UNIVERSIDADE
DE VIGO

Departamento de Estatística
e Investigación Operativa

Minería de Datos, 2008-2009

El Problema de Clasificación

El Problema de Clasificación

1. **Problema de clasificación.**

   a) Sea una población X con valores en $\mathcal{X} \subset \mathbb{R}^d$, cuyos elementos pertenecen a una de c posibles clases $\mathcal{C}_j$, $j = 1, .., c$.

   b) **Regla de clasificación**. Una regla de clasificación es una función $g : \mathcal{X} \to \{\mathcal{C}_1, ..., \mathcal{C}_c\}$ tal que:

   $$g(\boldsymbol{x}) = \left\{ \begin{array}{cc} \mathcal{C}_1 & \text{si } \boldsymbol{x} \in \mathcal{X}_1 \\ \vdots & \vdots \\ \mathcal{C}_c & \text{si } \boldsymbol{x} \in \mathcal{X}_c \end{array} \right| \text{ con } \{\mathcal{X}_j\}_{j=1}^c \text{ partición de } \mathcal{X} : \left\{ \begin{array}{l} \cup_{j=1}^c \mathcal{C}_j = \mathcal{X} \\ \mathcal{C}_j \cap \mathcal{C}_l = \varnothing \text{ si } j \neq l \end{array} \right. \quad (1)$$

   es decir, que asigna cada ejemplo $\mathsf{X} = \boldsymbol{x}$ a la clase $\mathcal{C}_j$ si $\boldsymbol{x} \in \mathcal{X}_j$.

   c) Se trata de determinar una regla de clasificación que proporcione la mejor clasificación.

   d) Es necesario definir algún **criterio de bondad de la clasificación** que permita elegir entre diversas reglas de clasificación alternativas.

2. El marco estadístico general del problema supone una variable $\mathsf{Z} = (\mathsf{X}, \mathsf{Y})$ donde $\mathsf{X} \in \mathcal{X} \subset \mathbb{R}^d$ es el vector de covariables e $\mathsf{Y} \in \mathcal{I}$ es la variable respuesta que identifica las clases, siendo $\mathcal{I}$ un conjunto de valores etiqueta de las clases cuya codificación se elige en función de las técnicas de clasificación a utilizar para resolver el problema.

   a) En el caso de dos clases, codificaciones frecuentes son $\mathcal{I} = \{0, 1\}$ ó $\mathcal{I} = \{-1, 1\}$.

   b) En el caso de $c > 2$ clases, es frecuente utilizar $\mathcal{I} = \{1, 2, ..., c\}$.

   c) Otra codificación frecuente válida para dos o más clases es la basada en una variable vectorial de c componentes: $\mathsf{Y} = (\mathsf{Y}_1, ..., \mathsf{Y}_c)^T \in \{0, 1\}^c$ de tal forma que $\mathcal{C}_j \equiv \mathsf{Y}_j = 1$ e $\mathsf{Y}_k = 0$ $\forall k \neq j$.

3. Se definen:

   a) Las **probabilidades a priori** de las clases $\mathbb{P}(\mathcal{C}_j)$, $j = 1, ..., c$, es decir, la proporción de ejemplos ($\boldsymbol{x} \in \mathcal{X}$) de cada clase $\mathcal{C}_j$, verificándose que $\sum_{j=1}^c \mathbb{P}(\mathcal{C}_j) = 1$.

   b) Las **distribuciones marginales** (funciones de masa o de densidad de probabilidad) de X en cada clase $\mathcal{C}_j$:

   $$\mathsf{p}_j(\boldsymbol{x}) = \mathbb{P}(\mathsf{X} = \boldsymbol{x}|\mathcal{C}_j) = \mathbb{P}(\mathsf{X} = \boldsymbol{x}|\mathsf{Y} = y_j),\ j = 1, ..., c$$

   c) Las **probabilidades a posteriori** de las clases $\mathbb{P}(\mathcal{C}_j|\boldsymbol{x})$, que especifican la probabilidad de pertenencia de cada $\boldsymbol{x} \in \mathcal{X}$ a cada una de las clases.

