Bloque 2
Aprendizaje Automático

Tema 4:
Aprendizaje No Supervisado:

Clustering. Algoritmo K-medias.

Tema 4 – Clustering. Algoritmo K-medias.

Bibliografía

- R.O. Duda, D.G. Stork, P.E. Hart. Pattern Classification. Wiley 2001.
- A.R. Webb, K.D. Copsey. Stastistical Pattern Recognition, Wiley, tercera edición, 2011.

1. Introducción

En este tema vamos a estudiar el aprendizaje NO SUPERVISADO o clustering, que utiliza muestras no etiquetadas.

Consiste en definir un agrupamiento sobre objetos no etiquetados.

Objetivo del clustering [Anderberg, 1973]:

> Agrupar los objetos en clases, de modo que los objetos pertenecientes a una misma clase tengan un alto grado de **asociación natural** entre sí, mientras que las clases sean relativamente distintas unas de otras.

Se trata, por tanto:

- encontrar *agrupamientos naturales* de un conjunto de objetos
- estos agrupamientos pueden describirse como grupos con fuerte semejanza interna (clases)
- los objetos dentro de un grupo (clase) están fuertemente relacionados entre sí
- los objetos en grupos (clases) diferentes son muy distintos.

Dos tipos de clustering: particional y jerárquico

2. Agrupamientos particionales

Problema general

Dado un conjunto de **N** objetos o muestras (x_1, x_2, …, x_n) donde cada objeto es un vector de D dimensiones, el objetivo es encontrar una partición Π de los **N** objetos en **C** clases o clusters, es decir:

$\Pi = \{X_1, \ldots, X_C\}$ donde $X_1, \ldots, X_C$ representan los **C** clústeres

2. Agrupamientos particionales

N=5 muestras $(x_1, x_2, x_3, x_4, x_5)$

Si queremos agrupar las N muestras en **C=1 clúster**

$\Pi = \{X_1\}$

$x_1\ x_2\ x_3\ x_4\ x_5$

una única partición

Si queremos agrupar las N muestras en **C=2 clusters**

$\Pi = \{X_1, X_2\}$

x_1 | $x_2\ x_3\ x_4\ x_5$

x_2 | $x_1\ x_3\ x_4\ x_5$

x_3 | $x_1\ x_2\ x_4\ x_5$

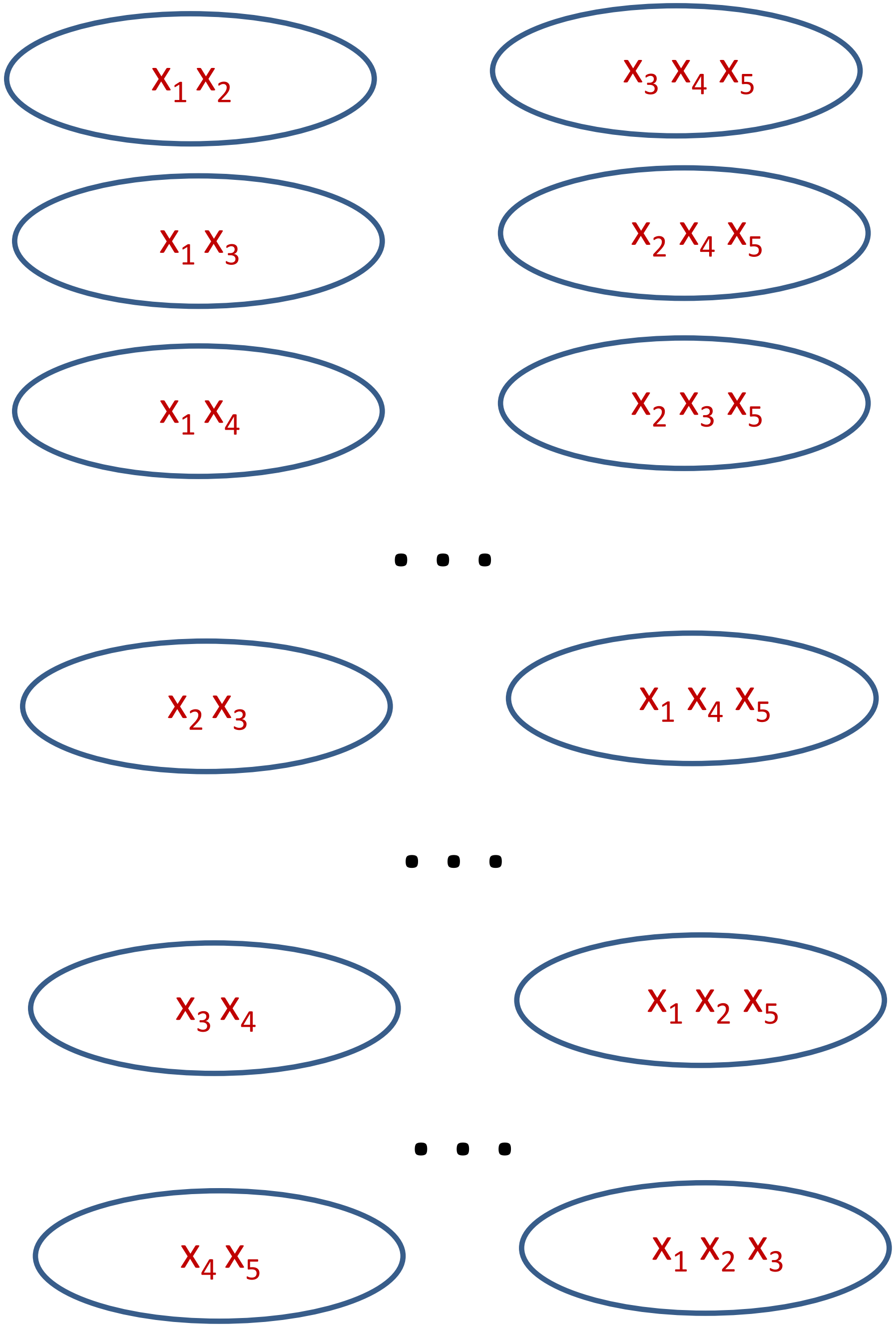
muchas particiones y más …

.

.

