

Clasificación Bayesiana

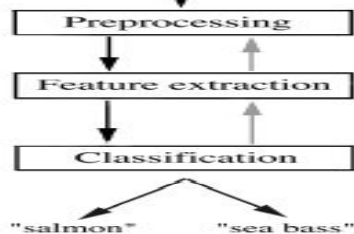
Dr. Álvaro Pardo

Universidad Católica del Uruguay

apardo@ucu.edu.uy

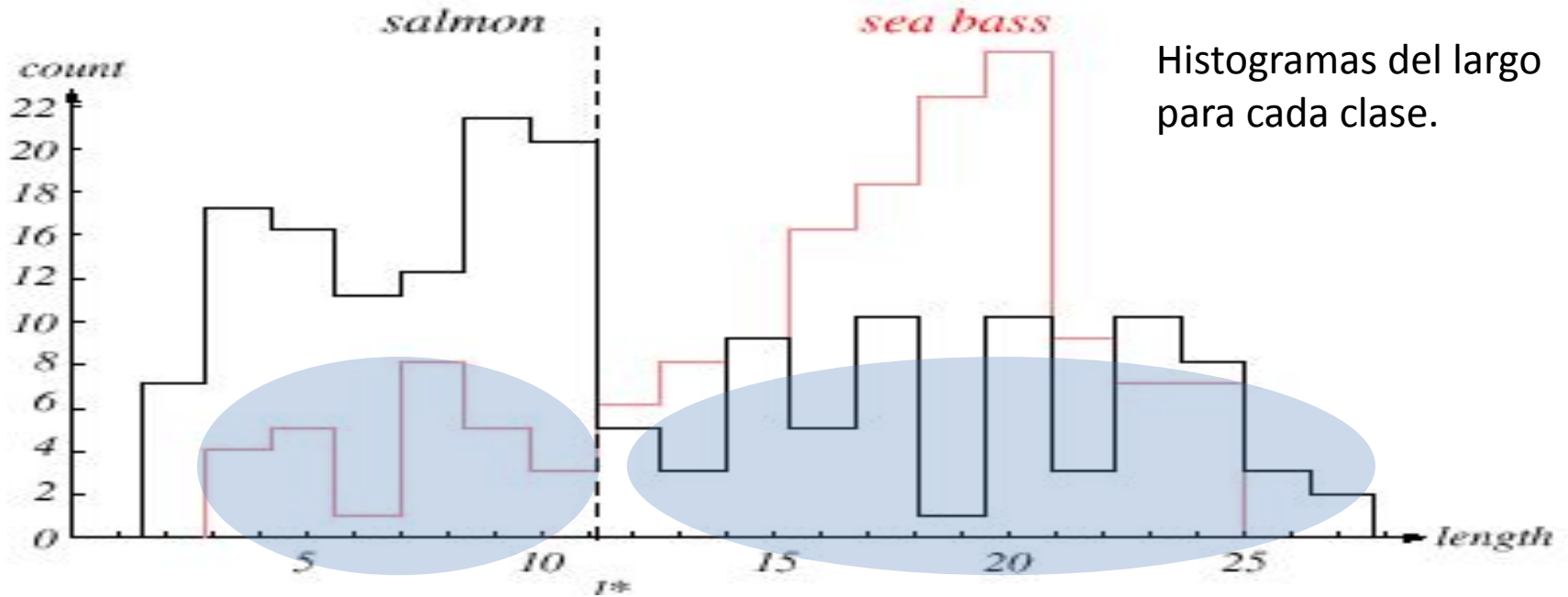
@AlvaroPardoUy

Ejemplo



- **Objetivo:** Reconocer la clase del pescado; clasificar en salom o sea bass.
- **Definición de características:** Se puede usar el largo para discriminar entre las dos clases.
- **Clases:** {salmón, sea bass}
- **Patrones, vector de características:** $x = \{\text{largo}(\text{cm})\}$

Ejemplo



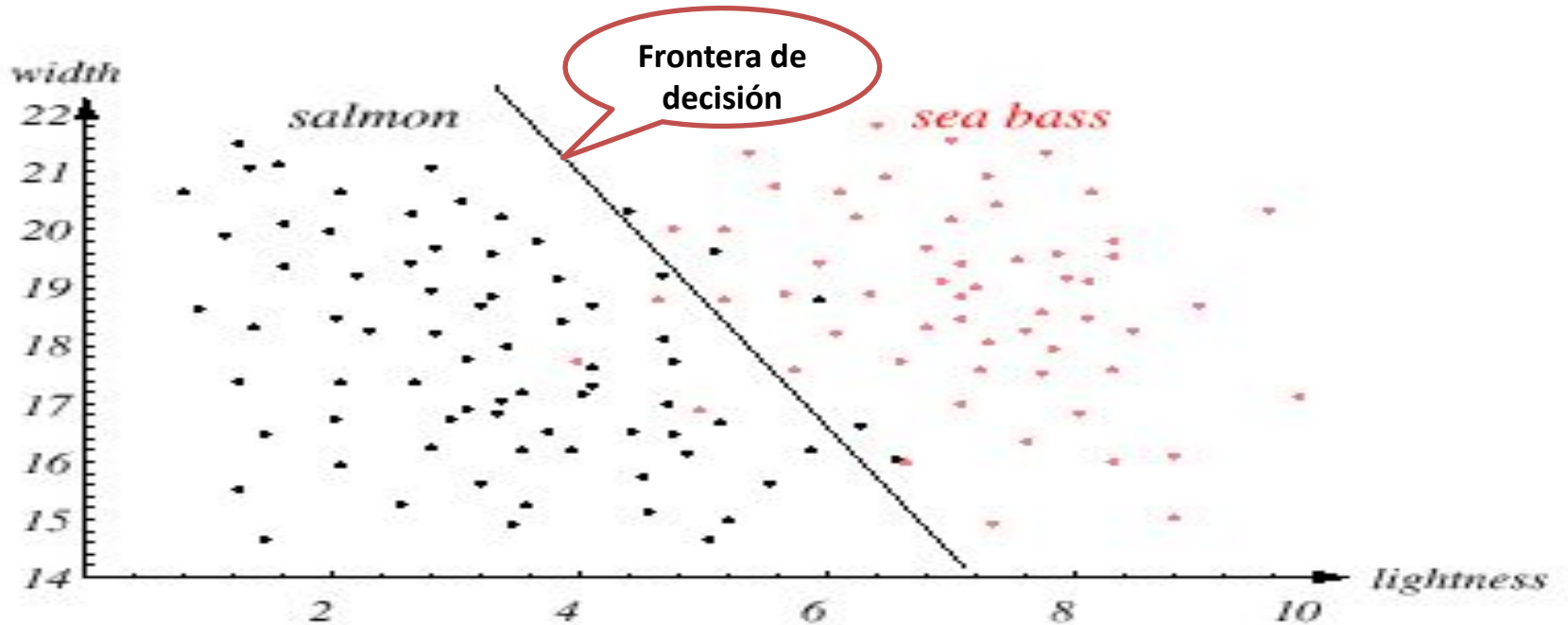
Errores al clasificar sea bass
como salmón.

