# Minería de Datos

## Técnicas de Minería de datos – Agrupación

## Agenda:

- **Tareas descriptivas**
- **Agrupación**
- **Algoritmo K-medias**

# *Tareas descriptivas*

***Conjunto de datos:***

| Atributo 1 | Atributo 2 | ..... | Atributo d |
|---|---|---|---|
| | | | |
| | | | |
| | | | |
| | | | |

***No existe o no se toma en cuenta una salida asociada a cada instancia***

**Modelos descriptivos**

- **Derivan patrones (correlaciones, grupos, anomalías, trayectorias) que sumarizan las relaciones fundamentales de los datos.**
- **Identifican patrones que explican o describen los datos.**
- **Sirven para explorar las propiedades de los datos**

# *Tareas descriptivas*

**Tareas de la minería de datos**

**Tareas predictivas**

**Clasificación**
**Regresión**

***Aprendizaje supervisado*:**

**Se dispone de un atributo que representa la respuesta del problema**

**Tareas descriptivas**

**Agrupación (clustering)**
**Análisis de asociación (descubrimiento de reglas de asociación)**
**Detección de anomalías**

***Aprendizaje no supervisado:***

**La respuesta del problema no está dada**

# *Agrupación*

***Una de las tareas más frecuentes en minería da datos***

**Divide los datos en grupos (*clusters*).**
**Los grupos deberían capturar la estructura natural de los datos**
**En un grupo los objetos comparten características comunes**

**Análisis de grupos:**

**Técnicas que automáticamente encuentran grupos en los datos.**

**Los datos son agrupados utilizando sólo la información que describe a los objetos y sus relaciones**

**Objetos dentro de un grupo deben ser similares (o estar relacionados), y diferentes (o no relacionados) a objetos de otros grupos**