2. Agrupamientos particionales

muchas particiones diferentes

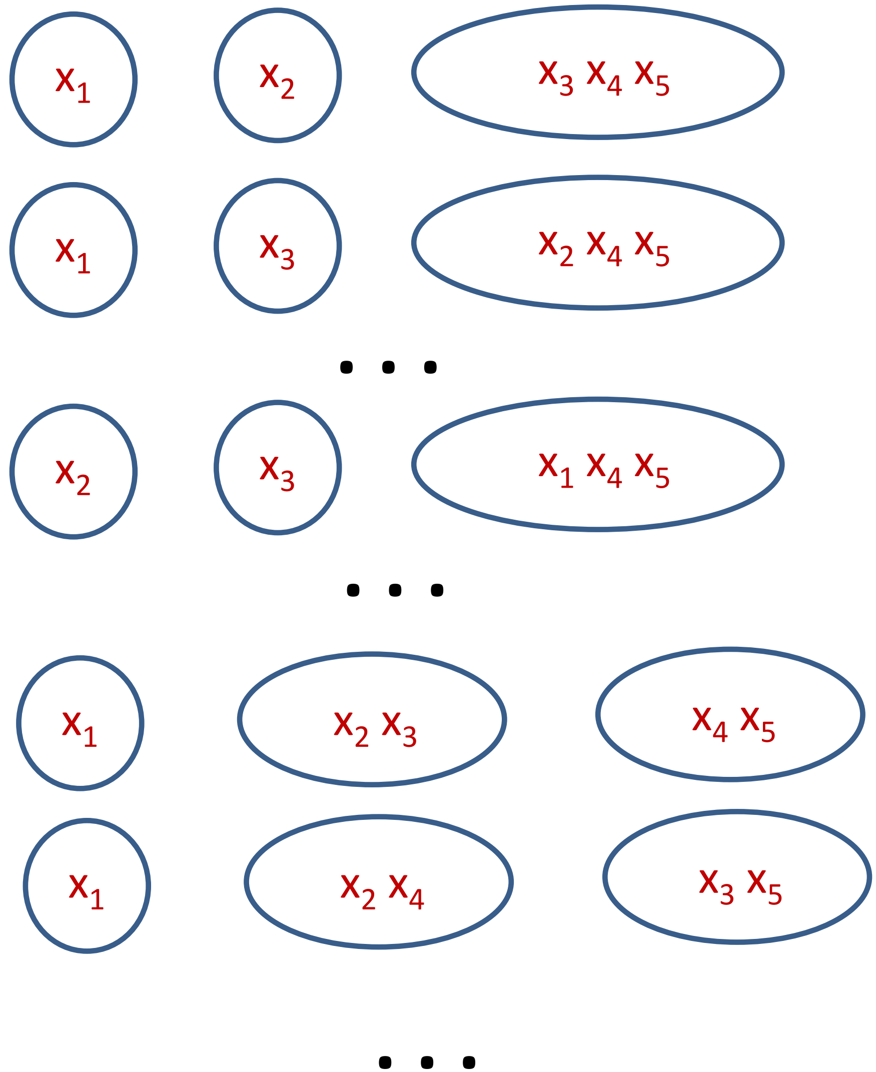


2. Agrupamientos particionales

Queremos agrupar las N muestras en **C=3 clusters**

$\Pi = \{X_1, X_2, X_3\}$

muchas particiones diferentes

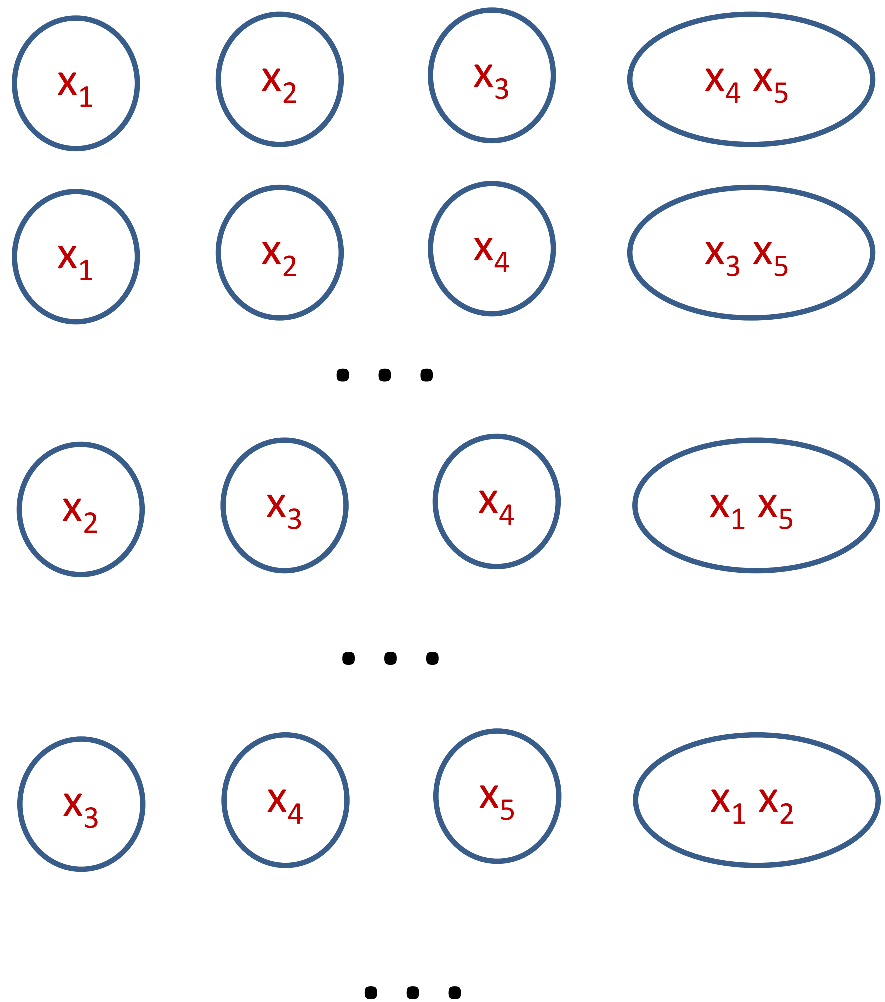


2. Agrupamientos particionales

Queremos agrupar las N muestras en **C=4 clusters**

$\Pi=\{X_1, X_2, X_3, X_4\}$

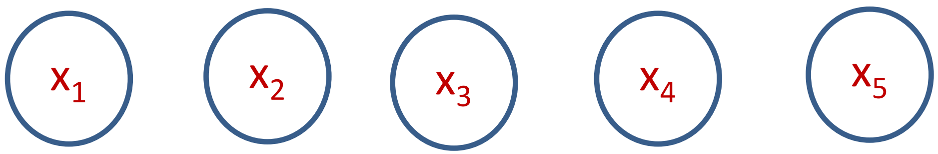
muchas particiones diferentes



2. Agrupamientos particionales

Queremos agrupar las N muestras en **C=5 clusters**

$\Pi = \{X_1 , X_2 , X_3 , X_4 , X_5 \}$

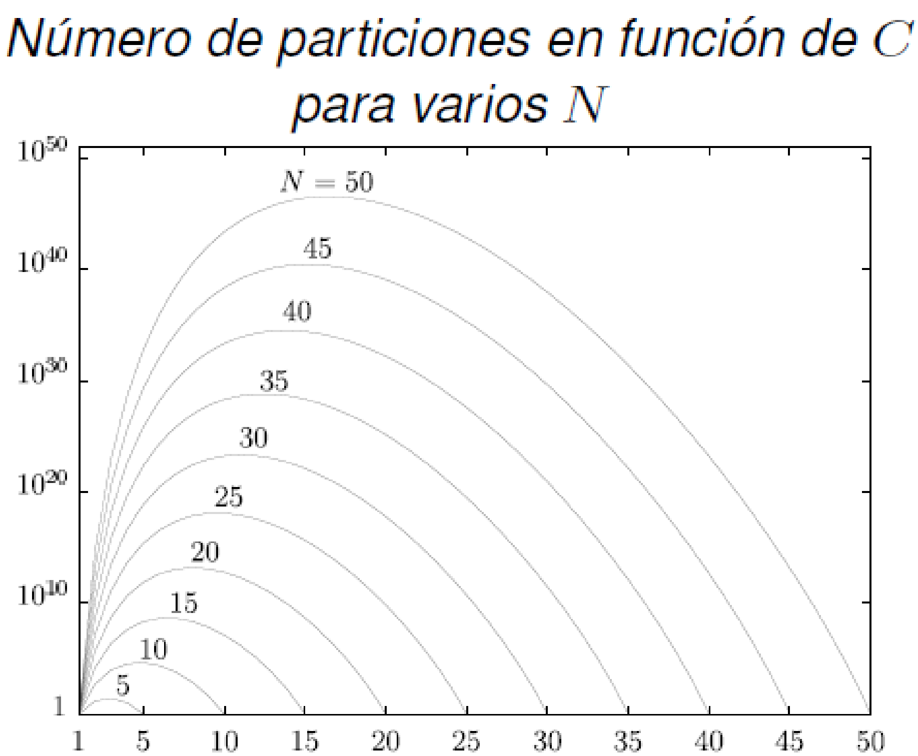


una única partición

2. Agrupamientos particionales

Dificultad

El número de particiones a evaluar es muy elevado, incluso para valores pequeños de N (muestras) y C (clusters). Buscar una solución óptima con técnicas de exploración exhaustiva (explícita o implícita) no es abordable excepto para algunos pocos casos particulares.



Solución

Buscar soluciones sub-óptimas mediante algoritmos de aproximación.

3. Suma de errores cuadráticos

Necesitamos un criterio para seleccionar la mejor partición Π de **N** objetos en **C** clusters.

Función criterio habitual: **suma de errores cuadráticos** (**SEC**)

Dada una partición de **N** datos en **C** clusters, $\Pi = \{X_1, \ldots, X_C\}$, su SEC (o función J) es:

$$J(X_1, \ldots, X_C) = \sum_c J_c$$

El SEC de una partición $\{X_1, \ldots, X_C\}$ es la suma de los SEC para cada clúster

$$\Pi^* = \underset{\Pi=\{X_1,\ldots,X_C\}}{\arg\min} J(\Pi)$$

La mejor partición, Π^*, es la que minimiza el SEC entre todas las posibles particiones

3. Suma de errores cuadráticos

La SEC de una partición Π de **N** objetos en **C** clusters, Π **= {X$_1$, …, X$_C$}**, se calcula del siguiente modo:

Paso 1

Cada clúster X_c contiene un subconjunto de los N objetos (conjuntos disjuntos). Para cada clúster X_c se calcula el error cuadrático para todas las muestras que pertenecen al cluster. Para ello, necesitamos calcular la media, m_c , del clúster X_c.

$$m_c = \frac{1}{|X_c|} \sum_{x \in X_c} x$$

m$_c$, media del clúster (centroide)
se interpreta como el prototipo natural de X_c

3. Suma de errores cuadráticos

Paso 2

Un objeto **x** se interpreta como una versión distorsionada de $\mathbf{m_c}$ y dicha distorsión se modela como el vector error $\mathbf{x} - \mathbf{m_c}$.
$\| \mathbf{x} - \mathbf{m_c} \|$ (norma Euclídea, distancia Euclídea entre dos vectores)

$$\|\boldsymbol{x} - \boldsymbol{m}_c\|^2$$

Error cuadrático de una muestra **x**

$$J_c = \sum_{\boldsymbol{x} \in X_c} \|\boldsymbol{x} - \boldsymbol{m}_c\|^2$$

La SEC de un clúster X_c (J_c) se calcula como la **S**uma de los **E**rrores **C**uadráticos para todas las muestras pertenecientes al cluster.

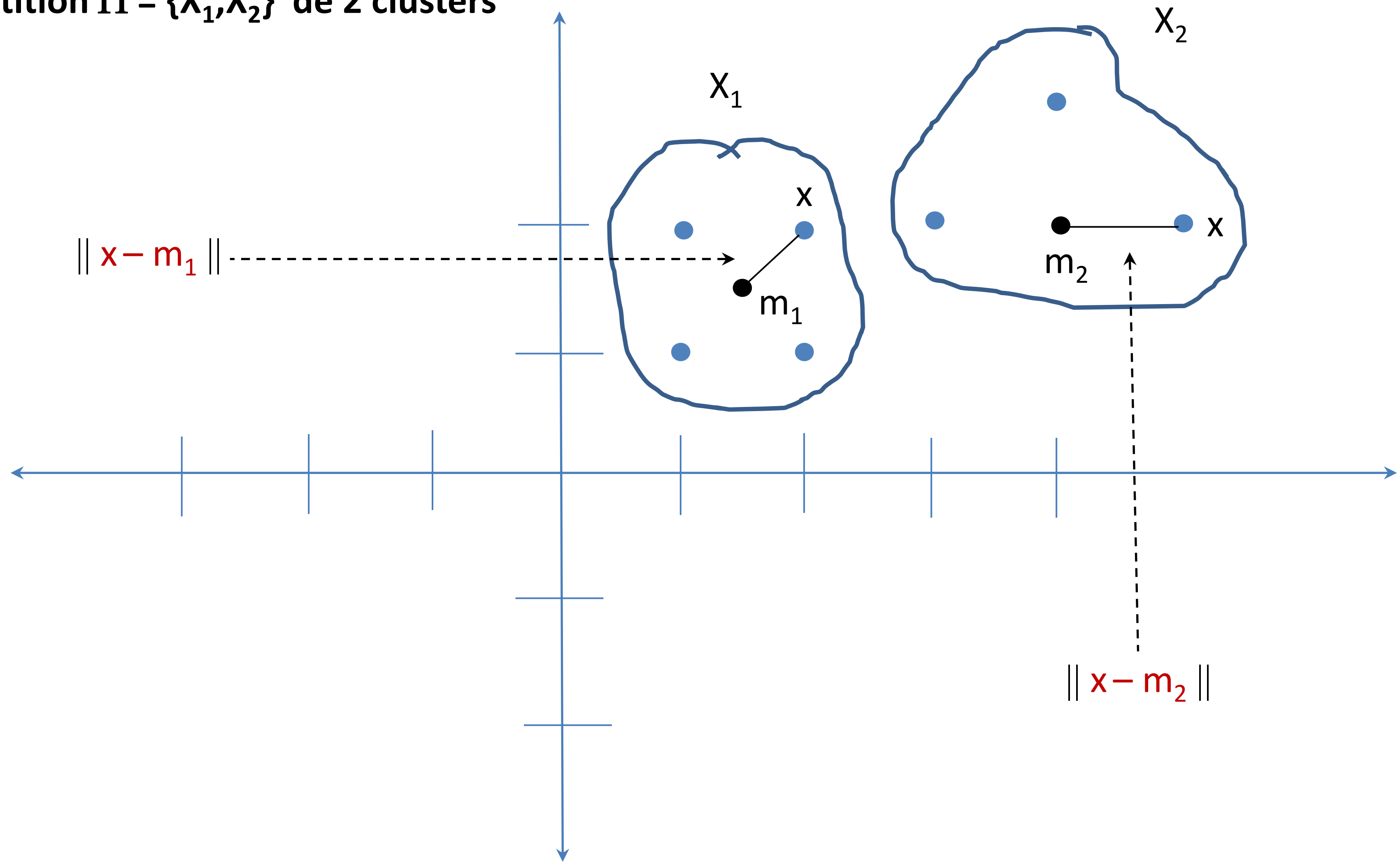
Paso 3

Finalmente, la SEC de Π = **{X_1 , …, X_C}** se calcula como

$$\textbf{SEC}\ (\Pi) = J(X_1, \ldots, X_C) = \sum_c J_c$$

3. Suma de errores cuadráticos

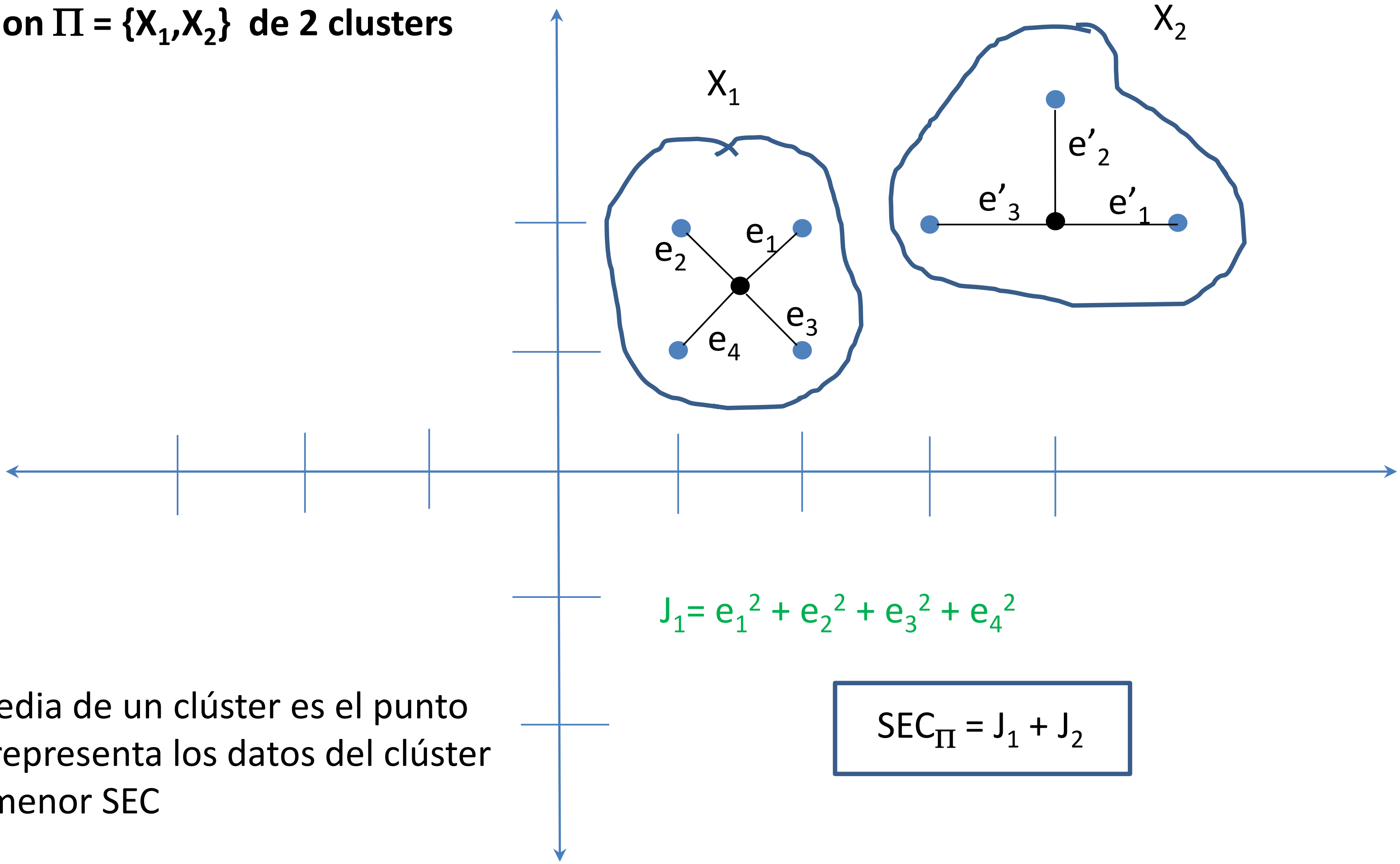
partition $\Pi = \{X_1, X_2\}$ de 2 clusters



3. Suma de errores cuadráticos

partition $\Pi = \{X_1, X_2\}$ de 2 clusters

$J_2 = e'^2_1 + e'^2_2 + e'^2_3$



$J_1 = e_1^2 + e_2^2 + e_3^2 + e_4^2$

$$\boxed{SEC_{\Pi} = J_1 + J_2}$$

La media de un clúster es el punto que representa los datos del clúster con menor SEC

3. Suma de errores cuadráticos

Para calcular la mejor partición, Π^* **,** tenemos que calcular la SEC de todas las particiones y quedarnos con la de menor SEC.

