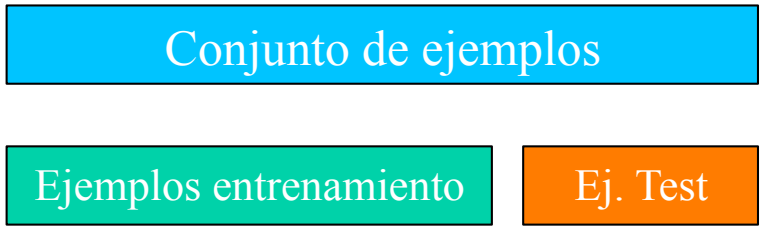
Etapas de las técnicas basadas en aprendizaje estadístico

En cualquier esquema basado en aprendizaje estadístico se distinguen 2 etapas:

- **Etapa de aprendizaje/entrenamiento (training)**
  - Permite realizar el diseño del sistema según un determinado algoritmo (algoritmo de entrenamiento/aprendizaje).
  - Si el aprendizaje es supervisado, es necesario disponer de un conjunto de ejemplos etiquetados (cuya clase es conocida): $\{(\underline{\mathbf{x}}^{(i)},t^{(i)})\}$, i=1,...,M
  - Si el aprendizaje es no supervisado, no se dispone (o no se hace uso) de la etiqueta de cada ejemplo: $\{\underline{\mathbf{x}}^{(i)}\}$, i=1,...,M

  Es deseable que el diseño ofrezca buena **capacidad de generalización**: capacidad para realizar correctamente la tarea ante ejemplos/observaciones no considerados previamente (y, por tanto, no utilizados durante el entrenamiento).

- **Etapa de evaluación (test)**. Permite evaluar las prestaciones del clasificador.



Partición aleatoria

Aprendizaje supervisado. Ejemplo (I)

Objetivo: identificación de coches en imágenes.

Conjunto de entrenamiento

Testing:

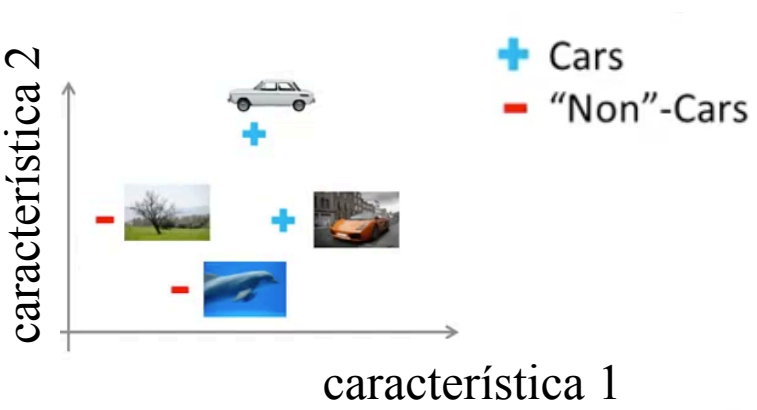
What is this?

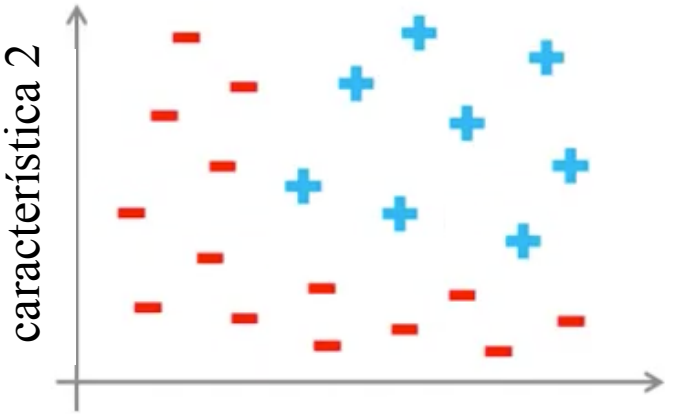
Aprendizaje supervisado. Ejemplo (II)

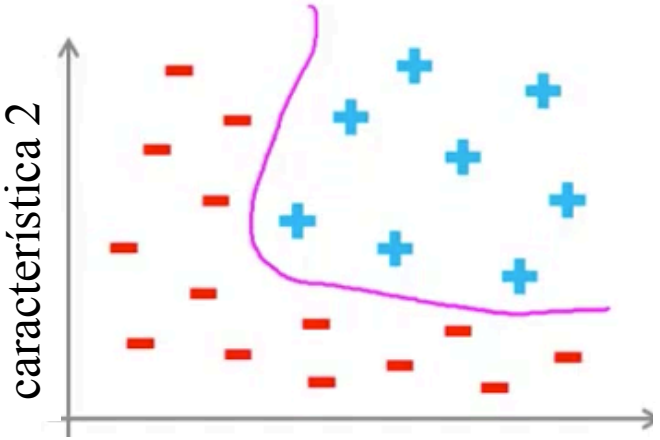
Extracción de características (*feature extraction*)

Algoritmo de aprendizaje (*machine learning algorithm*)

Coche
No coche







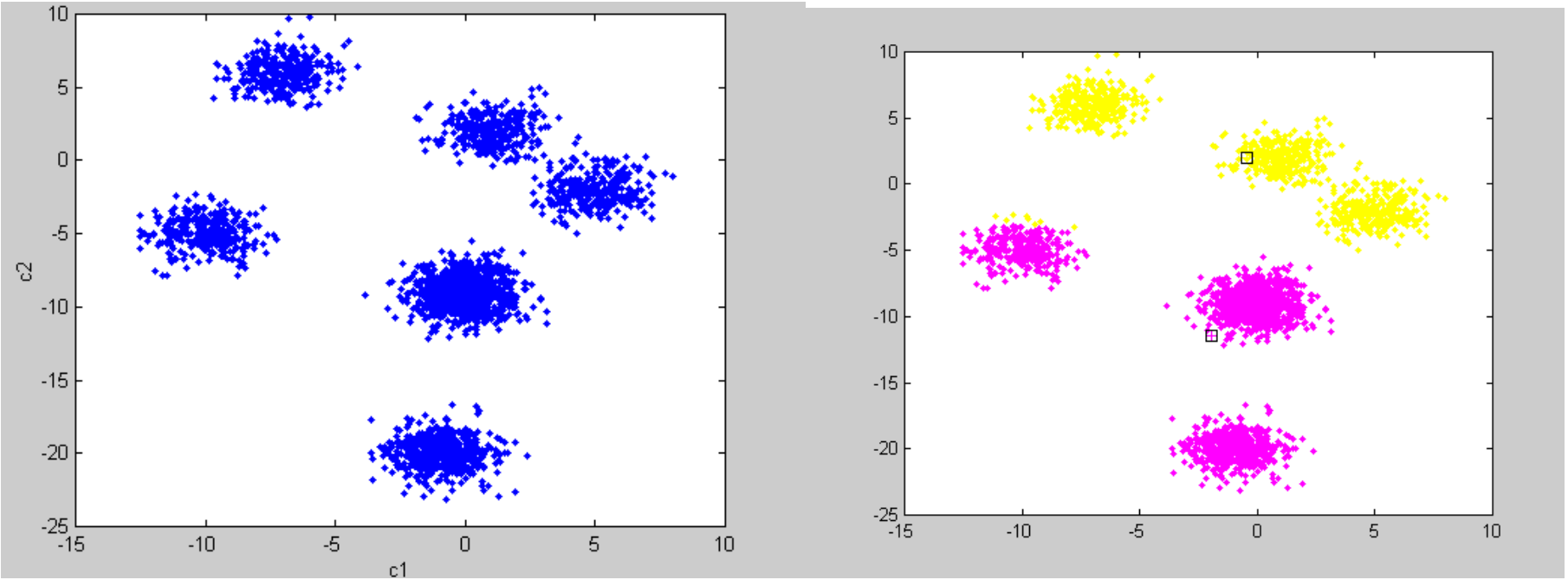
Solución no lineal
(frontera no lineal)

Aprendizaje no supervisado. Repaso algoritmo *k*-medias

Cada *cluster* está representado por un **centroide**.

