# Clasificación estadística de patrones

## Clasificador gaussiano

César Martínez
cmartinez _at_ sinc.unl.edu.ar

Tópicos Selectos en Aprendizaje Maquinal
Doctorado en Ingeniería, FICH-UNL

12 de setiembre de 2016

# Aprendizaje de un clasificador: métodos

- Supervisado:
  - ↪ patrones de entrenamiento *etiquetados*.
  - ↪ búsqueda de regiones de decisión.
    - Aproximación paramétrica: fronteras de decisión definidas por las funciones de densidad de probabilidad de las clases. Método: clasificador gaussiano.
    - Aproximación no paramétrica: fronteras de decisión definidas por los prototipos. Métodos: núcleos, vecinos, árboles de decisión, otros.
- No supervisado:
  - ↪ patrones de entrenamiento de *clase incierta*.
  - ↪ agrupamiento de patrones sin etiquetar.
    - Métodos: con cantidad de clases conocida ($K$-medias) o desconocida, grafos, etc.

sinc($i$)

# Aprendizaje supervisado

Se desconocen $P(\omega_i)$ y $P(\mathbf{x}|\omega_i)$, pero se disponen de $N$ muestras de etiqueta conocida (*prototipos*):

$$(x_1, c_1), \ldots, (x_N, c_N) \text{ i.i.d según } P(\mathbf{x})$$

- Estimación de $P(\omega_i)$: $\hat{P}(\omega_i) = \frac{N_i}{N}$, con $N_i$: número de muestras de $\omega_i$.
- Estimación de $P(\mathbf{x}|\omega_i)$: $\hat{P}(\mathbf{x}|\omega_i)$
  - Métodos paramétricos: suponen una f.d.p. de cierta forma paramétrica conocida (p.ej., gaussiana), de parámetros desconocidos.
  - Métodos no paramétricos: sin suposición sobre $P(\mathbf{x}|\omega_i)$.

*Supuesto de independencia funcional*: densidades condicionales independientes entre sí $\Rightarrow$ $c$ problemas de estimación independientes.

sinc($i$)

# La función de densidad normal

- Función unidimensional: variable aleatoria con *fdp*

$$N(\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left( -\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2 \right)$$

- Densidad normal multidimensional:

$$N_d(\boldsymbol{\mu}, \Sigma) = \frac{1}{\sqrt{(2\pi)^d \, |\Sigma|}} \exp\left( -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu}) \right)$$

sinc($i$)

# Clasificador gaussiano

- Clasificador de Bayes cuyas densidades condicionales son gaussianas. Si $P(\mathbf{x}|\omega_i) \approx N_d(\boldsymbol{\mu}, \Sigma_i)$, entonces la discriminante asociada a $\omega_i$ es:

$$
\begin{aligned}
g_i(\mathbf{x}) &= \log P(\mathbf{x}|\omega_i) + \log P(\omega_i) \\
&= -\tfrac{d}{2}\log 2\pi - \tfrac{1}{2}\log|\Sigma_i| - \tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^t\Sigma_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) + \log P(\omega_i) \\
&= -\tfrac{1}{2}\mathbf{x}^t\Sigma_i^{-1}\mathbf{x} \\
&\quad +\boldsymbol{\mu}_i^t\Sigma_i^{-1}\mathbf{x} \\
&\quad -\tfrac{1}{2}(\boldsymbol{\mu}_i^t\Sigma_i^{-1}\boldsymbol{\mu}_i + d\log 2\pi + \log|\Sigma_i|) + \log P(\omega_i)
\end{aligned}
$$

sinc(*i*)

# Clasificador gaussiano

- Caso 1: $\Sigma_i = \sigma^2 I$.
  - Matrices de covarianza cuadradas e iguales para todas las clases.
  - Variables no correlacionadas.
  - Las muestras forman nubes hiperesféricas del mismo tamaño.

