Minería de Datos

Técnicas de Minería de datos

Agrupación - K-medias

Algoritmo K-medias

Algoritmo:

{Entrada: k (Número de grupos), $C = (C_1, C_2, \ldots, C_k)$ = prototipos iniciales,
D = conjunto de datos = $\{x^i\}$, i = 1..N}

Inicializar centros

Repetir

Para j = 1 hasta k

$P_j \leftarrow \{\}$ % inicializar particiones

Fin_Para

Paso 1

Para i = 1 hasta N

$C_g = \underset{k}{\text{argmin}} \{d_i(x^i, C_j)\}$, j = 1 .. k *% Se determina el centro más cercano a x^i*

$P_g = P_g \cup \{x^i\}$ *% Se añade x^i a la partición asociada al prototipo C_g*

Fin_Para

Paso 2

Para j = 1 hasta k

$C_j = \frac{1}{m_j} \sum_{x^i \in P_j} x^i$ *% Se recalculan los centros de cada grupo, donde m_j = número de datos en la partición j*

Fin_Para

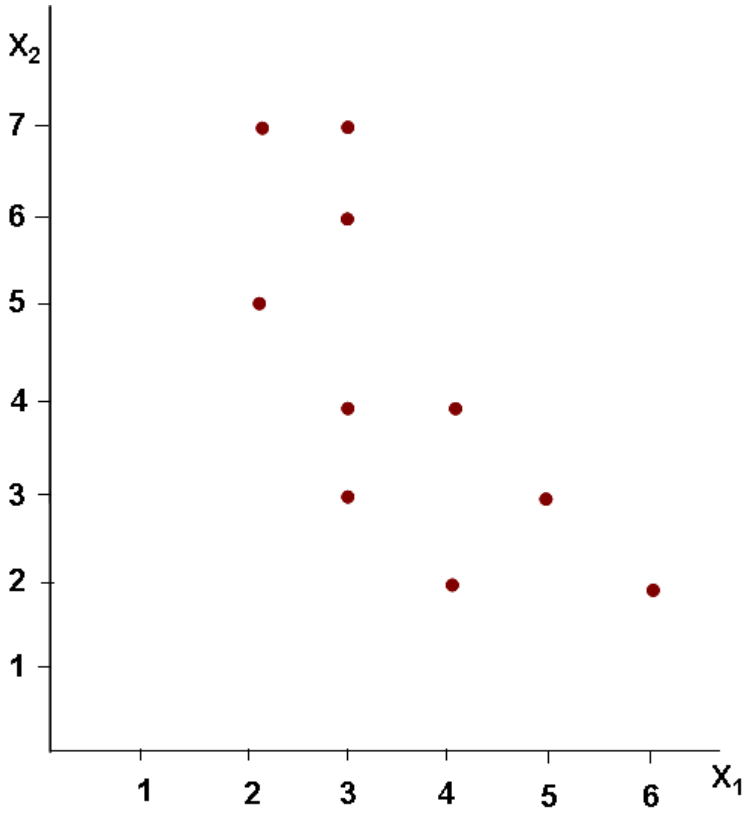
Hasta que los centros no cambien (o se alcance otra condición de parada)

{Salida: conjunto de k prototipos C, particiones P (=Grupos)}

Algoritmo K-medias

Ejemplo:

D = { (2,5), (2,7), (3,6), (3,7), (3,3), (3,4), (4,2), (4,4), (5,3), (6,2)}



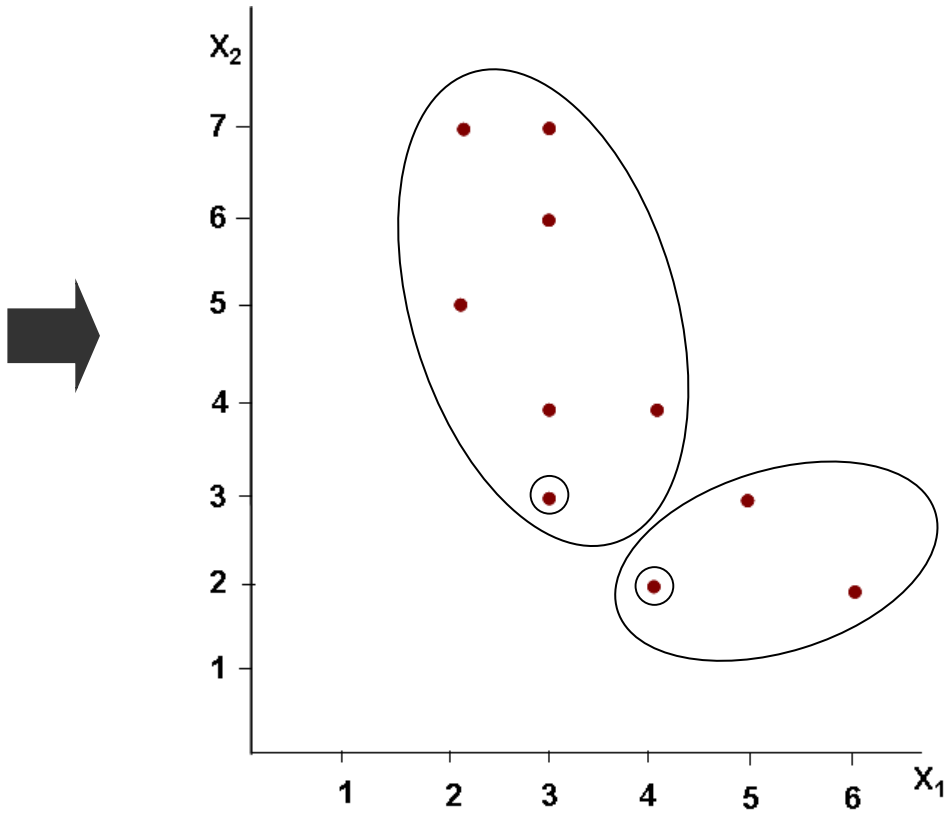
x_1 = ancho del caracter
x_2 = alto del caracter

Determinar 2 grupos en este conjunto de datos

Algoritmo K-medias

K = 2, C = {C_1, C_2}

Inicializar centros: C_1 = (3.3), C2 = (4,2)



Algoritmo K-medias

- Primera iteración:

Paso 1: determinar particiones asignando cada dato a su centro más cercano (medida de distancia = euclídea)

$P_1 = \{\ \}, \ P_2 = \{\ \}$

dato	$d(x^i, C_1)$	$d(x^i, C_2)$
x^1	2.23	3.60
x^2	4.12	5.38
x^3	3.00	4.12
x^4	4.00	5.09
x^5	0.00	1.41
x^6	1.00	2.23
x^7	1.41	0.00
x^8	1.41	2.00
x^9	2.00	1.41
x^{10}	3.16	2.00