Cada observación se asigna al *cluster* más próximo (criterio de similitud).

Ejemplo con un espacio de dos características



El color de cada observación es aleatorio e indica el *cluster* al que está asociada cada observación => salida del algoritmo de *clustering*.

En el algoritmo *k*-medias, el parámetro *k* hace referencia al número de grupos (clusters). No se debe confundir con el parámetro *k* del clasificador *k*-nn.

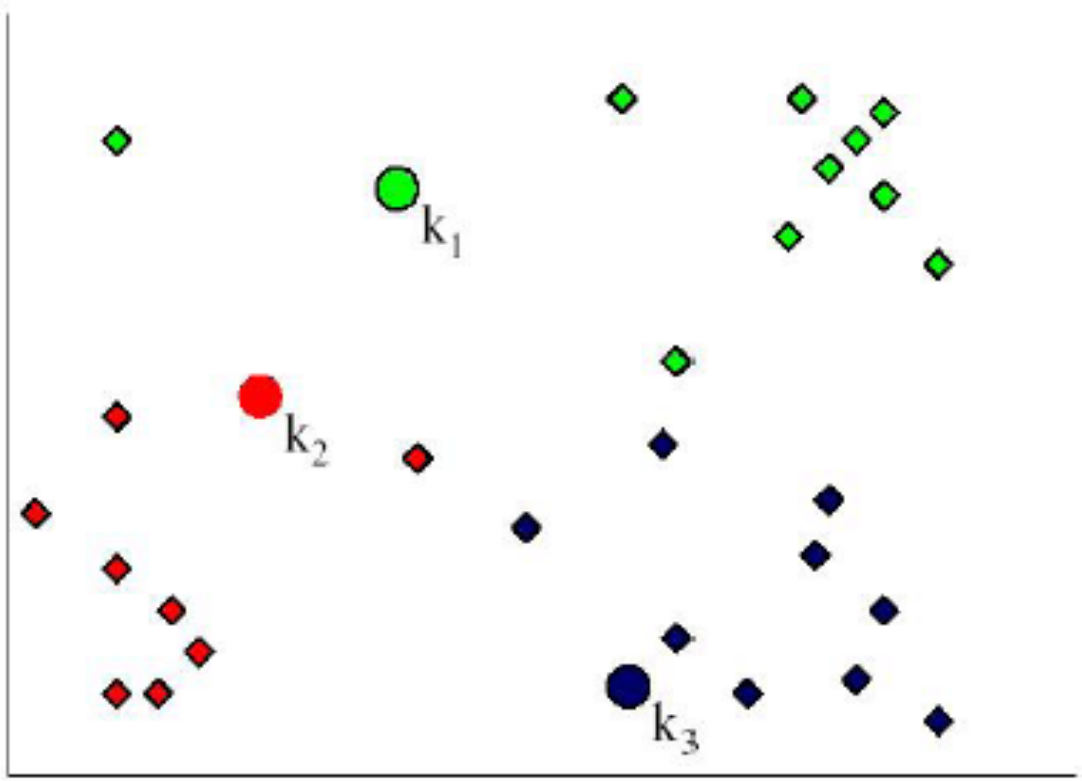
Algoritmo *k-means*:

1. Inicialización. En general, sin información a priori, se eligen aleatoriamente *k* ejemplos (reales o no) como centros de los *clusters* (centroides),

   k_j, j=1, 2, …, *k*.

   Nota: el resultado puede depender de los grupos iniciales.

2. Asignar cada ejemplo al *cluster* del centroide más próximo (la medida de similitud puede ser la distancia).



Se asigna cada muestra al centroide más cercano.
Cada color representa un *cluster* distinto.

En el algoritmo *k*-medias, el parámetro *k* hace referencia al número de grupos (clusters). No se debe confundir con el parámetro *k* del clasificador *k*-nn.

Algoritmo *k-means*:

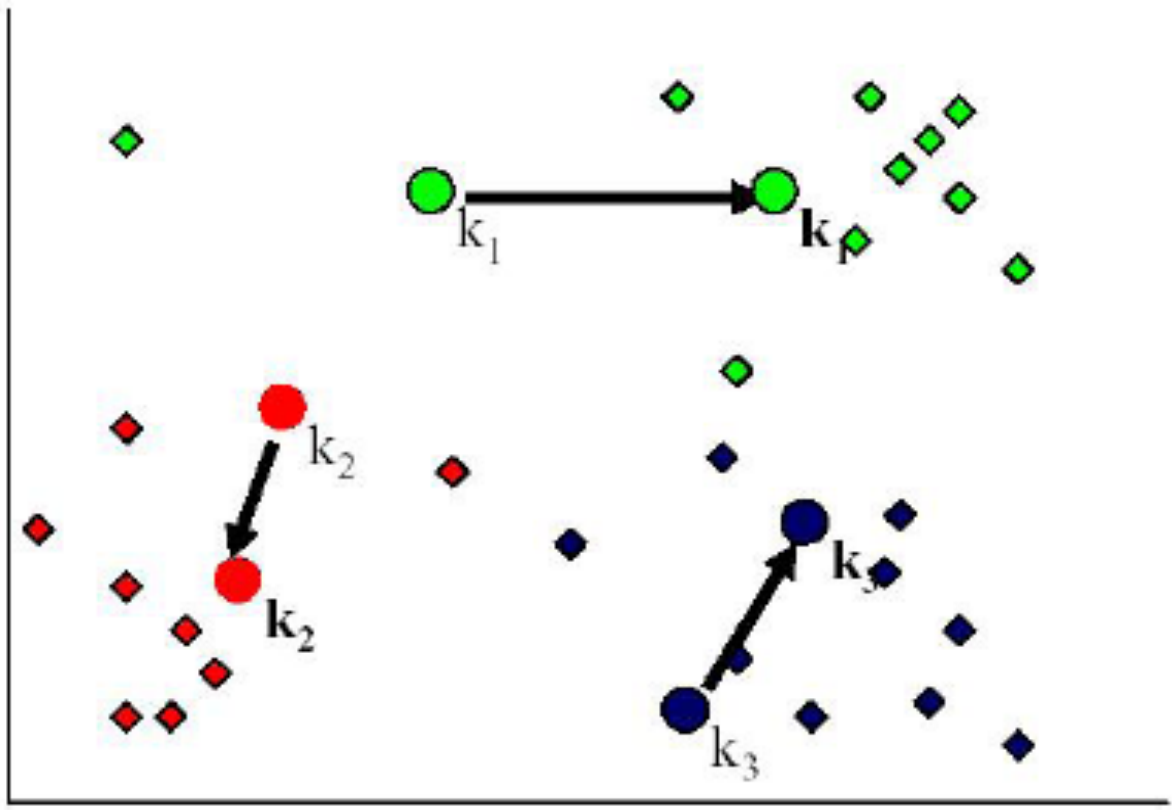
1. Inicialización. En general, sin información a priori, se eligen aleatoriamente *k* ejemplos como centros de los *clusters* (centroides),

   k_j, j=1, 2, …, *k*.

