Clasificación Supervisada

Clasificadores basados en distancia

Jesús Ariel Carrasco Ochoa
Instituto Nacional de Astrofísica, Óptica y Electrónica

Distancia

- Función $d: A \times A \rightarrow \Re$
  - $\forall a,b \in A \; d(a,b) \geq 0$
  - $\forall a,b \in A \; d(a,b) = d(b,a)$
  - $\forall a,b,c \in A \; d(a,b) \leq d(a,c) + d(c,b)$
  - $\forall a \in A \; d(a,a) = 0$
  - $\forall a,b \in A \; d(a,b) = 0$ ssi $a = b$

Ejemplos de distancias

$$x = (x_1, x_2, \ldots, x_n)$$

$$y = (y_1, y_2, \ldots, y_n)$$

- Euclidiana

$$d(x, y) = \sqrt{\sum_{j=1}^{n} (x_j - y_j)^2}$$

Ejemplos de distancias

$$x = (x_1, x_2, \ldots, x_n)$$

$$y = (y_1, y_2, \ldots, y_n)$$

- Manhattan

$$d(x, y) = \sum_{j=1}^{n} \left| x_j - y_j \right|$$

Ejemplos de distancias

$$x = (x_1, x_2, \ldots, x_n)$$

$$y = (y_1, y_2, \ldots, y_n)$$

- Mahalanobis

$$d(x, y) = \sqrt{(x - y)^T \Sigma^{-1} (x - y)}$$

Vecino más Cercano

- Sea T un conjunto de entrenamiento con m objetos $O_1,\ldots,O_m$
- Cada objeto O_i descrito por n variables $x_1,\ldots,x_n$
- Los objetos de T están agrupados en r clases $C_1,\ldots,C_r$ y se conoce a la clase a la que pertenece cada objeto
- Sea d una función de distancia que permite comparar objetos

Vecino más Cercano

- Dado un nuevo objeto O
- Se busca entre los objetos de ***T*** el objeto más cercano ($\boldsymbol{O_{NN}}$) a ***O*** de acuerdo a ***d***

$$O_{NN} = \min_{O_i \in T} \{ d(O_i, O) \}$$

- Se asigna a ***O*** la clase de $\boldsymbol{O_{NN}}$

K Vecinos más Cercanos

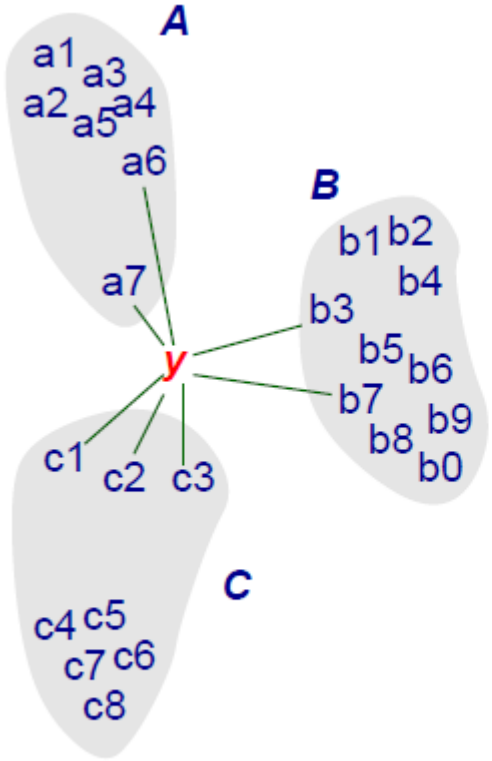
- Dado un nuevo objeto O
- Se busca entre los objetos de ***T*** los ***k*** objetos más cercanos $O_{NN1},\ldots,O_{NNk}$ a ***O*** de acuerdo a ***d***