Como: $|\Sigma_i| = \sigma^{2d}$ y $\Sigma_i^{-1} = \frac{1}{\sigma^2} I$:

$$g_i(\mathbf{x}) = -\frac{1}{2\sigma^d}\mathbf{x}^t\mathbf{x} + \frac{1}{\sigma^2}\boldsymbol{\mu}_i^t\mathbf{x} - \frac{1}{2}\left(\frac{1}{\sigma^2}\boldsymbol{\mu}_i^t\boldsymbol{\mu}_i + d\log 2\pi + d\log\sigma^2\right) + \log P(\omega_i)$$

$$\equiv \frac{1}{\sigma^2}\boldsymbol{\mu}_i^t\mathbf{x} - \frac{1}{2\sigma^2}\boldsymbol{\mu}_i^t\boldsymbol{\mu}_i + \log P(\omega_i)$$

$$= \mathbf{w}_i^t\mathbf{x} + w_{i0},$$

con

$$\mathbf{w}_i = \frac{1}{\sigma^2}\boldsymbol{\mu}_i$$

$$w_{i0} = -\frac{1}{2\sigma^2}\boldsymbol{\mu}_i^t\boldsymbol{\mu}_i + \log P(\omega_i)$$

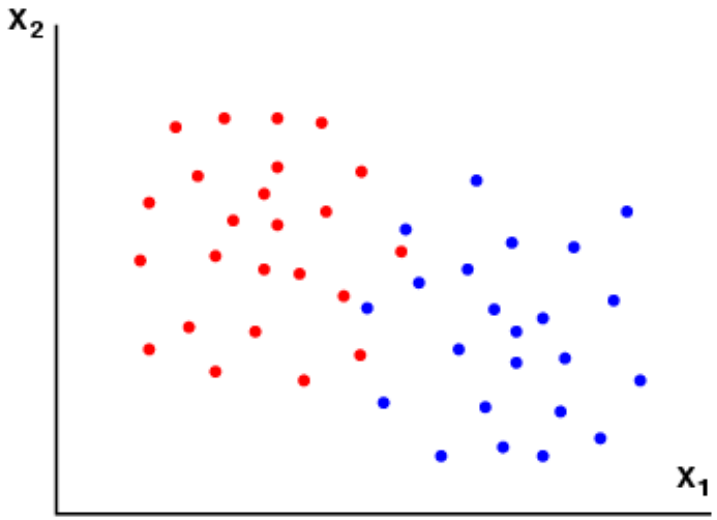
sinc(*i*)

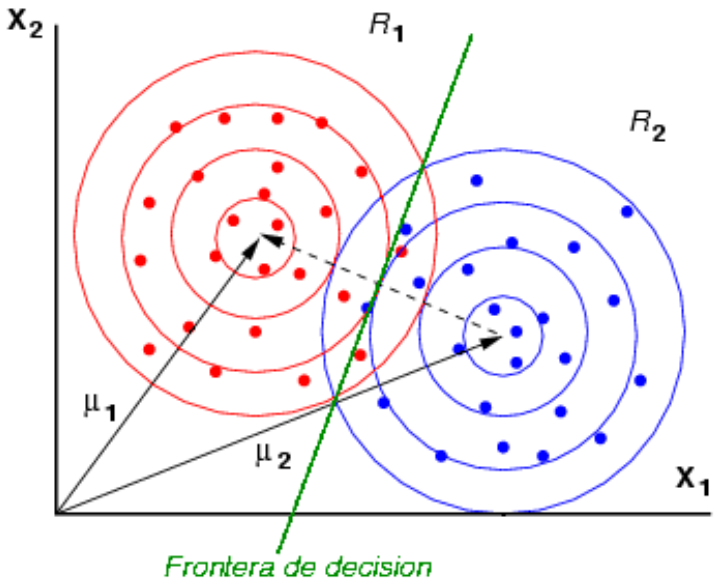
# Clasificador gaussiano

- Caso 1: $\Sigma_i = \sigma^2 I$.
  - Clasificador lineal por mínima distancia Euclídea, con discriminantes

$$g_i(\mathbf{x}) = -\frac{\|\mathbf{x} - \boldsymbol{\mu}_i\|^2}{2\sigma^2} + \log P(\omega_\mathbf{i})$$

  - Fronteras de decisión: hiperplanos.