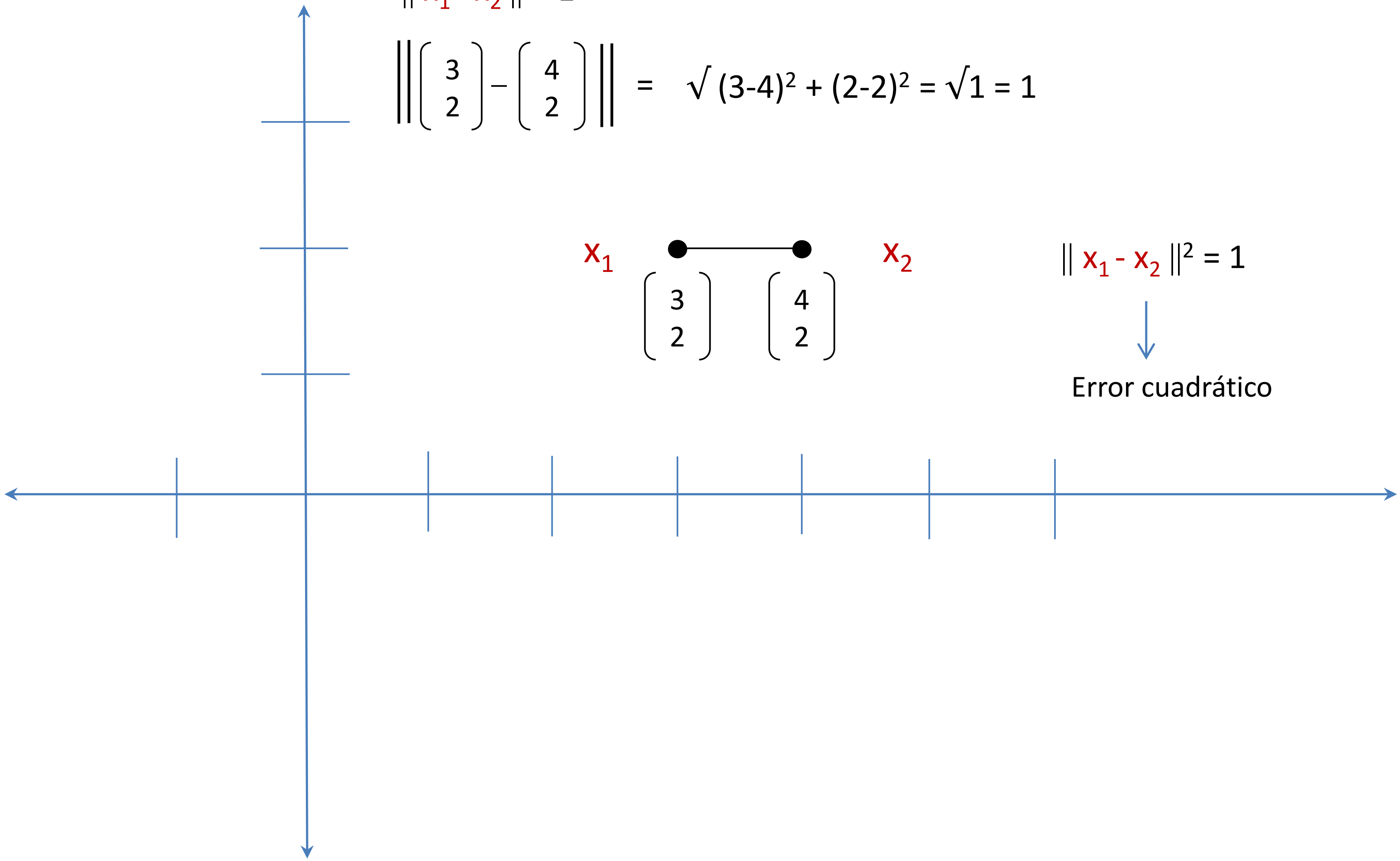
La SEC mide la suma (o media) de los cuadrados de las magnitudes de los vectores error y por tanto es un criterio a minimizar

La media de cada cluster, $\mathbf{m_c}$, es el punto que representa los datos del clúster con menor SEC.

3. Suma de errores cuadráticos (notas)

$\| x_1 - x_2 \| = 1 \longrightarrow$ Distancia Euclídea

$$\left\| \begin{pmatrix} 3 \\ 2 \end{pmatrix} - \begin{pmatrix} 4 \\ 2 \end{pmatrix} \right\| = \sqrt{(3-4)^2 + (2-2)^2} = \sqrt{1} = 1$$



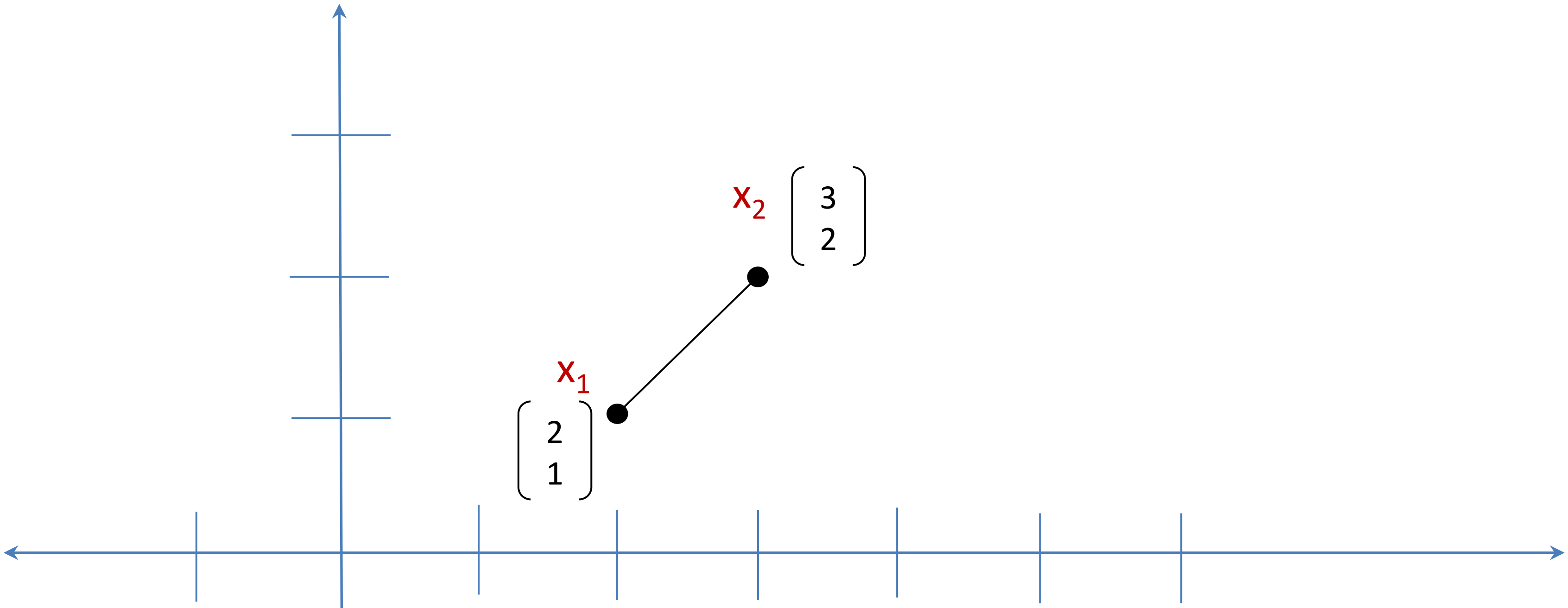
x_1 $\begin{pmatrix} 3 \\ 2 \end{pmatrix}$ x_2 $\begin{pmatrix} 4 \\ 2 \end{pmatrix}$

$\| x_1 - x_2 \|^2 = 1$

$\downarrow$

Error cuadrático

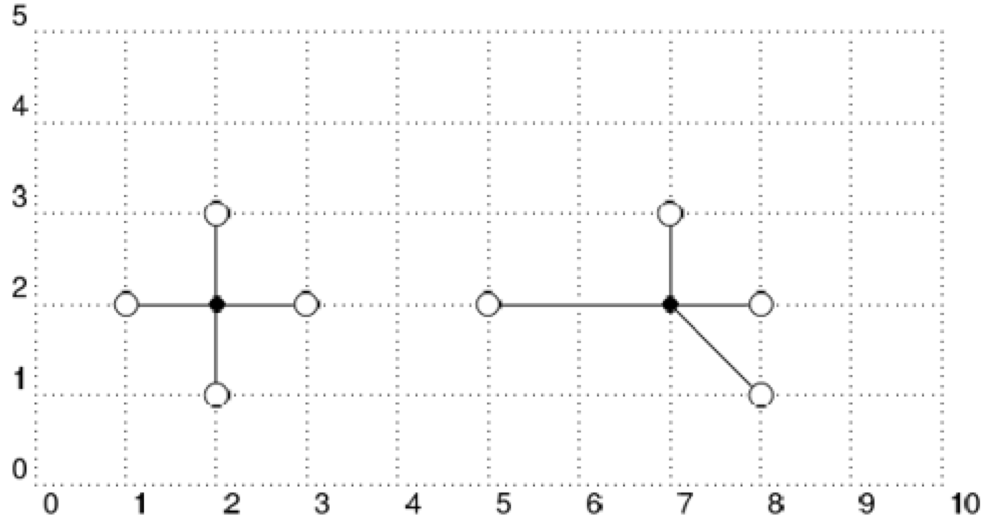
3. Suma de errores cuadráticos (notas)



$$\| x_1 - x_2 \| = \left\| \begin{pmatrix} 2 \\ 1 \end{pmatrix} - \begin{pmatrix} 3 \\ 2 \end{pmatrix} \right\| = \sqrt{(2-3)^2 + (1-2)^2} = \sqrt{2}$$

$$\| x_1 - x_2 \|^2 = 2$$

4. Ejemplo



Partición I (Π_1)

Partition I (Π_1): 8 samples $(x_1, \ldots, x_N)$, 2 clusters in partition $\Pi_1 = \{X_1, X_2\}$

$$X_1 = \{x_1, \ldots, x_4\} : x_1 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, x_2 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}, x_3 = \begin{pmatrix} 2 \\ 3 \end{pmatrix}, x_4 = \begin{pmatrix} 3 \\ 2 \end{pmatrix}$$

$$X_2 = \{x_5, \ldots, x_8\} : x_5 = \begin{pmatrix} 5 \\ 2 \end{pmatrix}, x_6 = \begin{pmatrix} 7 \\ 3 \end{pmatrix}, x_7 = \begin{pmatrix} 8 \\ 1 \end{pmatrix}, x_8 = \begin{pmatrix} 8 \\ 2 \end{pmatrix}$$

$$\boldsymbol{m}_1 = \frac{1}{4} \sum_{x \in X_1} x = \frac{1}{4} \begin{pmatrix} 8 \\ 8 \end{pmatrix} = \begin{pmatrix} 2 \\ 2 \end{pmatrix} \qquad \boldsymbol{m}_2 = \frac{1}{4} \sum_{x \in X_2} x = \frac{1}{4} \begin{pmatrix} 28 \\ 8 \end{pmatrix} = \begin{pmatrix} 7 \\ 2 \end{pmatrix}$$

4. Ejemplo

Cálculos para el clúster **X_1**

$$EC_{x1} = \|\boldsymbol{x1} - \boldsymbol{m}_1\|^2 = \left(\begin{pmatrix} 1 \\ 2 \end{pmatrix} - \begin{pmatrix} 2 \\ 2 \end{pmatrix}\right)^2 = \begin{pmatrix} -1 \\ 0 \end{pmatrix}^2 = 1$$

$$EC_{x2} = \|\boldsymbol{x2} - \boldsymbol{m}_1\|^2 = \left(\begin{pmatrix} 2 \\ 1 \end{pmatrix} - \begin{pmatrix} 2 \\ 2 \end{pmatrix}\right)^2 = \begin{pmatrix} 0 \\ -1 \end{pmatrix}^2 = 1$$

$$EC_{x3} = \|\boldsymbol{x3} - \boldsymbol{m}_1\|^2 = \ldots = 1$$

$$EC_{x4} = \|\boldsymbol{x4} - \boldsymbol{m}_1\|^2 = \ldots = 1$$

$$J_{X1} = \sum_{x \in X_1} \|\boldsymbol{x} - \boldsymbol{m}_1\|^2 = 4$$

4. Ejemplo

Cálculos para el clúster **X$_2$**

$$EC_{x5} = \|x5 - m_2\|^2 = \left(\begin{pmatrix} 5 \\ 2 \end{pmatrix} - \begin{pmatrix} 7 \\ 2 \end{pmatrix}\right)^2 = \begin{pmatrix} -2 \\ 0 \end{pmatrix}^2 = 4$$

$$EC_{x6} = \|x6 - m_2\|^2 = \left(\begin{pmatrix} 7 \\ 3 \end{pmatrix} - \begin{pmatrix} 7 \\ 2 \end{pmatrix}\right)^2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}^2 = 1$$

