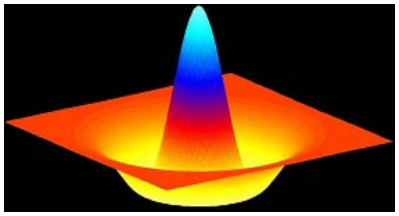


Procesamiento Digital de Imágenes

Clasificador de Mínima Distancia

Algoritmo de Clustering K-means

❑ Clasificación por comparación de patrones

- Cada clase es representada por un vector característico (**patrón**) que se obtiene luego de un proceso de entrenamiento con vectores representando a objetos pertenecientes a esa clase.
- La técnica de clasificación consiste de los siguientes pasos:

  - Entrenamiento de patrones

  - Comparación del vector característico correspondiente al objeto a clasificar con los patrones asociados a cada clase.

  - Criterio de decisión

El esquema general está representado en Fig. 1

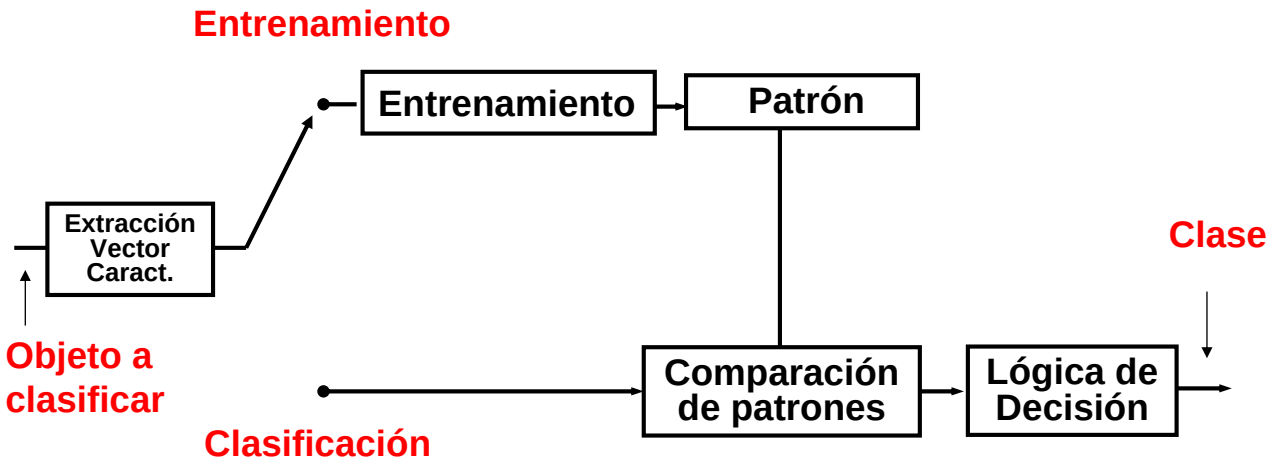


Fig. 1: Clasificación basada en Comparación de Patrones.

❑ Entrenamiento de patrones

En esta etapa se construye un **patrón de referencia** asociado a cada clase, basándose en los vectores característicos de todas las instancias de esa clase usadas para el entrenamiento. Hay varias formas en que puede realizarse el entrenamiento.

1. **Entrenamiento casual:** una única instancia de la clase es usada como patrón (muy poco robusto).
2. **Entrenamiento robusto:** se utilizan varias instancias de cada clase para construir un patrón de la misma.
3. **Entrenamiento por clustering**: se utiliza un gran número de instancias de cada clase para construir los patrones asociados.

❑ Comparación de patrones

En la etapa de comparación de patrones se realiza una comparación directa entre el vector característico asociado al objeto a clasificar y todos los posibles patrones aprendidos en la etapa de entrenamiento, de manera de determinar el mejor ajuste de acuerdo a algún criterio. Surge la necesidad de definir una **medida de similaridad (distancia)** entre vectores característicos que permita determinar cuál es el patrón de referencia que mejor se ajusta al vector característico asociado al objeto a clasificar.