Errores al clasificar salmón como
sea bass

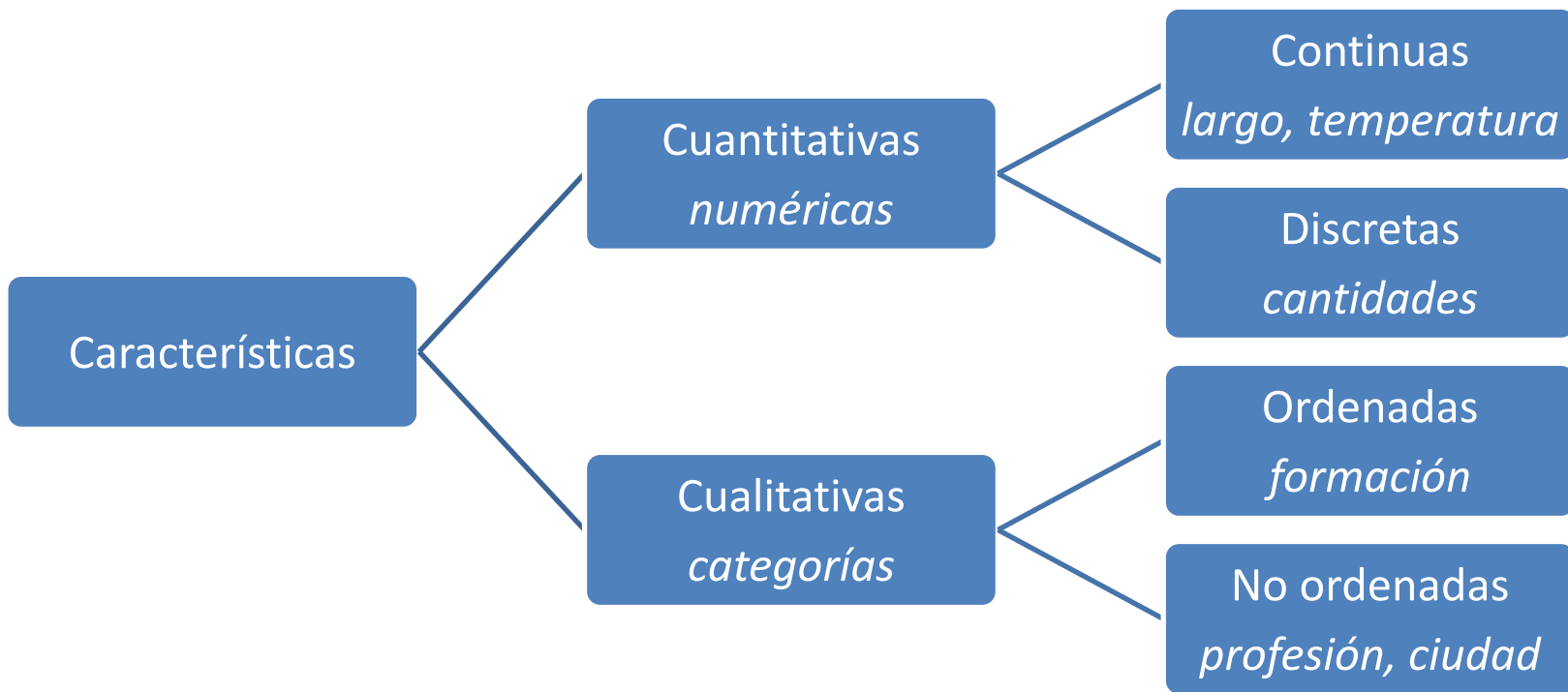
Ejemplo

Para mejorar la clasificación se introduce el brillo como nueva característica.

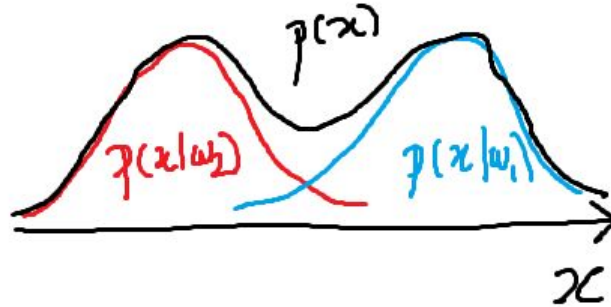
Patrón, vector de características $x = \{\text{largo}, \text{brillo}\}$



Características / Atributos



Ejemplo



- **Característica:** x (largo)
- **Clases:** salmón, sea bass
- Dado un objeto con largo x se asigna a la clase que maximiza la **probabilidad a posteriori** .