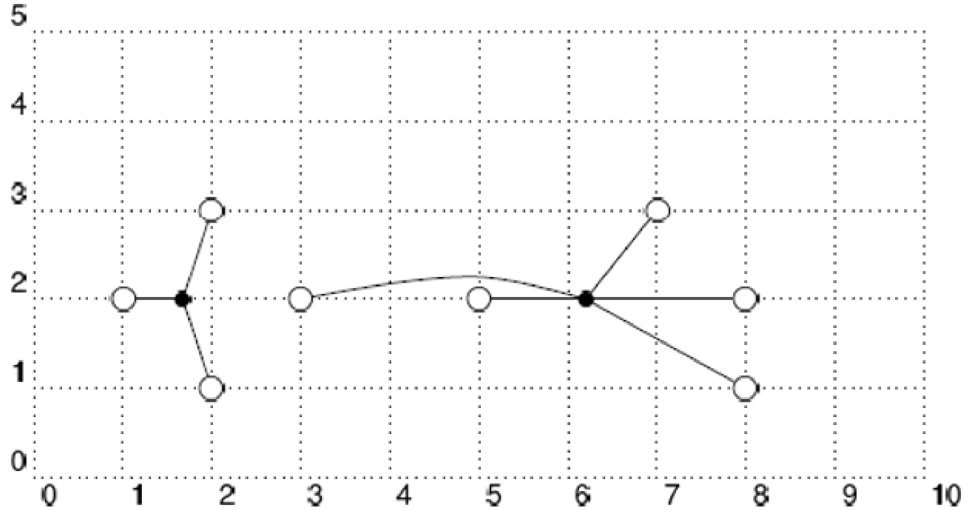
$$EC_{x7} = \|x7 - m_2\|^2 = \ldots = 2$$

$$EC_{x8} = \|x8 - m_2\|^2 = \ldots = 1$$

$$J_{X2} = \sum_{x \in X_2} \|x - m_2\|^2 = 8$$

$$\mathbf{J(\Pi_1) = SCE_{\Pi 1} = J(X_1, X_2) = 4+8 = 12}$$

4. Ejemplo



Partición II (Π_2)

Partition II (Π_2): 8 samples $(x_1, \ldots, x_N)$, 2 clusters in partition $\Pi_2 = \{X_1, X_2\}$

$$X_1 = \{x_1, \ldots, x_3\} : x_1 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, x_2 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}, x_3 = \begin{pmatrix} 2 \\ 3 \end{pmatrix}$$

$$X_2 = \{x_4, \ldots, x_8\} : x_4 = \begin{pmatrix} 3 \\ 2 \end{pmatrix}, x_5 = \begin{pmatrix} 5 \\ 2 \end{pmatrix}, x_6 = \begin{pmatrix} 7 \\ 3 \end{pmatrix}, x_7 = \begin{pmatrix} 8 \\ 1 \end{pmatrix}, x_8 = \begin{pmatrix} 8 \\ 2 \end{pmatrix}$$

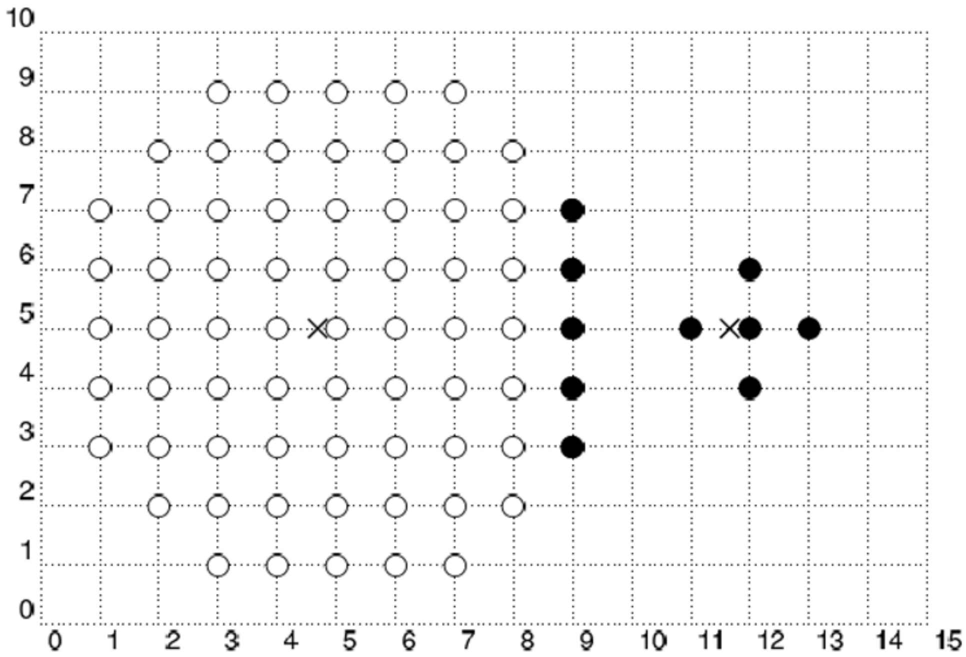
$$\boldsymbol{m}_1 = \frac{1}{3} \sum_{x \in X_1} x = \frac{1}{3} \begin{pmatrix} 5 \\ 6 \end{pmatrix} = \begin{pmatrix} 1.66 \\ 2 \end{pmatrix} \qquad \boldsymbol{m}_2 = \frac{1}{5} \sum_{x \in X_2} x = \frac{1}{5} \begin{pmatrix} 31 \\ 10 \end{pmatrix} = \begin{pmatrix} 6.2 \\ 2 \end{pmatrix}$$

Ejercicio: calcula la SEC de Π_2 y comprueba que se obtiene un valor mayor que 12.

Bondad del criterio SEC

SEC es apropiado solo si los datos forman **clústeres hiperesféricos de tamaño similar**.

Si el tamaño de los clústeres es muy distinto es posible que la agrupación natural no tenga el mínimo SEC y por tanto no se encontrarían clústeres naturales.



Si los clústeres tienen forma alargada u otras formas que no sean hiperesféricas, el criterio SEC podría no ser suficiente.

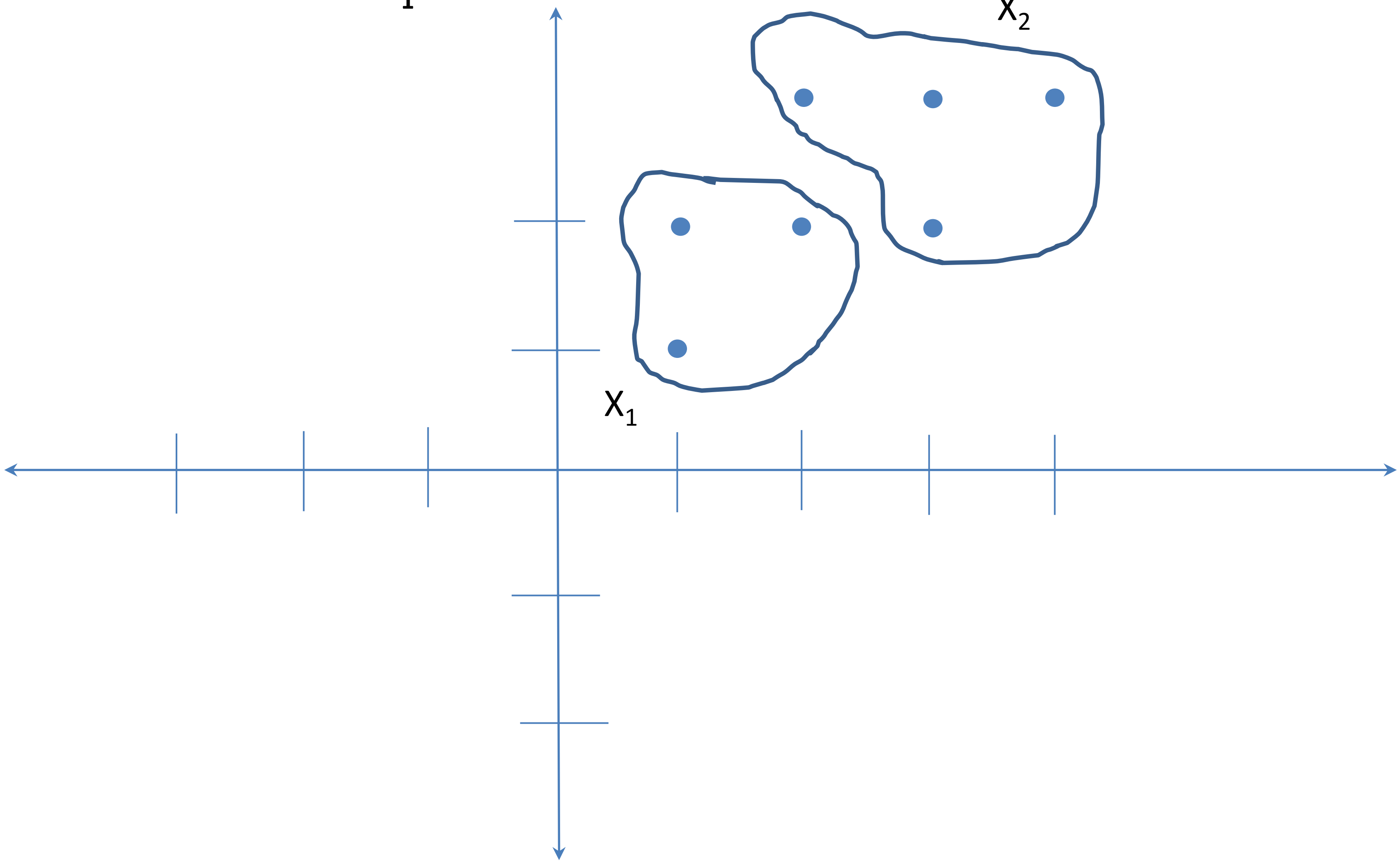
5. Algoritmo K-medias

Algoritmo que calcula la mejor partición para un conjunto de muestras.

Dada una partición inicial de las muestras, el algoritmo determina el mejor clúster para cada objeto, estudiando el SEC que se produce cuando las muestra se transfieren (mueven) a otros clusters.

