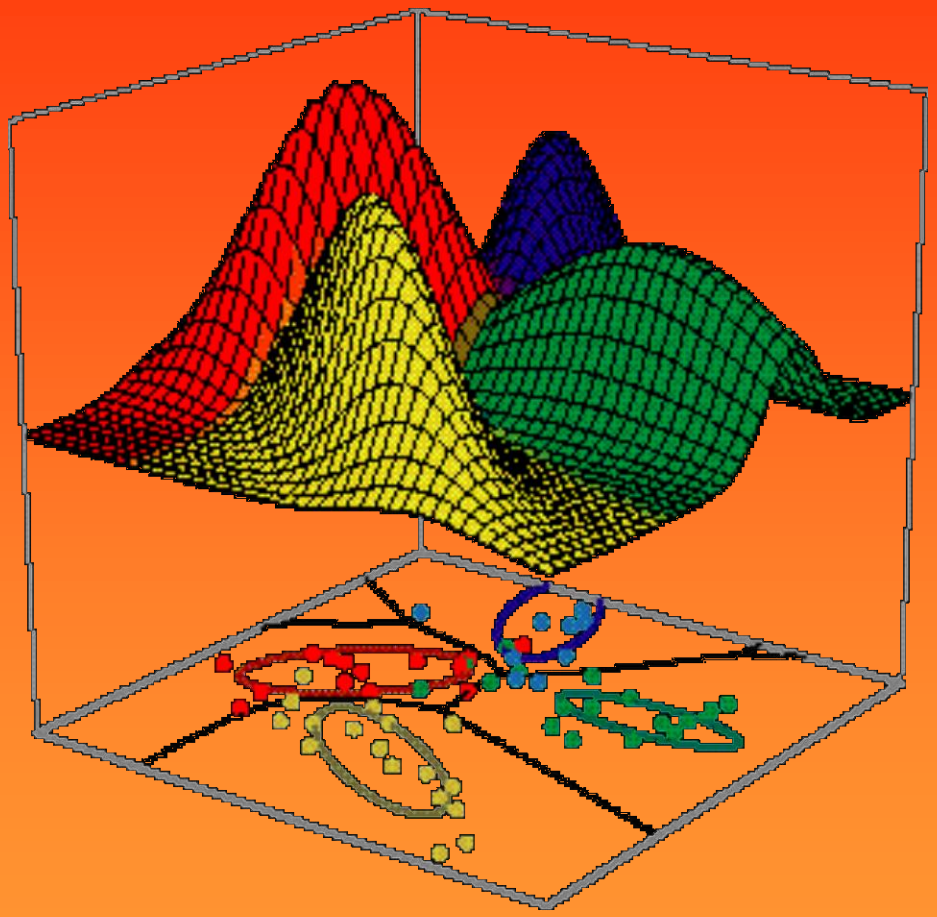


Pattern Classification

All materials in these slides were taken from *Pattern Classification (2nd ed) by R. O. Duda, P. E. Hart and D. G. Stork, John Wiley & Sons, 2000*
with the permission of the authors and the publisher

Capitulo 4 (Parte 1): Clasificacion no parametrica (Secciones 4.1-4.3)