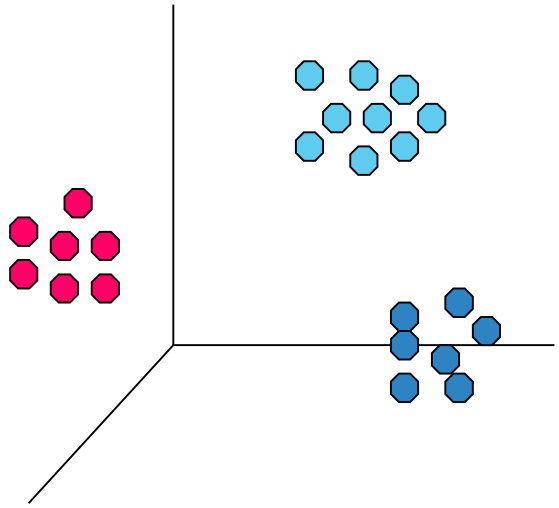
# *Agrupación*

**- Dada una colección de registros o instancias**

***Donde cada registro tiene asociado un conjunto de atributos, no hay salida definida.***

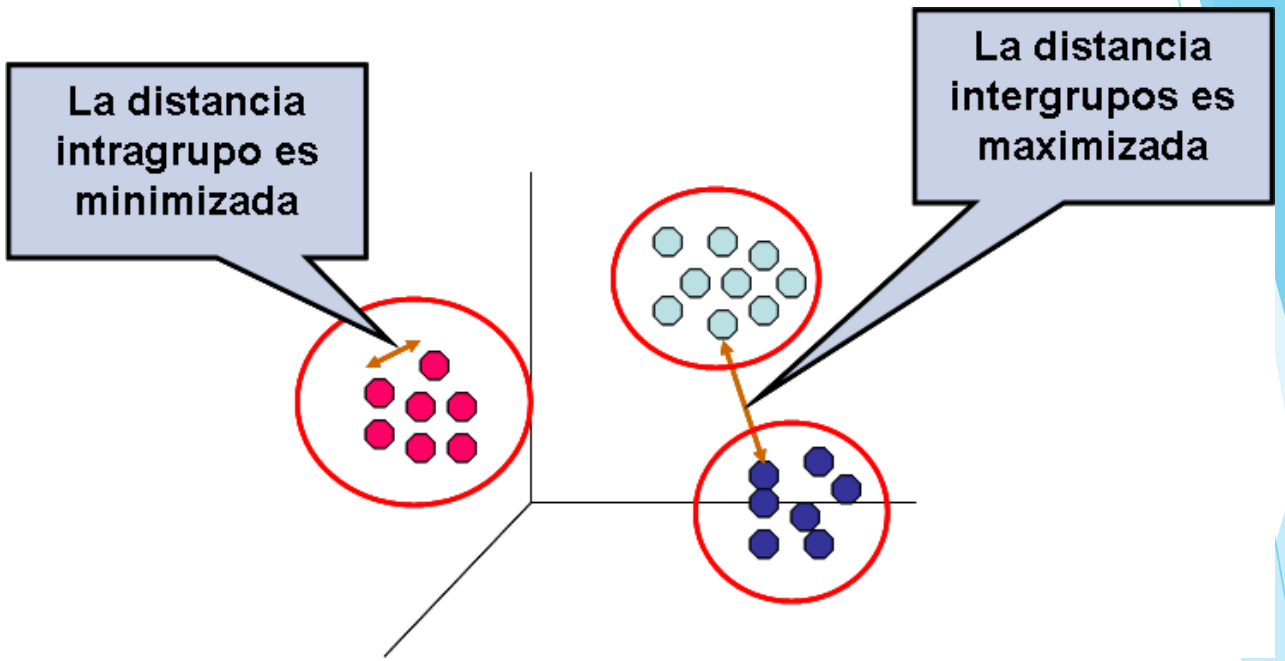
**- Encontrar grupos naturales a partir de los datos**

**- Objetivo: Los objetos de un grupo son muy similares entre sí y muy diferentes a los objetos de otros grupos.**



***Se utilizan medidas de similitud, que dependerán del tipo de variable presente en el conjunto de datos***

# *Agrupación*




***A mayor la similitud (u homogeneidad) dentro de un grupo y mayor la diferencia entre grupos, mejor será la agrupación.***

***La agrupación o clustering puede ser considerada como una forma de clasificación, pero no supervisada***

# *Agrupación*

***¿Para qué? Ejemplos de aplicaciones:***

- **En medicina: Encontrar grupos de pacientes con síntomas similares**
- **En economía: Determinar los países con economías similares**
- **En mercadeo: Encontrar clientes con comportamientos de compra similares**
- **En aplicaciones de recuperación de documentos: Encontrar documentos con contenidos relacionados.**
- **En e-learning: Determinar grupos de estudiantes con estilos de aprendizaje similares**

→ ***En general, es posible encontrar o descubrir patrones de comportamiento en cualquier área.***