   Nota: el resultado puede depender de los grupos iniciales.

2. Asignar cada ejemplo al *cluster* al que corresponda el centroide más próximo.

3. Recalcular las posiciones de los centroides como el promedio de las observaciones asignadas a cada *cluster*. ***¿Qué dimensión tiene entonces cada centroide?***



Se recalcula la posición de cada centroide como el promedio de las muestras/observaciones/ejemplos/instancias/casos de cada *cluster*.

En el algoritmo *k*-medias, el parámetro *k* hace referencia al número de grupos (clusters). No se debe confundir con el parámetro *k* del clasificador *k*-nn.

Algoritmo *k-means*:

1. Inicialización. En general, sin información a priori, se eligen aleatoriamente *k* ejemplos como centros de los *clusters* (centroides), k_j, j=1, 2, …, *k*.

   Nota: el resultado puede depender de los grupos iniciales.

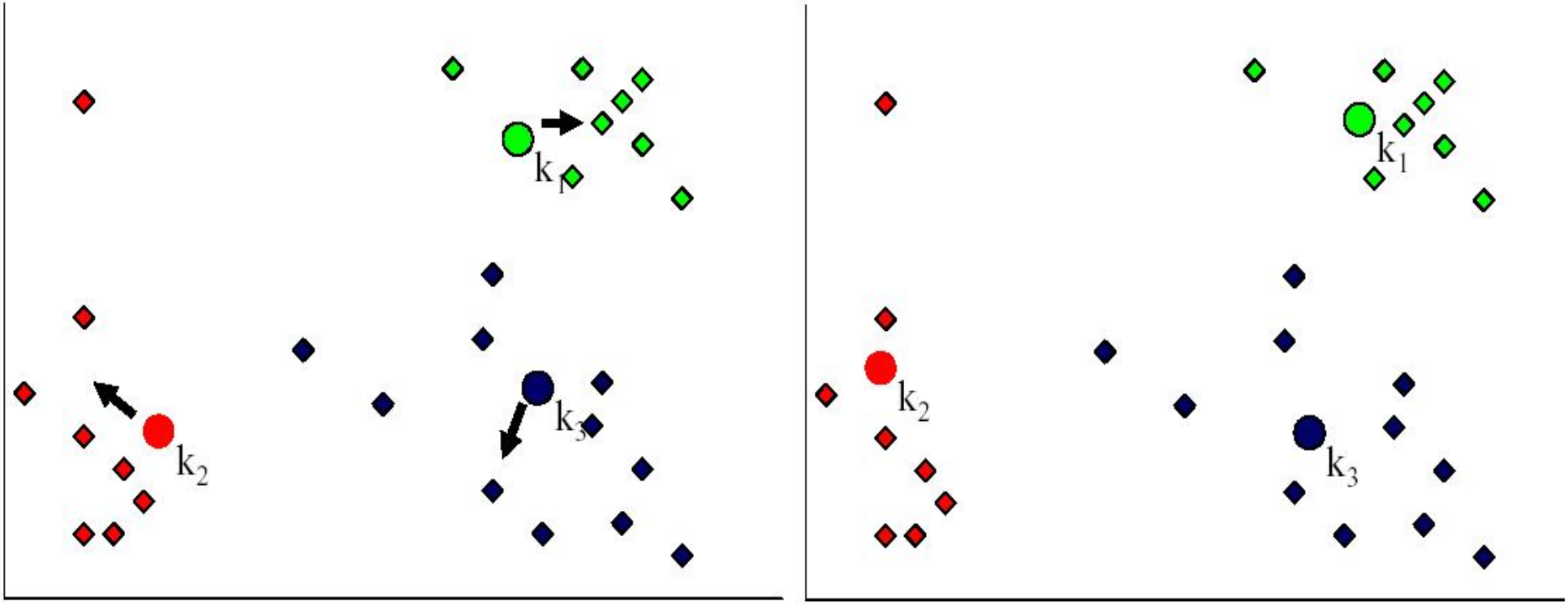
2. Asignar cada ejemplo al *cluster* al que corresponda el centroide más próximo (usando, p.e., la distancia Euclídea).

3. Recalcular las posiciones de los centroides como el promedio de las observaciones asignadas a cada *cluster.*

4. Volver al paso 2 hasta que se cumpla el criterio de parada.

   - número máximo de iteraciones
   - estabilización en la posición de los centroides

Recalcular los centros de los *clusters*



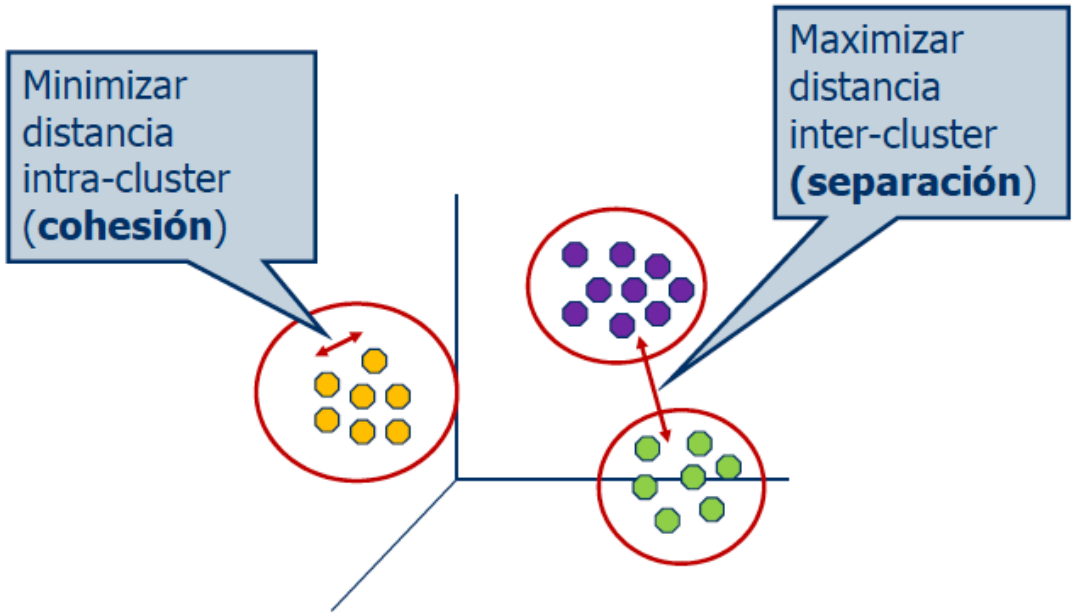
Reasignar las muestras al cluster más cercano ...