$P_1 = \{ x^1, x^2, x^3, x^4, x^5, x^6, x^8\}$

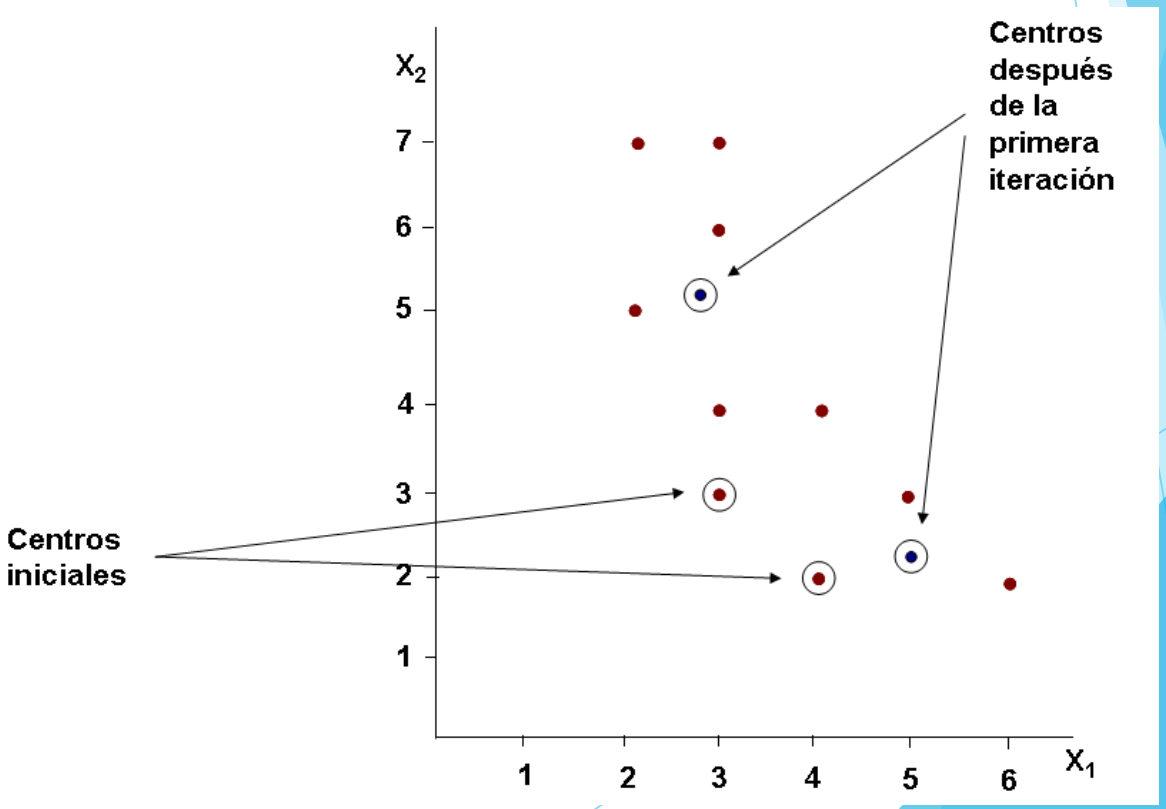
$P_2 = \{ x^7, x^9, x^{10}\}$

Algoritmo K-medias

Paso 2: Recalcular los centros como el punto medio de las particiones

$$C_1 = \frac{1}{7}\{(2,5) + (2,7) + (3,6) + (3,7) + (3,3) + (3,4) + (4,4)\} = (2.85, 5.14)$$

$$C_2 = \frac{1}{3}\{(4,2) + (5,3) + (6,2)\} = (5.00, 2.33)$$



Algoritmo K-medias

Como los centros de la iteración 1 no son iguales a los centros de la iteración anterior, repetir.

- 2da. Iteración:

C_1 = (2.83, 5.50)
C_2 = (4.50, 2.50)

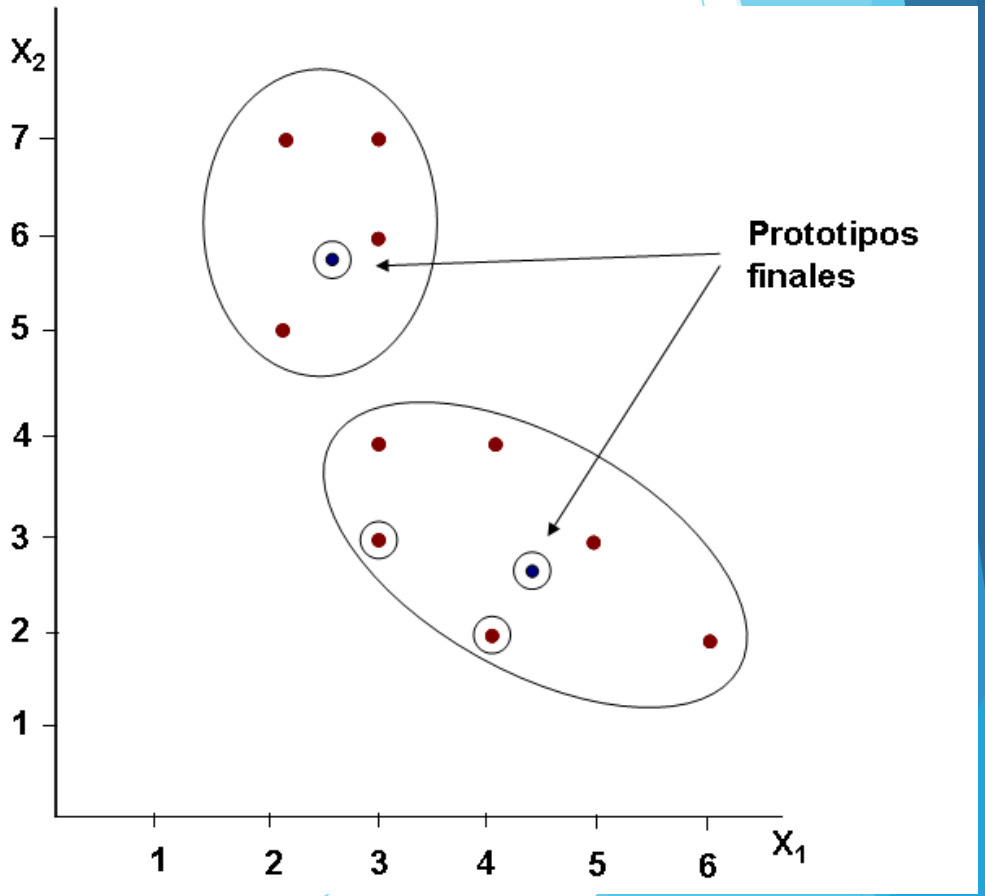
- 3ra. Iteración:

C_1 = (2.60, 5.80)
C_2 = (4.40, 2.80)

- 4ta. Iteración:

C_1 = (2.60, 5.80)
C_2 = (4.40, 2.80)

Convergencia



Algoritmo K-medias

Una vez determinados los grupos, se pueden asignar etiquetas a éstos utilizando conocimiento del dominio. Esto se logra examinando las características de los objetos dentro de cada grupo (por ejemplo: G1 es “b”, G2 es “a”)

¿Cómo clasificar?

- **Cuando llega un nuevo dato, se calcula la distancia de éste a cada centro o prototipo.**
- **Se asigna el nuevo dato al grupo con centro más cercano**

Algoritmo K-medias

Ejemplo:

Si el dato a clasificar es:

Z = (3.5, 4.6)

- Distancia de Z a los prototipos:

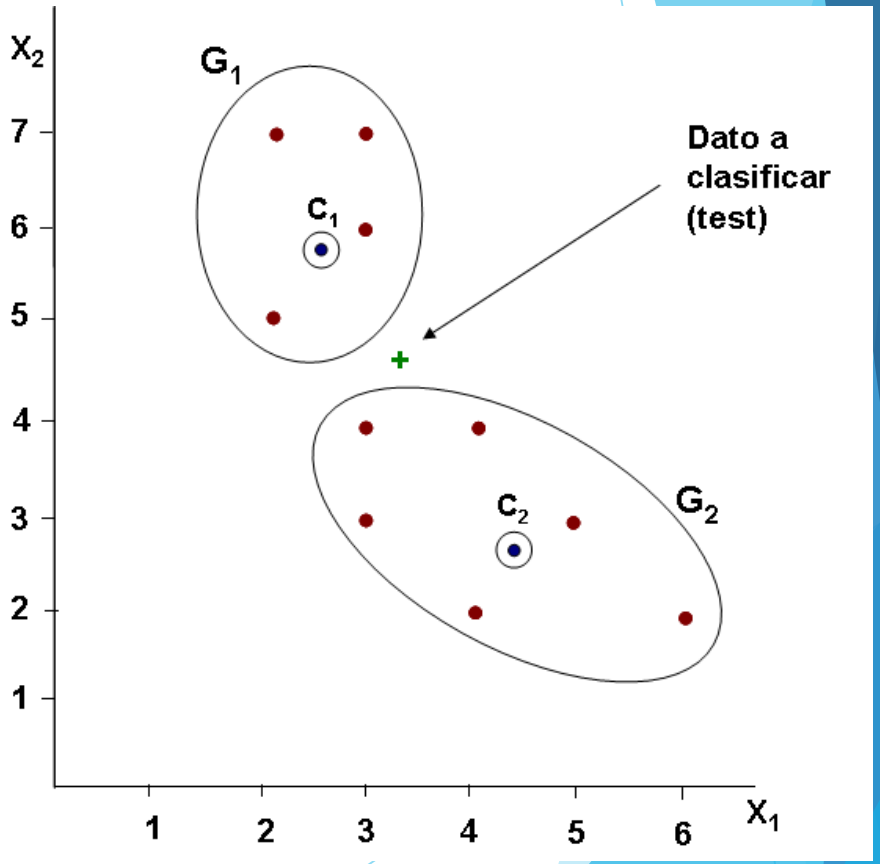
C_1 = (2.60, 5.80)
C_2 = (4.40, 2.80)

$d(z,C_1)$ = 1.50
$d(z,C_2)$ = 2.01

- Se asigna el nuevo dato al grupo con centro más cercano:

Z ⟶ Grupo G_1

➡ Letra b



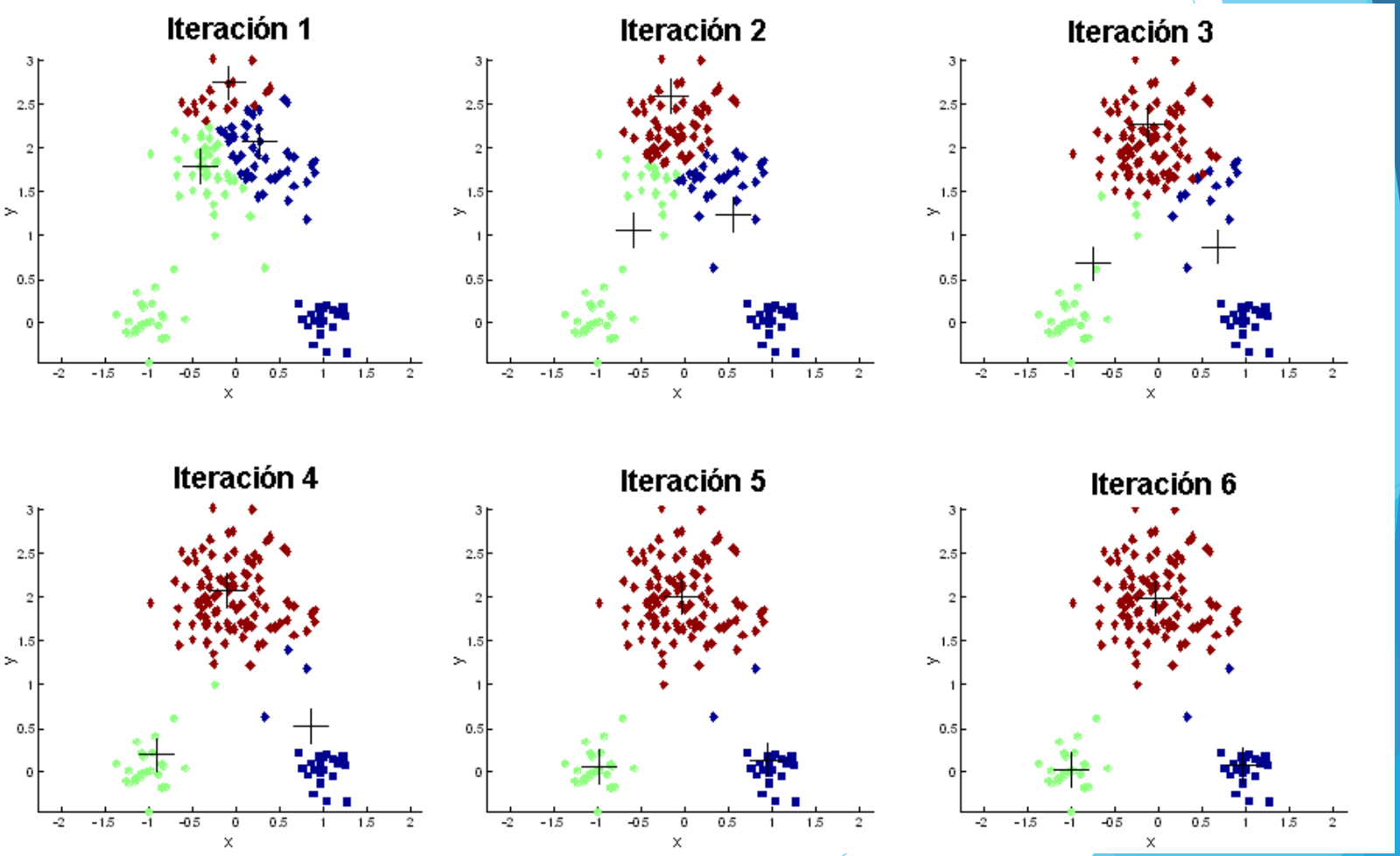
Algoritmo K-medias

Para resumir:

El centroide es la media de los datos en el grupo.
La "cercanía" se puede determinar por medidas de distancia (ejemplo: distancia Euclídea) o de similitud (ejemplo: coseno).
La convergencia puede ser obtenida con pocas iteraciones
Se pueden establecer otros criterios de parada (como el no. de iteraciones).
A priori no se conoce el número de grupos (k)
Como los centros iniciales se seleccionan de manera aleatoria, los grupos pueden variar entre diferentes corridas