$$O_{NN_1} = \min_{O_i \in T} \{d(O_i, O)\}$$

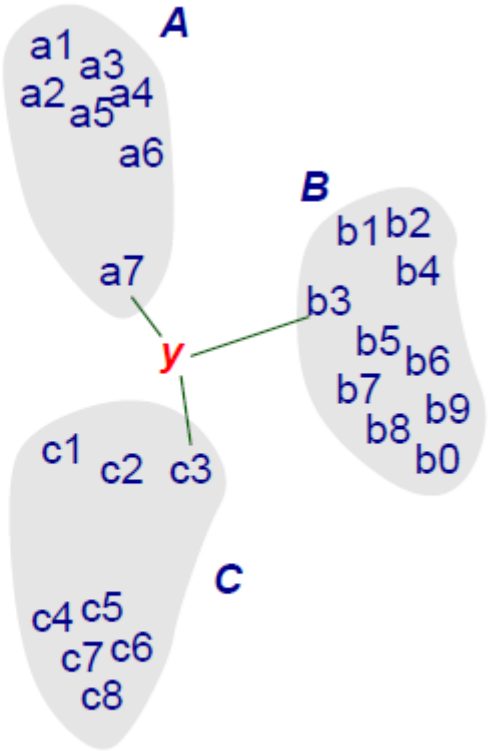
$$O_{NN_k} = \min_{O_i \in T - \{O_{NN_1},\ldots,O_{NN_{k-1}}\}} \{d(O_i, O)\}$$

- Se asigna a ***O*** la clase mas frecuente entre las clases de $O_{NN1},\ldots,O_{NNk}$

K Vecinos más Cercanos



7-NN(y)={c1,c2,c3,a7,a6,b3,b7}



NN(y)=a7

Problemas

- Falla si existen objetos muy parecidos en clases diferente
- Si ***T*** tiene muchos objetos, la búsqueda de los vecinos más cercanos puede ser muy costosa
- Se han desarrollado estrategias para encontrar el vecino más cercano de manera rápida
- La mayoría basada en el hecho de que una distancia cumple la desigualdad triangular
- Si la función de comparación no cumple la desigualdad triangular hay que buscar métodos alternativos para buscar los vecinos más cercanos

Clasificación basada en la distancia a la media de la clase

- Sea T un conjunto de entrenamiento con m objetos $O_1,\ldots,O_m$
- Cada objeto O_j descrito por n variables $x_1,\ldots,x_n$
- Los objetos de T están agrupados en r clases $C_1,\ldots,C_r$ y se conoce la clase de cada objeto
- Sea n_i el número de objetos de T en C_i
- Se tiene una función d que evalúa la distancia entre objetos

Clasificación basada en la distancia a la media de la clase

Se calcula la media μ_i para cada clase C_i

Dado un nuevo objeto O se le asigna la clase C_O si

$$C_O = \min_{C_i} \{ d(O, \mu_i) \}$$

Clasificación basada en la distancia a la media de la clase

Si ***d*** es la distancia Euclidiana

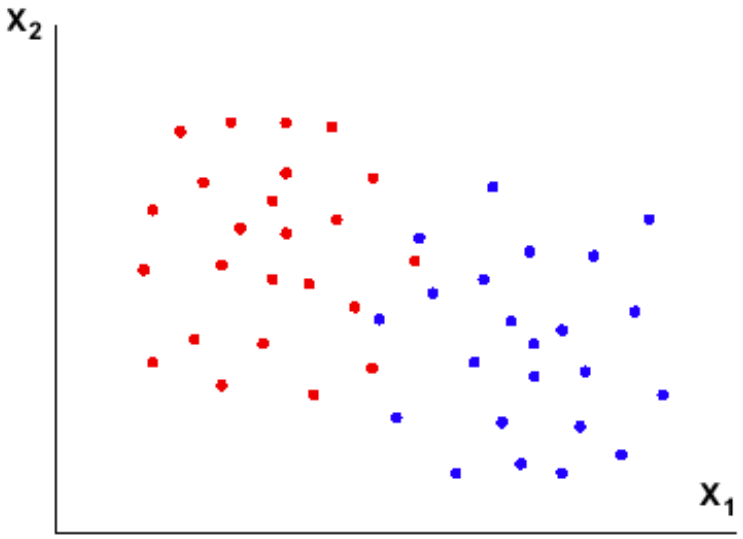
$$C_O = \min_{C_i} \left\{ \sqrt{\sum_{j=1}^{m} (x_j(O) - \mu_{ij})^2} \right\}$$

