

# Aprendizaje no supervisado

## VC02: Agrupamiento por particiones: K-means

Félix José Fuentes Hurtado

[felixjose.fuentes@campusviu.es](mailto:felixjose.fuentes@campusviu.es)

Universidad Internacional de Valencia

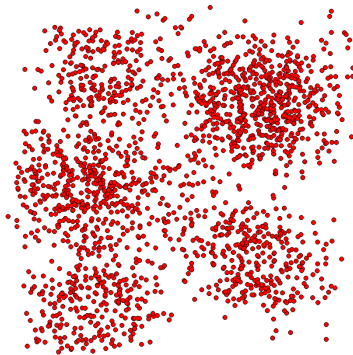
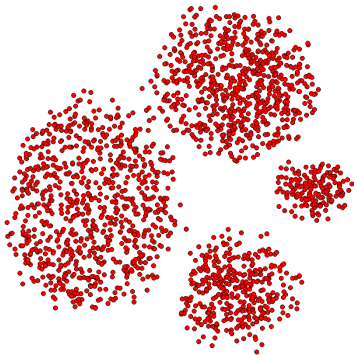
## Definición

Dado un conjunto de datos, el agrupamiento trata de identificar subgrupos homogéneos de ejemplos que manifiestan diferencias relevantes con los otros subgrupos que se formen.

# Agrupamiento

## Definición

Dado un conjunto de datos, el agrupamiento trata de identificar subgrupos homogéneos de ejemplos que manifiestan diferencias relevantes con los otros subgrupos que se formen.



## Definición

Dado un conjunto de datos, el agrupamiento trata de identificar subgrupos homogéneos de ejemplos que manifiestan diferencias relevantes con los otros subgrupos que se formen.

- ▶ Vectores descriptores de los ejemplos
- ▶ Conjunto de datos
- ▶ No existe ninguna variable “especial” respuesta
- ▶ Formar grupos:
  - \* No se conoce el número de grupos
  - \* No se conocen las pertenencias de ejemplos a grupos

Dos instrucciones:

## Definición

Dado un conjunto de datos, el agrupamiento trata de identificar **subgrupos homogéneos** de ejemplos que manifiestan **diferencias relevantes con los otros subgrupos** que se formen.

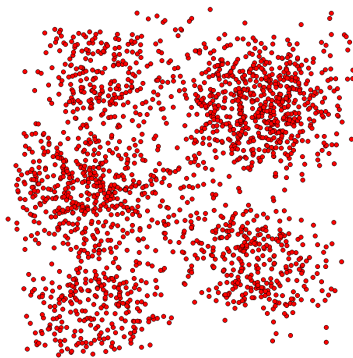
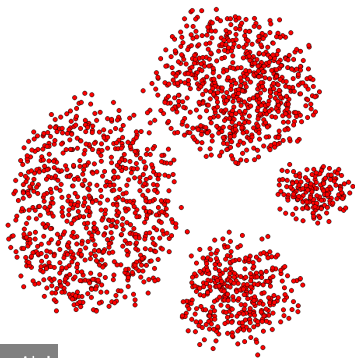
- ▶ Dispersión intraclúster
- ▶ Dispersión interclúster

# Agrupamiento

Dos instrucciones:

## Definición

Dado un conjunto de datos, el agrupamiento trata de identificar **subgrupos homogéneos** de ejemplos que manifiestan **diferencias relevantes con los otros subgrupos** que se formen.



## Objetivo

Encontrar un agrupamiento que maximice la dispersión interclúster y minimice la dispersión intraclúster:

- Dispersión intraclúster

$$I(C) = \frac{1}{2} \sum_{k=1}^K \sum_{i: C(x_i)=k} \sum_{i': C(x_{i'})=k} d(x_i, x_{i'})$$

- Dispersión interclúster

$$O(C) = \frac{1}{2} \sum_{k=1}^K \sum_{i: C(x_i)=k} \sum_{i': C(x_{i'}) \neq k} d(x_i, x_{i'})$$

## Objetivo

Encontrar un agrupamiento que maximice la dispersión interclúster y minimice la dispersión intraclúster:

¡Ambos objetivos son equivalentes!