El criterio para detener el algoritmo puede ser un número máximo de iteraciones, la estabilización en la posición de los centroides, …

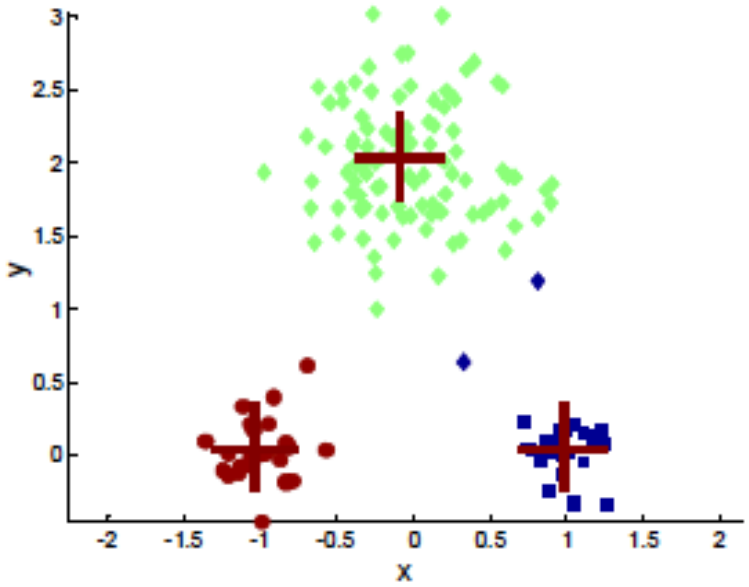
¿Qué criterio se está optimizando en *k*-medias?

Al considerar la distancia Euclídea y la media aritmética, se está **minimizando** la **variación *intra-cluster***. Esta medida se usa, por tanto, como criterio del grado de ajuste (cohesión y separación) de los centroides.

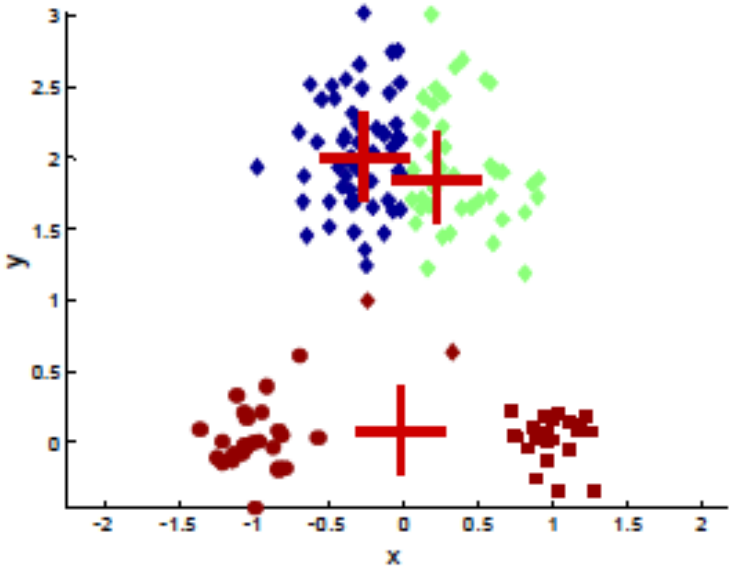
$$\sum_{i=1}^{k} \frac{1}{|C_i|} \sum_{\underline{x} \in C_i} d^2\left(\underline{x}, \underline{c}_i\right)$$



¿Qué criterio se está optimizando en *k*-medias?



Solución óptima



Posible resultado proporcionado por k-means

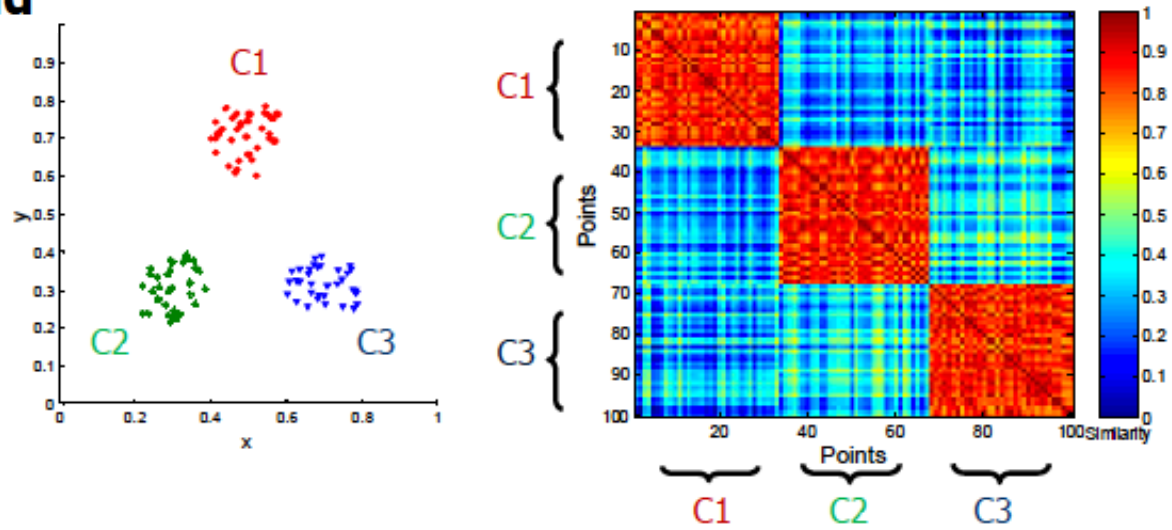
¿Cree que los centroides finales proporcionan un mínimo global de la variación *intra-cluster*?

¿A qué es sensible el algoritmo k-medias?

- A la **elección inicial de los centroides**

  - Se puede considerar la realización de varias ejecuciones con distintos conjuntos de centroides iniciales, y comparar los resultados

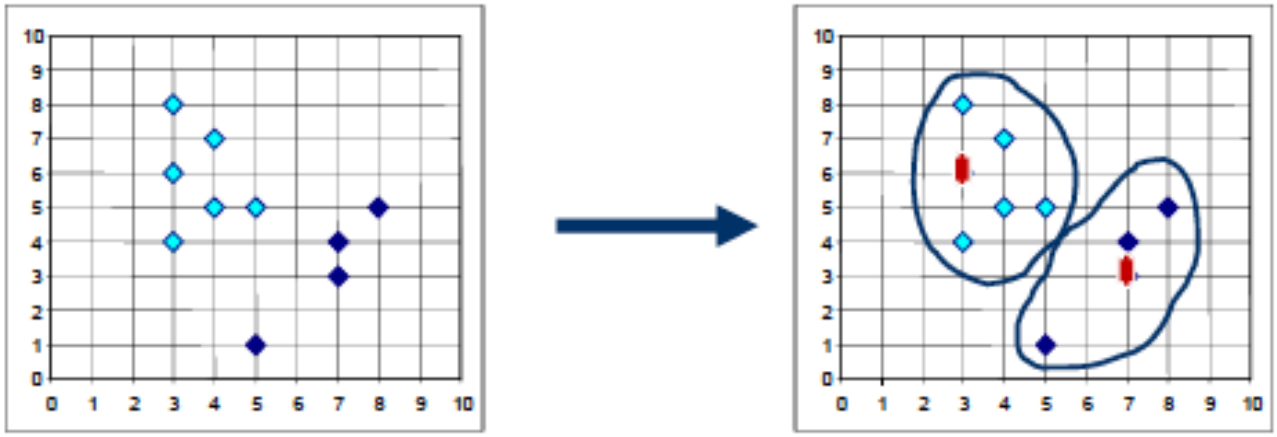
Matriz de similitud



  - Se puede realizar una elección “inteligente” (no aleatoria) de los centroides iniciales

¿A qué es sensible el algoritmo k-medias?