# Clasificador gaussiano

- Caso 2: $\Sigma_i = \Sigma$.
  - Matrices de covarianza arbitrarias e iguales para todas las clases.
  - Variables correlacionadas.
  - Las muestras forman nubes hiperelípticas del mismo tamaño y forma.

$$g_i(\mathbf{x}) =$$

$$-\tfrac{1}{2}\mathbf{x}^t\Sigma^{-1}\mathbf{x} + \boldsymbol{\mu}_i^t\Sigma^{-1}\mathbf{x} - \tfrac{1}{2}\left(\boldsymbol{\mu}_i^t\Sigma^{-1}\boldsymbol{\mu}_i + d\log 2\pi + \log|\Sigma|\right) + \log P(\omega_i)$$

$$\equiv \boldsymbol{\mu}_i^t\Sigma^{-1}\mathbf{x} - \tfrac{1}{2}\boldsymbol{\mu}_i^t\Sigma^{-1}\boldsymbol{\mu}_i + \log P(\omega_i)$$

$$= \mathbf{w}_i^t\mathbf{x} + w_{i0},$$

con

$$\mathbf{w}_i = \Sigma^{-1}\boldsymbol{\mu}_i$$

$$w_{i0} = -\tfrac{1}{2}\boldsymbol{\mu}_i^t\Sigma^{-1}\boldsymbol{\mu}_i + \log P(\omega_i)$$

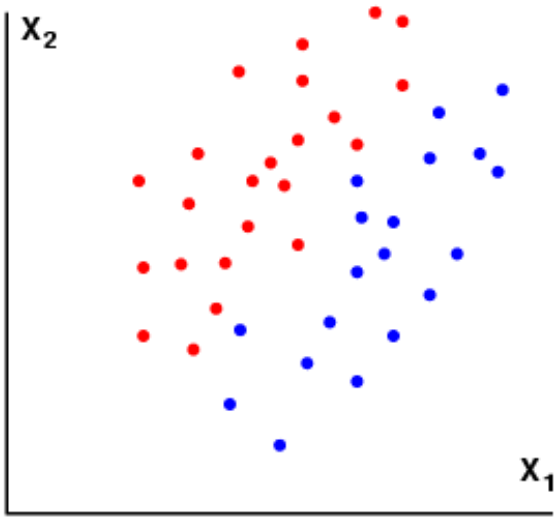
sinc(*i*)

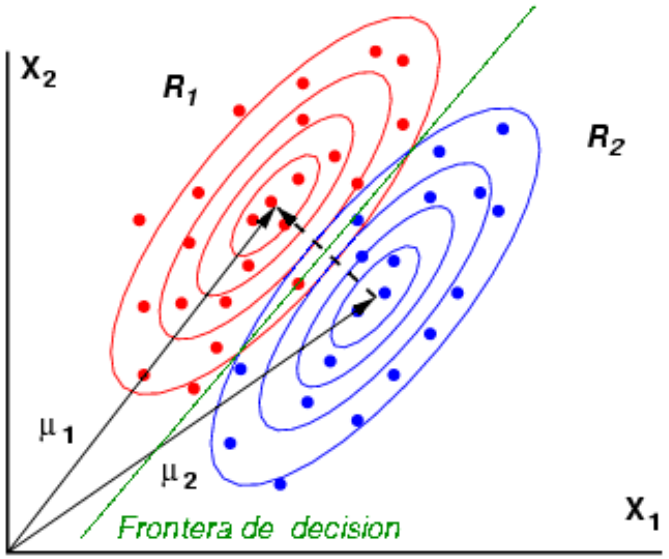
# Clasificador gaussiano

- Caso 2: $\Sigma_i = \Sigma$.
  - Clasificador lineal por mínima distancia de Mahalanobis:

$$g_i(\mathbf{x}) = -\frac{1}{2}\boldsymbol{\mu}_i^T \Sigma^{-1} \boldsymbol{\mu}_i + \boldsymbol{\mu}_i^T \Sigma^{-1} \mathbf{x} + \log P(\omega_i)$$

  - Fronteras de decisión: hiperplanos.