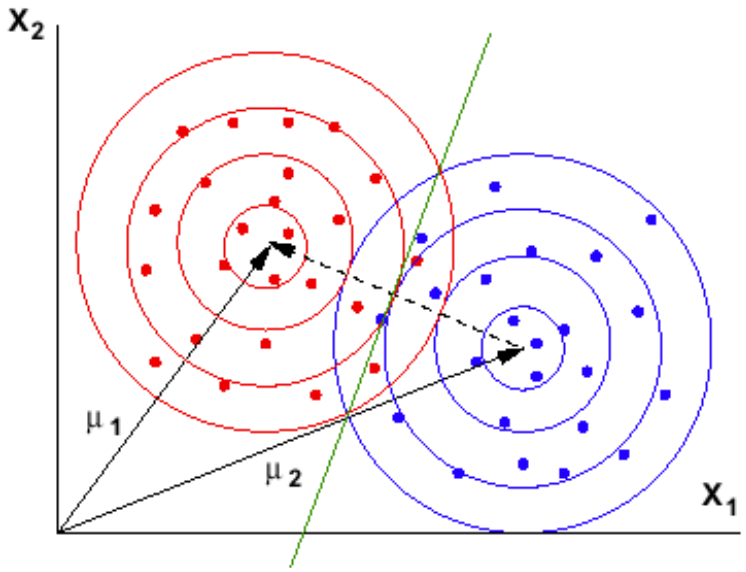
Como solamente se quiere encontrar el mínimo