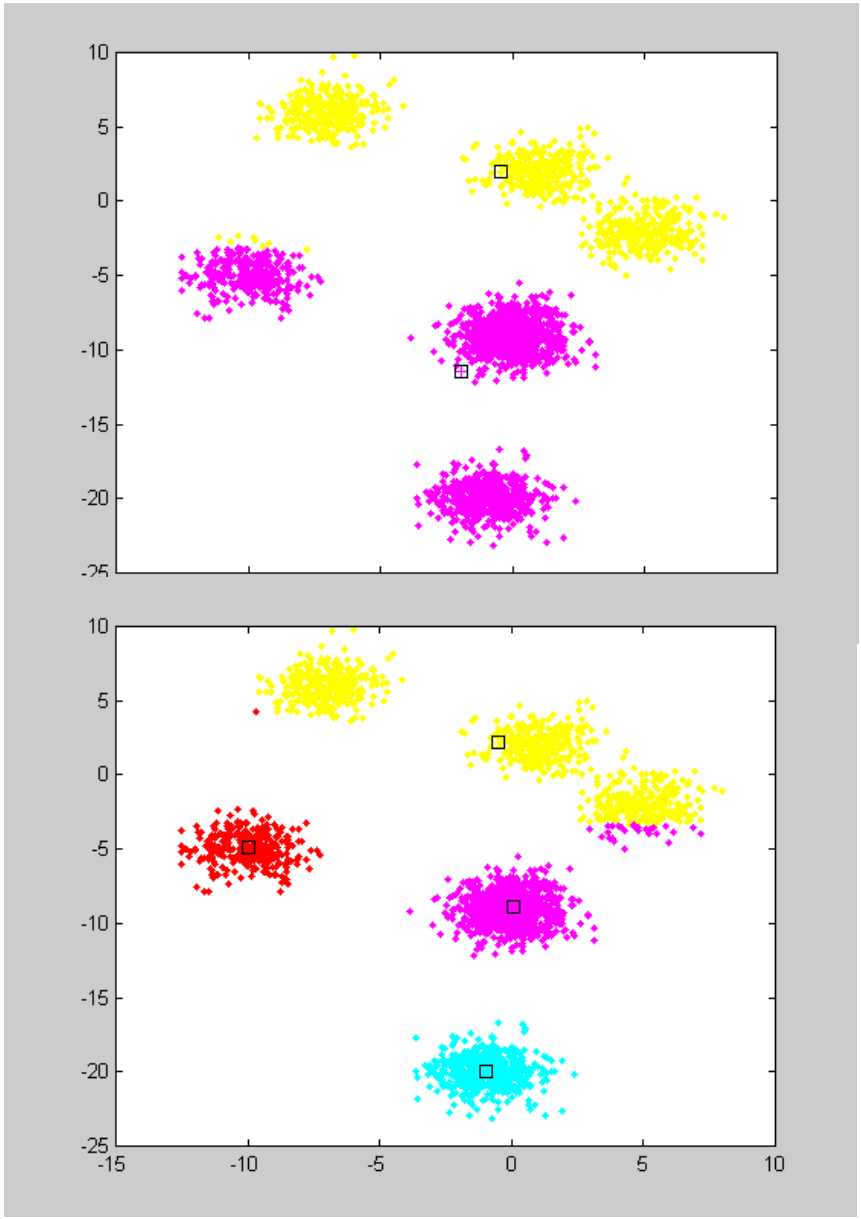
- A los **valores atípicos**
  - Se puede considerar la mediana en lugar de la media (para el cálculo de los centroides) => *k*-medoids, donde el centroide es siempre una de las observaciones del clúster.