# Clasificador gaussiano

- Caso 3: $\Sigma_i$ arbitraria.
  - Matrices de covarianza arbitrarias y diferentes para cada clase, sin restricciones.
  - Las muestras forman nubes hiperelípticas cualquier tamaño y orientación.

$$g_i(\mathbf{x}) =$$

$$-\tfrac{1}{2}\mathbf{x}^t\Sigma_i^{-1}\mathbf{x}+\boldsymbol{\mu}_i^t\Sigma_i^{-1}\mathbf{x}-\tfrac{1}{2}\left(\boldsymbol{\mu}_i^t\Sigma_i^{-1}\boldsymbol{\mu}_i+d\log 2\pi+\log|\Sigma_i|\right)+\log P(\omega_i)$$

$$\equiv -\tfrac{1}{2}\mathbf{x}^t\Sigma_i^{-1}\mathbf{x}+\boldsymbol{\mu}_i^t\Sigma_i^{-1}\mathbf{x}-\tfrac{1}{2}\left(\boldsymbol{\mu}_i^t\Sigma_i^{-1}\boldsymbol{\mu}_i+\log|\Sigma_i|\right)+\log P(\omega_i)$$

$$=\mathbf{x}^tW_i\mathbf{x}+\mathbf{w}_i^t\mathbf{x}+w_{i0},$$

con

$$W_i=-\tfrac{1}{2}\Sigma_i^{-1}$$

$$\mathbf{w}_i=\Sigma_i^{-1}\boldsymbol{\mu}_i$$

$$w_{i0}=-\tfrac{1}{2}\boldsymbol{\mu}_i^t\Sigma_i^{-1}\boldsymbol{\mu}_i-\tfrac{1}{2}\log|\Sigma_i|+\log P(\omega_i)$$

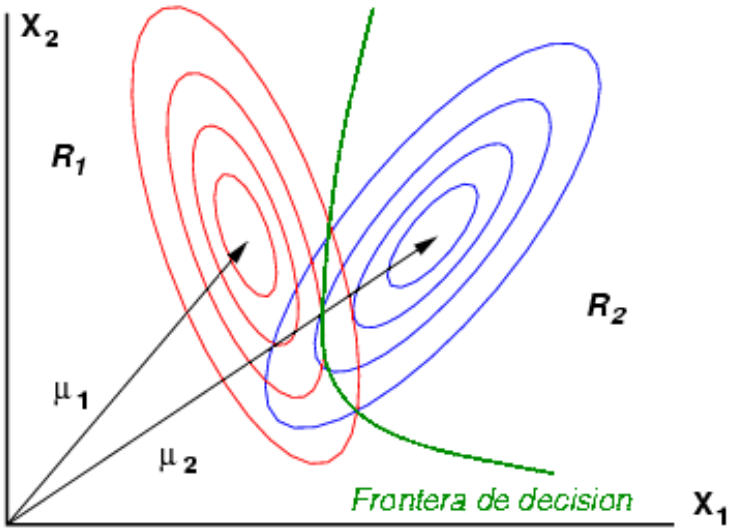
sinc(*i*)

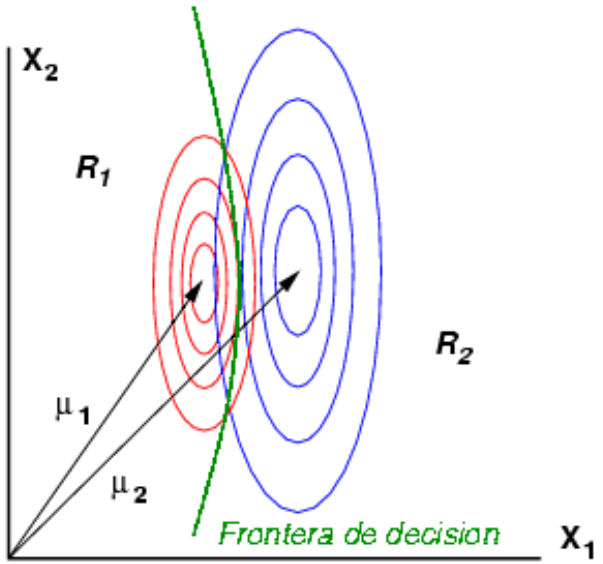
# Clasificador gaussiano

- Caso 3: $\Sigma_i$ arbitraria.
  - Clasificador cuadrático, con discriminantes

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T\Sigma_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) - \frac{1}{2}\log|\Sigma_i| + \log P(\omega_i)$$

  - Fronteras de decisión: hipercuádricas.