4. En lo que sigue, nos centramos en un problema con sólo dos clases $\mathcal{C}_1, \mathcal{C}_2$.

Criterios de Bondad de la Clasificación

Criterio 1. Minimización de la Probabilidad de Error

1. Para una regla de clasificación g, la probabilidad de error es (teorema de la probabilidad total):

$$\mathbb{P}(\text{error}) = \mathbb{P}(\text{error}|\boldsymbol{x} \text{ está en } \mathcal{C}_1)\mathbb{P}(\mathcal{C}_1) + \mathbb{P}(\text{error}|\boldsymbol{x} \text{ está en } \mathcal{C}_2)\mathbb{P}(\mathcal{C}_2) \quad (2)$$

2. Se cumple el siguiente lema (la integral de una función es mínima sobre el conjunto en el que la función toma valores negativos):

   Lema 1 *Para cualquier función $h : \mathcal{X} \subset \mathbb{R}^d \to \mathbb{R}$, se verifica que:*

$$\Omega_0 = \{\boldsymbol{x} : h(\boldsymbol{x}) < 0\} = \arg\min_{\Omega} \int_{\Omega} h(\boldsymbol{x}) d\boldsymbol{x}.$$

   es decir, el valor mínimo de la integral de una función se obtiene sobre el conjunto en el que esa función es negativa.

3. Para una regla de clasificación (1), la probabilidad de error es (utilizando (2)):

$$\begin{aligned}
\mathbb{P}(\text{error}) &= \int_{\mathcal{X}_2} \mathsf{p}_1(\boldsymbol{x}) d\boldsymbol{x} \cdot \mathbb{P}(\mathcal{C}_1) + \int_{\mathcal{X}_1} \mathsf{p}_2(\boldsymbol{x}) d\boldsymbol{x} \cdot \mathbb{P}(\mathcal{C}_2) \\
&= \left(1 - \int_{\mathcal{X}_1} \mathsf{p}_1(\boldsymbol{x}) d\boldsymbol{x}\right) \cdot \mathbb{P}(\mathcal{C}_1) + \int_{\mathcal{X}_1} \mathsf{p}_2(\boldsymbol{x}) d\boldsymbol{x} \cdot \mathbb{P}(\mathcal{C}_2) \\
&= \mathbb{P}(\mathcal{C}_1) + \int_{\mathcal{X}_1} \left[\mathbb{P}(\mathcal{C}_2)\mathsf{p}_2(\boldsymbol{x}) - \mathbb{P}(\mathcal{C}_1)\mathsf{p}_1(\boldsymbol{x})\right] d\boldsymbol{x}
\end{aligned}$$

   Aplicando el lema anterior, dicha probabilidad se minimiza si se elige $\mathcal{X}_1 = \Omega_0 = \{\boldsymbol{x} : [\mathbb{P}(\mathcal{C}_2)\mathsf{p}_2(\boldsymbol{x}) - \mathbb{P}(\mathcal{C}_1)\mathsf{p}_1(\boldsymbol{x})] < 0\}$, es decir, cuando se utiliza la regla de clasificación:

$$g(\boldsymbol{x}) = \begin{cases} \mathcal{C}_1 & \text{si } \dfrac{\mathsf{p}_1(\boldsymbol{x})}{\mathsf{p}_2(\boldsymbol{x})} > \dfrac{\mathbb{P}(\mathcal{C}_2)}{\mathbb{P}(\mathcal{C}_1)} \\ \mathcal{C}_2 & \text{caso contrario} \end{cases}$$

   (en caso de igualdad vale una asignación arbitraria).

   Por tanto, con el criterio de minimización de la probabilidad de error, la regla óptima de clasificación es la que se basa en que el ratio de verosimilitud supere al ratio inverso de probabilidades a priori.

Criterio 2. Maximización de la Probabilidad A Posteriori. Regla de Bayes

1. El criterio consiste en elegir la regla que maximiza la probabilidad a posteriori $\mathbb{P}(\mathcal{C}_j|\boldsymbol{x})$, $j = 1, 2$. Es decir, asigna $\boldsymbol{x}$ a la clase más verosímil.