## Objetivo

Encontrar un agrupamiento que maximice la dispersión interclúster y minimice la dispersión intraclúster:

$$T = \frac{1}{2} \sum_{i=1}^n \sum_{i'=1}^n d(x_i, x_{i'})$$

## Objetivo

Encontrar un agrupamiento que maximice la dispersión interclúster y minimice la dispersión intraclúster:

$$T = \frac{1}{2} \sum_{k=1}^K \sum_{i: C(x_i)=k} \left( \sum_{i': C(x_{i'})=k} d(x_i, x_{i'}) + \sum_{i': C(x_{i'}) \neq k} d(x_i, x_{i'}) \right)$$

## Objetivo

Encontrar un agrupamiento que maximice la dispersión interclúster y minimice la dispersión intraclúster:

$$T = I(C) + O(C)$$

$$\arg \min_C I(C) = \arg \min_C T - O(C) = \arg \max_C O(C)$$

El objetivo es buscar el mejor agrupamiento  $C$  que maximiza (minimiza) la dispersión intraclúster (interclúster)

$$\arg \min_C I(C) = \arg \min_C T - O(C) = \arg \max_C O(C)$$

El objetivo es buscar el mejor agrupamiento  $C$  que maximiza (minimiza) la dispersión intraclúster (interclúster)

$$\arg \min_C I(C) = \arg \min_C T - O(C) = \arg \max_C O(C)$$

¡Número inabarcable de posibles combinaciones!

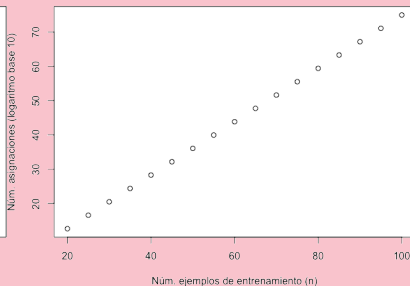
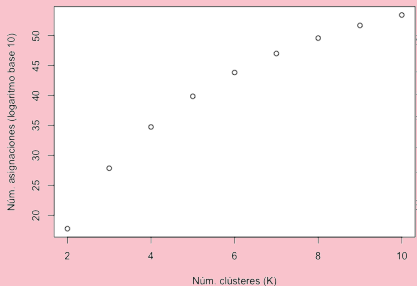
$$S(n, K) = \frac{1}{K!} \sum_{k=1}^K (-1)^{K-k} \binom{K}{k} k^n$$

# Agrupamiento

El objetivo es buscar el mejor agrupamiento  $C$  que maximiza (minimiza) la dispersión intraclúster (interclúster)

$$\arg \min_C I(C) = \arg \min_C T - O(C) = \arg \max_C O(C)$$

¡Número inabarcable de posibles combinaciones!



¡Necesario recurrir a heurísticas de búsqueda!

¡Necesario recurrir a heurísticas de búsqueda!

## Heurística

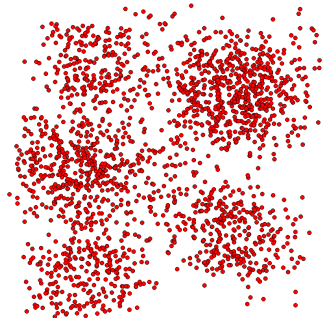
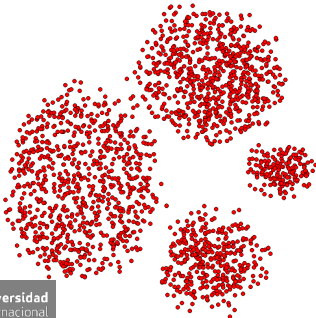
En informática, se trata de técnicas diseñadas para resolver un problema de manera rápida cuando la aproximación exhaustiva es muy lenta y/o para encontrar una solución aproximada cuando encontrar la solución exacta es muy difícil o imposible.

Se puede expresar como un *trade-off* (balance) entre velocidad y optimalidad-completitud.