***La agrupación también ayuda a sumarizar los datos***

# *Agrupación*

***¿Qué no es agrupación? Ejemplos:***

**Aprendizaje supervisado**

***Se tiene información de la etiqueta de clase***

**Segmentación simple**

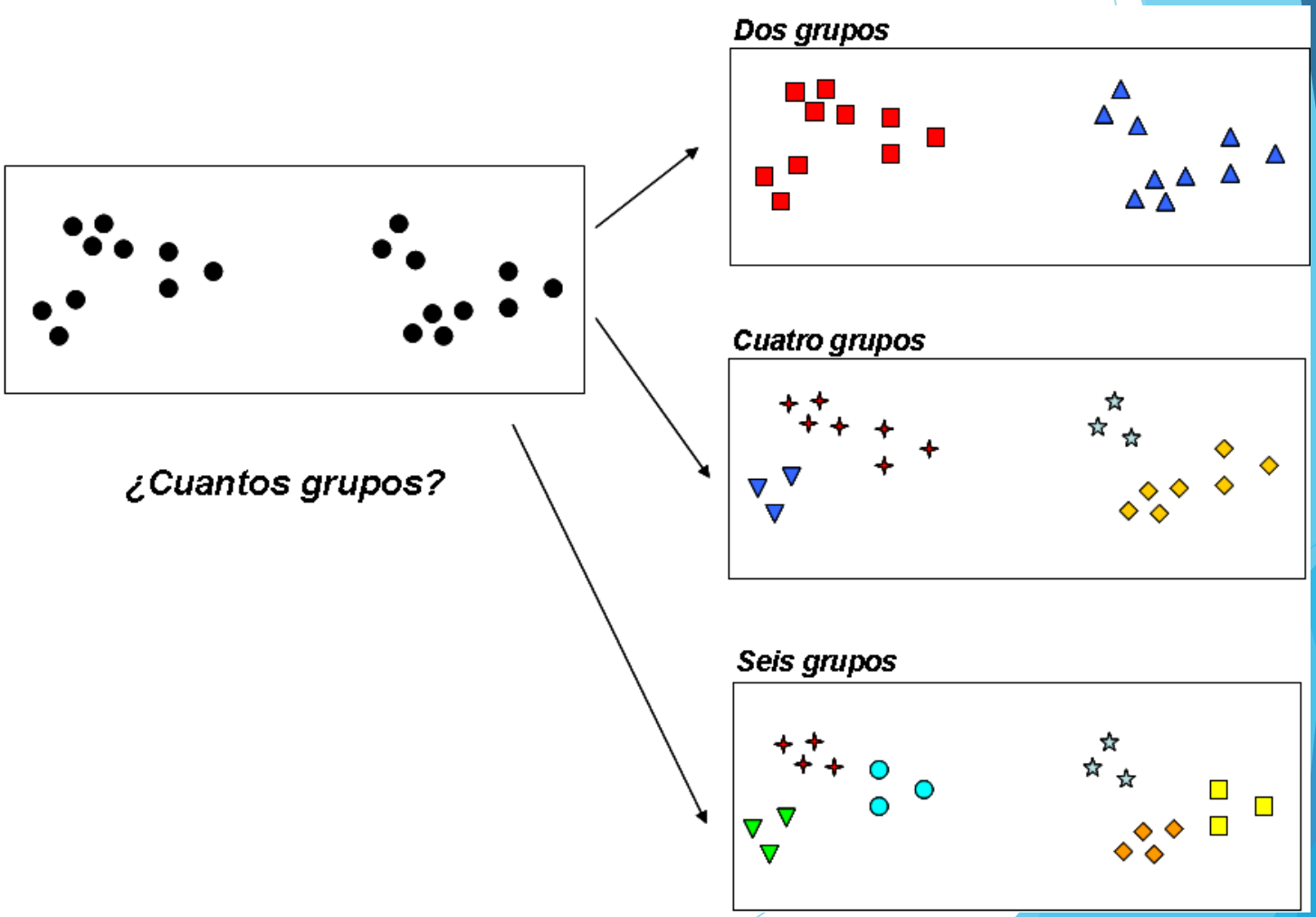
***Por ejemplo, asignar estudiantes a grupos de registros de manera alfabética según el apellido***

