, aceptando uno de ellos como siguiente punto semilla siempre y cuando esté, por lo menos, a alguna distancia, *d*, de todos los puntos elegidos anteriormente Se continúa de esta forma hasta completar los *k* puntos deseados o el conjunto de datos se agota.

	X_1	X_2
I_1	1	5
I_2	2	10
I_3	5	5
I_4	8	10
I_5	10	1
I_6	12	2
I_7	6	4

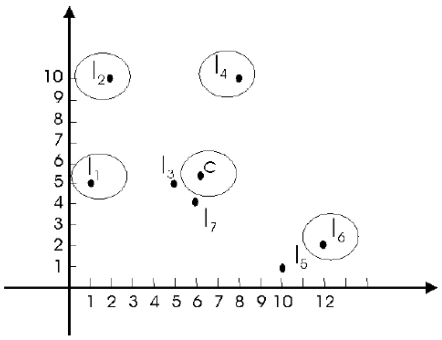
Distancia euclídea $\Longrightarrow$

	I_1	I_2	I_3	I_4	I_5	I_6	I_7
I_1	0						
I_2	26	0					
I_3	16	34	0				
I_4	74	36	34	0			
I_5	97	145	41	85	0		
I_6	130	164	58	80	5	0	
I_7	26	50	2	40	25	40	0

•$C(\bar{x_1}, \bar{x_2}) = (6{,}28, 5{,}28)$
1º punto semilla

$$\begin{cases} d(I_1, C) = 27{,}95 & d(I_2, C) = 40{,}59 & d(I_3, C) = 1{,}71 \\ d(I_4, C) = 23{,}99 & d(I_5, C) = 32{,}15 & d(I_6, C) = 44{,}16 \\ d(I_7, C) = 1{,}71 \end{cases}$$

Ordenación por distancias a C $(I_6 I_2 I_5 I_1)(I_4)(I_3 I_7)$



1º punto semilla C $d_1 = 40$

2º punto semilla I_6

3º punto semilla I_2 $d_2 = 30$

4º punto semilla I_1 $d_3 = 15$

5º punto semilla I_4 $d_4 = 5$