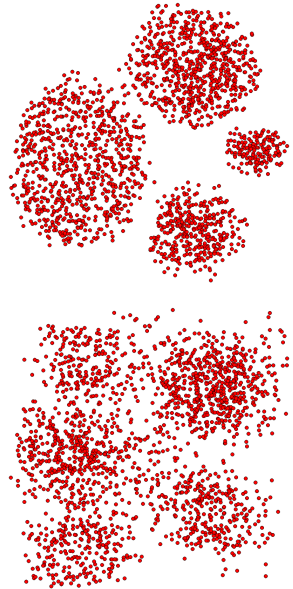
## Heurísticas de búsqueda del mejor agrupamiento

1. Encontrar un agrupamiento válido
2. Plantear diferentes alternativas a ese agrupamiento
3. Escoger la mejor alternativa
4. Volver al paso 2



## Tipos de algoritmos de agrupamiento

- ▶ Basados en particiones
- ▶ Jerárquicos
- ▶ Espectrales
- ▶ Basados en densidad
- ▶ Probabilísticos



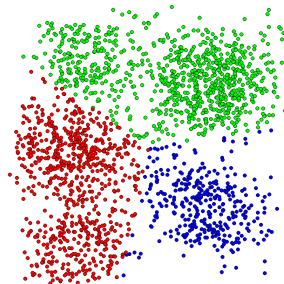
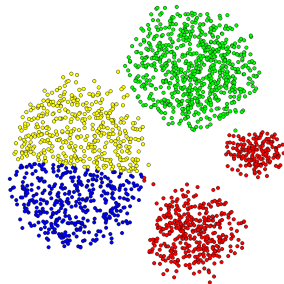
# Agrupamiento basado en particiones

## Búsqueda de la mejor partición de los datos

Se particiona el dataset según criterios basados en distancia

\*\* Uso de centro(ide)s

\*\* ¿Fijar el número de clústeres (K)?



# Agrupamiento basado en particiones

## Búsqueda de la mejor partición de los datos

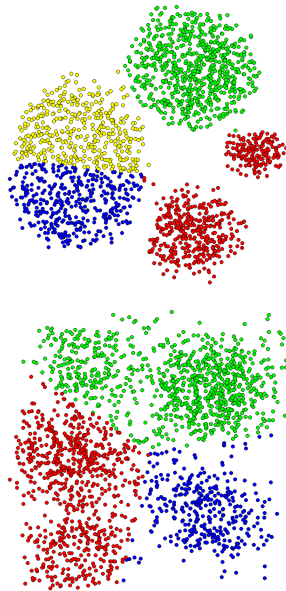
Se particiona el dataset según criterios basados en distancia

\*\* Uso de centro(ide)s

\*\* ¿Fijar el número de clústeres (K)?

Algoritmos:

- ▶  $K$ -means
- ▶  $K$ -medoids



# Agrupamiento basado en particiones

## K-means

### Intuición

Los clústeres homogéneos se agrupan alrededor de un centro. Por lo tanto, se puede calcular:

1. **Centro:** El centro de un clúster es la medio de los elementos que pertenecen al él
2. **Pertenencia:** Un ejemplo pertenece al clúster cuyo centro le es más cercano

**\*\*** La combinación de ambos conceptos permite construir el agrupamiento

### Dispersión intraclúster

$$I(C) = \sum_{k=1}^K N_k \cdot \sum_{x_i: C(x_i)=k} \|x_i - \bar{x}_k\|^2$$

### Objetivo a minimizar

$$\arg \min_{C: (\bar{x}_1, \dots, \bar{x}_K)} I(C)$$

# Agrupamiento basado en particiones

K-means

## Heurística

Partiendo de un conjunto de centros aleatorio, buscar la pertenencia más probable de los ejemplos a los clústeres y obtener un nuevo conjunto de centros (agrupamiento)

¡Naturaleza iterativa!