**Resultados de un query**

***Los grupos son el resultado de una especificación externa***

# Agrupación

***Importante: la noción de un grupo puede ser ambigua***



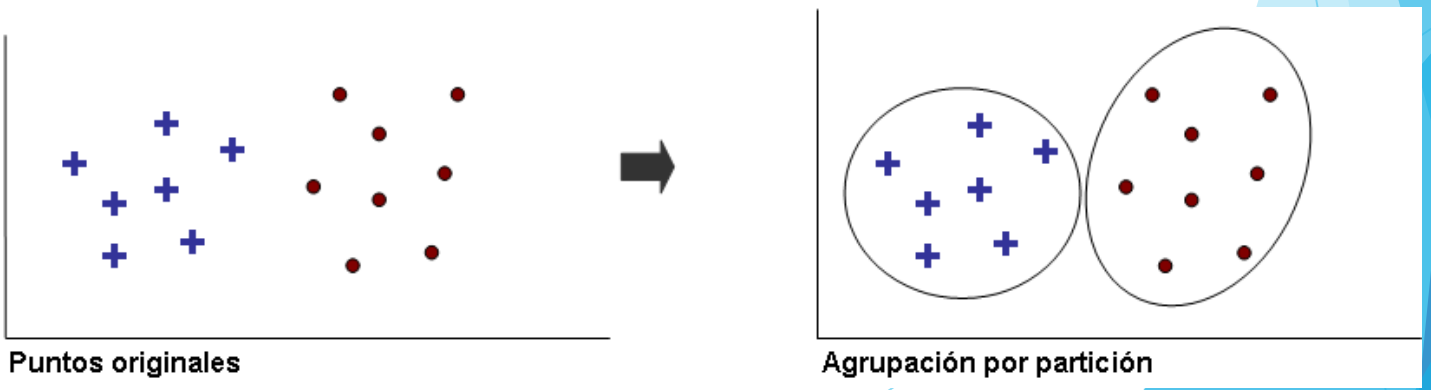
# *Agrupación*

***- ¿Cómo pueden ser las agrupaciones?***

***Una agrupación es un conjunto de grupos***
***Las agrupaciones pueden ser (entre otras):***

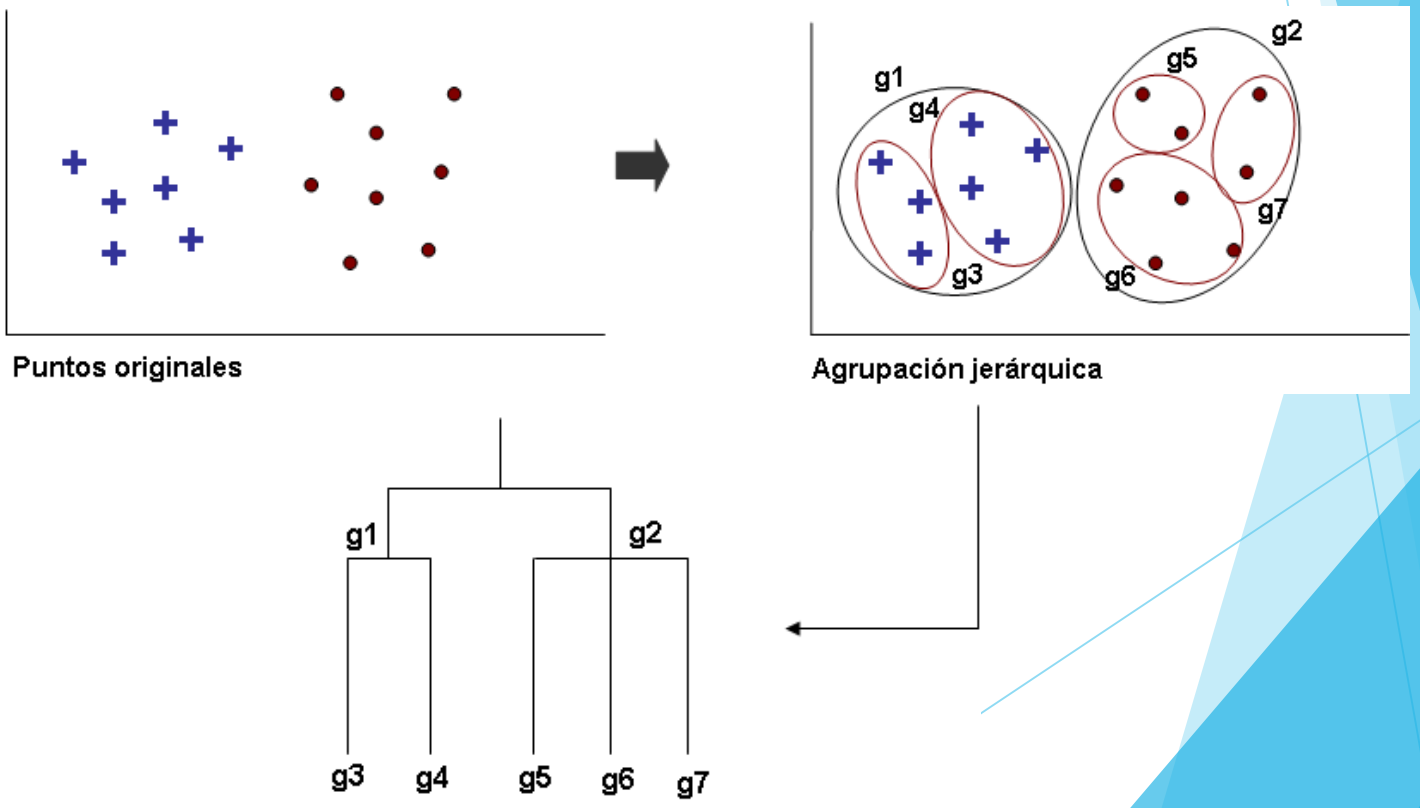
## a) Por particiones o jerárquico

***<u>Por particiones:</u>*** **Se divide el conjunto de datos en grupos no solapados, de tal forma que cada objeto se encuentra en un solo grupo.**