Introducción a la Decisión Bayesiana

- **Problema:**

- **Clases:** $\{\omega_1, \dots, \omega_N\}$
- **Características:** vector x
- **Objetivo:** dado un nuevo patrón (con características x) definir la clase

- Se modelan las características como variables aleatorias.
- Para cada clase ω_i se asume conocida (o estimable) la **distribución de probabilidad** de la característica x : $p(x|\omega_i)$.
- Adicionalmente se asumen conocidas las **probabilidades a priori** $P(\omega_i)$. Estas miden la probabilidad de observar ω_i independientemente de cualquier medida.
- En base a esta información el objetivo es obtener las **probabilidades a posteriori** $P(\omega_i|x)$ para poder realizar la clasificación.

Bayes

Verosimilitud (likelihood)

Probabilidad a prior (prior)

$$P(\omega_i|x) = \frac{p(x|\omega_j)P(\omega_j)}{p(x)}$$

Evidencia

Si hay c clases:

$$p(x) = \sum_{i=1}^c p(x|\omega_i)P(\omega_i)$$

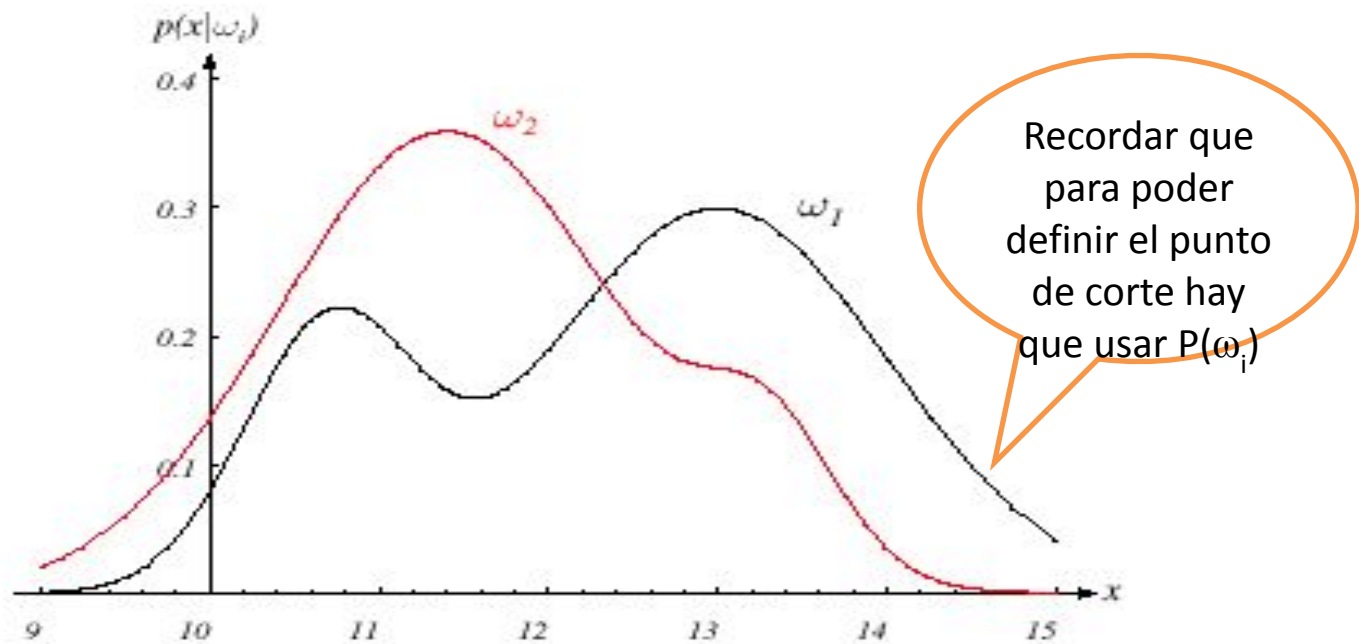


FIGURE 2.1. Hypothetical class-conditional probability density functions show the probability density of measuring a particular feature value x given the pattern is in category ω_i . If x represents the lightness of a fish, the two curves might describe the difference in lightness of populations of two types of fish. Density functions are normalized, and thus the area under each curve is 1.0. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