# Agrupamiento basado en particiones

## K-means

---

### K-means

Recibe: Conjunto de entrenamiento,  $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ ; número de clústeres,  $K$

- 
1. Elección (aleatoria) de  $K$  puntos del conjunto de entrenamiento como centros,  $\{\bar{\mathbf{x}}_1, \dots, \bar{\mathbf{x}}_K\}$ .
  2. Asignar cada ejemplo  $\mathbf{x}_i$  al clúster del centro más cercano:  
$$C(\mathbf{x}_i) = \operatorname{argmin}_{k \in \{1, \dots, K\}} \|\mathbf{x}_i - \bar{\mathbf{x}}_k\|^2$$
  3. Para cada clúster  $k$ , recalcular su centro:  $\bar{\mathbf{x}}_k = \operatorname{argmin}_{\mathbf{x}} \sum_{\mathbf{x}_i: C(\mathbf{x}_i)=k} \|\mathbf{x}_i - \mathbf{x}\|^2$
  4. Los pasos 2 y 3 se iteran hasta que los centros no cambian.

---

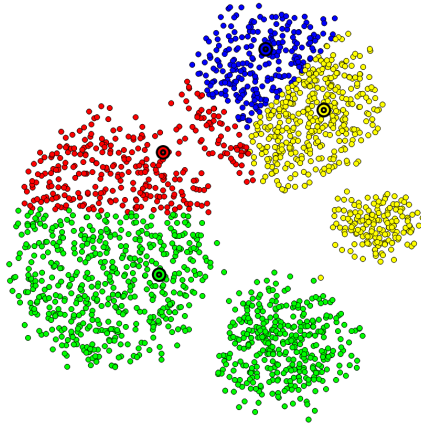
Devuelve: Conjunto de centros,  $\{\bar{\mathbf{x}}_1, \dots, \bar{\mathbf{x}}_K\}$ ; Asignación,  $C$