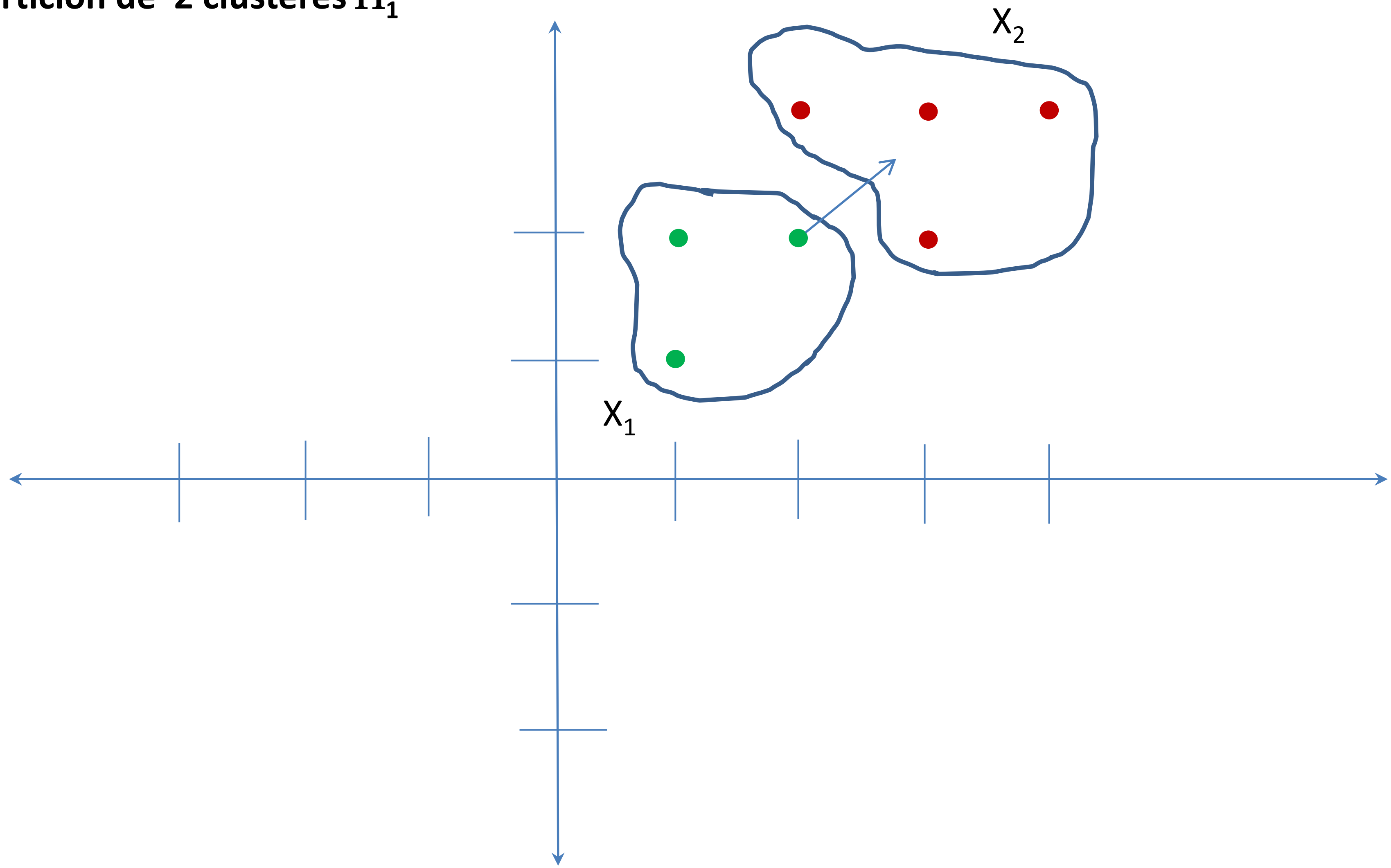
5. Algoritmo K-medias

partición de 2 clústeres Π_1



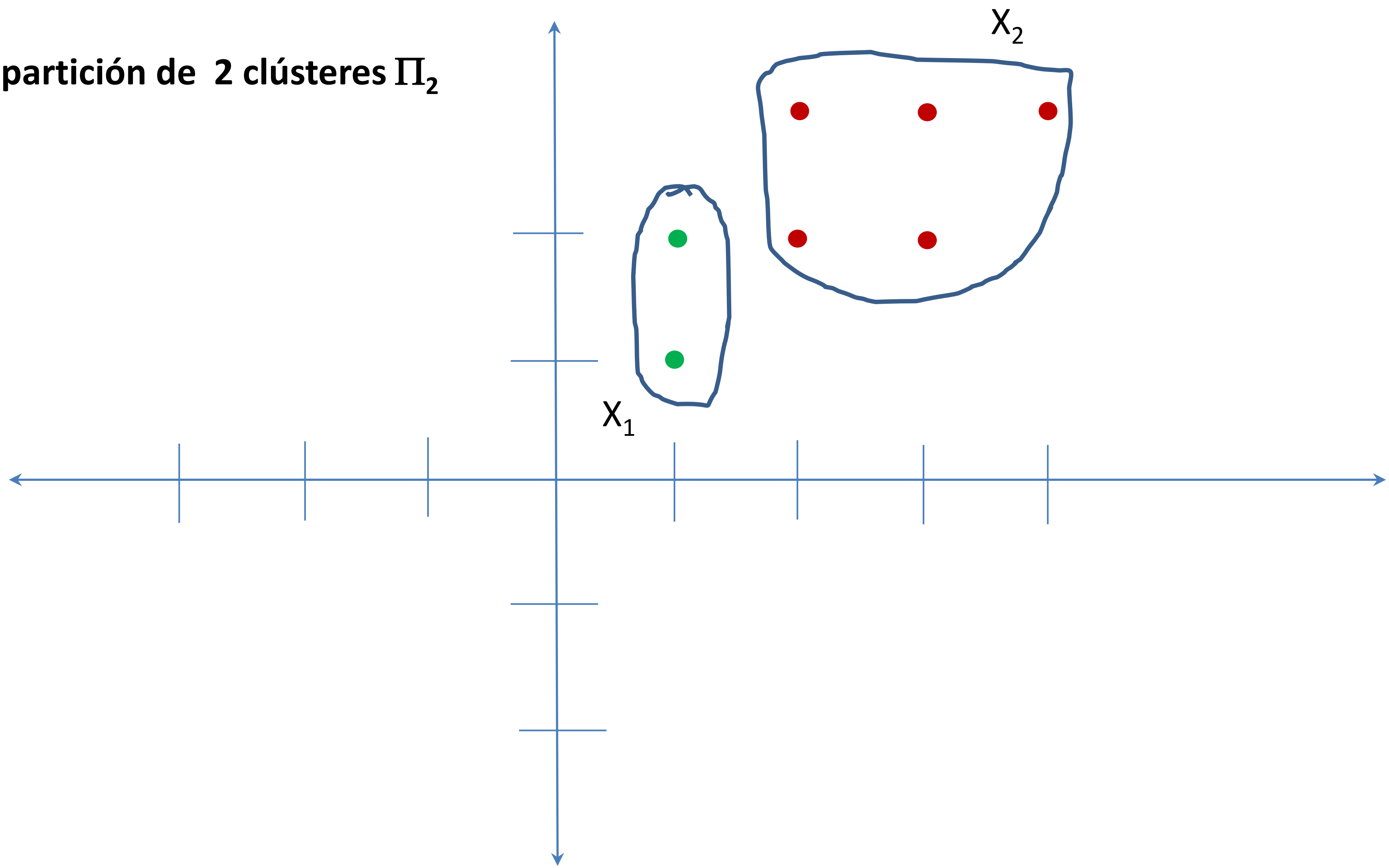
5. Algoritmo K-medias

partición de 2 clústeres Π_1



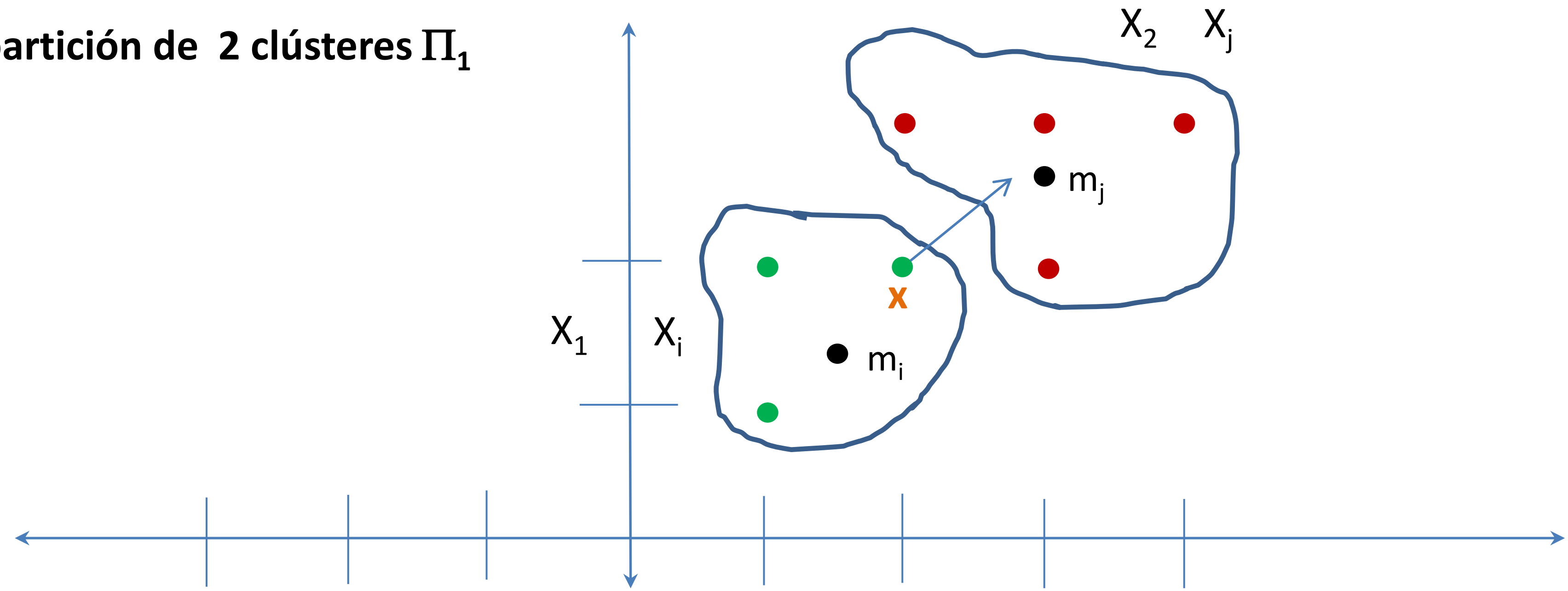
5. Algoritmo K-medias

partición de 2 clústeres Π_2



5. Algoritmo K-medias

partición de 2 clústeres Π_1



$$\triangle J = \underbrace{\frac{n_j}{n_j+1}\|\boldsymbol{x}-\boldsymbol{m}_j\|^2}_{\text{clúster destino de } \mathbf{x}} - \underbrace{\frac{n_i}{n_i-1}\|\boldsymbol{x}-\boldsymbol{m}_i\|^2}_{\text{clúster origen de } \mathbf{x}}$$

$$\Delta J = \frac{4}{4+1}\|\boldsymbol{x}-\boldsymbol{m}_j|^2 - \frac{3}{3-1}\|\boldsymbol{x}-\boldsymbol{m}_i\|^2$$

5. Algoritmo K-medias

Cálculo incremental de la SEC al transferir una muestra **x** del clúster $\mathbf{X_i}$ al clúster $\mathbf{X_j}$

$$X_i' = X_i - \{\boldsymbol{x}\} \qquad X_j' = X_j + \{\boldsymbol{x}\}$$

$$\boldsymbol{m}_i' = \boldsymbol{m}_i - \frac{\boldsymbol{x} - \boldsymbol{m}_i}{n_i - 1} \qquad \boldsymbol{m}_j' = \boldsymbol{m}_j + \frac{\boldsymbol{x} - \boldsymbol{m}_j}{n_j + 1}$$

$$J_i' = J_i - \frac{n_i}{n_i - 1} \|\boldsymbol{x} - \boldsymbol{m}_i\|^2 \qquad J_j' = J_j + \frac{n_j}{n_j + 1} \|\boldsymbol{x} - \boldsymbol{m}_j\|^2$$

$$\triangle J = \frac{n_j}{n_j + 1} \|\boldsymbol{x} - \boldsymbol{m}_j\|^2 - \frac{n_i}{n_i - 1} \|\boldsymbol{x} - \boldsymbol{m}_i\|^2$$

La transferencia es favorable si el incremento de SEC es negativo.

$$\frac{n_j}{n_j + 1} \|\boldsymbol{x} - \boldsymbol{m}_j\|^2 < \frac{n_i}{n_i - 1} \|\boldsymbol{x} - \boldsymbol{m}_i\|^2$$

Estas ecuaciones permiten minimizar el SEC mediante refinamientos sucesivos desde una partición inicial dada.

5. Algoritmo K-medias

C-medias (bibliografía en castellano) **K-medias** (bibliografía en inglés)

Input:

- $\mathbf{X = (x_1, x_2, \ldots, x_n)}$, las N muestras/objetos
- **C**, el número de clusters
- $\Pi = \{\mathbf{X_1}, \ldots, \mathbf{X_C}\}$, una partición inicial aleatoria que clasifica los N objetos en C clústeres

Output:

- Π^*, la mejor partición
- $\mathbf{m_1, \ldots, m_C}$, la media de cada clúster de Π^*
- **J** (opcionalmente, según la versión del algoritmo K-medias)

Objetivo: conseguir una partición de los N objetos en C clústeres tal que la SEC sea mínima

Dos versiones del algoritmo K-medias:

- versión correcta o de Duda & Hart
- versión popular

5. Algoritmo K-medias

Algorithm C*-means* (versión "correcta" [Duda & Hart])

Input: $X; C; \Pi = \{X_1, \dots, X_C\};$ (partición inicial aleatoria)

Output: $\Pi^* = \{X_1, \dots, X_C\}; \boldsymbol{m}_1, \dots, \boldsymbol{m}_C; J$

for $c = 1$ **to** C **do** $\boldsymbol{m}_c = \frac{1}{n_c} \sum_{\boldsymbol{x} \in X_c} \boldsymbol{x}$ **endfor** (se calcula las medias de los clústeres de la partición de entrada)

repeat

$transfers =$ false

forall $\boldsymbol{x} \in X$ (let $i : \boldsymbol{x} \in X_i$) **do** (para todos los objetos … tomamos objeto **x** que pertenece a clúster $\mathbf{X_i}$)

if $n_i > 1$ **then** (si el clúster $\mathbf{X_i}$ tiene más de una muestra, no solo el objeto x)

$$j^* = \underset{j \neq i}{\text{argmin}} \frac{n_j}{n_j + 1} \|\boldsymbol{x} - \boldsymbol{m}_j\|^2$$

(se estudia **x** en todos los otros clústeres distintos de $\mathbf{X_i}$ tomamos el min clúster $\mathbf{X_j^*}$)

$$\triangle J = \frac{n_{j^*}}{n_{j^*} + 1} \|\boldsymbol{x} - \boldsymbol{m}_{j^*}\|^2 - \frac{n_i}{n_i - 1} \|\boldsymbol{x} - \boldsymbol{m}_i\|^2$$