❑ Medidas de Distancia (ó Distorsión)

- Una característica fundamental de los sistemas de clasificación con el enfoque de **comparación de patrones** es la forma en que los vectores característicos son combinados y comparados con los patrones de referencia.
- Para poder realizar estas operaciones es necesario definir una **medida de distancia** entre vectores característicos.

Definición: Una distancia entre dos vectores $\boldsymbol{x}$ e $\boldsymbol{y}$ de un espacio vectorial X es una función a valores reales $d(\mathbf{x}, \mathbf{y})$ sobre el producto Cartesiano $X \times X$, que verifica las propiedades:

$$(\mathrm{a})\ 0 \le d(\mathbf{x},\mathbf{y}) < \infty, \quad \forall\ \mathbf{x},\mathbf{y} \in X,\ \text{y}\ \ d(\mathbf{x},\mathbf{y}) = 0\ \text{si y solo si}\ \mathbf{x} = \mathbf{y}$$

$$(\mathrm{b})\ d(\mathbf{x},\mathbf{y}) = d(\mathbf{y},\mathbf{x}),\ \forall\ \mathbf{x},\mathbf{y} \in X$$

$$(\mathrm{c})\ d(\mathbf{x},\mathbf{y}) \le d(\mathbf{x},\mathbf{z}) + d(\mathbf{y},\mathbf{z}),\ \forall\ \mathbf{x},\mathbf{y},\mathbf{z} \in X$$

- Algunas de las medidas de distancia más utilizadas son las distancias o **métricas inducidas por las normas en espacios** L_p.

  Por ejemplo, si f_i, f_i' con $i = 1,2,\cdots,D$ son las componentes de dos vectores característicos f y f', pueden definirse las siguientes métricas inducidas por las normas L_p

$$d_1(f, f') = \sum_{i=1}^{D} \left| f_i - f_i' \right|$$ **distancia** L_1

$$d_2(f, f') = \sqrt{\sum_{i=1}^{D} \left(f_i - f_i' \right)^2}$$ **distancia Euclidea (o** L_2 **)**

- Una medida de distancia muy utilizada que tiene en cuenta la variabilidad de cada componente del vector característico es la **distancia Euclidea ponderada**, definida como

$$d_{2w}(f, f') = \sqrt{\sum_{i=1}^{D} \left(w_i \left(f_i - f_i' \right) \right)^2}$$

donde

$$w_i = \frac{1}{\sigma_i}$$

siendo σ_i^2 una estima de la varianza de la *i*-ésima componente f_i. Aquí, los datos que son menos confiables (con mayor varianza o dispersión) son pesados menos.

• Distancia de Mahalanobis

Cuando el vector característico no es homogéneo, en el sentido de que está compuesto por características de distinta naturaleza, puede suceder que las distintas componentes estén en distintas escalas. Al usar la norma Euclidea, o la norma 1, para calcular la distancia entre vectores, las componentes en las escalas más pequeñas casi no tendrán influencia en el valor de la distancia, a pesar de que esas componentes puedan diferir mucho. Surge entonces la necesidad de usar distancias normalizadas. Una posible forma de normalizar la distancia entre un vector x y la media μ del conjunto de vectores al cual pertenece x, para el caso escalar, es definiendo

$$r = \left| \frac{x - \mu}{\sigma} \right| \qquad (1)$$

donde σ es el desvio estándar del conjunto de vectores, y r es la distancia normalizada.

De esta forma un vector que está separado un desvio estándar de la media, estará a una distancia normalizada igual a 1 (independientemente de las unidades en que se mida la distancia).

Esta idea se puede generalizar para el caso de un vector de dimensión *d,* definiendo

$$r^2 = \left(\frac{x_1 - \mu_1}{\sigma_1} \right)^2 + \left(\frac{x_2 - \mu_2}{\sigma_2} \right)^2 + \cdots + \left(\frac{x_d - \mu_d}{\sigma_d} \right)^2$$

Esta definición tiene sentido si cada componente del vector característico mide propiedades independientes del objeto a clasificar.