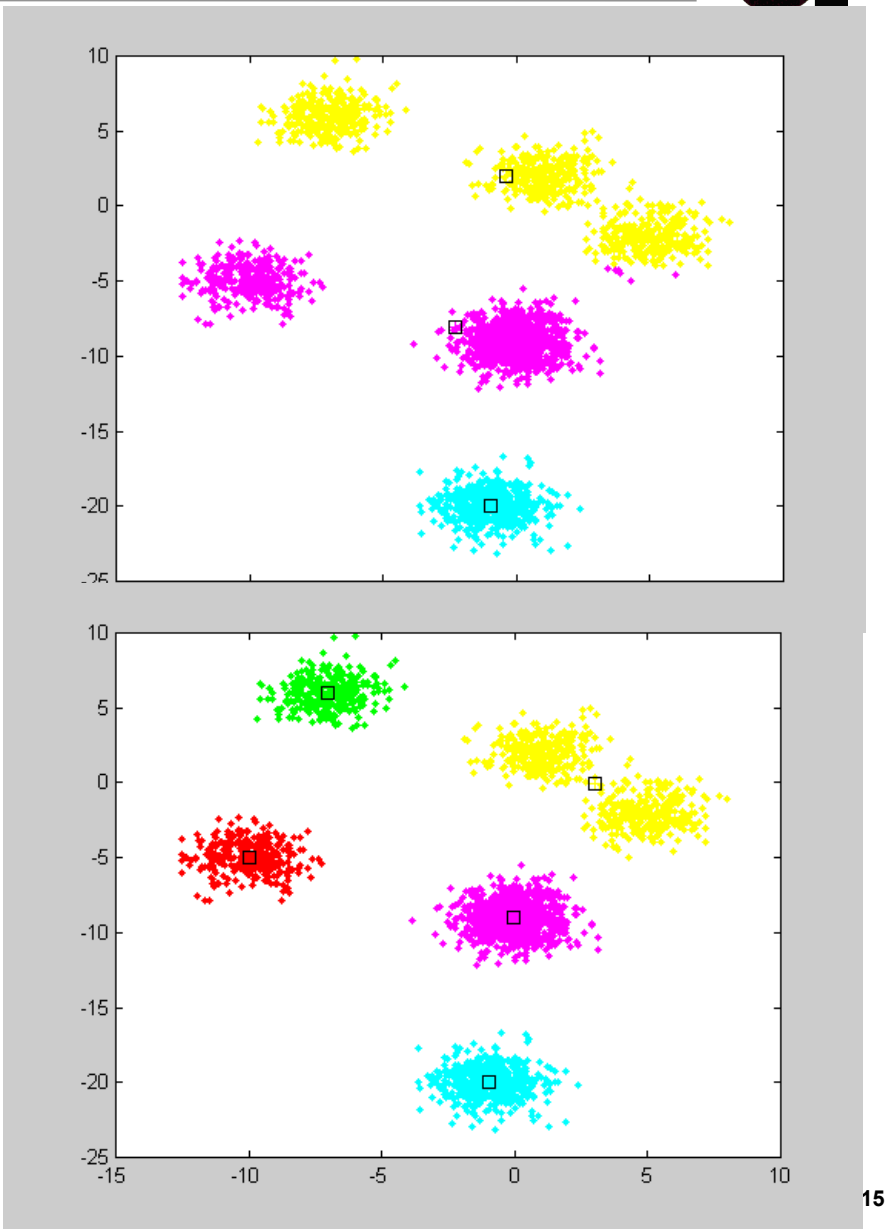


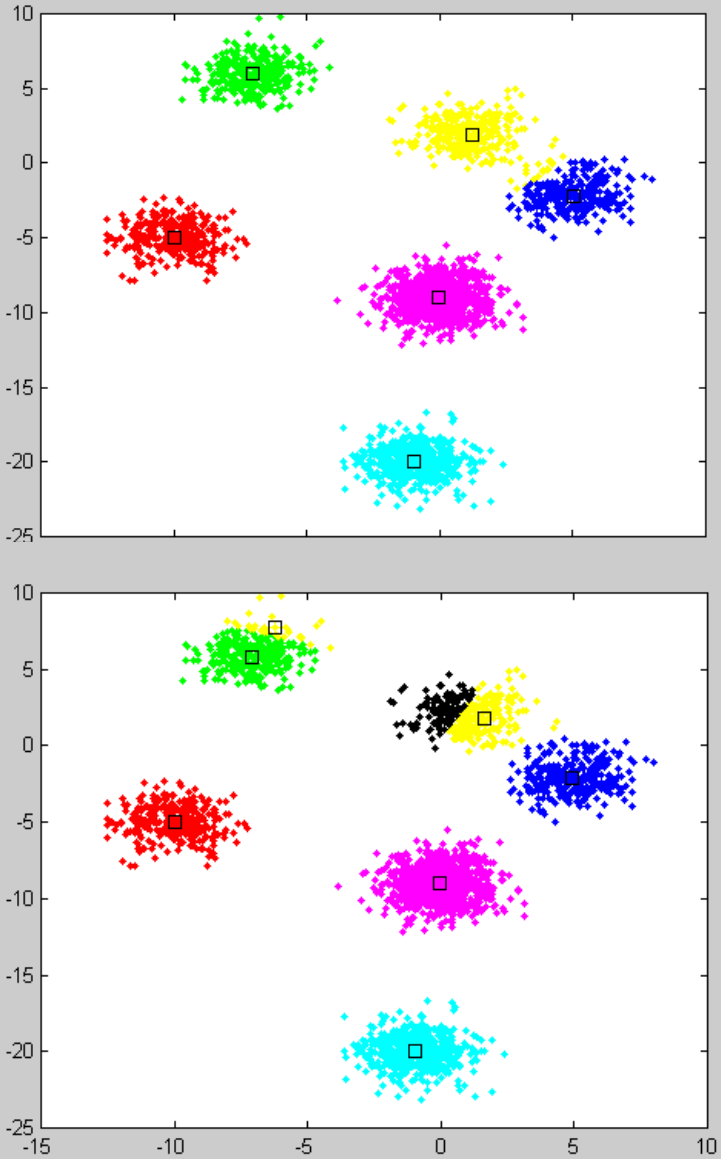
  - Se pueden eliminar previamente los outliers, aunque a veces pueden resultar ser valores interesantes

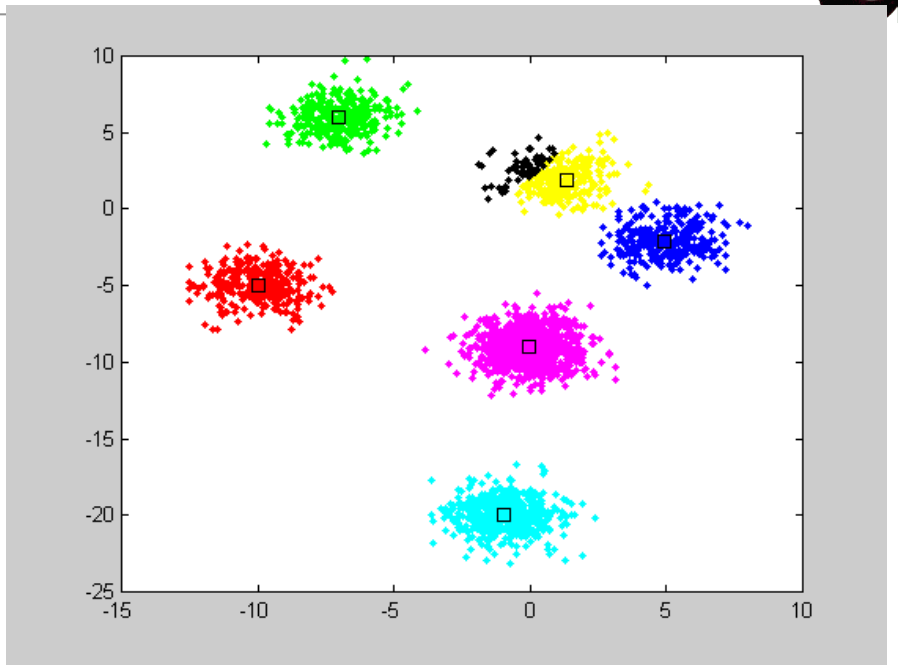
¿Cómo escoger el valor de *k*?