# *Agrupación*

***Jerárquico***: **Conjuntos de grupos anidados que pueden organizarse como un árbol jerárquico (los grupos tienen subgrupos ). Cada grupo en el árbol, excepto para los nodos hojas, es la unión de sus hijos y la raíz de árbol es el grupo que contiene a todos los objetos.**



## b) Exclusivo, solapado o difuso (*fuzzy*)

***Exclusivo***: **Cada objeto se asigna a un solo grupo.**

***Solapado***: **Un objeto puede pertenecer simultáneamente a más de un grupo. Ejemplo: una persona en un universidad puede ser estudiante y empleado a la vez.**

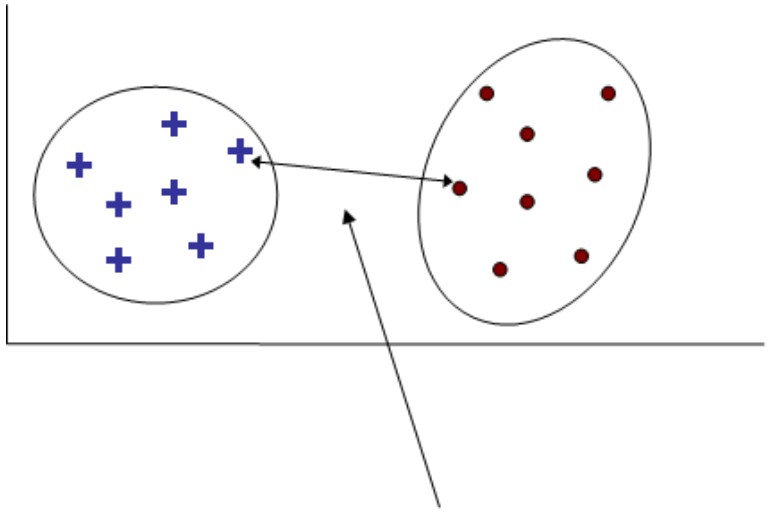
***Difuso***: **Cada objeto pertenece a un grupo o a varios grupos, pero con un grado de pertenencia que está entre 0 y 1. Los grupos pueden ser tratados como conjuntos difusos.**

# Agrupación

- *¿Cómo pueden ser los grupos?*

## a) Grupos bien separados:

***Conjunto de objetos donde cada dato está más cercano a todos los objetos de su grupo, que a cualquier otro dato de otro grupo***

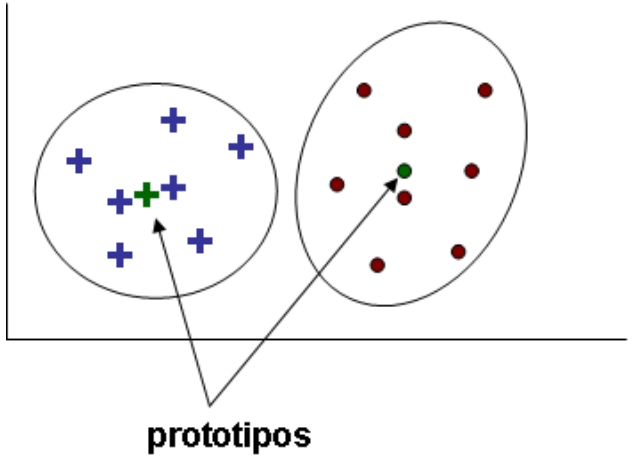


***La distancia entre dos puntos de diferentes grupos es mas grande que la distancia entre cualesquiera dos puntos de un grupo***

***Definición idealista que sólo se satisface cuando los datos contiene grupos naturales que son bastante diferentes***

## b) Grupos basados en prototipos:

***Conjunto de objetos en el cual cada dato está más cercano al prototipo que define o representa al grupo, que al prototipo de cualquier otro grupo***




***Cada dato está más cercano al centro de su grupo que al centro de cualquier otro grupo***

***Para datos continuos:***
***Prototipo (también llamado centroide) = Punto medio de todos los puntos dentro del grupo***