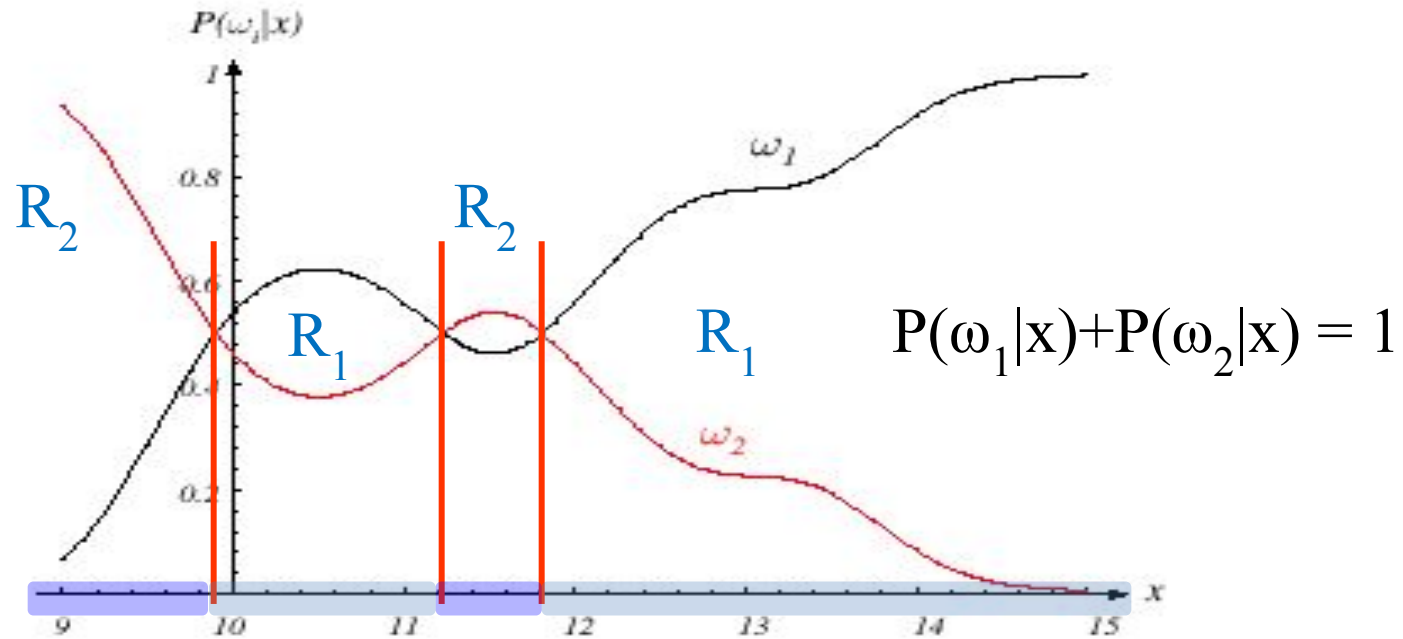




FIGURE 2.2. Posterior probabilities for the particular priors $P(\omega_1) = 2/3$ and $P(\omega_2) = 1/3$ for the class-conditional probability densities shown in Fig. 2.1. Thus in this case, given that a pattern is measured to have feature value $x = 14$, the probability it is in category ω_2 is roughly 0.08, and that it is in ω_1 is 0.92. At every x , the posteriors sum to 1.0. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Decisión - Clasificación

- Clasificación en función de la observación x :

Si $P(\omega_1 | x) > P(\omega_2 | x)$  La clase verdadera es ω_1

Si $P(\omega_1 | x) < P(\omega_2 | x)$  La clase verdadera es ω_2

Entonces las probabilidades de error son:

$P(\text{error} | x) = P(\omega_1 | x)$ si se decide ω_2

$P(\text{error} | x) = P(\omega_2 | x)$ si se decide ω_1

Fronteras de decisión

$$g_i(x) = P(w_i|x) = \frac{p(x|w_i)P(w_i)}{p(x)}$$

Eliminando $p(x)$:

$$g_i(x) = p(x|w_i)P(w_i).$$

En Log:

$$g_i(x) = \log p(x|w_i) + \log P(w_i)$$

Para dos clases se puede usar:

$$\begin{aligned} g(x) &= g_1(x) - g_2(x) \\ &= P(w_1|x) - P(w_2|x) \\ &= \log \frac{p(x|w_1)}{p(x|w_2)} - \log \frac{P(w_1)}{P(w_2)} \end{aligned}$$

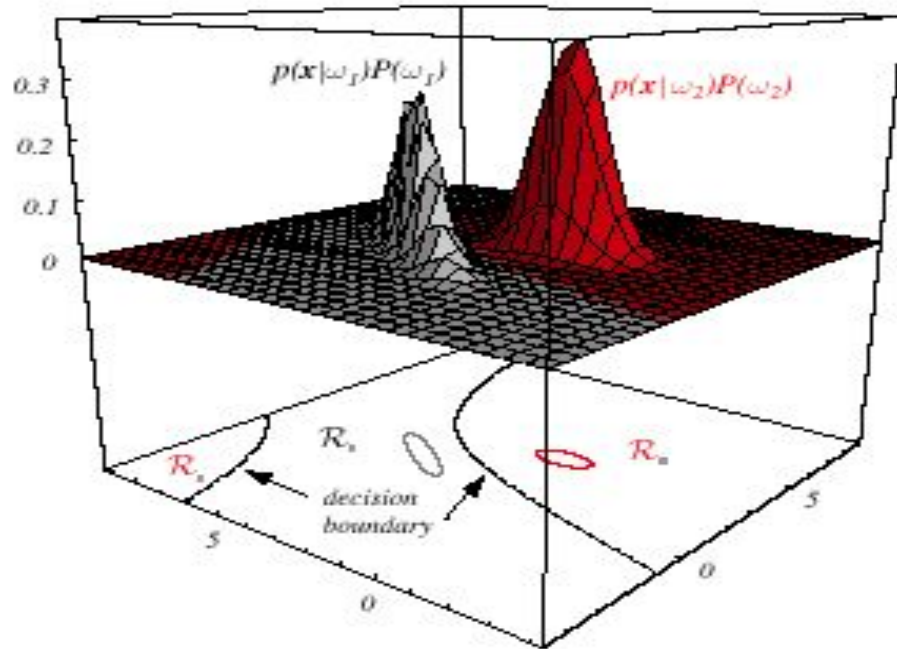


FIGURE 2.6. In this two-dimensional two-category classifier, the probability densities are Gaussian, the decision boundary consists of two hyperbolas, and thus the decision region \mathcal{R}_2 is not simply connected. The ellipses mark where the density is $1/e$ times that at the peak of the distribution. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Distribución Normal