Introduccion

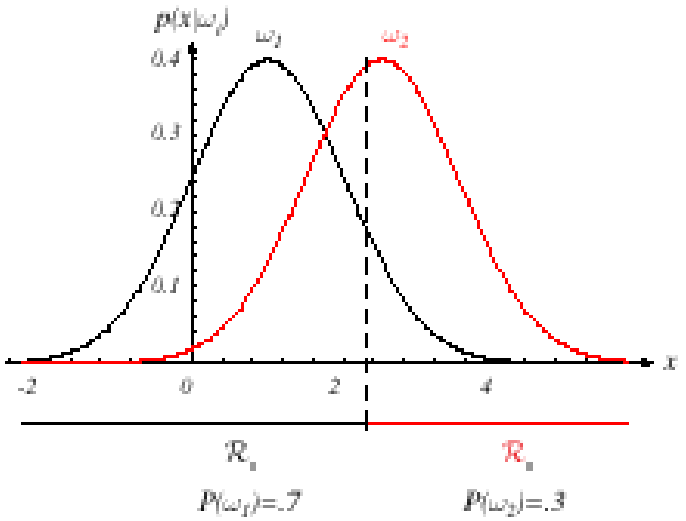
Estimation de densidades

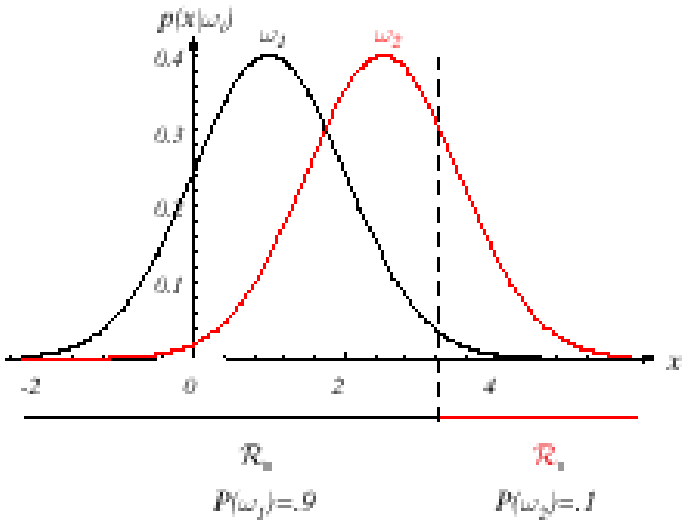
Ventanas de Parzen

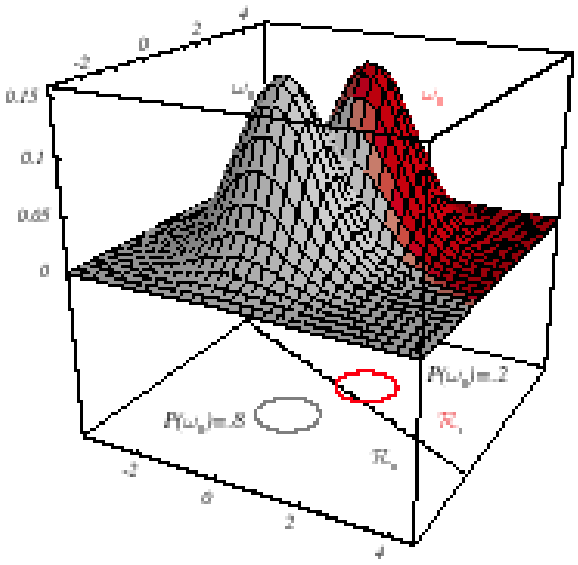
Introduccion

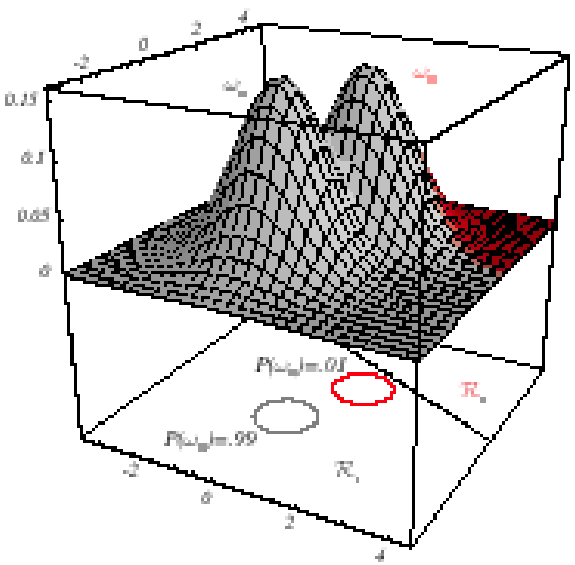
- Las densidades parametricas son unimodales (tienen un unico maximo), mientras que muchos problemas practicos involucran densidades multimodales
- Procedimientos no parametricos pueden ser usados con distribuciones arbitrarias y sin la hipotesis de que la forma de las densidades es conocida

- Hay dos tipos de metodos no parametricos:
  - Estimacion de $P(x \mid \omega_j)$
  - Saltear la probabilidad e ir directamente a la estimacion de la probabilidad a posteriori