---



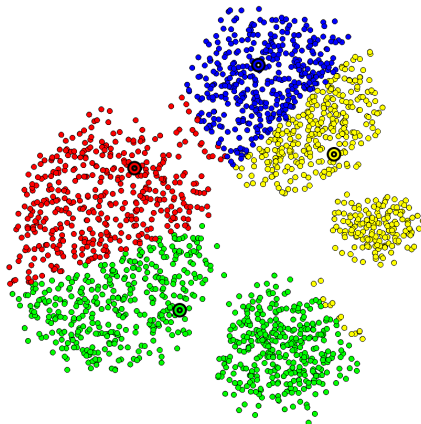
# Agrupamiento basado en particiones

K-means



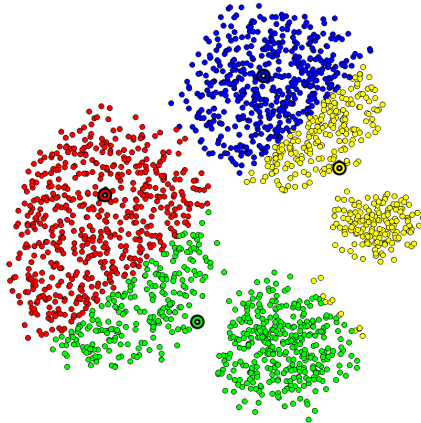
# Agrupamiento basado en particiones

K-means



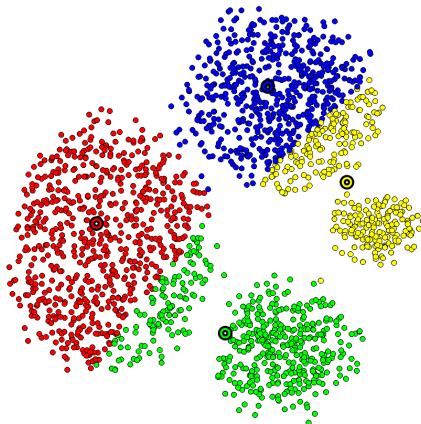
# Agrupamiento basado en particiones

K-means



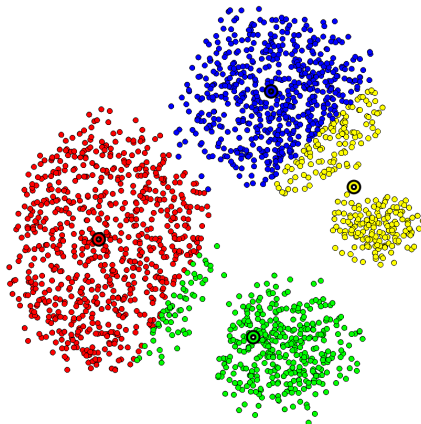
# Agrupamiento basado en particiones

K-means



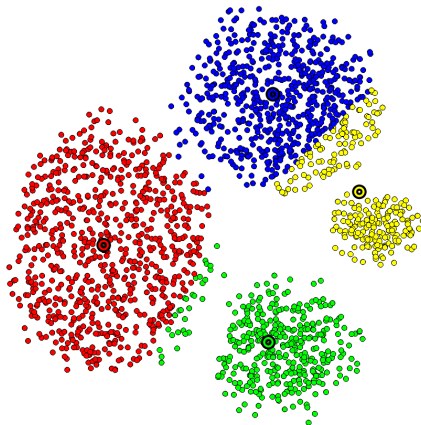
# Agrupamiento basado en particiones

K-means



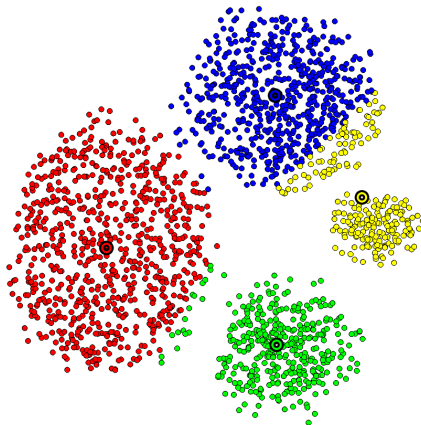
# Agrupamiento basado en particiones

K-means



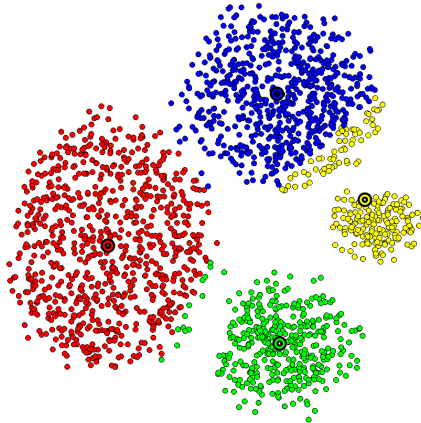
# Agrupamiento basado en particiones

K-means



# Agrupamiento basado en particiones

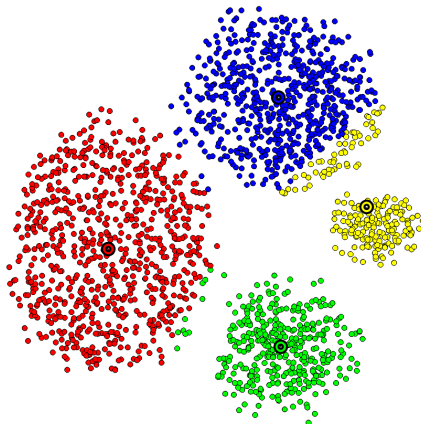
K-means





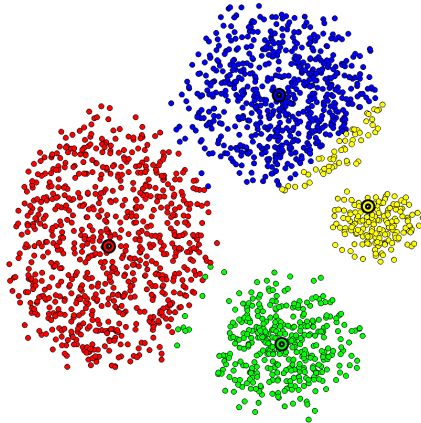
# Agrupamiento basado en particiones

K-means



# Agrupamiento basado en particiones

K-means



# Agrupamiento basado en particiones

## K-means

### Ventajas

- ▶ Intuitivo
- ▶ Rápido
- ▶ Sencillo
- ▶ Mejorable

¡Algoritmo de *clustering* más popular!