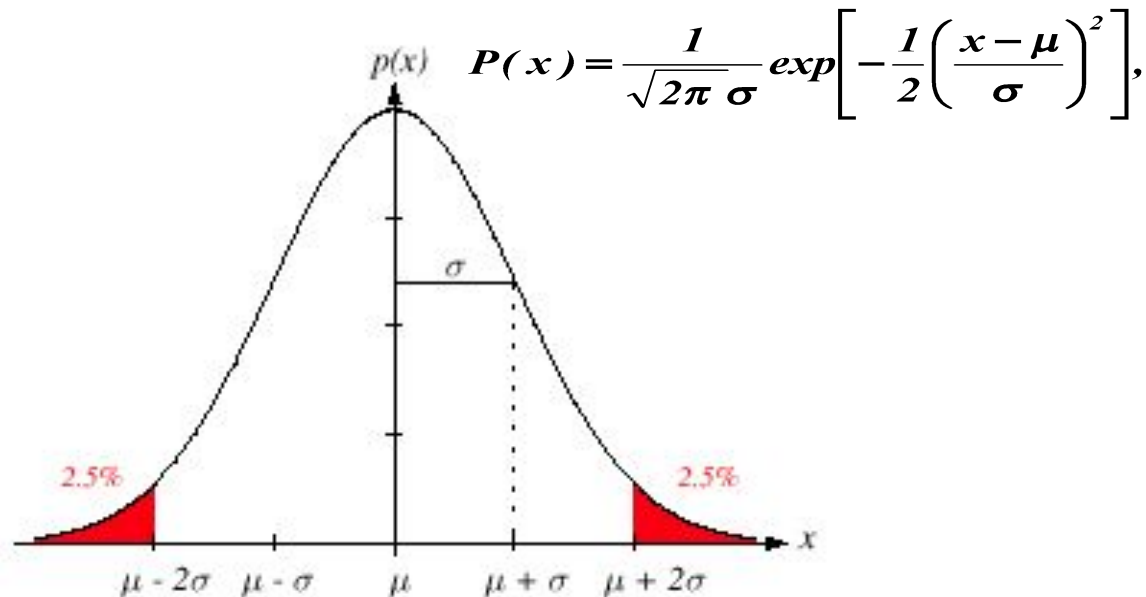


FIGURE 2.7. A univariate normal distribution has roughly 95% of its area in the range $|x - \mu| \leq 2\sigma$, as shown. The peak of the distribution has value $p(\mu) = 1/\sqrt{2\pi}\sigma$. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Clasificador Naive Bayes

Las n medidas x_i se asumen independientes:

$$P(w_i|x_1, \dots, x_n) = \frac{P(w_i)P(x_1, \dots, x_n|w_i)}{P(x_1, \dots, x_n)} = \frac{P(w_i)\prod_{j=1}^n P(x_j|w_i)}{P(x_1, \dots, x_n)}$$

Es un método simple y rápido (al modelar cada medida por separado). Da buenos resultados en muchos problemas de clasificación.

Ejercicio: Detección de Fraude

Distribución Normal Multivariada

$$P(x) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp \left[-\frac{1}{2} (x - \mu)^t \Sigma^{-1} (x - \mu) \right]$$

- Parámetros:
 - Vector de medias: (μ_1, \dots, μ_d)
 - Matriz de Covarianza (dxd): Σ
- Estimación de parámetros:
 - Se obtienen a partir del conjunto de entrenamiento usando estimaciones estadísticas.
 - Observación: Cantidad de puntos de entrenamiento vs dimensión.

Funciones discriminantes para la Distribución Normal

- La función discriminante para la clase i es:

$$g_i(x) = \ln P(x \mid \omega_i) + \ln P(\omega_i)$$

- Para la Distribución Normal:

$$g_i(x) = -\frac{1}{2}(x - \mu_i)^t \sum_i^{-1} (x - \mu_i) - \frac{d}{2} \ln 2\pi - \frac{1}{2} \ln |\Sigma_i| + \ln P(\omega_i)$$

Caso 1: Características independientes

$$\Sigma_i = \sigma^2 Id$$

$$g_i(x) = w_i^t x + w_{i0}$$

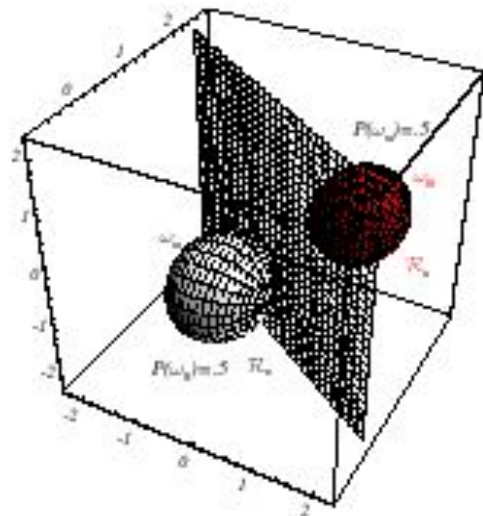
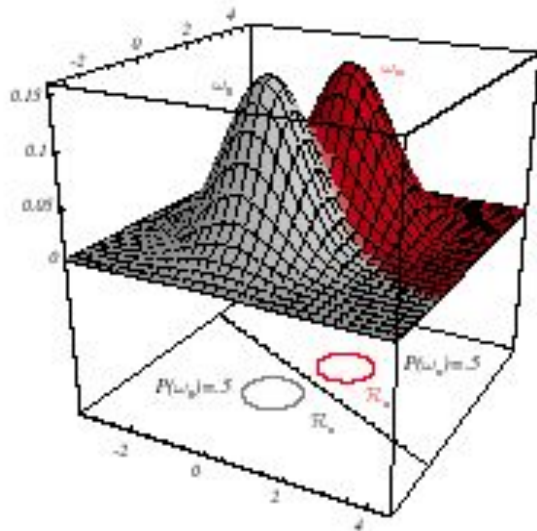
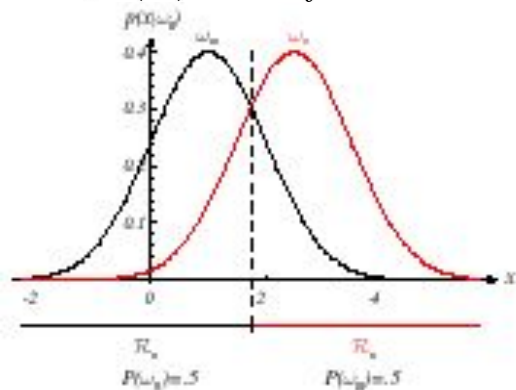


FIGURE 2.10. If the covariance matrices for two distributions are equal and proportional to the identity matrix, then the distributions are spherical in d dimensions, and the boundary is a generalized hyperplane of $d - 1$ dimensions, perpendicular to the line separating the means. In these one-, two-, and three-dimensional examples, we indicate $p(\mathbf{x}|\omega_i)$ and the boundaries for the case $P(\omega_1) = P(\omega_2)$. In the three-dimensional case, the grid plane separates \mathcal{R}_1 from \mathcal{R}_2 . From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Figura extraída de Pattern Classification de Duda, Hart, Stork.