***Para datos nominales:***
***Prototipo (también llamado medoide) = Punto más representativo del grupo***

# *Agrupación*

**También se pueden encontrar grupos que comparten propiedades = grupos conceptuales.**

**Se necesita un concepto muy específico para poder encontrar con éxito los grupos.**

**El proceso de encontrar estos grupos se conoce como *clustering* conceptual**

☞ ***Independiente del tipo de agrupación o el tipo de grupos es necesaria una medida de similitud, que dependerán del tipo de variable presente en el conjunto de datos***

☞ ***La agrupación (clustering) ayuda a encontrar grupos útiles de objetos, donde la utilidad se define en función del objetivo del análisis de datos***

# *Agrupación*

*- ¿Cómo evaluar una agrupación?*

**Una de las medidas más utilizadas es la suma de las distancias al cuadrado:**

$$SSE = \sum_{i=1}^{K} \sum_{X \in G_i} d^2(m_i, x)$$

**Donde x es un dato en el grupo $G_i$ y $m_i$ es el punto o dato más representativo del grupo $G_i$**

**$m_i$ se corresponde con el centro, prototipo o media del grupo.**
**Entonces, dadas dos agrupaciones, se seleccionará aquella con el error más pequeño.**
**También es posible utilizar otras medidas basadas en la cohesión y separación entre grupos**

# *Agrupación*

***- ¿Cómo determinar un modelo de agrupación?***

**Algunas técnicas:**

- ☞ **Métodos basados en prototipos (ejemplo: algoritmo k-medias)**
- ☞ **Métodos jerárquicos (ejemplo: algoritmo jerárquico por aglomeración)**
- ☞ **Redes neuronales**
- ☞ **Métodos estadísticos**
- ☞ **Algoritmos genéticos**
- ☞ **Otros**

# *Algoritmo K-medias*

**Técnica de *clustering* por partición basada en prototipos, que intenta encontrar un número K de grupos (especificado por el usuario).**

**Uno de los métodos más utilizados**

**Puede ser aplicado a un amplio rango de datos, ya que requiere sólo una medida de proximidad entre un par de objetos**

**Versiones:**

***a) K- medias:*** ***Define un prototipo en términos de un centroide, el cual generalmente es la media de un conjunto de puntos.***

***El prototipo casi nunca corresponde a un punto actual.***

***Se aplica a espacios continuos n-dimensionales***

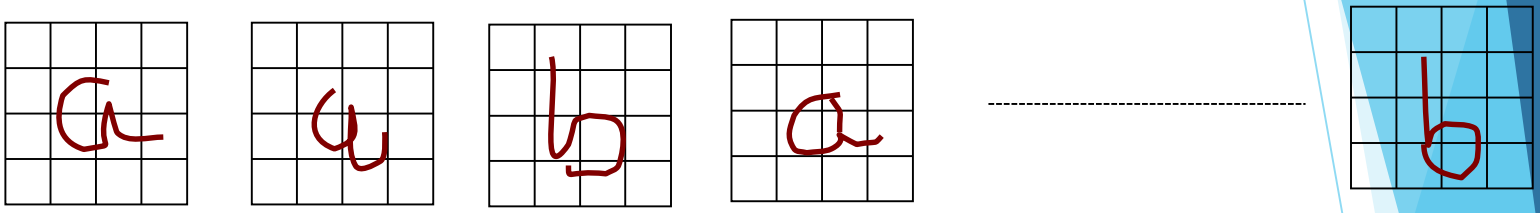
***b) K – medoide:*** ***Define un prototipo en términos de un medoide, el cual es el punto más representativo de los datos para un grupo de puntos.***

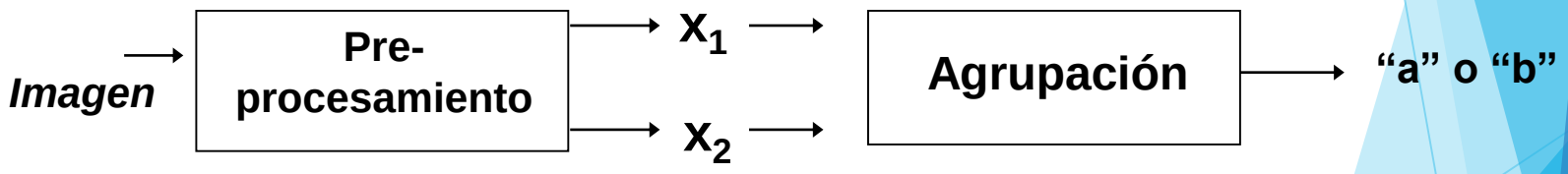
***El prototipo, por definición, corresponde a un objeto del conjunto de datos***