(incremento de SEC, diferencia entre $\mathbf{X_j^*}$ y el clúster actual de **x**)

if $\triangle J < 0$ **then**

$transfers =$ true (mover **x** a $\mathbf{X_j^*}$)

$$\boldsymbol{m}_i = \boldsymbol{m}_i - \frac{\boldsymbol{x} - \boldsymbol{m}_i}{n_i - 1} \qquad \boldsymbol{m}_{j^*} = \boldsymbol{m}_{j^*} + \frac{\boldsymbol{x} - \boldsymbol{m}_{j^*}}{n_{j^*} + 1}$$

$$X_i = X_i - \{\boldsymbol{x}\} \qquad X_{j^*} = X_{j^*} + \{\boldsymbol{x}\}$$

$J = J + \triangle J$ (valor de SEC de la mejor partición hasta el momento)

endif

endif

endforall

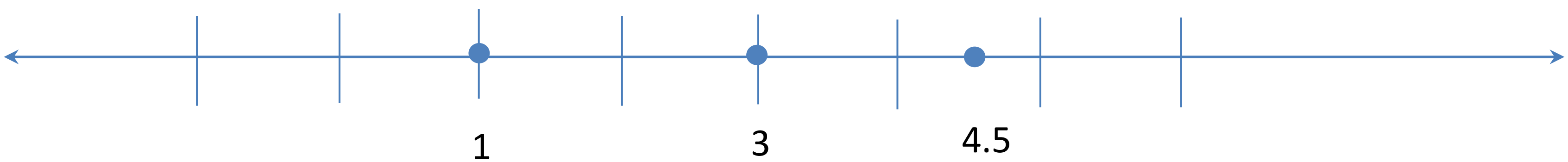
until $\neg transfers$ // Coste por iteración: $O(N \cdot C \cdot D)$, $N = |X|$, D = dimensión

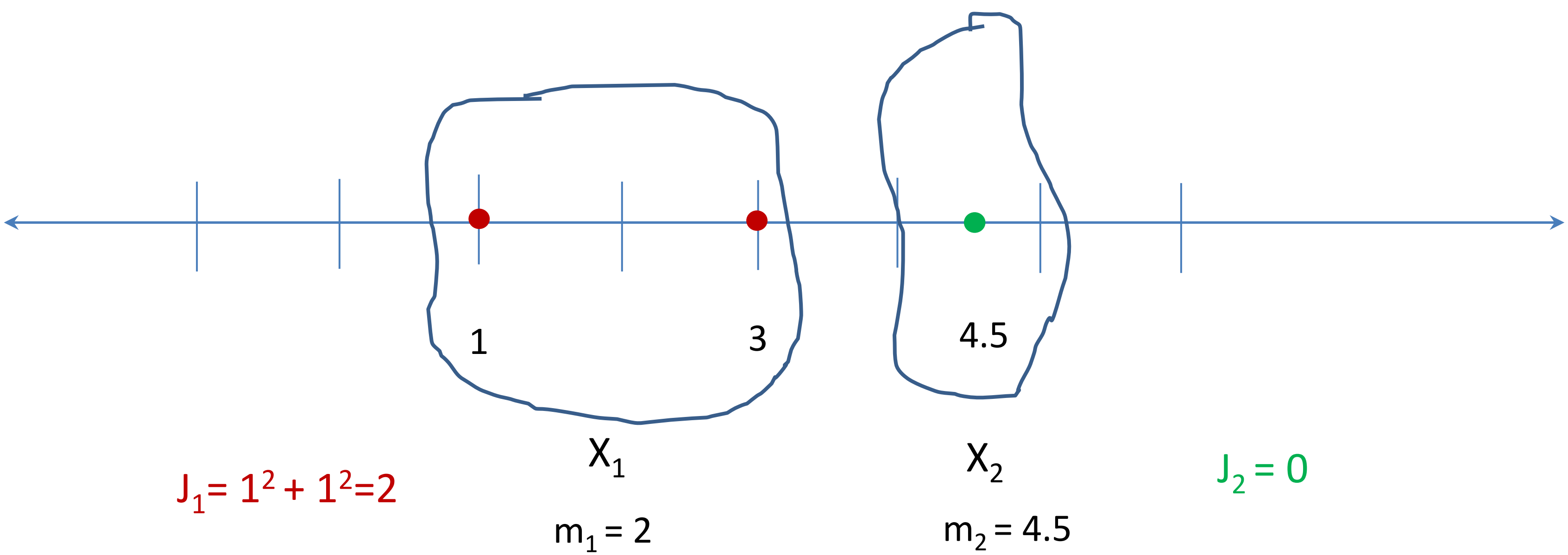
5. Algoritmo K-medias: ejemplo aplicación K-medias (Duda & Hart)

$X = \{1, 3, 4.5\} \subset \mathbb{R}$; $\quad x_1 = (1), x_2 = (3), x_3 = (4.5)$

$C = 2$ clusters; $\quad X_1 = \{x_1, x_2\}, \quad X_2 = \{x_3\} \quad \Pi^0 = \{\{1, 3\}, \{4.5\}\}$

$\boldsymbol{m}_1 = 2 \quad J_1 = 2 \quad \boldsymbol{m}_2 = 4.5 \quad J_2 = 0 \quad J(\Pi^0) = 2$





5. Algoritmo K-medias: ejemplo aplicación K-medias (Duda & Hart)

$x = x_1 = 1$ (Cluster X_1, $i = 1$, $n_1 = 2$: two objects in X_1)

A) we have only one more cluster (X_2) so $\arg\min_{j \neq i}$ only analyzes cluster $j = 2$ ($n_j = n_2 = 1$).

B)

$$j^* = \arg\min_{j \neq i} \frac{n_j}{n_j + 1} \|\boldsymbol{x} - \boldsymbol{m}_j\|^2 = \arg\min_{j \neq i} \frac{n_2}{n_2 + 1} \|\boldsymbol{x}_1 - \boldsymbol{m}_2\|^2 =$$
$$\frac{1}{2} \|1 - 4.5\|^2 = 6.125$$

C)

$$\triangle J = \frac{n_{j*}}{n_{j*} + 1} \|\boldsymbol{x} - \boldsymbol{m}_{j*}\|^2 - \frac{n_i}{n_i - 1} \|\boldsymbol{x} - \boldsymbol{m}_i\|^2 =$$
$$\frac{n_2}{n_2 + 1} \|\boldsymbol{x}_1 - \boldsymbol{m}_2\|^2 - \frac{n_1}{n_1 - 1} \|\boldsymbol{x}_1 - \boldsymbol{m}_1\|^2 =$$
$$6.125 - \frac{2}{1}\|1 - 2\|^2 = 6.125 - 2 = 4.125$$

D) x_1 is not transferred to X_2

5. Algoritmo K-medias: ejemplo aplicación K-medias (Duda & Hart)

$x = x_2 = 3$ (Cluster X_1, $i = 1$, $n_1 = 2$: two objects in X_1)

A) we have only one more cluster (X_2) so $\arg\min_{j \neq i}$ only analyzes cluster $j = 2$ ($n_j = n_2 = 1$).

B)

$$j^* = \arg\min_{j \neq i} \frac{n_j}{n_j + 1} \|\boldsymbol{x} - \boldsymbol{m}_j\|^2 = \arg\min_{j \neq i} \frac{n_2}{n_2 + 1} \|\boldsymbol{x}_2 - \boldsymbol{m}_2\|^2 = \frac{1}{2} \|3 - 4.5\|^2 = 1.125$$

C)

$$\triangle J = \frac{n_{j*}}{n_{j*} + 1} \|\boldsymbol{x} - \boldsymbol{m}_{j*}\|^2 - \frac{n_i}{n_i - 1} \|\boldsymbol{x} - \boldsymbol{m}_i\|^2 = \frac{n_2}{n_2 + 1} \|\boldsymbol{x}_2 - \boldsymbol{m}_2\|^2 - \frac{n_1}{n_1 - 1} \|\boldsymbol{x}_2 - \boldsymbol{m}_1\|^2 = 1.125 - \frac{2}{1}\|3 - 2\|^2 = 1.125 - 2 = -0.875$$

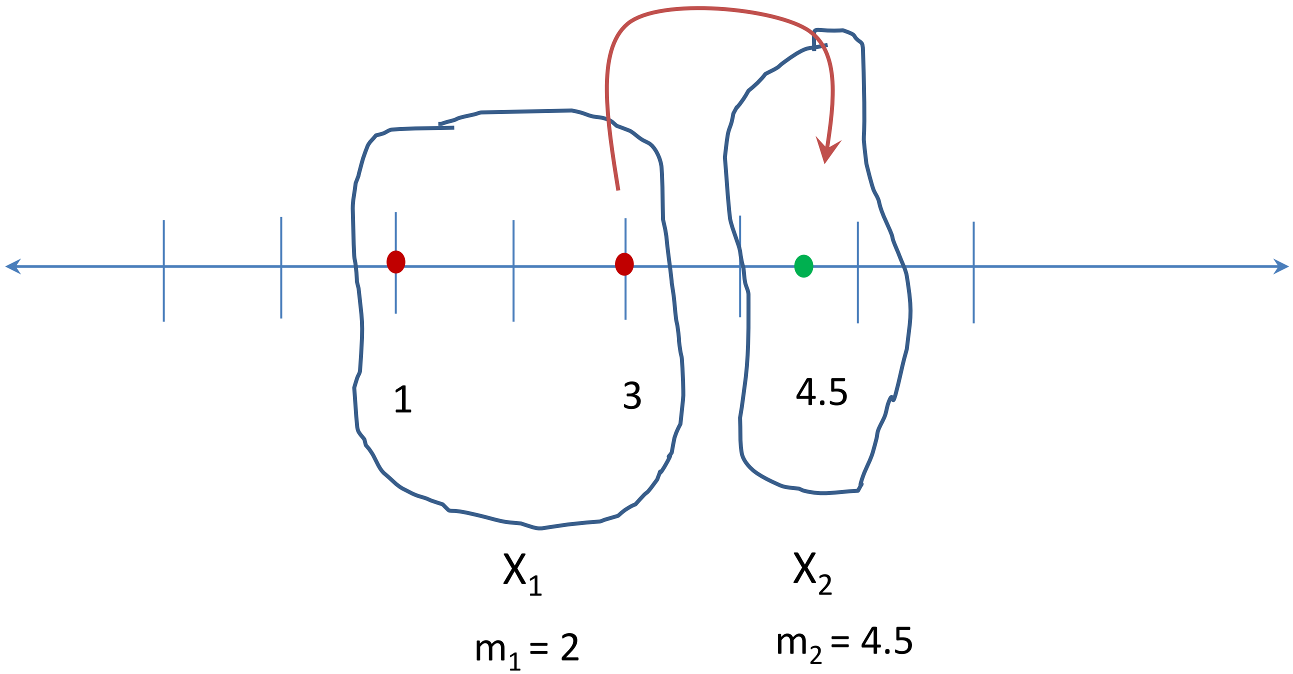
D) $transfers = true$, we move x_2 to cluster X_2

E)

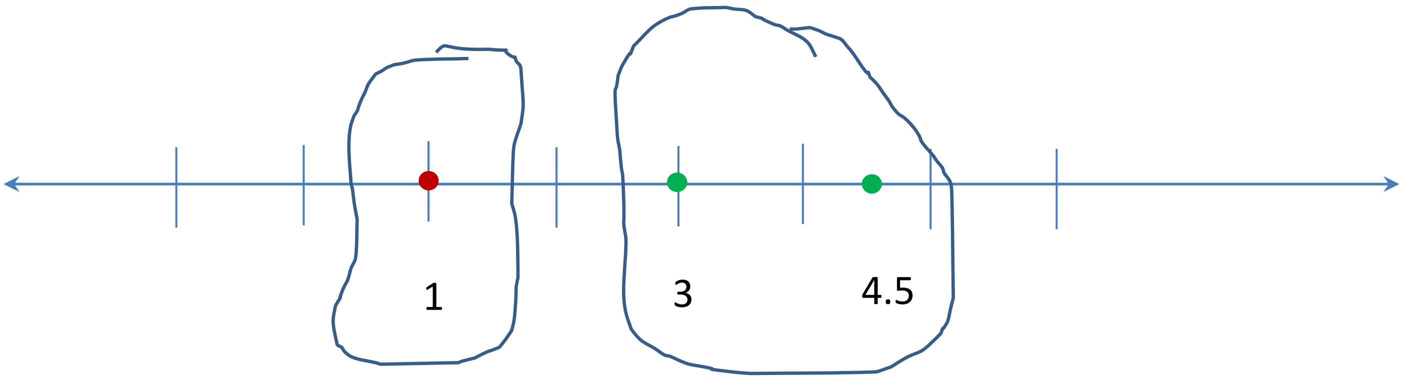
$$\boldsymbol{m}_i = \boldsymbol{m}_i - \frac{\boldsymbol{x} - \boldsymbol{m}_i}{n_i - 1} = \boldsymbol{m}_1 - \frac{\boldsymbol{x}_2 - \boldsymbol{m}_1}{n_1 - 1} = 2 - \frac{3 - 2}{2 - 1} = 1 \quad (m_1 = 1)$$

$$\boldsymbol{m}_{j*} = \boldsymbol{m}_{j*} + \frac{\boldsymbol{x} - \boldsymbol{m}_{j*}}{n_{j*} + 1} = \boldsymbol{m}_2 + \frac{\boldsymbol{x}_2 - \boldsymbol{m}_2}{n_2 + 1} = 4.5 + \frac{3 - 4.5}{1 + 1} = 3.75 \quad (m_2 = 3.75)$$