Frontera de Decisión

- Hiperplano que separa regiones:

$$x_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\sigma^2}{\|\mu_i - \mu_j\|^2} \ln \frac{P(\omega_i)}{P(\omega_j)} (\mu_i - \mu_j)$$

es perpendicular (normal) a la recta que une las medias. Si $P(\omega_i) = P(\omega_j)$ entonces $x_0 = (\mu_i + \mu_j)$

Dependencia de las probabilidades a priori

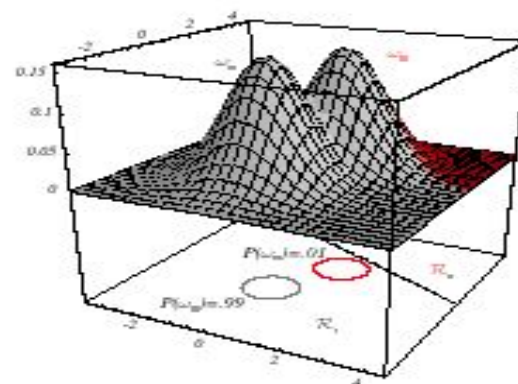
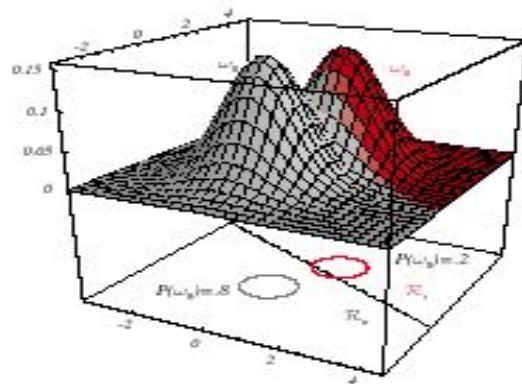
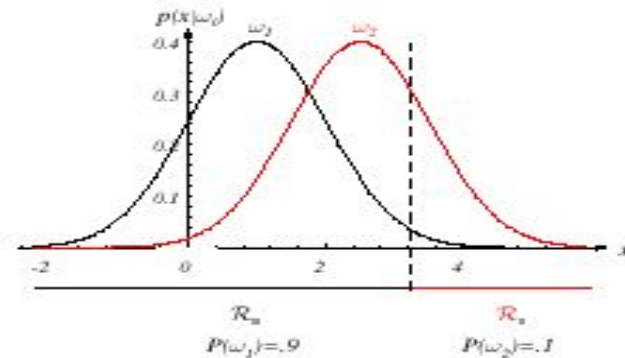
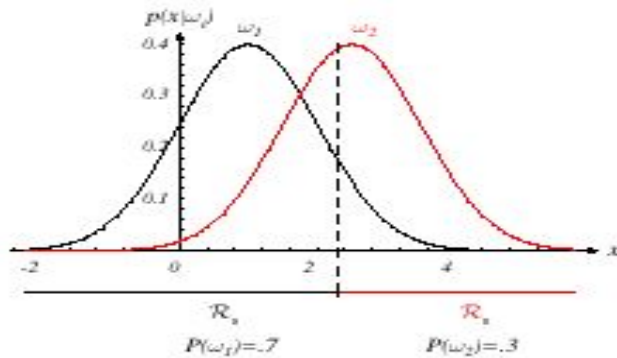


Figura extraída de Pattern Classification de Duda, Hart, Stork.