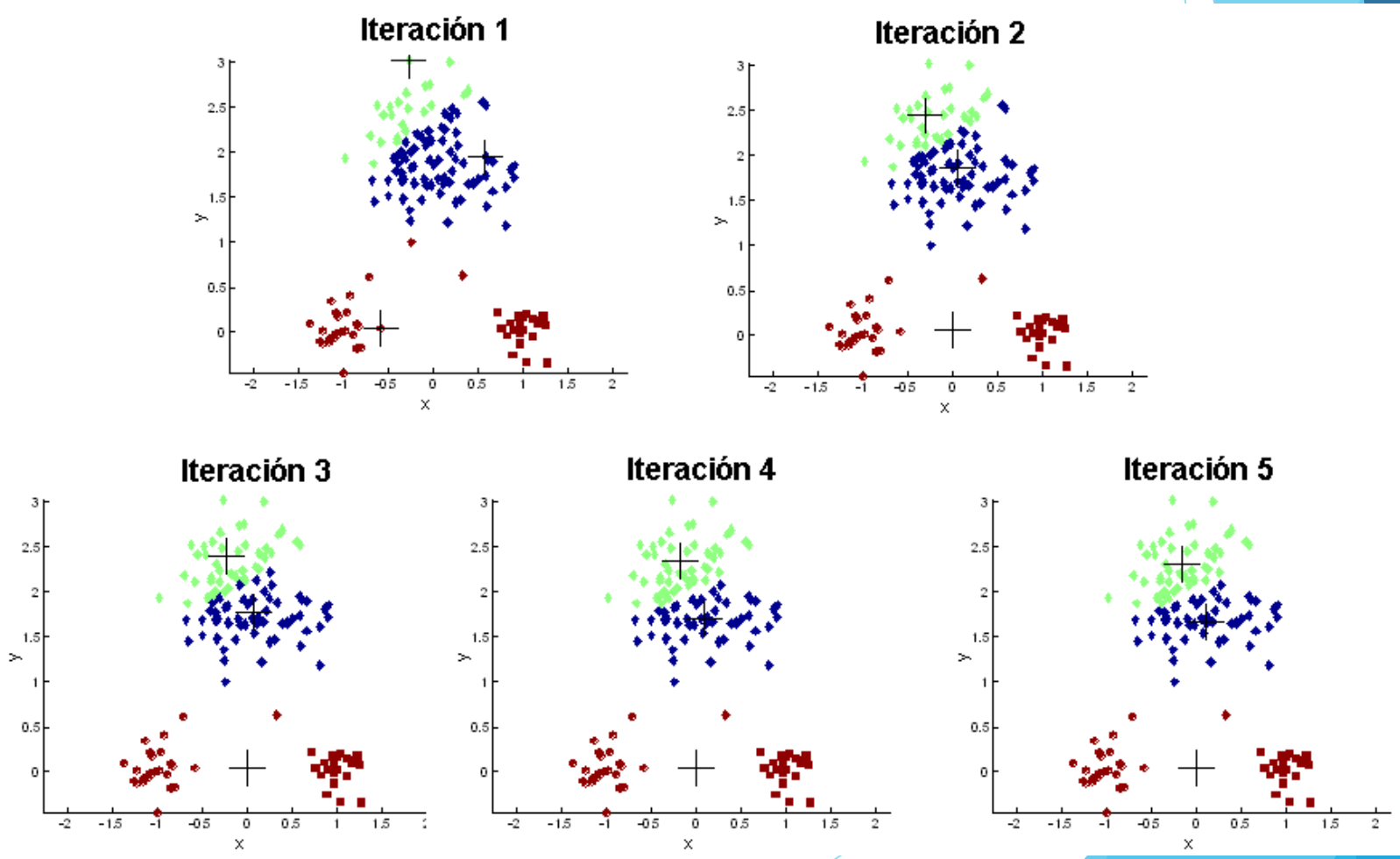
Algoritmo K-medias

Importancia de los centros iniciales:



Algoritmo K-medias

Para el mismo conjunto de datos con otros centros iniciales:



Aplicaciones

Ejemplo: Detección de patrones de comportamiento en sistemas e-learning

La utilización de plaraformas virtuales de enseñanza y aprendizaje, como Moodle, ha alcanzado un gran auge en muchos ámbitos que van desde el estrictamente académico hasta el comercial por su característica principal de alcanzar a aprendices geográficamente distribuidos.