Algoritmo iterativo de división de regiones





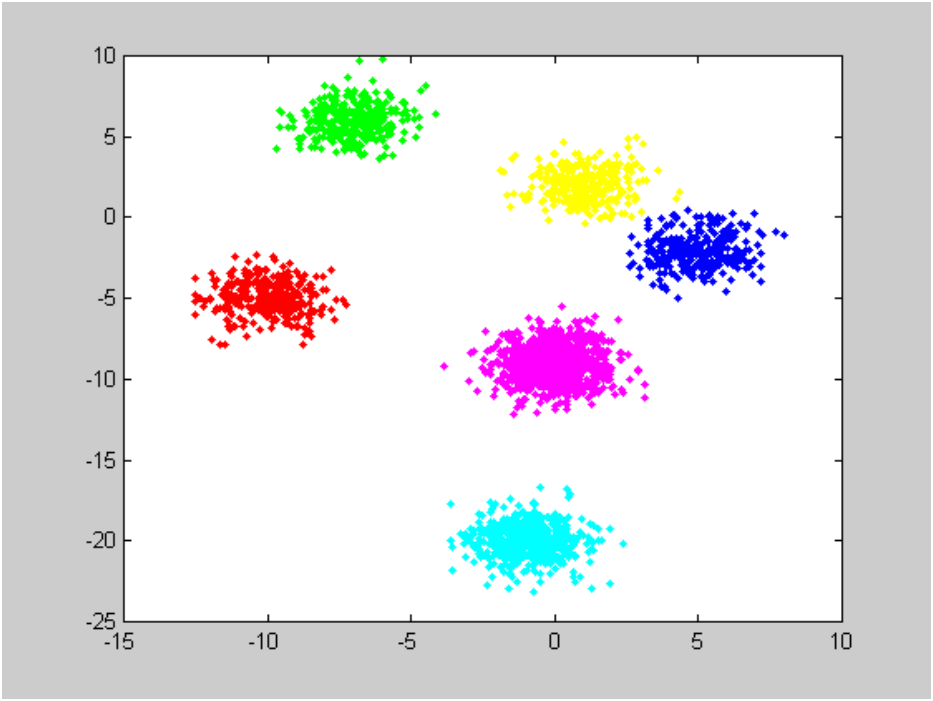


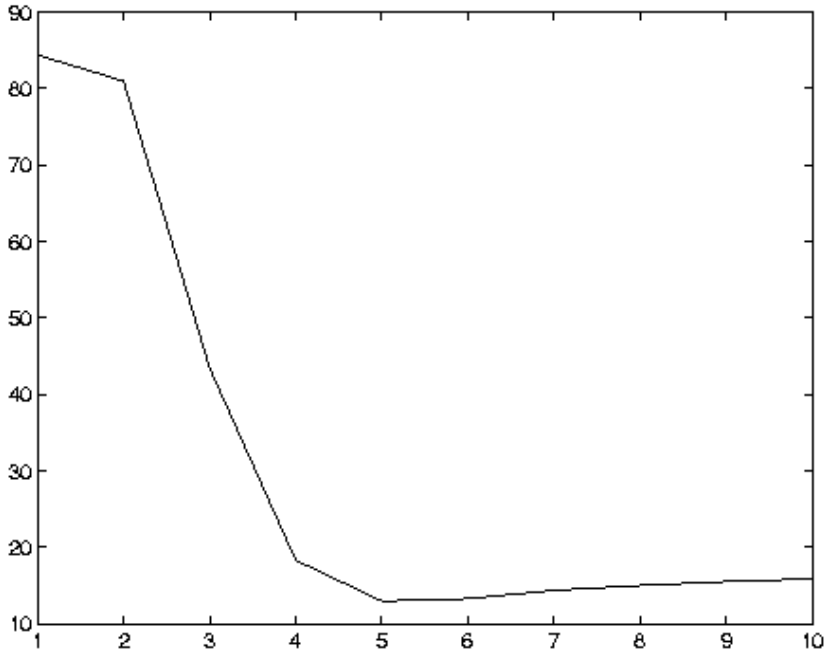


¿Hasta cuándo seguimos aumentando el valor de k?

Un posible criterio de parada: mínimo de la media de la varianza de cada región*número de regiones

Agrupación final





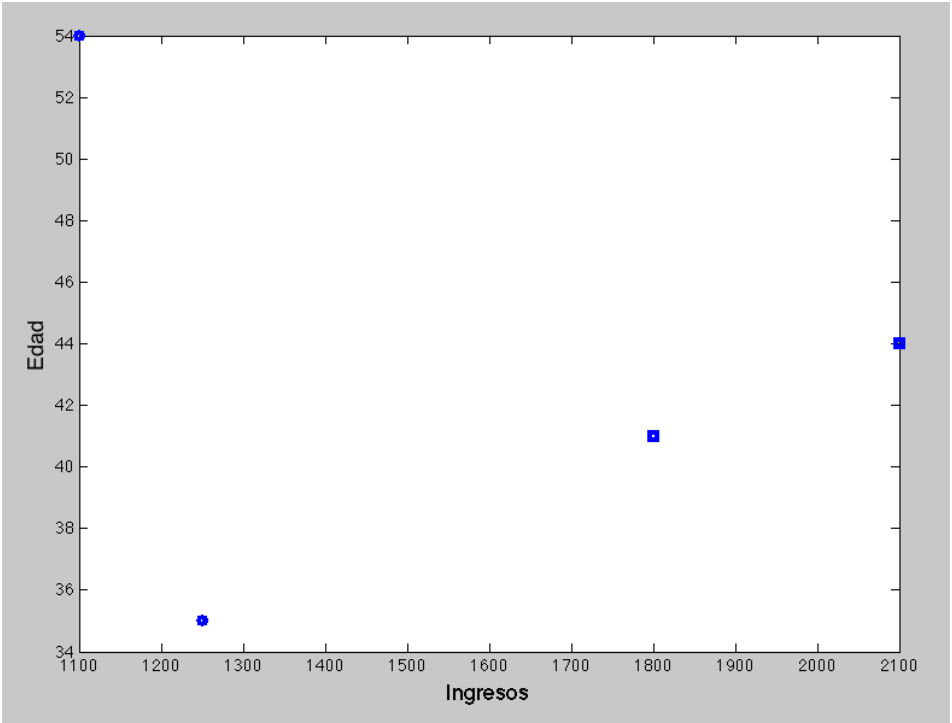
Número de iteraciones

Transformación del rango de cada atributo

Las herramientas que se basan en una medida de distancia son muy sensibles a la presencia de atributos con distinto rango dinámico.

Ejemplo con 2 atributos: ingresos mensuales (en euros), y edad.

	Ingresos mes	Edad
Ana	1100	54
Luis	1250	35
Laura	2100	44
Isabel	1800	41



La distancia Euclídea pondera de la misma forma una variación de 10 euros en el salario que una variación de 10 años en la edad.

Transformación del rango de cada atributo

Escalar los valores de los atributos para que estén en un rango específico.

La transformación de rango más común es la **transformación lineal uniforme**, que escala el atributo a una escala genérica entre 0 y 1.

$$x' = \frac{x - min_x}{max_x - min_x}$$

	Ingresos mes	Edad
Ana	1100	54
Luis	1250	35
Laura	2100	44
Isabel	1800	41

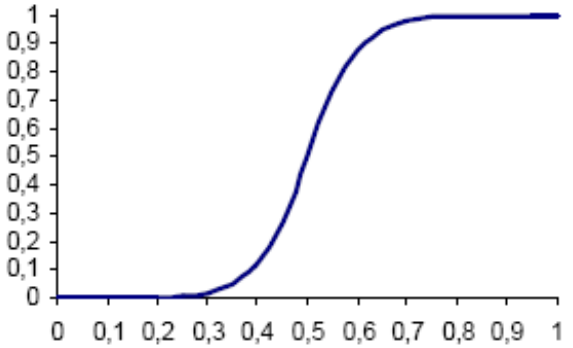
	Ingresos mes normaliz	Edad normaliz
Ana	0	1
Luis	0.15	0
Laura	1	0.47
Isabel	0.7	0.31

La transformación lineal uniforme es muy sensible a la presencia de valores anómalos.

Ejemplo: medidas de temperatura (rango normal entre 20º y 40º). Dato con temperatura a 1000º
Si transformamos para obtener valores en [0,1], la mayoría de valores estarán en [0.02, 0.04] ⇒ poca precisió

¿Posible solución? Escalado **no lineal**

P.e., el **escalado sigmoidal** realiza una transformación no lineal a una escala genérica entre 0 y 1. Utiliza una transformación que es más pronunciada en el centro y más aplanada en los bordes.