5. Algoritmo K-medias: ejemplo aplicación K-medias (Duda & Hart)



5. Algoritmo K-medias: ejemplo aplicación K-medias (Duda & Hart)



X_1 $\qquad$ X_2

$m_1 = 1$ $\qquad$ $m_2 = 3.75$

$J_1 = 0$ $\qquad$ $J_2 = 0.75^2 + 0.75^2 = 1.125$

$$\boldsymbol{m}_i = \boldsymbol{m}_i - \frac{\boldsymbol{x} - \boldsymbol{m}_i}{n_i - 1} = \boldsymbol{m}_1 - \frac{\boldsymbol{x}_2 - \boldsymbol{m}_1}{n_1 - 1} = 2 - \frac{3 - 2}{2 - 1} = 1 \quad (m_1 = 1)$$

$$\boldsymbol{m}_{j*} = \boldsymbol{m}_{j*} + \frac{\boldsymbol{x} - \boldsymbol{m}_{j*}}{n_{j*} + 1} = \boldsymbol{m}_2 + \frac{\boldsymbol{x}_2 - \boldsymbol{m}_2}{n_2 + 1} = 4.5 + \frac{3 - 4.5}{1 + 1} = 3.75 \quad (m_2 = 3.75)$$

5. Algoritmo K-medias: ejemplo aplicación K-medias (Duda & Hart)

$x = x_3 = 4.5$ (Cluster X_2, $i = 2$, $n_2 = 2$: two objects in X_2)

A) we have only one more cluster (X_1) so $\arg\min_{j \neq i}$ only analyzes cluster $j = 1$ ($n_j = n_1 = 1$).

B)

$$j^* = \arg\min_{j \neq i} \frac{n_j}{n_j + 1} \|\boldsymbol{x} - \boldsymbol{m}_j\|^2 = \arg\min_{j \neq i} \frac{n_1}{n_1 + 1} \|\boldsymbol{x_3} - \boldsymbol{m}_1\|^2 =$$

$$\frac{1}{2} \|4.5 - 1\|^2 = 6.125$$

C)

$$\triangle J = \frac{n_{j^*}}{n_{j^*} + 1} \|\boldsymbol{x} - \boldsymbol{m}_{j^*}\|^2 - \frac{n_i}{n_i - 1} \|\boldsymbol{x} - \boldsymbol{m}_i\|^2 =$$

$$\frac{n_1}{n_1 + 1} \|\boldsymbol{x_3} - \boldsymbol{m}_1\|^2 - \frac{n_2}{n_2 - 1} \|\boldsymbol{x_3} - \boldsymbol{m}_2\|^2 =$$

$$6.125 - \frac{2}{1} \|4.5 - 3.75\|^2 = 6.125 - 1.125 = 5$$

D) x_3 is not transferred to X_1

El algoritmo hace una iteración más del bucle 'repeat … until' porque la variable ***transfers*** es TRUE. En la siguiente iteración del algoritmo no se producen modificaciones (convergencia).

Resultado final: Π^* **= {X$_1$, X$_2$}**, $X_1 = \{x_1\}$, $X_2 = \{x_2, x_3\}$, $J(\Pi^*) = SCE_{\Pi^*} = 1.125$

6. Algoritmo K-medias convencional o popular

Algorithm C-*means* (versión "popular")

Input: $X; C; \Pi = \{X_1, \ldots, X_C\};$

Output: $\Pi^* = \{X_1, \ldots, X_C\}; \boldsymbol{m}_1, \ldots, \boldsymbol{m}_C$

repeat

$\quad transfers = \text{false}$

$\quad$ **for** $c = 1$ **to** C **do** $\boldsymbol{m}_c = \frac{1}{n_c} \sum_{\boldsymbol{x} \in X_c} \boldsymbol{x}$ **endfor**

$\quad$ **forall** $\boldsymbol{x} \in X$ (let $i : \boldsymbol{x} \in X_i$) **do**

$\quad\quad$ **if** $n_i > 1$ **then**

$\quad\quad\quad j^* = \underset{1 \le j \le C}{\text{argmin}}\; d(\boldsymbol{x}, \boldsymbol{m}_j)$ (distancia Euclídea de x al punto medio del cluster: $d(x,m_j) = \| x - m_j \|$)

(analizamos **x** en todos los clusters, incluyendo su propio cluster, y tomamos el menor)

$\quad\quad\quad$ **if** $j^* \neq i$ **then**

$\quad\quad\quad\quad transfers = \text{true}$ (si el min clúster es diferente del clúster actual de **x**, movemos **x** al min cluster)

$\quad\quad\quad\quad X_i = X_i - \{\boldsymbol{x}\}; X_{j^*} = X_{j^*} + \{\boldsymbol{x}\}$

$\quad\quad\quad$ **endif**

$\quad\quad$ **endif**

$\quad$ **endforall**

until $\neg transfers$

// Coste por iteración: $O(N \cdot C \cdot D)$, $N = |X|$, D = coste de $d(\cdot, \cdot)$

6. Algoritmo K-medias convencional o popular

La versión ‘popular’ no calcula el SEC sino que toma la distancia Euclídea (no la distancia Euclídea al cuadrado) de cada muestra **x** al punto medio, mueve la muestra **x** al clúster que devuelve menor valor y actualiza los objetos del clúster actual y el clúster mínimo.

Una vez se han estudiado todas las muestras entonces se re-calculan las medidas de todos los clústeres de nuevo.

Las tres diferencias respecto al algoritmo Duda&Hart son:

1. considera la distancia Euclídea en lugar del error cuadrático (distancia Euclídea al cuadrado)
2. para una muestra dada, mide la distancia a los puntos medios de todos los clústeres incluido el suyo propio
3. no actualiza las medias de los clústeres hasta que se ha completado una vuelta del bucle que analiza todas las muestras

Propiedades de los algoritmos K-medias:

- ninguna de las dos versiones garantiza la obtención de un mínimo global de la SEC
- La versión Duda&Hart obtiene *un mínimo local*
- la versión ‘popular’ no garantiza la minimización local en algunos casos

7. Otras interpretaciones de SEC

El criterio SEC se puede reescribir sin incluir las medias de los clusters:

$$J(X_1, \ldots, X_C) = \frac{1}{2} \sum_c n_c \, \bar{s}_c \tag{7}$$

- n_c: número de datos en X_c
- $\bar{s}_c$: media de las distancias euclídeas al cuadrado entre todos estos datos:

$$\bar{s}_c = \frac{1}{n_c^2} \sum_{\boldsymbol{x}, \boldsymbol{x}' \in X_c} \|\boldsymbol{x} - \boldsymbol{x}'\|^2 \tag{8}$$

SEC es la suma ponderada de medias de distancias cuadráticas “intra-cluster”

Se puede redefinir $\bar{s}_c$ y dar criterios similares a SEC con distancia genérica d:

$$\bar{s}_c = \frac{1}{n_c^2} \sum_{x, x' \in X_c} d(x, x') \qquad \bar{s}_c = \frac{1}{n_c^2} \max_{x, x' \in X_c} d(x, x') \tag{9}$$

Permite criterio SEC incluso en datos *no-vectoriales*

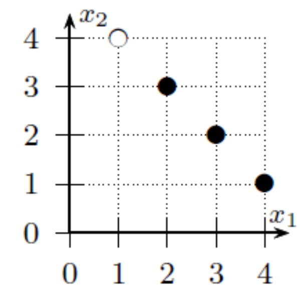
41 ☐ La figura a la derecha muestra una partición de 4 puntos bidimensionales en 2 clústers (representados mediante los símbolos $\bullet$ y $\circ$). La transferencia del punto $(2,3)^t$ del clúster $\bullet$ al $\circ$ conduce a una variación de la SEC, ΔJ, tal que:

A) $\Delta J > 0$.

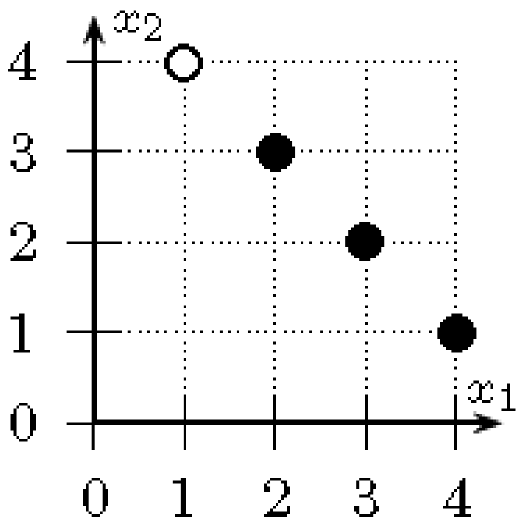
B) $0 \geq \Delta J > -1$.

C) $-1 \geq \Delta J > -2$.

D) $-2 \geq \Delta J$.



$$m_b = \begin{pmatrix} 1 \\ 4 \end{pmatrix}$$



$$\Delta J = \frac{n_j}{n_j+1}\|\boldsymbol{x}-\boldsymbol{m}_j\|^2 - \frac{n_i}{n_i-1}\|\boldsymbol{x}-\boldsymbol{m}_i\|^2$$

$$\Delta J = \frac{1}{1+1}\left\|\begin{pmatrix} 2 \\ 3 \end{pmatrix} - \begin{pmatrix} 1 \\ 4 \end{pmatrix}\right\|^2 - \frac{3}{3-1}\left\|\begin{pmatrix} 2 \\ 3 \end{pmatrix} - \begin{pmatrix} 3 \\ 2 \end{pmatrix}\right\|^2 =$$

$$\frac{1}{2}\left\|\begin{pmatrix} 1 \\ -1 \end{pmatrix}\right\|^2 - \frac{3}{2}\left\|\begin{pmatrix} -1 \\ 1 \end{pmatrix}\right\|^2 =$$

$$0.5 \cdot 2 - 1.5 \cdot 2 = -2$$

$$m_n = \frac{1}{3}\begin{pmatrix} 2 \\ 3 \end{pmatrix} + \begin{pmatrix} 3 \\ 2 \end{pmatrix} + \begin{pmatrix} 4 \\ 1 \end{pmatrix} = \begin{pmatrix} 3 \\ 2 \end{pmatrix}$$

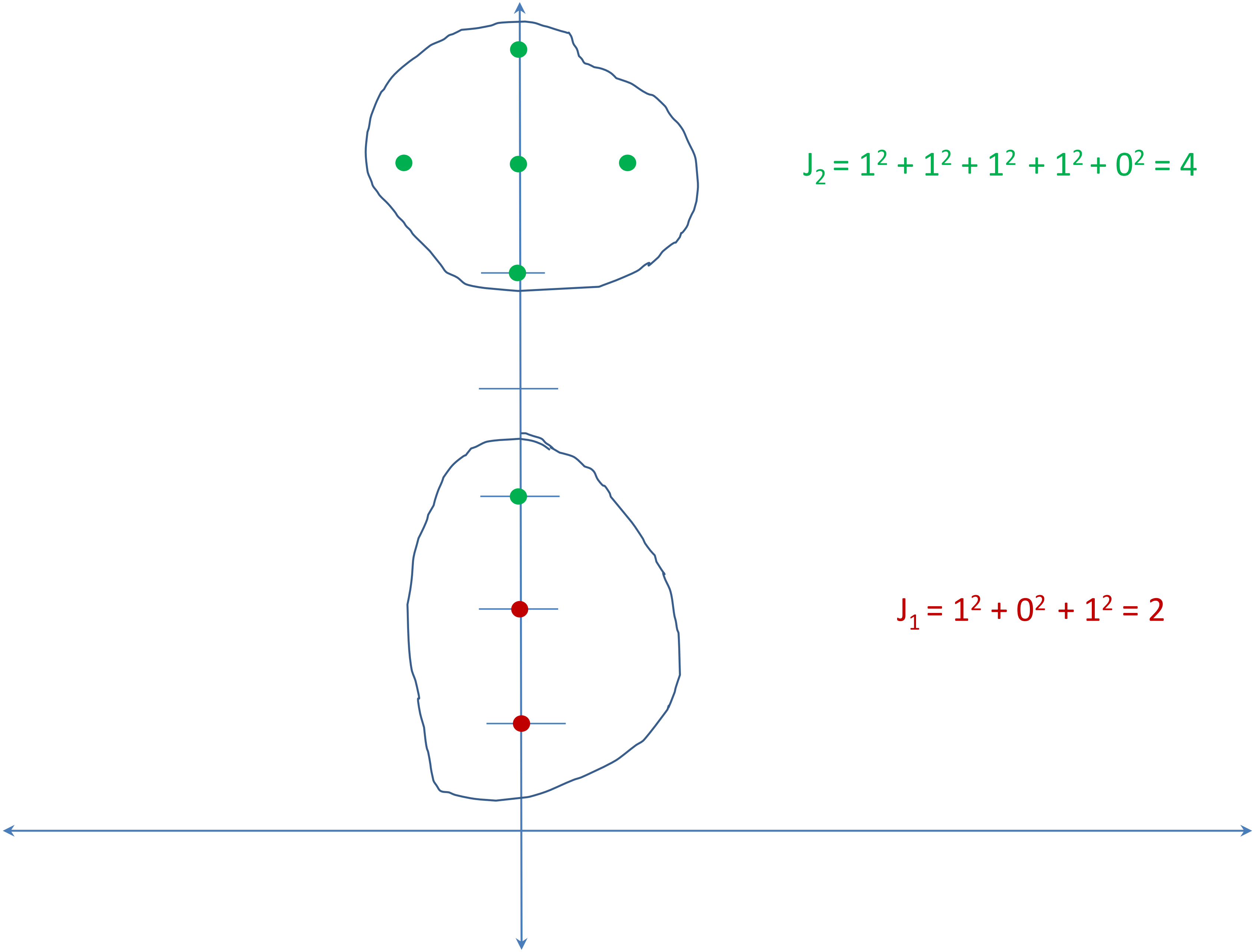
31 ☐ Supongamos que hemos aplicado el algoritmo C-medias a un conjunto de puntos bi-dimensionales para obtener un agrupamiento (partición) en dos *clústers*. Tras una serie de iteraciones del algoritmo C-medias tenemos el agrupamiento: $\{\ \{(0,1)^{\mathbf{t}}, (0,2)^{\mathbf{t}}\},\ \{(0,3)^{\mathbf{t}}, (0,5)^{\mathbf{t}}, (0,6)^{\mathbf{t}}, (0,7)^{\mathbf{t}}, (1,6)^{\mathbf{t}}, (-1,6)^{\mathbf{t}}\}\ \}$. Indica la respuesta *correcta*.