Estas plataformas son capaces de almacenar un informe detallado de las actividades de los usuarios, donde se recogen estadísticas asociadas al uso de las herramientas de colaboración, participación y evaluación activadas en el curso en el que participan (foros, pizarras, chats, Wikis, cuestionarios, entre otros).

Aplicaciones

De cara al profesor, sería de gran utilidad brindarle información acerca de la utilización por parte de los estudiantes del sistema virtual, con la idea de evaluar y posteriormente mejorar el rendimiento de los aprendices en la interacción o participación en los cursos a distancia.

Sin embargo, ¿Cómo modelar y explicar el comportamiento de los usuarios?

Aplicaciones

➡ **Estudio descriptivo del nivel de interacción de los estudiantes de postgrado en la plataforma Moodle de la Facultad de Ciencias**
(Luis Arredondo)

PROBLEMA:

☞**¿Cómo los estudiantes utilizan estas herramientas?**

☞**¿Este uso puede dar evidencias de un perfil de los estudiantes?**

Objetivo: ***Determinar diferentes perfiles de los estudiantes de postgrado en cursos dictados a través de la plataforma Moodle***

Tarea de minería de datos: Agrupación

Aplicaciones

Recolección de los datos:

Para cada usuario matriculado en un curso dado, se logró recopilar su registro de actividades en la plataforma Moodle.

Preparación de los datos:

Se construyeron variables en función de los porcentajes de utilización en Foros, Tareas, Recursos, Chats, SCORMS, Wiki y Cuestionario.

Se encontraron ausencias en algunas variables; sin embargo, tenían una interpretación asociada con la interacción del estudiante en la plataforma (no utilizó el recurso).

Selección de variables: Algoritmo InfoGainAttributeEval con el método Ranker. Se seleccionaron las variables:

%Foros

%Tareas

%Recursos

%Chats

Aplicaciones

Minería de datos:

Tarea de minería de datos: Agrupación

Lenguaje de representación: no es un requerimiento

Algoritmo: K - medias

Medida de rendimiento: Suma de las distancias al cuadrado

Aplicaciones

Se seleccionó el valor de k=4. Esta decisión se tomó en virtud de balancear el error obtenido y la excesiva segmentación del espacio de variables, lo cual generaría muchos grupos formados por muy pocos individuos dificultándose así la interpretación de la información.

Los cuatros grupos encontrados fueron:

Cluster	%Foro		%Tareas		%Recursos		%Chats	
	Media	Desviación Estándar	Media	Desviación Estándar	Media	Desviación Estándar	Media	Desviación Estándar
0	0.0421	0.086	0.892	0.124	0.7689	0.2636	0.4671	0.1247
1	0.1991	0.2622	0.1761	0.2673	0.186	0.1838	0.3579	0.277
2	0.0733	0.1723	0.2972	0.2203	0.6933	0.1271	0.5271	0.175
3	0.5425	0.1949	0.7685	0.1903	0.7469	0.198	0.5447	0.2662

Aplicaciones

Para realizar la interpretación de los grupos, se efectuó una discretización de las variables. A cada intervalo se le asignó una etiqueta asociada al nivel de uso de la herramienta. La discretización se realizó con intervalos equidistantes.

Intervalo	Etiqueta
[0, 0.2]	Muy bajo
(0.2, 0.4]	Bajo
(0.4, 0.6]	Medio
(0.6, 0.8]	Alto
(0.8, 1]	Muy alto

La interpretación que se le dio a cada grupo fue la siguiente:

- **Grupo 1: muy bajo foros, muy alto tareas, alto recursos y medio chats (*reactivos*).**
- **Grupo2: muy bajo foros, muy bajo tareas, de igual manera, muy bajo recursos y bajo chats (*no participan*).**
- **Grupo3: muy bajo foros, bajo tareas, alto recursos y medio chats (*descargan*).**
- **Grupo4: medio foros, alto tareas, alto recursos y medio chats (*modelo*).**

Se logró determinar cuatro (4) tipos de usuarios, cuyos centroides permiten caracterizarlos de acuerdo al uso de las herramientas presentes en los cursos en Moodle