# Estimación de parámetros

- Aprendizaje: casos

  $P(\mathbf{x}|\omega_i) \approx N(\boldsymbol{\mu}_i, \Sigma_i)$, con $\boldsymbol{\mu}_i$ desconocida: $\boldsymbol{\Theta}_i = \boldsymbol{\mu}_i$

$$P(\mathbf{x}|\omega_i, \boldsymbol{\Theta}_i) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma_i|^{\frac{1}{2}}} \exp\left( -\frac{1}{2}(\mathbf{x} - \boldsymbol{\Theta}_i)^t \Sigma_i^{-1} (\mathbf{x} - \boldsymbol{\Theta}_i)\right)$$

  $P(\mathbf{x}|\omega_i) \approx N(\boldsymbol{\mu}_i, \Sigma_i)$, con $\boldsymbol{\mu}_i$ y $\Sigma_i$ desconocidas: $\boldsymbol{\Theta}_{i1} = \boldsymbol{\mu}_i$, $\boldsymbol{\Theta}_{i2} = \Sigma_i$

$$P(\mathbf{x}|\omega_i, \boldsymbol{\Theta}_i) = \frac{1}{(2\pi)^{\frac{d}{2}} |\boldsymbol{\Theta}_{i2}|^{\frac{1}{2}}} \exp\left( -\frac{1}{2}(\mathbf{x} - \boldsymbol{\Theta}_{i1})^t \boldsymbol{\Theta}_{i2}^{-1} (\mathbf{x} - \boldsymbol{\Theta}_{i1})\right)$$

- Estimación por máxima verosimilitud: proceso iterativo sobre los datos.

sinc($i$)

# Estimación de parámetros

- Principio general de estimación por máxima verosimilitud:
  Sean $X = \mathbf{x}_1, \ldots, \mathbf{x}_N$: muestras i.i.d. según $P(\mathbf{x}|\boldsymbol{\Theta})$.
  Como las muestras fueron extraídas independientemente:

$$P(X|\boldsymbol{\Theta}) = \prod_{k=1}^{N} P(\mathbf{x}_k|\boldsymbol{\Theta}),$$

  función denominada *verosimilitud* de $\boldsymbol{\Theta}$ respecto a $X$.
  La *estimación de máxima verosimilitud* de $\boldsymbol{\Theta}$ se define como:

$$\hat{\boldsymbol{\Theta}} = \arg\max_{\boldsymbol{\Theta}} \prod_{k=1}^{N} P(\mathbf{x}_k|\boldsymbol{\Theta}) = \arg\max_{\boldsymbol{\Theta}} \sum_{k=1}^{N} \log P(\mathbf{x}_k|\boldsymbol{\Theta})$$

  Se calcula mediante: $\boldsymbol{\nabla}_{\hat{\boldsymbol{\Theta}}} \left( \sum_{k=1}^{N} \log P(\mathbf{x}_k|\boldsymbol{\Theta}) \right) = \mathbf{0}$

sinc($i$)

# Estimación de parámetros

- Caso 1D: $\mu$ desconocida

  Tenemos que $\Theta = \mu$, entonces

$$P(x_k|\Theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x_k - \Theta)^2\right)$$

$$\log P(x_k|\Theta) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(x_k - \Theta)^2$$

$$\sum_{k=1}^{N} \nabla_\Theta \log P(x_k|\hat{\Theta}) = 0 \;\Rightarrow\; \hat{\Theta} = \hat{\mu} = \frac{1}{N}\sum_{k=1}^{N} x_k, \text{ (media muestral)}$$