### Desventajas

- ▶ El número de clústeres es un parámetro ( $K$ )
- ▶ Dependencia de la inicialización
- ▶ Dependencia de los *outliers*
- ▶ Funciona con variables descriptivas continuas
- ▶ Dificultades cuando los clústeres son de diferente tamaño y/o densidad, o no son convexos

# Agrupamiento basado en particiones

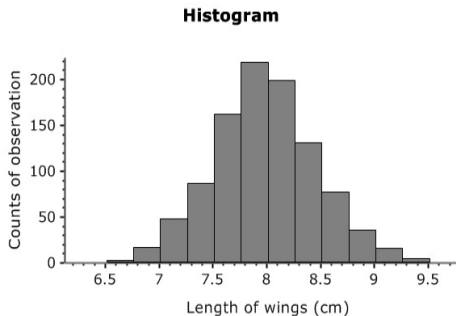
## K-means

### Desventajas

- ▶ El número de clústeres es un parámetro ( $K$ )  
*Validación cruzada*
- ▶ Dependencia de la inicialización  
*Múltiples ejecuciones del algoritmo, K-means++*
- ▶ Dependencia de los *outliers*  
*Preproceso*
- ▶ Funciona con variables descriptivas continuas  
*K-medoids*
- ▶ Dificultades cuando los clústeres son de diferente tamaño y/o densidad, o no son convexos

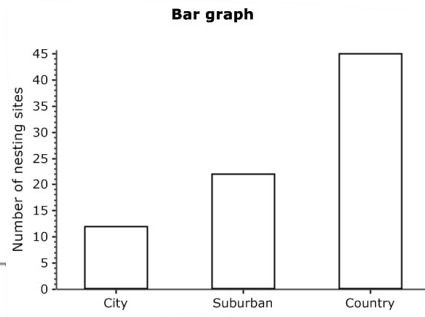
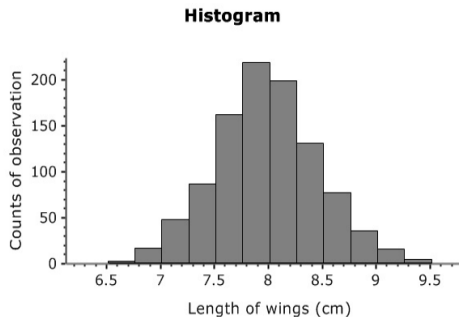
# Agrupamiento basado en particiones

K-medoids



# Agrupamiento basado en particiones

## K-medoids



# Agrupamiento basado en particiones

K-medoids

## Intuición

La idea iterativa de  $K$ -means

Se cambia el centro por el centriode:

Los centros son, en todo momento, ejemplos del conjunto de entrenamiento



# Agrupamiento basado en particiones

## K-medoids

---

### K-means

Recibe: Conjunto de entrenamiento,  $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ ; número de clústeres,  $K$

- 
1. Elección (aleatoria) de  $K$  puntos del conjunto de entrenamiento como centros,  $\{\bar{\mathbf{x}}_1, \dots, \bar{\mathbf{x}}_K\}$ .
  2. Asignar cada ejemplo  $\mathbf{x}_i$  al clúster del centro más cercano:  
$$C(\mathbf{x}_i) = \operatorname{argmin}_{k \in \{1, \dots, K\}} \|\mathbf{x}_i - \bar{\mathbf{x}}_k\|^2$$
  3. Para cada clúster  $k$ , recalcular su centro:  $\bar{\mathbf{x}}_k = \operatorname{argmin}_{\mathbf{x}} \sum_{\mathbf{x}_i: C(\mathbf{x}_i)=k} \|\mathbf{x}_i - \mathbf{x}\|^2$
  4. Los pasos 2 y 3 se iteran hasta que los centros no cambian.