Considerando la expresión (1) para el caso escalar, la distancia normalizada puede escribirse

$$r^2 = (x-\mu)^T \frac{1}{\sigma^2}(x-\mu)$$

donde σ^2 es la varianza del conjunto de vectores, que para el caso multidimensional resulta

$$r^2 = (x-\mu)^T \Sigma^{-1}(x-\mu)$$

donde Σ es la matriz de covarianza del conjunto de vectores.

Esta formulación de la distancia es más general ya que tiene en cuenta la interacción entre coeficientes a través de la matriz de covarianza y se la denominada **Distancia de Mahalanobis.**

Si se tienen N_j vectores de entrenamiento para cada una de las M clases, la **distancia de Mahalanobis** entre un vector a clasificar x, y el vector patrón representativo de la clase j, resulta

$$d_M(x,\mu_j) = \sqrt{(x-\mu_j)^T \Sigma_j^{-1}(x-\mu_j)}$$

donde

$$\mu_j = \frac{1}{N_j}\sum_{k=1}^{N_j} x_{kj}^e$$

Media (muestral) de los vectores de entrenamiento x_{kj}^e

$$\Sigma_j = \frac{1}{N_j - 1}\sum_{k=1}^{N_j}(x_{kj}^e - \mu_j)(x_{kj}^e - \mu_j)^T$$

Matriz de Covarianza (muestral) de los vectores de entrenamiento x_{kj}^e

x : vector característico a clasificar

❑ Clasificador de Mínima Distancia

Denotemos con $\mu_1, \mu_2, \ldots, \mu_M$ a los vectores promedio representativos de los patrones 1, 2, ..., M, respectivamente, obtenidos computando la media muestral de los datos de entrenamiento correspondientes a cada clase. Es decir

$$\mu_j = \frac{1}{N_j} \sum_{k=1}^{N_j} x_{kj}^e \qquad j=1, 2, \ldots, M$$

Donde $\{x_{kj}^e\}_{k=1}^{N_j}$ es el conjunto de datos de entrenamiento de la clase j .

La Fig. 1, representa los datos de entrenamiento agrupados en clusters para el caso de vectores característicos de dimensión 3.

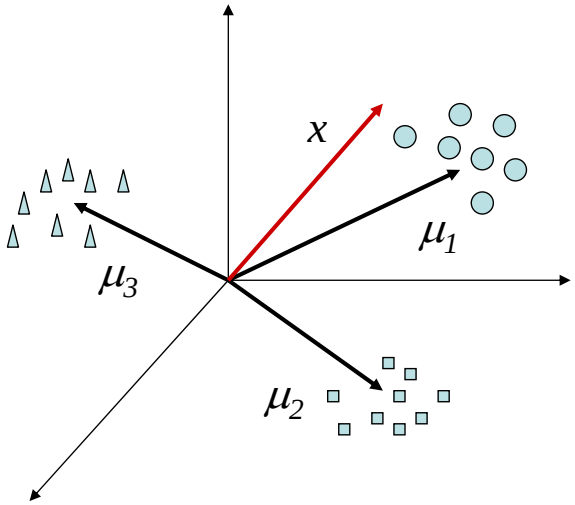


Fig. 1: Datos de entrenamiento agrupados en clusters correspondientes a las distintas clases.

Una manera obvia de clasificar un objeto representado por el vector x es computar la distancia de x a cada uno de los patrones $\mu_1, \mu_2, \ldots, \mu_M$ y asignar a x la clase cuyo patrón está a la mínima distancia. Por este motivo este clasificador se denomina de **mínima distancia.**