# Estimación de parámetros

- Caso 1D: $\mu$ y $\sigma^2$ desconocidas

  Tenemos $\boldsymbol{\Theta} = \begin{pmatrix}\Theta_1 \\ \Theta_2\end{pmatrix}$, con $\Theta_1 = \mu$ y $\Theta_2 = \sigma^2$, así:

$$P(x_k|\boldsymbol{\Theta}) = \frac{1}{\sqrt{2\pi\Theta_2}} \exp\left(-\frac{1}{2\Theta_2}(x_k - \Theta_1)^2\right)$$

$$\log P(x_k|\boldsymbol{\Theta}) = -\frac{1}{2}\log(2\pi\Theta_2) - \frac{1}{2\Theta_2}(x_k - \Theta_1)^2$$

$$\sum_{k=1}^{N} \nabla_{\boldsymbol{\Theta}} \log P(x_k|\hat{\boldsymbol{\Theta}}) = \mathbf{0} \;\Rightarrow\; \begin{cases} \hat{\Theta}_1 = \hat{\mu} = \frac{1}{N}\sum_{k=1}^{N} x_k \\ \hat{\Theta}_2 = \hat{\sigma}^2 = \frac{1}{N}\sum_{k=1}^{N}(x_k - \hat{\mu})^2 \end{cases}$$

sinc($i$)

# Estimación de parámetros

- Caso $N$-D: $\boldsymbol{\mu}$ desconocido

$$\hat{\boldsymbol{\Theta}} = \hat{\boldsymbol{\mu}} = \frac{1}{N}\sum_{k=1}^{N}\mathbf{x}_k$$

- Caso $N$-D: $\boldsymbol{\mu}$ y $\Sigma$ desconocidos

$$\begin{cases} \hat{\boldsymbol{\Theta}}_1 = \hat{\boldsymbol{\mu}} = \frac{1}{N}\sum_{k=1}^{N}\mathbf{x}_k \\ \hat{\Theta}_2 = \hat{\sigma}^2 = \frac{1}{N}\sum_{k=1}^{N}(\mathbf{x}_k - \hat{\boldsymbol{\mu}})(\mathbf{x}_k - \hat{\boldsymbol{\mu}})^t \end{cases}$$

sinc($i$)

# Manos a la obra!

1. Sea un clasificador geométrico lineal definido por:
   $g_1(\mathbf{x}) = -x_1$
   $g_2(\mathbf{x}) = x_1 + x_2 - 1$
   $g_3(\mathbf{x}) = x_1 - x_2 - 1$
   - Calcule y grafique las fronteras y regiones de decisión.
   - Clasifique los puntos: (2,1); (2,-1); (-1,1); (-1,-1).
2. Sean $A$ y $B$ dos clases de igual probabilidad a priori definidas por:
   $P(\mathbf{x}|A) \rightsquigarrow N\left(\begin{pmatrix}3\\0\end{pmatrix}, \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\right)$
   $P(\mathbf{x}|B) \rightsquigarrow N\left(\begin{pmatrix}0\\3\end{pmatrix}, \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\right)$
   - Construya un clasificador gaussiano y encuentre la frontera.
   - Realice una representación gráfica del problema.
3. Implemente un clasificador gaussiano por ML para el mini-corpus de muestras proporcionado, realizando una representación gráfica de la situación.
   Calcule el desempeño del sistema al clasificar todas las muestras del corpus.

sinc($i$)

# Fin

- Bibliografía:


  - Duda R., Hart P, Stork D., Pattern Classification, Second Edition. Wiley-Interscience, 2000. Capítulos 1, 2.1 a 2.6, 3.1 a 3.2.

  - Theodoridis S., Koutroumbas K., Pattern Recognition, Fourth Edition. Elsevier, 2009. Capítulo 2 (2.1 a 2.5).

  - Murphy K., Machine Learning-A Probabilistic Perspective. The MIT Press, 2012. Capítulos 2 a 4 (secciones salteadas).


sinc(*i*)