# *Algoritmo K-medias*

**Ejemplo:** **Reconocimiento de caracteres manuscritos a partir de imágenes**

**Conjunto de datos:**






**Donde:** $x_1$ **= ancho del caracter**

$x_2$ **= alto del caracter**

# 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

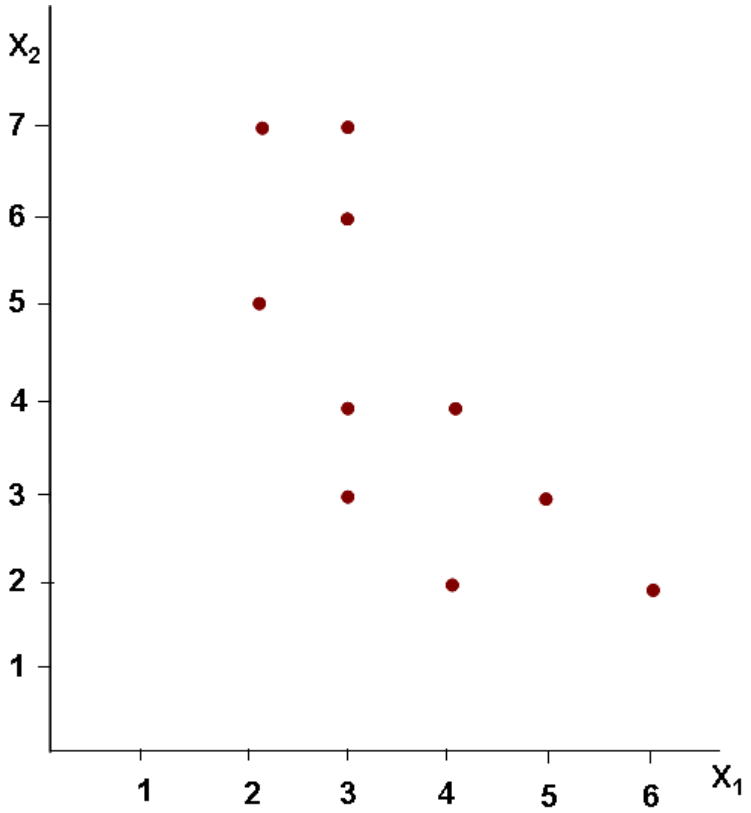
Hasta que los centros no cambien (o se alcance otra condición de parada)