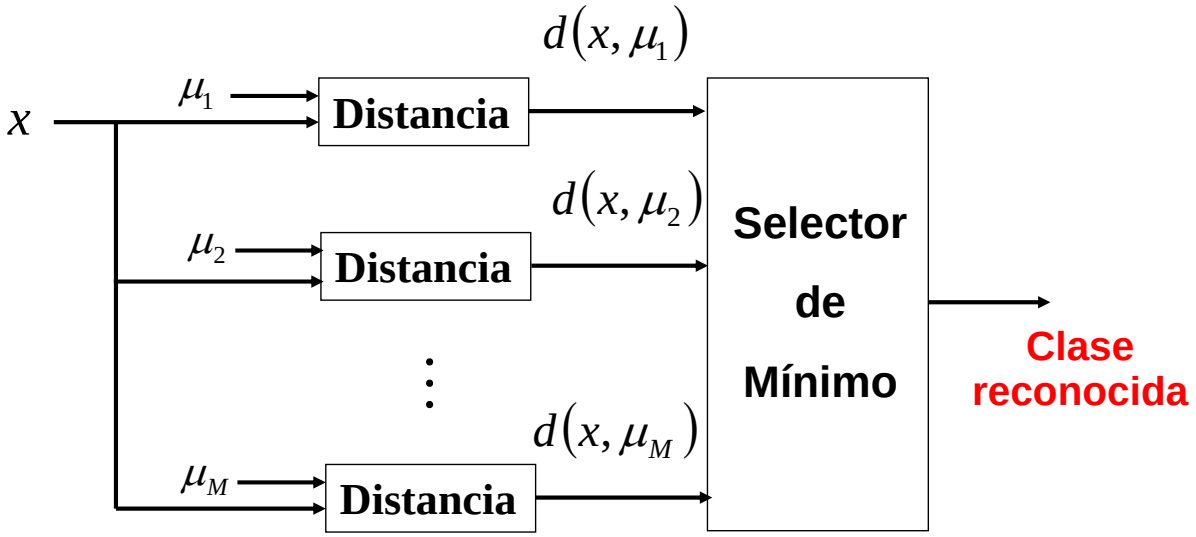


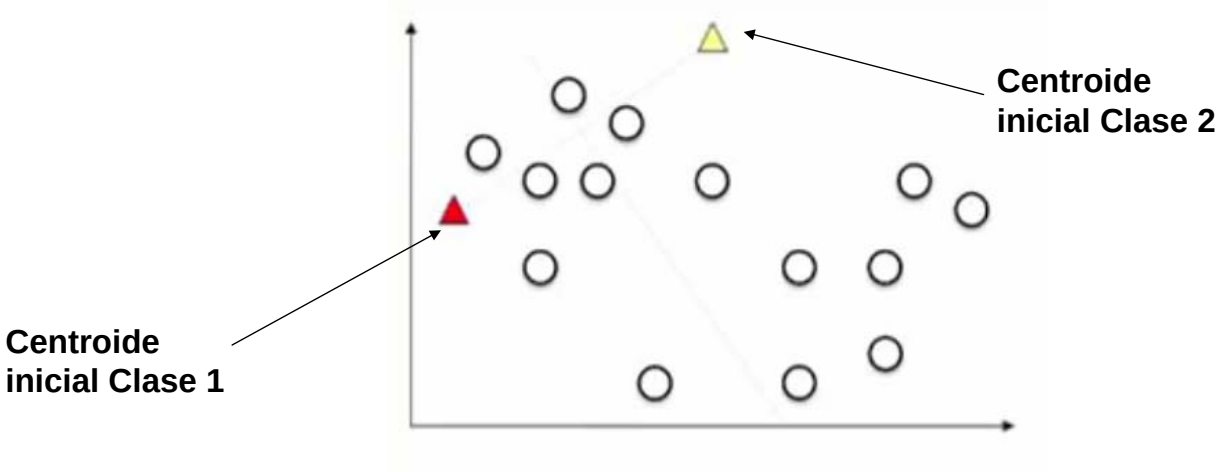
Figura 2: Diagrama en bloques de un clasificador de mínima distancia.