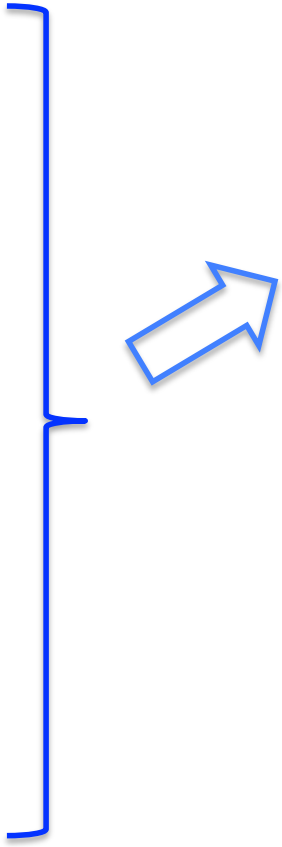
Transformación del rango de cada atributo

Otra opción es **escalar** los valores de los atributos para que tengan determinados estadísticos, por ejemplo media nula y desviación típica uno (tipificación).

	Ingresos mes	Edad
Ana	1100	54
Luis	1250	35
Laura	2100	44
Isabel	1800	41

	Ingresos mes	Edad
media	1562.5	43.5
desv típica	467.9	7.9

$$x' = \frac{x - media_x}{desviacion\_tipica_x}$$



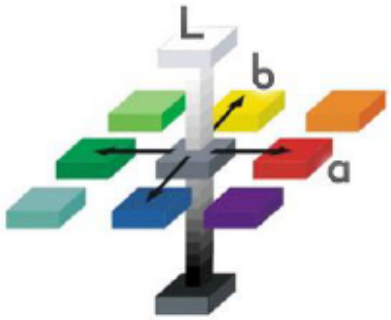
	Ingresos mes tipific	Edad tipific
Ana	-0.9884	1.32
Luis	-0.6678	-1.07
Laura	1.1487	0.06
Isabel	0.5076	-0.31

CUESTIÓN 7 (1.9 p)

Universidad Rey Juan Carlos

5-Abril-2018

Considere la interpretación de las componentes de color *ab* del espacio CIE-Lab (véase la Figura C7-1(a)) y la imagen en color de la Figura C7-1(b).



(a)

Figura C7-1. (a) Esquema para interpretar los colores en el espacio CIE-Lab; (b) Imagen en color.

(b)

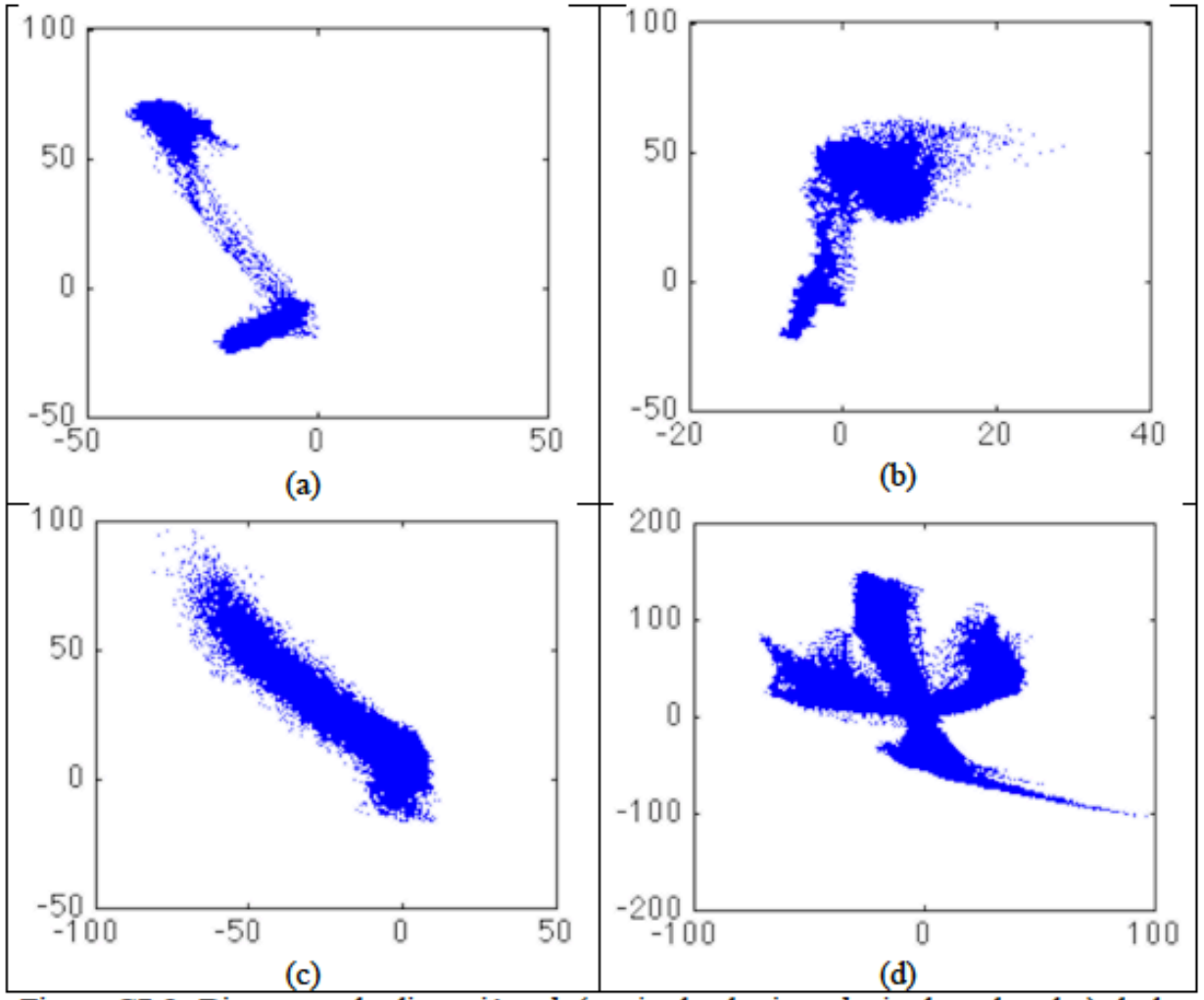


Figura C7-2. Diagramas de dispersión *ab* (*a*, eje de abscisas, *b* eje de ordenadas) de los píxeles correspondientes a cuatro imágenes.

Si se desea segmentar la imagen de la Figura C7-1(b) aplicando el algoritmo *k*-medias (*k-means*) sobre el espacio ab seleccionado, responda a las siguientes preguntas:

c) Desde un punto de vista metodológico, ¿qué etapas seguiría para realizar la segmentación? Describa brevemente cada una de ellas (1p)

Justifique razonadamente todas sus respuestas.

- Conversión al espacio Lab => extraer componentes a y b (plano imagen).
- Redimensionar cada componente para que sea un vector columna, de manera que cada píxel sea una fila en la matriz ab.
- Estandarización de cada característica (a y b) para que tenga media nula y desviación típica 1 [indicar expresión]. La razón es porque el algoritmo k-medias hace uso, por defecto, de la distancia Euclídea.
- Luego, sobre el espacio ab, se aplicaría el algoritmo 2-medias => se conoce la disposición de los centroides.
- A cada centroide se le asignaría una "etiqueta" diferente. Cada etiqueta identifica un objeto de interés
- Para cada píxel, se determina el centroide más próximo y se le asignara la misma etiqueta que la del centroide más próximo.

De esa manera, habríamos segmentado la imagen.