{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)}**



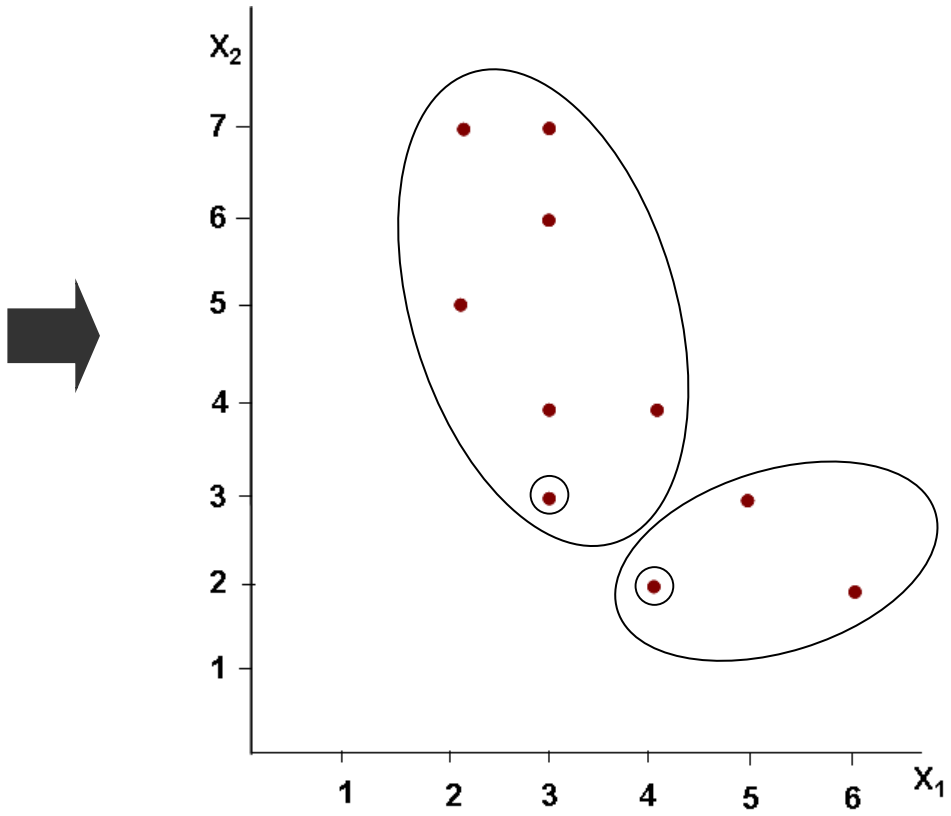

***$x_1$ = ancho del caracter***
***$x_2$ = alto del caracter***

**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 |

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

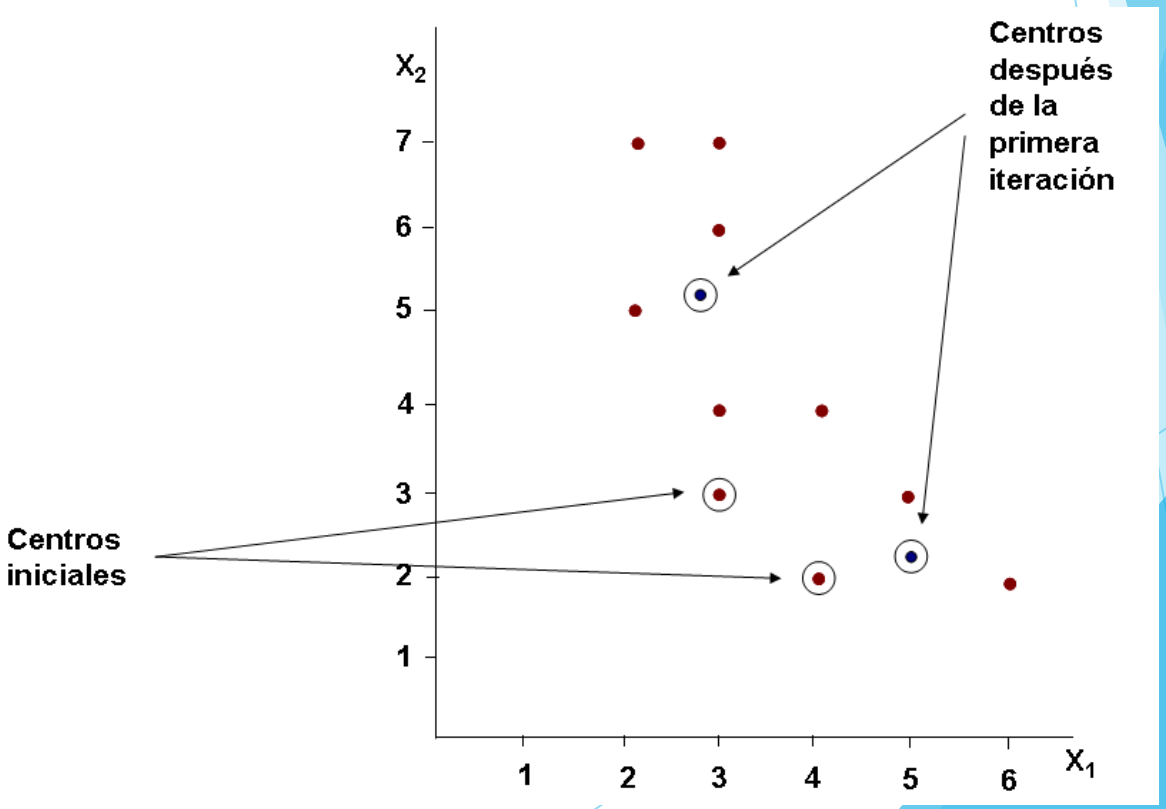
$P_2 = \{ x^7, x^9, x^{10}\}$

# 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***