2. Dicha regla se denomina **regla de Bayes** que, por tanto, asigna $\boldsymbol{x} \in \mathcal{X}$ a la clase $\mathcal{C}_1$ si $\mathbb{P}(\mathcal{C}_1|\boldsymbol{x}) > \mathbb{P}(\mathcal{C}_2|\boldsymbol{x})$, y a la clase $\mathcal{C}_2$ en caso contrario. Pero, si aplicamos la definición de probabilidad condicionada, se obtiene:

$$\mathbb{P}(\mathcal{C}_1|\boldsymbol{x}) > \mathbb{P}(\mathcal{C}_2|\boldsymbol{x}) \Leftrightarrow \frac{\mathbb{P}(\mathcal{C}_1)\mathsf{p}(\boldsymbol{x}|\mathcal{C}_1)}{\mathsf{p}_{\mathsf{X}}(\boldsymbol{x})} > \frac{\mathbb{P}(\mathcal{C}_2)\mathsf{p}(\boldsymbol{x}|\mathcal{C}_2)}{\mathsf{p}_{\mathsf{X}}(\boldsymbol{x})} \Leftrightarrow \frac{\mathsf{p}_1(\boldsymbol{x})}{\mathsf{p}_2(\boldsymbol{x})} > \frac{\mathbb{P}(\mathcal{C}_2)}{\mathbb{P}(\mathcal{C}_1)}$$

   con lo que la regla de Bayes es equivalente a la que minimiza la probabilidad de error.

Criterio 3. Maximización del Ratio de Verosimilitud

1. El ratio de verosimilitud es $\mathsf{p}_1(\boldsymbol{x})/\mathsf{p}_2(\boldsymbol{x})$, luego la regla que maximiza dicho ratio es un caso particular de la anterior cuando $\mathbb{P}(\mathcal{C}_1) = \mathbb{P}(\mathcal{C}_2) = 1/2$.

   Por tanto, dicha regla asigna $\boldsymbol{x} \in \mathcal{X}$ a la clase $\mathcal{C}_1$ si $\mathsf{p}_1(\boldsymbol{x})/\mathsf{p}_2(\boldsymbol{x}) > 1$.

2. **Ejemplo**. En hipótesis de normalidad en ambas clases con igual matriz de covarianzas Σ:

$$\mathsf{p}_j(\boldsymbol{x}) = (2\pi)^{-d/2}\,|\Sigma|^{-1/2}\exp[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_j)^T\Sigma(\boldsymbol{x}-\boldsymbol{\mu}_j)],\ j=1,2$$

   se tiene:

$$\begin{aligned}\frac{\mathsf{p}_1(\boldsymbol{x})}{\mathsf{p}_2(\boldsymbol{x})} &= \exp\{\frac{1}{2}[(\boldsymbol{x}-\boldsymbol{\mu}_2)^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2) - (\boldsymbol{x}-\boldsymbol{\mu}_1)^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)]\}\\ \Rightarrow\ \ln\frac{\mathsf{p}_1(\boldsymbol{x})}{\mathsf{p}_2(\boldsymbol{x})} &= \frac{1}{2}[(\boldsymbol{x}-\boldsymbol{\mu}_2)^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2) - (\boldsymbol{x}-\boldsymbol{\mu}_1)^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)]\end{aligned}$$

   que es la diferencia de las **distancias de Mahalanobis**[1] del vector $\boldsymbol{x}$ a clasificar, con respecto a las medias de las clases, y puede escribirse:

$$\begin{aligned}&= (\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T\Sigma^{-1}\boldsymbol{x} - \frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T\Sigma^{-1}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)\\ &= \mathbf{w}^T[\boldsymbol{x} - \frac{1}{2}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)]\\ &= \mathbf{w}^T(\boldsymbol{x}-\boldsymbol{\mu}) = \mathbf{w}^T\boldsymbol{x} - \mathbf{w}^T\boldsymbol{\mu}\\ &= \mathbf{w}^T\boldsymbol{x} - \mu\end{aligned}$$

   donde $\boldsymbol{\mu} = \frac{1}{2}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)$ es la media global, $\mathbf{w} = \Sigma^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$ se denomina **dirección discriminante de Fisher** (en su versión poblacional), y:

$$\begin{aligned}\mu &= \mathbf{w}^T\boldsymbol{\mu} = \mathbf{w}^T\frac{1}{2}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2) = \frac{1}{2}(\mu_1+\mu_2)\\ \mu_j &= \mathbf{w}^T\boldsymbol{\mu}_j, j=1,2\end{aligned}$$

   son, respectivamente, la media global y las de cada clase proyectadas sobre la dirección discriminante.