---

Devuelve: Conjunto de centros,  $\{\bar{\mathbf{x}}_1, \dots, \bar{\mathbf{x}}_K\}$ ; Asignación,  $C$

---

# Agrupamiento basado en particiones

## K-medoids

---

### K-medoids

Recibe: Conjunto de entrenamiento,  $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ ; número de clústeres,  $K$

- 
1. Elección (aleatoria) de  $K$  puntos del conjunto de entrenamiento como medoides,  $\{\tilde{\mathbf{x}}_1, \dots, \tilde{\mathbf{x}}_K\}$ .
  2. Asignar cada ejemplo  $\mathbf{x}_i$  al clúster del medoide más cercano:  
$$C(\mathbf{x}_i) = \operatorname{argmin}_{k \in \{1, \dots, K\}} d(\mathbf{x}_i, \tilde{\mathbf{x}}_k)$$
  3. Para cada clúster  $k$ , recalcular su medoide:  $\tilde{\mathbf{x}}_k = \operatorname{argmin}_{\mathbf{x}: C(\mathbf{x})=k} \sum_{\mathbf{x}_l: C(\mathbf{x}_l)=k} d(\mathbf{x}_l, \mathbf{x})$
  4. Los pasos 2 y 3 se iteran hasta que los centros no cambian.

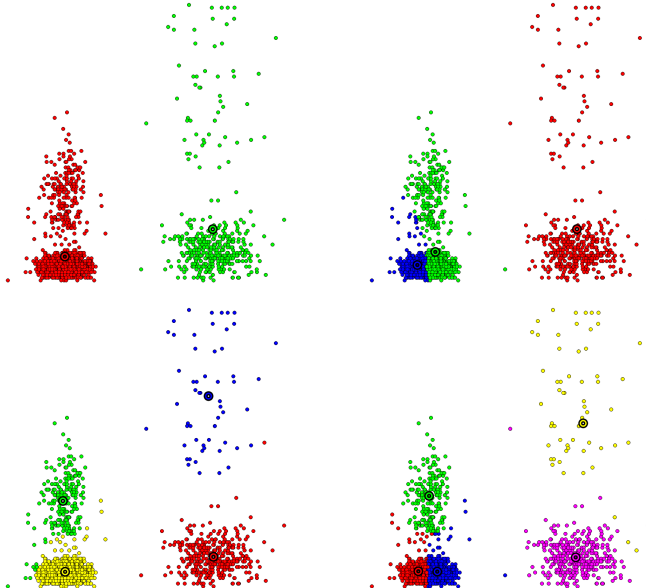
---

Devuelve: Conjunto de centros,  $\{\tilde{\mathbf{x}}_1, \dots, \tilde{\mathbf{x}}_K\}$ ; Asignación,  $C$

---

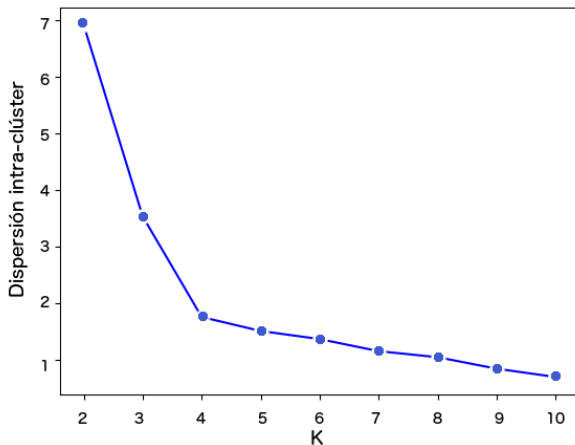
# Agrupamiento basado en particiones

Elegir el número de clústeres (K)



# Agrupamiento basado en particiones

Elegir el número de clústeres (K)



# Agrupamiento basado en particiones

Elegir el número de clústeres ( $K$ )

## Ideas:

- ▶ La dispersión intraclúster siempre se reduce
- ▶ Elegir el punto donde el cambio de tendencia es más pronunciado

# Aprendizaje no supervisado

## VC02: Agrupamiento por particiones: K-means

Félix José Fuentes Hurtado

[felixjose.fuentes@campusviu.es](mailto:felixjose.fuentes@campusviu.es)

Universidad Internacional de Valencia