Dependencia de las probabilidades a priori

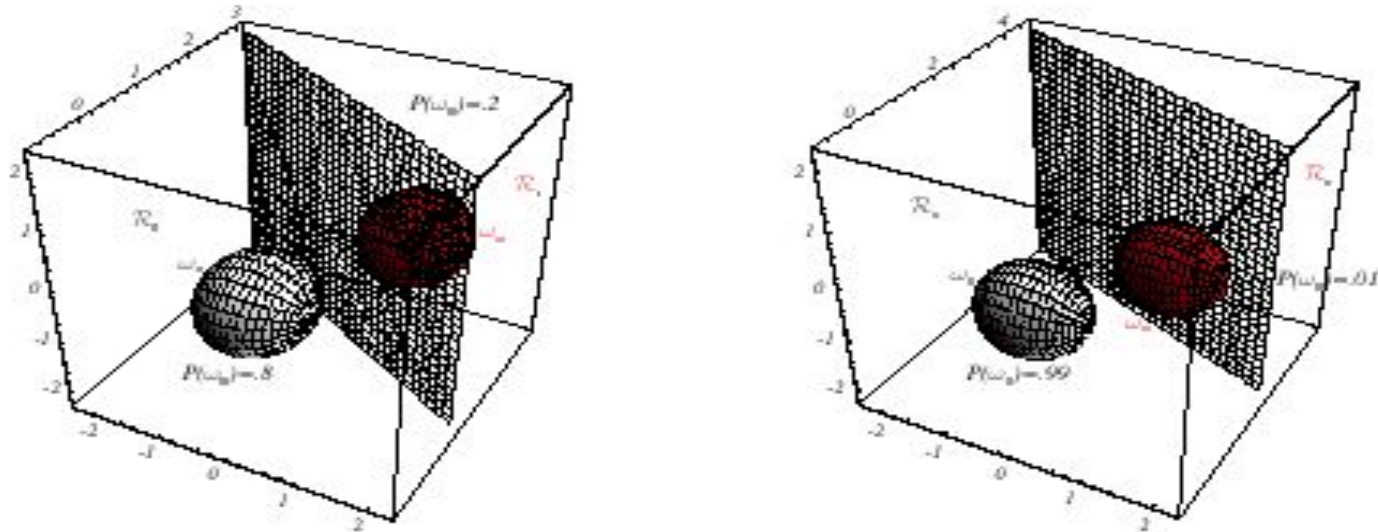
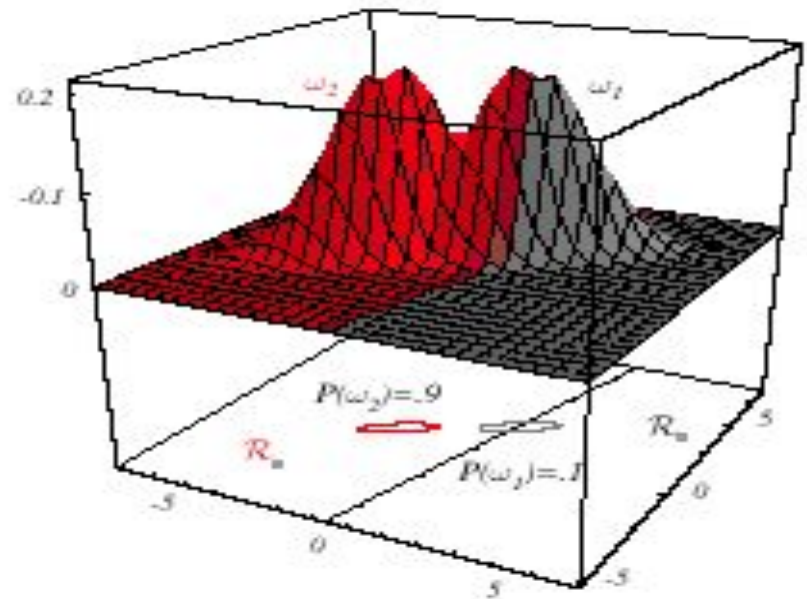
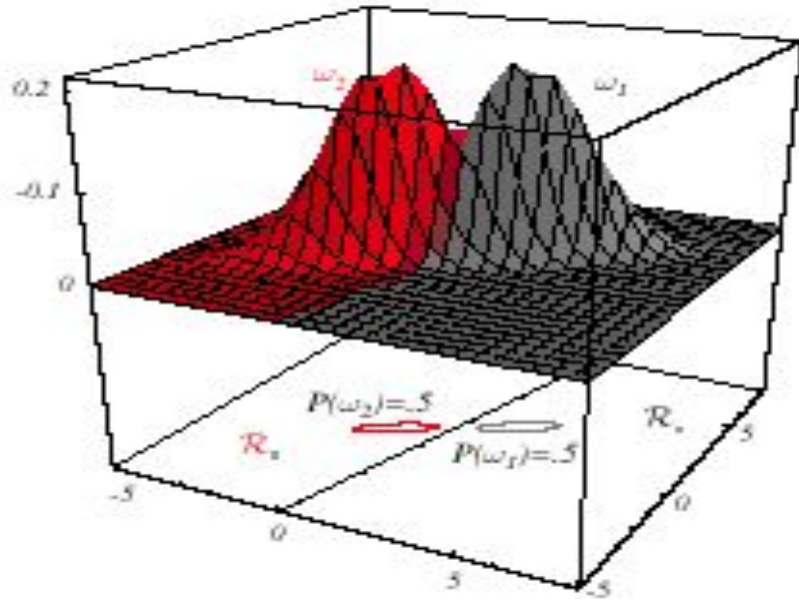


FIGURE 2.11. As the priors are changed, the decision boundary shifts; for sufficiently disparate priors the boundary will not lie between the means of these one-, two- and three-dimensional spherical Gaussian distributions. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Caso 2: Matrices de covarianza iguales



Caso 2: Matrices de covarianza iguales

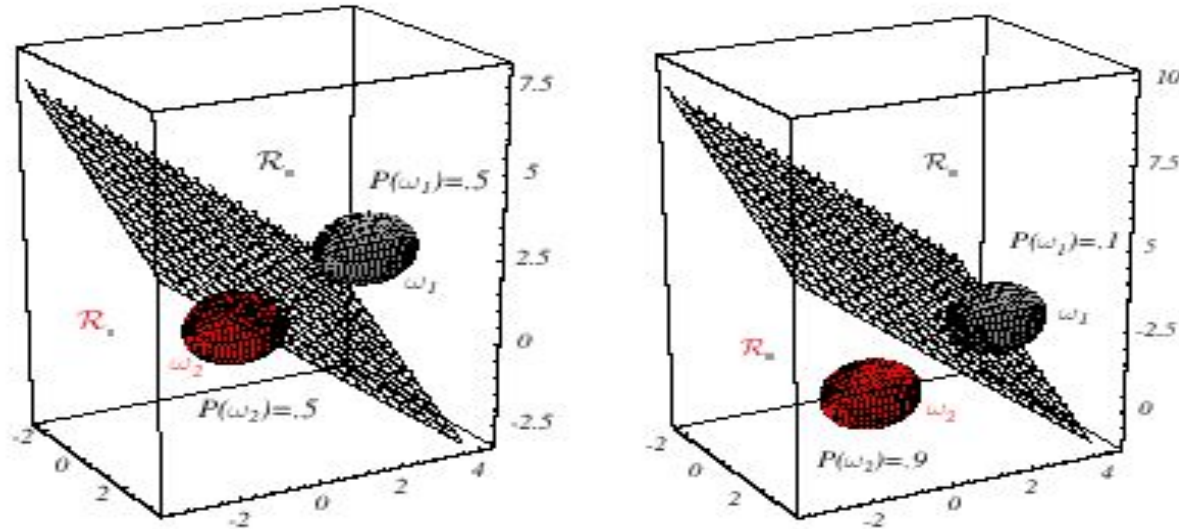


FIGURE 2.12. Probability densities (indicated by the surfaces in two dimensions and ellipsoidal surfaces in three dimensions) and decision regions for equal but asymmetric Gaussian distributions. The decision hyperplanes need not be perpendicular to the line connecting the means. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Caso 3: Matrices de covarianza arbitrarias