   a) La regla resultante es[2] (véase Figura 1):

$$g(\boldsymbol{x}) = \begin{cases}\mathcal{C}_1 & \text{si } \ln\dfrac{\mathsf{p}_1(\boldsymbol{x})}{\mathsf{p}_2(\boldsymbol{x})} = \mathbf{w}^T\boldsymbol{x} - \mu > 0\\ \mathcal{C}_2 & \text{en caso contrario}\end{cases}$$

   Con ello, la regla de clasificación asigna un vector $\boldsymbol{x} \in \mathcal{X}$ a la clase $\mathcal{C}_j$ si su proyección sobre la dirección discriminante cae a un lado u otro del punto medio μ de los puntos proyectados en dicha dirección (cae del lado más cercano a μ_j).

[1] Si $\mathbf{X}$ es una v.a. vectorial, se define la distancia de Mahalanobis entre dos puntos $\mathbf{x}, \mathbf{x}'$ como $d(\mathbf{x},\mathbf{x}') = (\mathbf{x}-\mathbf{x}')^T\Sigma^{-1}(\mathbf{x}-\mathbf{x}'))$. Esta distancia modifica la distancia euclídea ponderando **inversamente** las direcciones de máxima variabilidad. En una dirección de gran varianza, la distancia de Mahalanobis entre dos puntos resulta *menor* que la euclídea. Si todas las componentes aleatorias de X poseen la misma varianza, la distancia de Mahalanobis y la euclídea son la misma.

[2] Nótese que $\ln(u) < 0$ si $0 < u < 1$, $\ln(1) = 0$ y $\ln(u) > 0$ si $u > 1$.

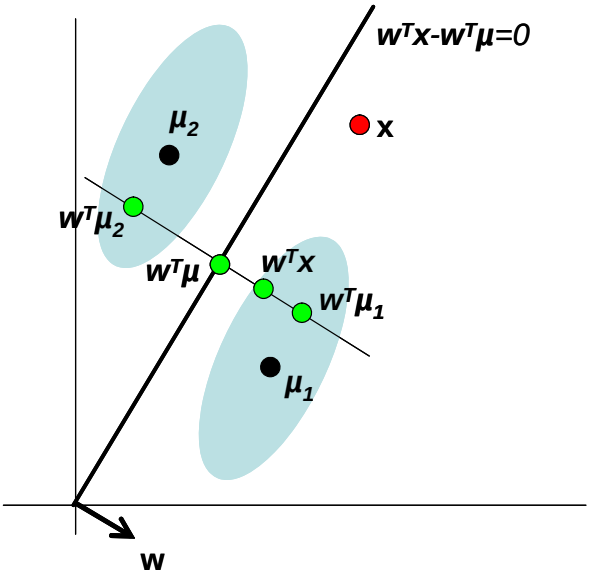

Figura 1: Discriminante Lineal de Fisher. Nótese como el punto $\mathbf{x}$ se considera más cercano a $\mathcal{C}_1$ (distancia de Mahalanobis) al estar en la dirección de mayor variabilidad de $\mathsf{p}_1(\mathbf{x})$. Así, el hiperplano discriminante de Fisher, asigna $\mathbf{x}$ a $\mathcal{C}_1$.

3. En el ejemplo anterior, la regla que minimiza el error de clasificación y su equivalente, la regla de Bayes, producen como función discriminante:

$$\mathbf{w}^T \boldsymbol{x} - \mu = \ln \frac{\mathbb{P}(\mathcal{C}_2)}{\mathbb{P}(\mathcal{C}_1)} \tag{3}$$

que, si $\mathbb{P}(\mathcal{C}_1) = \mathbb{P}(\mathcal{C}_2) = 1/2$, coincide con el discriminante de Fisher poblacional.

Regresión (Discriminación) Logística

1. Cuando las distribuciones condicionadas $\mathsf{p}(\boldsymbol{x}|\mathcal{C}_j)$ son normales y poseen igual matriz de varianzas-covarianzas Σ, vimos que el discriminante de Fisher (poblacional) con probabilidades a priori $\mathbb{P}(\mathcal{C}_j)$ iguales, minimiza los tres criterios de bondad estudiados, con lo que **su versión muestral** produce buenos resultados.