$$C_O = \min_{C_i} \left\{ \sum_{j=1}^{m} (x_j(O) - \mu_{ij})^2 \right\}$$

Clasificación basada en la distancia a la media de la clase

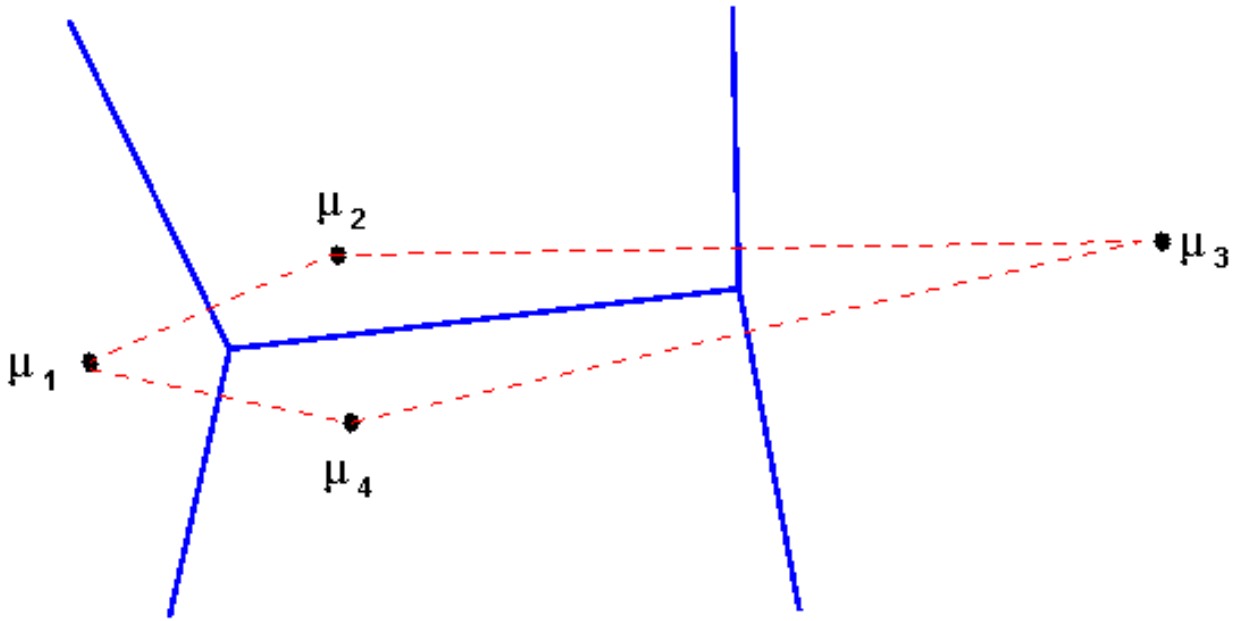
Si $m=2$ y $r=2$





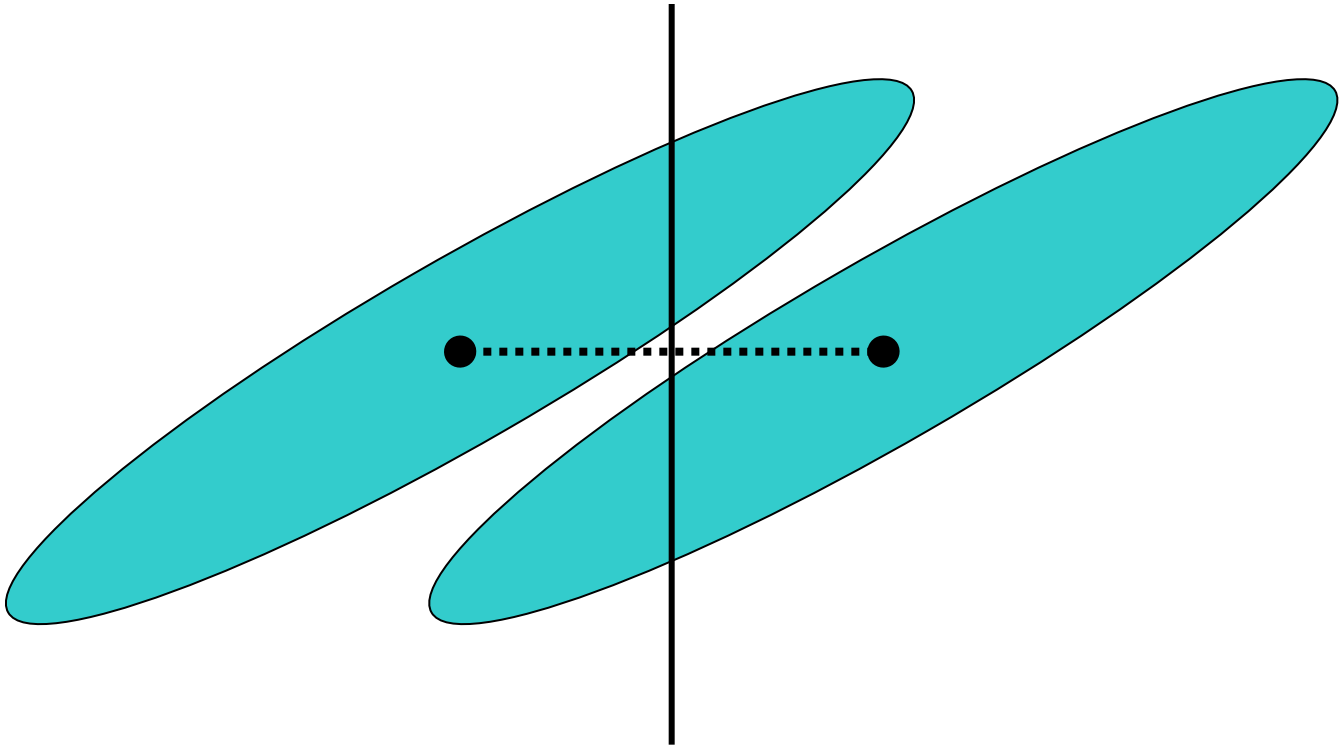
Clasificación basada en la distancia a la media de la clase