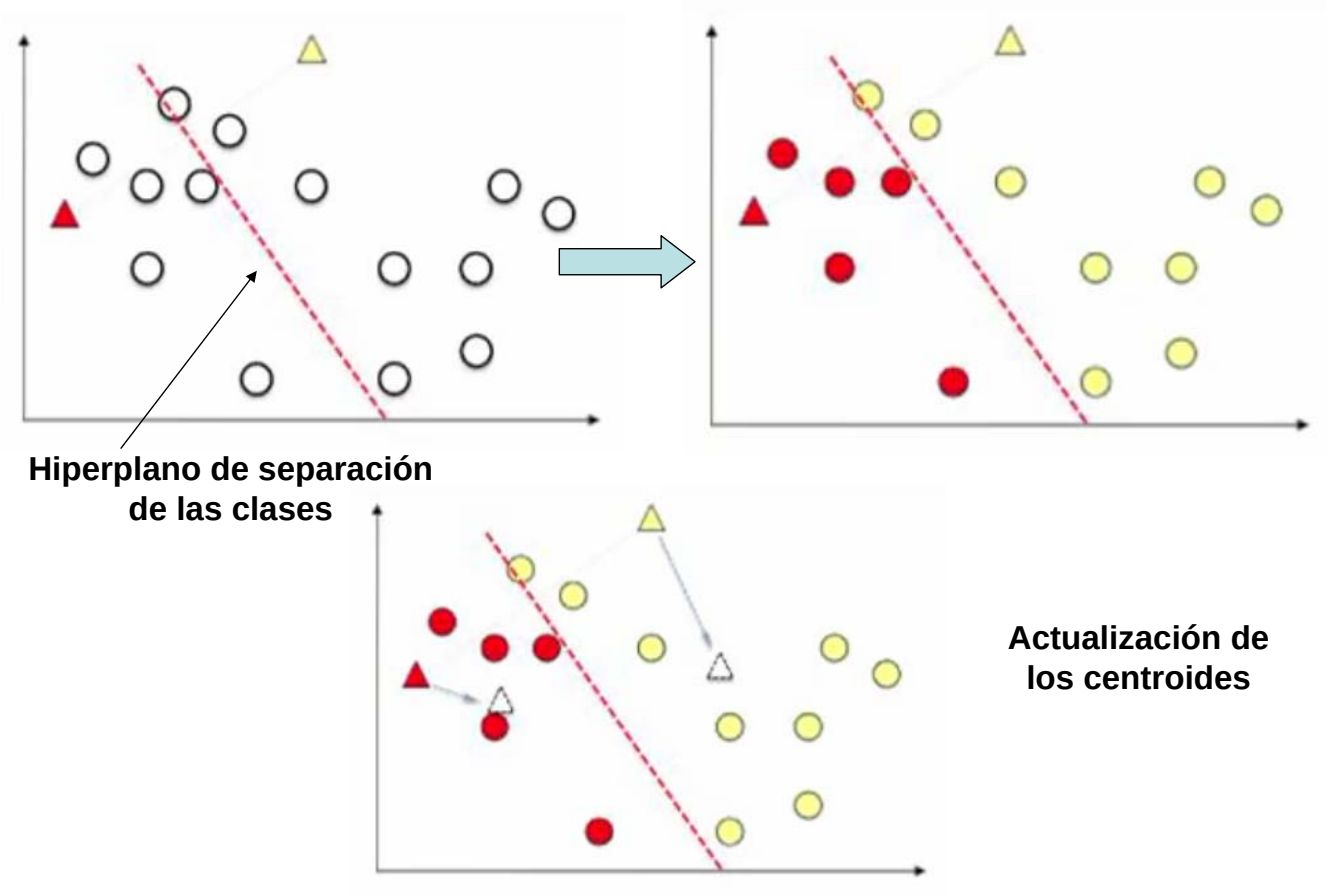
❑ Algoritmo de Clustering K-means

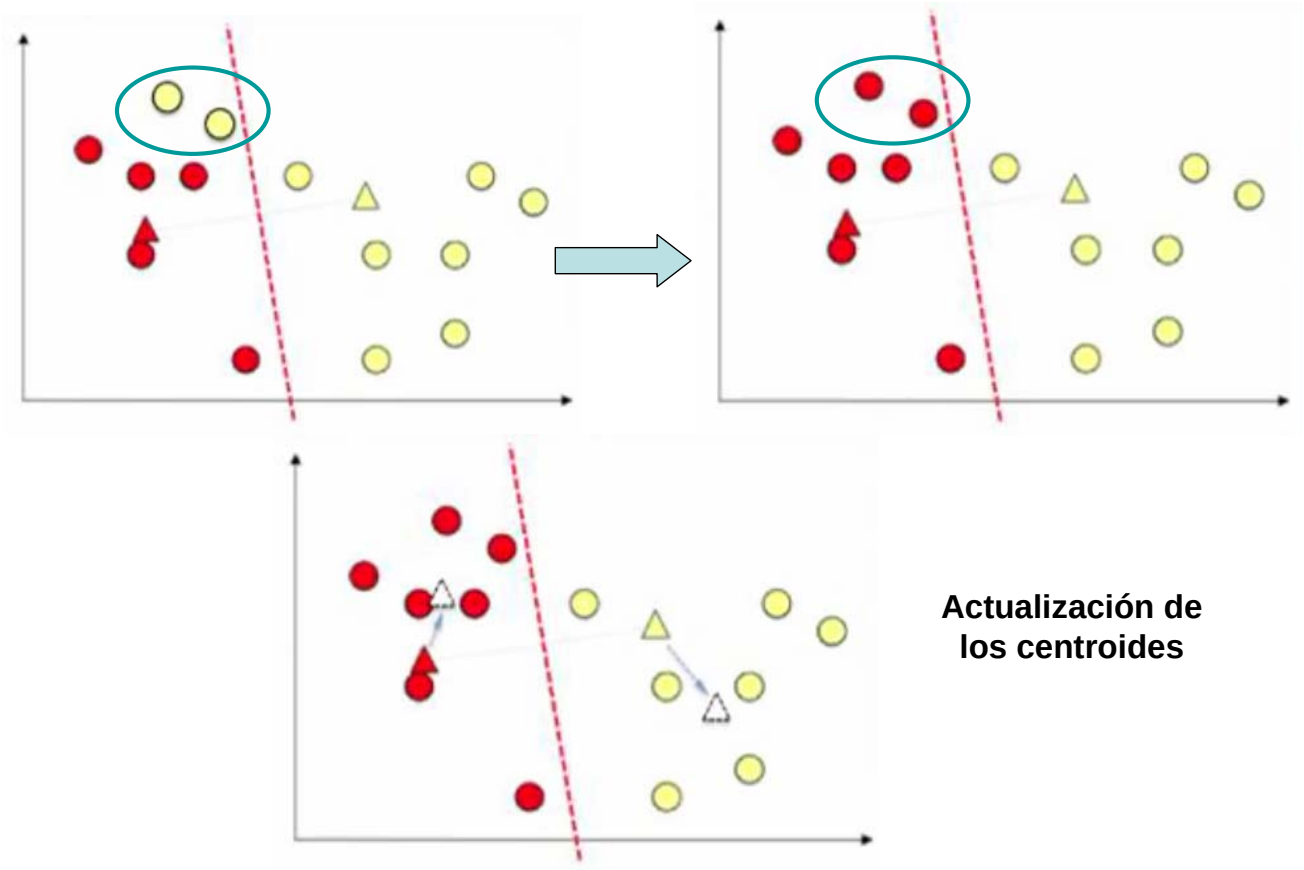
1. **Inicialización:** Se eligen arbitrariamente *M* vectores del conjunto de *L* vectores de entrenamiento como patrones iniciales de cada clase.
2. **Búsqueda del vecino más cercano:** para cada vector de entrenamiento se busca el patrón que sea el más cercano (de acuerdo a la medida de distancia que se esté empleando) y se asigna el vector de entrenamiento a la celda correspondiente a esa clase.
3. **Actualización de centroides:** se actualiza el vector patrón asociado a cada clase por el centroide de los vectores de entrenamiento asignados a esa celda.
4. **Iteración:** Se repiten los pasos 2 y 3 hasta que la distancia promedio cae por debajo de un umbral predeterminado.