2. Sin embargo, cuando este no es el caso, el discriminante de Fisher (muestral) puede producir peores resultados.

   Una alternativa es estimar la probabilidad a posteriori $\mathbb{P}(\mathcal{C}_1|\boldsymbol{x}) \doteq \mathsf{f}(\boldsymbol{x})$ (la otra es $1-\mathsf{f}(\boldsymbol{x})$) mediante alguna función $\hat{f}(\boldsymbol{x})$ y utilizar la denominada **regla de clasificación *plug-in*** :

$$g_{\hat{f}}(\boldsymbol{x}) = \begin{cases} \mathcal{C}_1 & \text{si } \hat{f}(\boldsymbol{x}) > 1/2 \\ \mathcal{C}_2 & \text{si } \hat{f}(\boldsymbol{x}) \leq 1/2 \end{cases}$$

3. **Enfoques posibles** para estimar $\mathsf{f}(\boldsymbol{x}) = \mathbb{P}(\mathcal{C}_1|\boldsymbol{x})$:

a) **Regresión por mínimos cuadrados:** Si etiquetamos las clases por medio de una variable aleatoria $\mathsf{Y} \in \{0, 1\}$ de tal forma que la clase $\mathcal{C}_1$ se corresponde con $\mathsf{Y} = 1$ y $\mathcal{C}_2$ con $\mathsf{Y} = 0$, entonces:

$$\mathbb{E}(\mathsf{Y}|\boldsymbol{x}) = 1 \cdot \mathbb{P}(\mathsf{Y} = 1|\boldsymbol{x}) + 0 \cdot \mathbb{P}(\mathsf{Y} = 0|\boldsymbol{x}) = \mathbb{P}(\mathsf{Y} = 1|\boldsymbol{x})$$

con lo que la función de regresión $\mathbb{E}(\mathsf{Y}|\boldsymbol{x})$ coincide con $\mathbb{P}(\mathsf{Y} = 1|\boldsymbol{x})$ y cualquier método de regresión que produzca una estimación $\hat{f}(\boldsymbol{x}) = \widehat{\mathbb{E}}(\mathsf{Y}|\boldsymbol{x}) = \widehat{\mathbb{P}}(\mathsf{Y} = 1|\boldsymbol{x})$ permite construir una regla de clasificación tipo *plug-in.*

Problema: pueden resultar valores estimados mayores que 1 o menores que cero.

b) **Estimación mediante máxima verosimilitud**. Como $\mathsf{Y}|\boldsymbol{x} \sim Be(p)$ con $p = \mathsf{f}(\boldsymbol{x}) = \mathbb{P}(\mathsf{Y} = 1|\boldsymbol{x})$, la distribución a posteriori es:

$$\begin{array}{ll} \text{si } \mathsf{Y} \in \{0, 1\} & \mathsf{p}_{\mathsf{Y}|\boldsymbol{x}}(y) = \mathsf{f}(\boldsymbol{x})^y[1 - \mathsf{f}(\boldsymbol{x})]^{1-y} \\ \text{si } \mathsf{Y} \in \{-1, 1\} & \mathsf{p}_{\mathsf{Y}|\boldsymbol{x}}(y) = \mathsf{f}(\boldsymbol{x})^{(1+y)/2}[1 - \mathsf{f}(\boldsymbol{x})]^{(1-y)/2} \end{array}$$

pues se pasa de la codificación $\mathsf{Y} \in \{0, 1\}$ a la codificación $\mathsf{Y} \in \{-1, 1\}$ mediante la transformación $y' = 2y - 1$ y a la inversa, mediante $y = \frac{1}{2}(1 + y')$.