A) La suma de errores cuadráticos (SEC) es 15 y puede llegar a ser 8.

B) La SEC es 15 y cuando el algoritmo converja será 12.

C) La SEC es 12 y cuando el algoritmo converja será 10.

D) La SEC es 12 y cuando el algoritmo converja será 6.



$J_2 = 1^2 + 1^2 + 1^2 + 1^2 + 0^2 = 4$

$J_1 = 1^2 + 0^2 + 1^2 = 2$

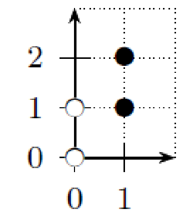
34 ☐ En la figura de la derecha se muestra una partición de 4 puntos bidimensionales de 2 clústers. La transferencia del punto $(1,1)^t$ del clúster $\bullet$ al clúster $\circ$

A) produce un incremento en la SEC.

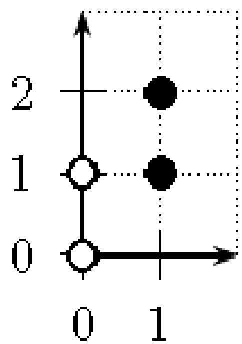
B) produce un decremento en la SEC.

C) no altera la SEC.

D) produce una SEC negativa.



$$m_n = \begin{pmatrix} 1 \\ 1.5 \end{pmatrix}$$



$$m_b = \begin{pmatrix} 0 \\ 0.5 \end{pmatrix}$$

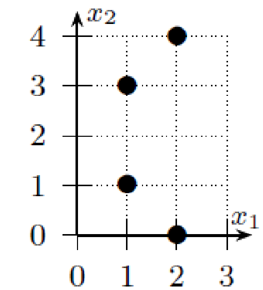
$$\triangle J = \frac{n_j}{n_j+1}\|\boldsymbol{x} - \boldsymbol{m}_j\|^2 - \frac{n_i}{n_i-1}\|\boldsymbol{x} - \boldsymbol{m}_i\|^2$$

$$\Delta J = \frac{2}{2+1}\left\|\begin{pmatrix} 1 \\ 1 \end{pmatrix} - \begin{pmatrix} 0 \\ 0.5 \end{pmatrix}\right\|^2 - \frac{2}{2-1}\left\|\begin{pmatrix} 1 \\ 1 \end{pmatrix} - \begin{pmatrix} 1 \\ 1.5 \end{pmatrix}\right\|^2 =$$

$$\frac{2}{3}\left\|\begin{pmatrix} 1 \\ 0.5 \end{pmatrix}\right\|^2 - \frac{2}{1}\left\|\begin{pmatrix} 0 \\ -0.5 \end{pmatrix}\right\|^2 =$$

$$0.666 \cdot 1.25 - 2 \cdot 0.25 \quad = 0.33$$

32 ☐ En la figura de la derecha se representan 4 muestras de bidimensionales. ¿Cuál es el número de clústers que minimiza la suma de errores cuadráticos para dicho conjunto de muestras?



A) 1

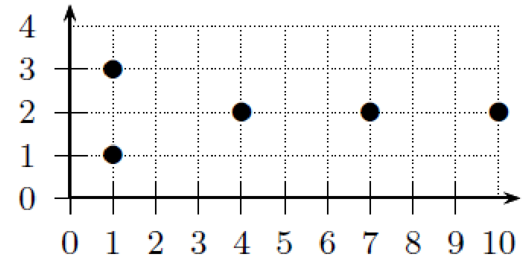
B) 2

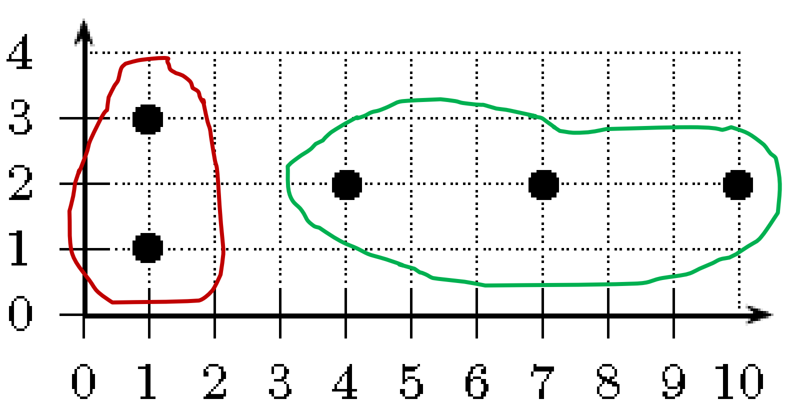
C) 3

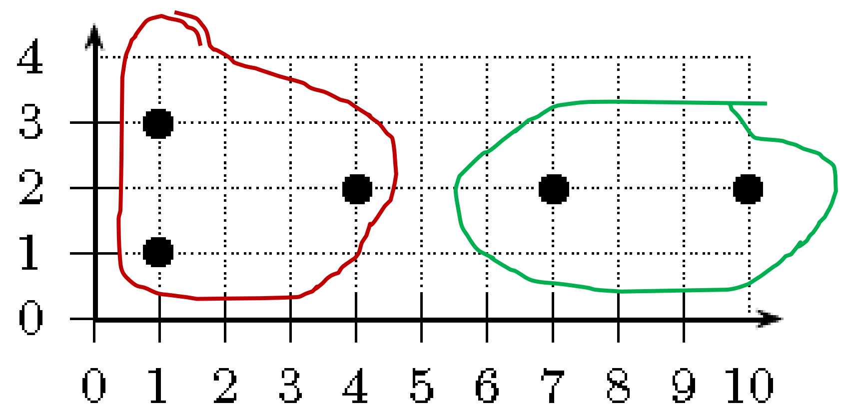
D) 4

3 ☐ La menor suma de errores cuadráticos con la que los puntos de la figura a la derecha pueden agruparse en dos clústers es:

A) Menor que 10
B) Entre 10 y 15
C) Entre 15 y 20
D) Mayor que 20







$J_2 = 1.5^2 + 1.5^2 = 4.5$

$J_1 = 1^2 + 1^2 = 2$

$J_2 = 3^2 + 0 + 3^2 = 18$

$$m_b = \frac{1}{3}\left(\begin{pmatrix}1\\1\end{pmatrix} + \begin{pmatrix}1\\3\end{pmatrix} + \begin{pmatrix}4\\2\end{pmatrix}\right) = \begin{pmatrix}2\\2\end{pmatrix}$$

$J_1 = \sqrt{2}^2 + \sqrt{2}^2 + 2^2 = 8$

44 ☐ Se tiene una partición de un conjunto de datos 3-dimensionales en un número de clústers dado, $C \geq 2$. Considérese la transferencia del dato $\mathbf{x} = (3, 2, 9)^t$ de un clúster i a otro j, $j \neq i$. Se sabe que el clúster i contiene 3 datos (contando $\mathbf{x}$) y el j 4. Asimismo, se sabe que la media del clúster i es $\mathbf{m}_i = (7, 3, 3)^t$ y la del j $\mathbf{m}_j = (7, 6, 7)^t$. Si se realiza dicha transferencia, se producirá un incremento de la suma de errores cuadráticos, ΔJ, tal que:

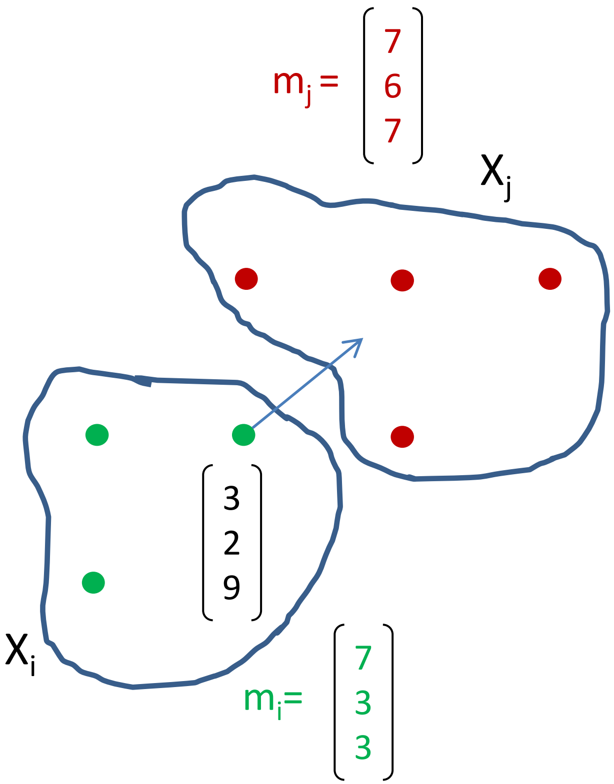
A) $\Delta J < -70$

B) $-70 \leq \Delta J < -30$

C) $-30 \leq \Delta J < 0$

D) $\Delta J \geq 0$

$$\triangle J = \frac{n_j}{n_j + 1} \|\boldsymbol{x} - \boldsymbol{m}_j\|^2 - \frac{n_i}{n_i - 1} \|\boldsymbol{x} - \boldsymbol{m}_i\|^2$$



$$\Delta J = \frac{4}{4+1} \left\| \begin{pmatrix} 3 \\ 2 \\ 9 \end{pmatrix} - \begin{pmatrix} 7 \\ 6 \\ 7 \end{pmatrix} \right\|^2 - \frac{3}{3-1} \left\| \begin{pmatrix} 3 \\ 2 \\ 9 \end{pmatrix} - \begin{pmatrix} 7 \\ 3 \\ 3 \end{pmatrix} \right\|^2 =$$

$$\frac{4}{5} \left\| \begin{pmatrix} -4 \\ -4 \\ 2 \end{pmatrix} \right\|^2 - \frac{3}{2} \left\| \begin{pmatrix} -4 \\ -1 \\ 6 \end{pmatrix} \right\|^2 =$$

$$0.8 \cdot 36 - 1.5 \cdot 53 = -50.7$$

52 $\square$ Se tiene una partición de un conjunto de datos 3-dimensionales en un número de clústers dado, $C \geq 2$. Considérese la transferencia del dato $\mathbf{x} = (3, 6, 4)^t$ de un clúster j a otro i, $j \neq i$. Se sabe que el clúster j contiene 3 datos (contando $\mathbf{x}$) y el i 3. Asimismo, se sabe que la media del clúster j es $\mathbf{m}_j = (3, 3, 2)^t$ y la del i $\mathbf{m}_i = (7, 6, 9)^t$. Si se realiza dicha transferencia, se producirá un incremento de la suma de errores cuadráticos, ΔJ, tal que:

A) $\Delta J < -70$

B) $-70 \leq \Delta J < -30$

C) $-30 \leq \Delta J < 0$

D) $\Delta J \geq 0$

Bloque 2
Aprendizaje Automático

Tema 4:
Aprendizaje No Supervisado:

Clustering. Algoritmo K-medias.