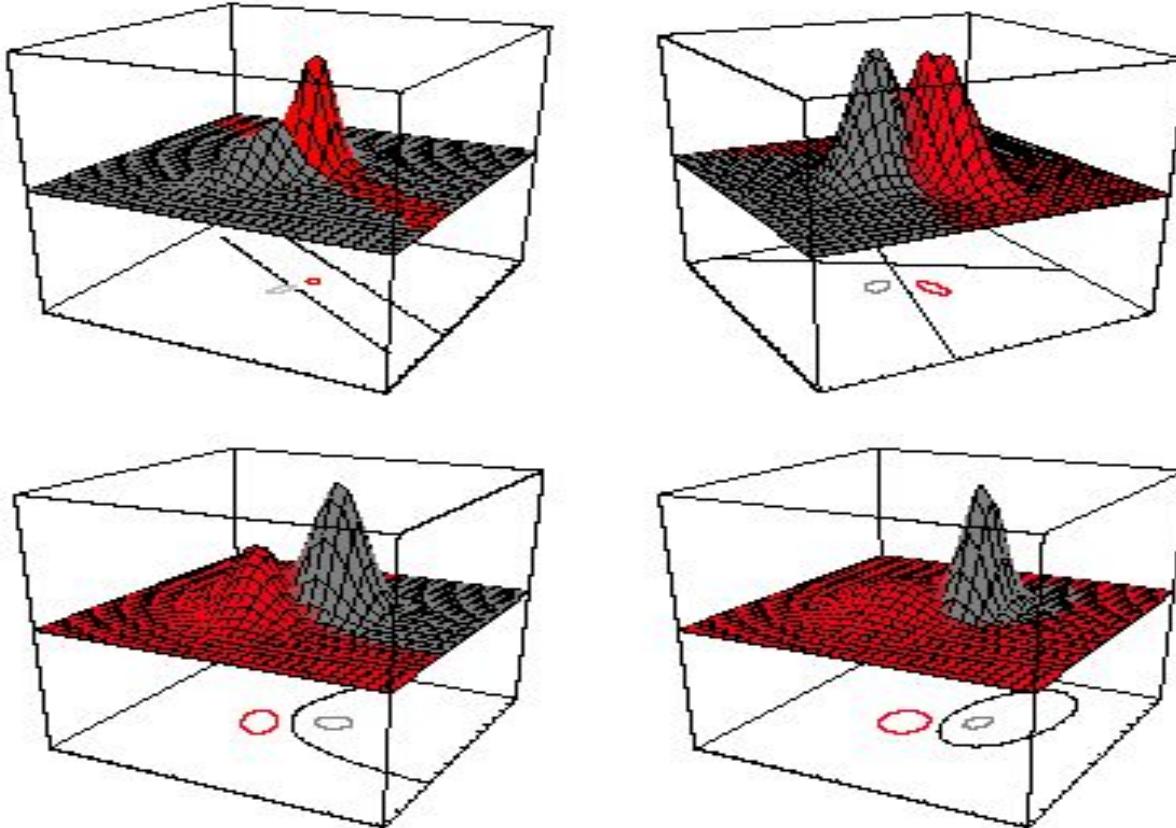


Figura extraída de Pattern Classification de Duda, Hart, Stork.

MÉTODOS NO PARAMÉTRICOS

Clasificación con Ventanas de Parzen

- Se estima la densidad para cada clase y se encuentra el punto de corte correspondiente.
- Los resultados de clasificación dependen de la selección del tipo de ventana y sus parámetros.

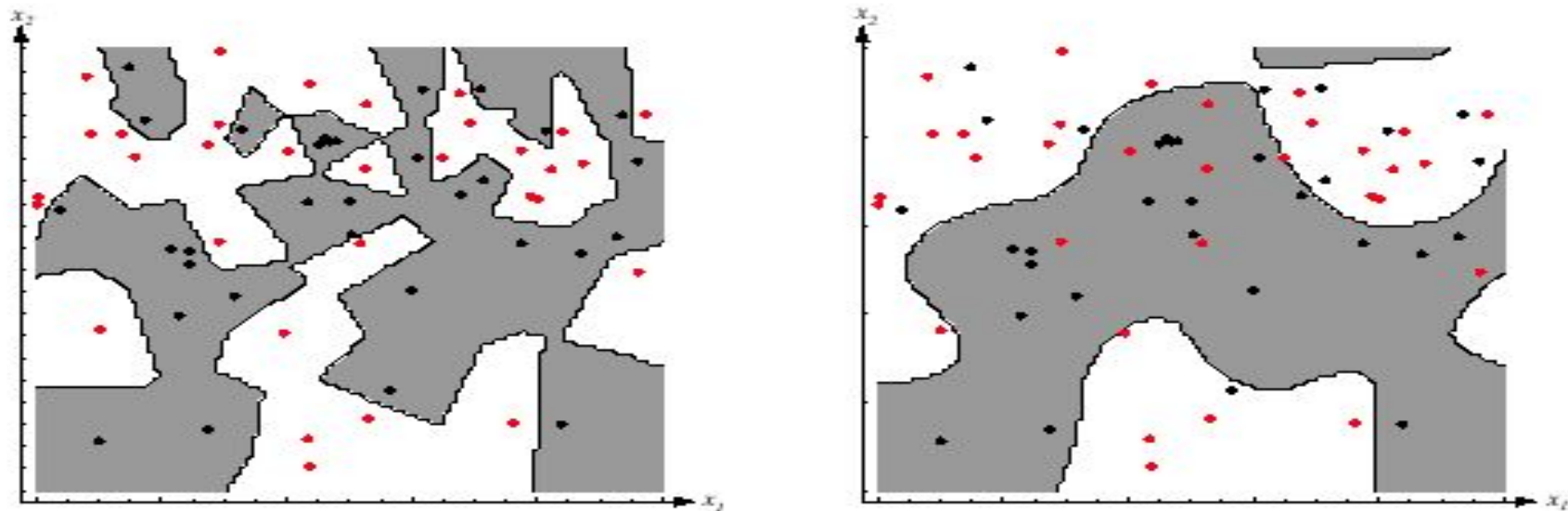


FIGURE 4.8. The decision boundaries in a two-dimensional Parzen-window dichotomizer depend on the window width h . At the left a small h leads to boundaries that are more complicated than for large h on same data set, shown at the right. Apparently, for these data a small h would be appropriate for the upper region, while a large h would be appropriate for the lower region; no single window width is ideal overall. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.