Si ***m***=2 y ***r***=4



Clasificación basada en la distancia a la media de la clase

En algunos casos la distancia Euclidiana puede ser inadecuada



Clasificación basada en la distancia a la media de la clase

Si ***d*** es la distancia de Mahalanobis

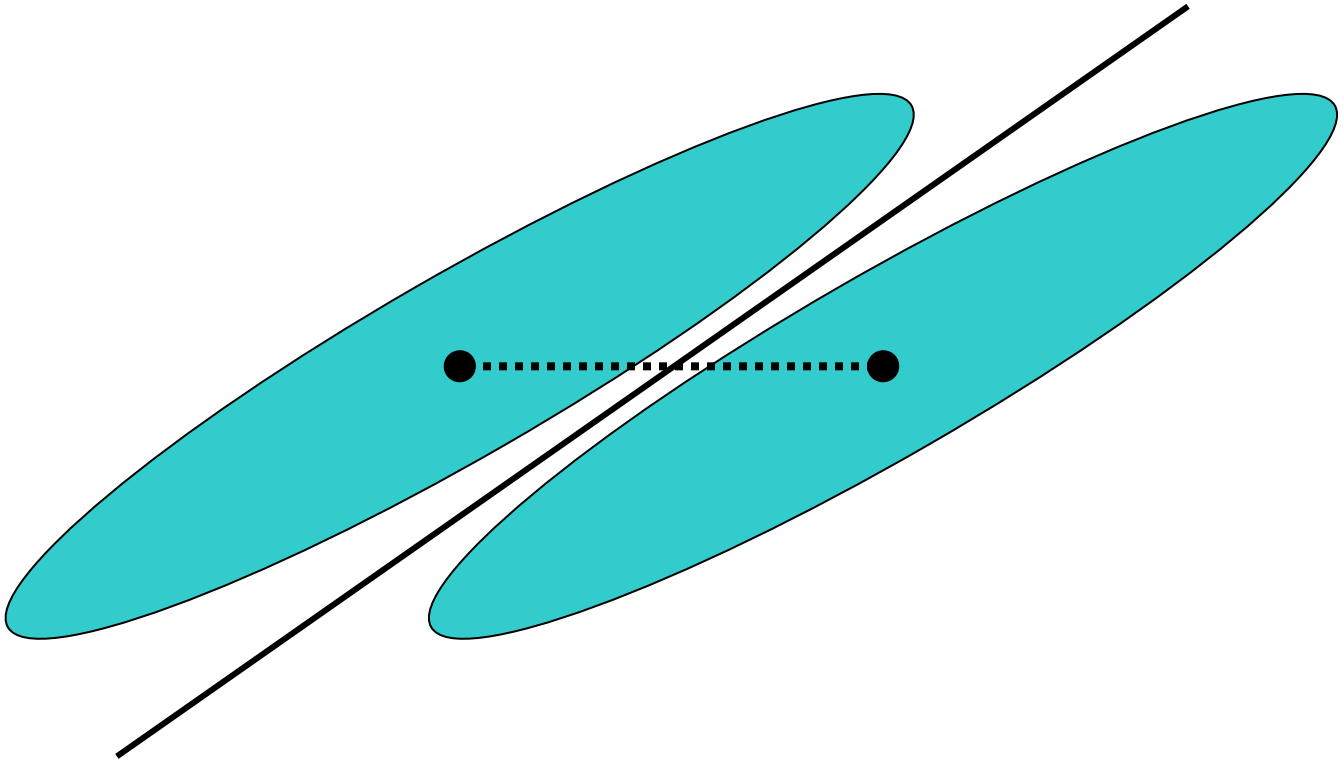
$$C_O = \min_{C_i} \left\{ \sqrt{(O-\mu_i)^T \Sigma_i^{-1} (O-\mu_i)} \right\}$$

Como solamente se quiere encontrar el mínimo

$$C_O = \min_{C_i} \left\{ (O-\mu_i)^T \Sigma_i^{-1} (O-\mu_i) \right\}$$

Clasificación basada en la distancia a la media de la clase

Con la distancia de Mahalanobis



Clasificación basada en la distancia a la media de la clase

Si

$$\Sigma_i = I \quad \rightarrow \quad \Sigma_i^{-1} = I$$

Como solamente se quiere encontrar el mínimo

$$C_O = \min_{C_i} \left\{ (O - \mu_i)^T (O - \mu_i) \right\}$$

Clasificación Supervisada

Clasificadores basados en distancia

Jesús Ariel Carrasco Ochoa
Instituto Nacional de Astrofísica, Óptica y Electrónica