Así, en lugar de mínimos cuadrados, se utiliza la verosimilitud. Por ejemplo, en el caso $\mathsf{Y} \in \{0, 1\}$:

$$-\ln \mathsf{p}_{\mathsf{Y}|\boldsymbol{x}}(y) = -[y \ln f(\boldsymbol{x}) + (1 - y) \ln(1 - f(\boldsymbol{x}))]$$

conocida como **función de error de entropía cruzada**, con lo que se minimiza en f:

$$-\ln L(f) = -\sum_{i=1}^{n} \ln \mathsf{p}_{\mathsf{Y}|\boldsymbol{x}_i}(y_i) = -\frac{1}{n}\sum_{i=1}^{n}[y_i \ln f(\boldsymbol{x}_i) + (1 - y_i) \ln(1 - f(\boldsymbol{x}_i))]$$

y se evita el problema anterior pues ahora se obtiene $\hat{f}(\boldsymbol{x}) = \hat{\mathbb{P}}(\mathsf{Y} = 1|\boldsymbol{x}) \in [0, 1]$.

4. **Regresión logística**.

a) En lo anterior, no hemos especificado el modelo de f.

b) Si se utilizan modelos lineales $f(\boldsymbol{x}) = \mathbf{w}^T\boldsymbol{x} + w_0$ pueden obtenerse estimaciones pobres si la verdadera probabilidad a posteriori $\mathsf{f}(\boldsymbol{x}) = \mathbb{P}(\mathcal{C}_1|\boldsymbol{x})$ es no lineal.

c) Para evitarlo, se utiliza el ratio $\mathsf{p}(\mathcal{C}_1|\boldsymbol{x})/\mathsf{p}(\mathcal{C}_2|\boldsymbol{x}))$ (**odds ratio** – ratio de posibilidades) cuyo logaritmo neperiano ya toma valores en todo $\mathbb{R}$ (y no presenta el problema de tener que restringirse a $[0, 1]$) y se ajusta linealmente[3]:

$$\ln \frac{\mathsf{p}(\mathcal{C}_1|\boldsymbol{x})}{1 - \mathsf{p}(\mathcal{C}_1|\boldsymbol{x})} = \mathbf{w}^T\boldsymbol{x} + w_0 \Leftrightarrow \mathsf{p}(\mathcal{C}_1|\boldsymbol{x}) = \frac{1}{1 + e^{-(\mathbf{w}^T\boldsymbol{x}+w_0)}} \in [0, 1] \tag{4}$$

d) Ello se conoce como **regresión logística** (izquierda). Su versión equivalente (derecha) es una **red neuronal para clasificación de una sóla neurona** (Adaline sigmoide).

e) Ahora, debido a (4), los minimos cuadrados producen un resultado válido $\hat{f}(\boldsymbol{x}) = \hat{\mathsf{p}}(\mathcal{C}_1|\boldsymbol{x}) \in [0, 1]$, aunque lo mejor es estimar mediante máxima verosimilitud binomial (entropía cruzada).

f) A pesar de que el modelo f de la derecha de la ecuación (4) es no lineal, sus curvas de nivel son lineales, con lo que la frontera discriminante que produce la correspondiente regla de clasificación *plug-in* sigue siendo lineal. Por esa razón, **la regresión logística se considera un método lineal de clasificación**.

[3] La función $\ln \frac{a}{1-a}$ se denomina **función link**.

g) En hipótesis de normalidad:

$$\mathsf{p}(\boldsymbol{x}|\mathcal{C}_j) = (2\pi)^{-d/2}\,|\boldsymbol{\Sigma}|^{-1/2}\exp\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_j)\},\ j=1,2$$

se obtiene:

$$\begin{aligned}\mathbf{w} &= \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\\ w_0 &= -\frac{1}{2}\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1+\frac{1}{2}\boldsymbol{\mu}_2^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2+\ln\frac{\mathbb{P}(\mathcal{C}_1)}{\mathbb{P}(\mathcal{C}_2)}\end{aligned}$$

expresión poblacional que coincide con la expresión[4] (3).

[4] A pesar de que la expresión anterior coincide con el discriminante de Fisher a nivel poblacional, en la práctica, no produce los mismos resultados, pues mientras aquél se estima mediante la distribución conjunta, éste se estima con la distribución condicionada, por lo que es menos eficiente al utilizar una menor cantidad de información.

Bibliografía

Duda R O, Hart P E y Stork D G [2001]. *Pattern Classification*, John Wiley.

Fukunaga K [1990]. *Introduction to Statistical Pattern Recognition*, Morgan Kaufmann.

Seber G A F [1984]. *Multivariate Observations*, John Wiley.