• A la etapa de asignación se la denomina **Expectation,** en tanto que a la etapa de actualización de centroides se la denomina **Maximization**. De esta forma, el algoritmo de K-means es una variante del algoritmo **EM (Expectation-Maximization).**

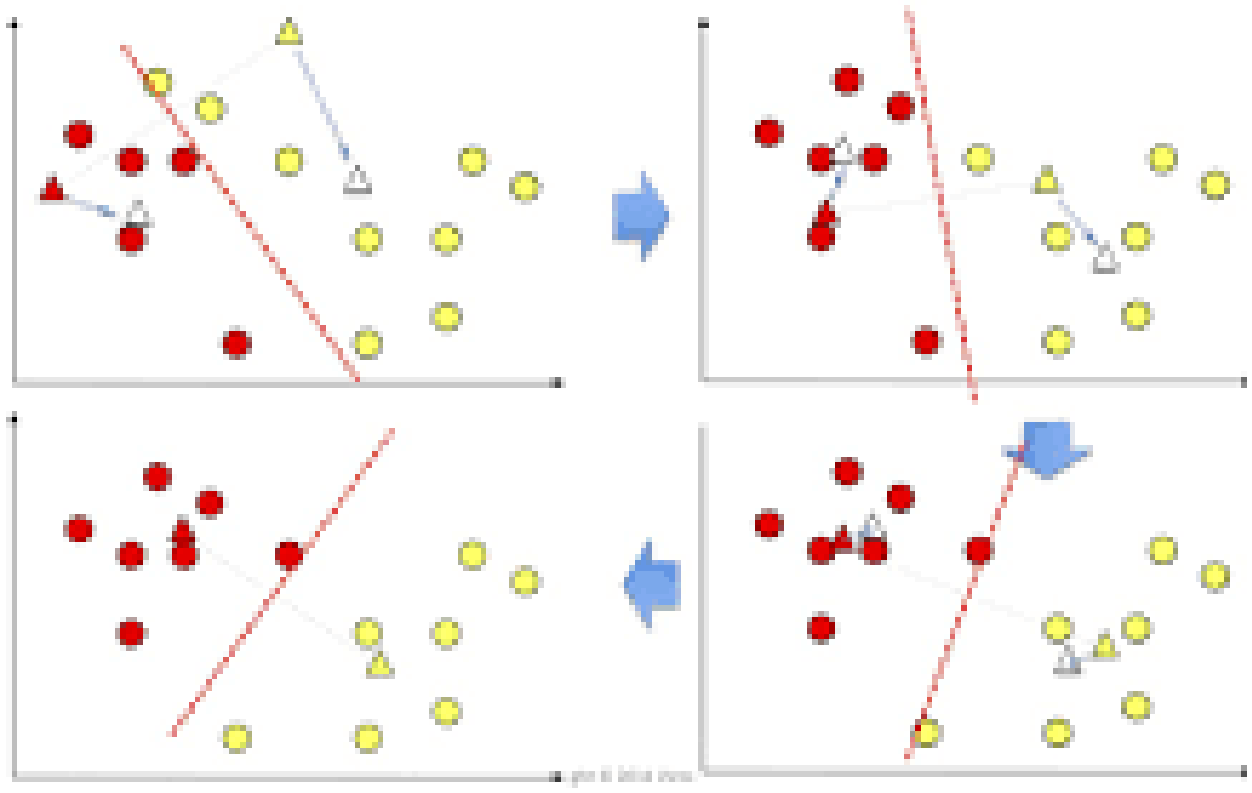
Ejemplo: 2 clases, 2 atributos, distancia Euclidea







Actualización de los centroides



Ejemplo: 3 clases