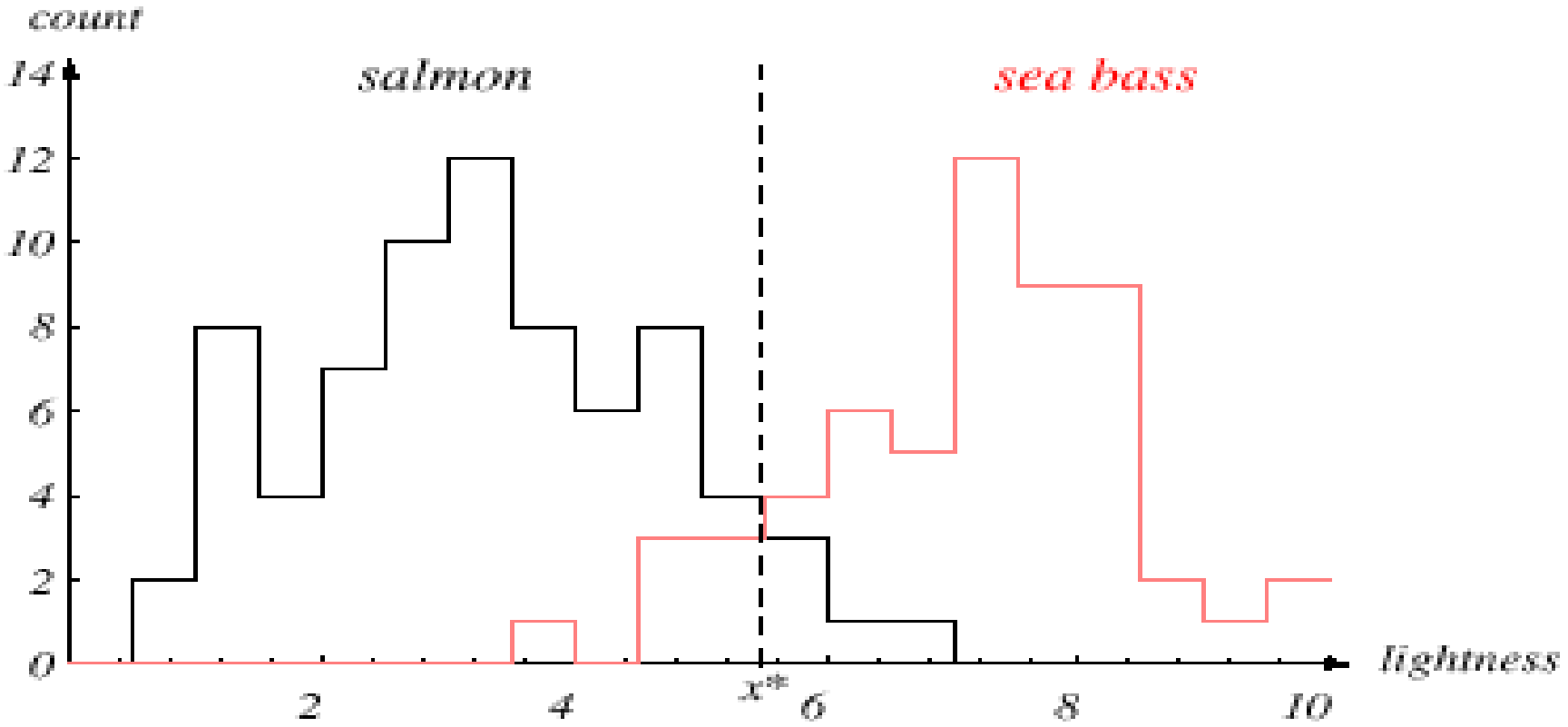

p(x|ωj)
ω1
ω2
0.4
0.3
0.2
0.1
x
-2
0
2
4
P(ω1)=.7
P(ω2)=.3


p(x|ωj)
ω1
ω2
0.4
0.3
0.2
0.1
x
-2
0
2
4
P(ω1)=.9
P(ω2)=.1


0.15
0.1
0.05
0
P(ω2)=.2
P(ω1)=.8


0.15
0.1
0.05
0
P(ω2)=.01
P(ω1)=.99

$$P(w_i) \neq P(w_j)$$


count
14
12
10
8
6
4
2
0
salmon
sea bass
2
4
x*
6
8
10
lightness

Estimacion de densidades

- Idea básica, estimar p en el punto x:

- La probabilidad de que un vector caiga en la region R es:

$$P = \int_{\Re} p(x')dx' \qquad (1)$$

- P es una version suavizada o promediada de la densidad p(x).
- Si se tiene una muestra de talla n, la probabilidad de que k puntos caigan en R es :

$$P_k = \binom{n}{k} P^k (1-P)^{n-k} \qquad (2)$$

y el valor esperado de k es :

$$E(k) = nP \qquad (3)$$

Estimacion ML de $P = \theta$

$$\max_{\theta}(P_k \mid \theta)$$ es alcanzada por $$\hat{\theta} = \frac{k}{n} \cong P$$

Por lo cual, el radio k/n es un buen estimador para la probabilidad P y por lo tanto de la densidad p.

si p(x) es continua y la region R es tan pequeña que p no varia significantemente en ella , podemos escribir:

$$\int_{\Re} p(x')dx' \cong p(x)V \qquad (4)$$

donde x es un punto en R y V es el volumen dentro de R.

Combinando las ecuaciones (1) , (3) y (4) se obtiene:

$$p(x) \cong \frac{k/n}{V}$$

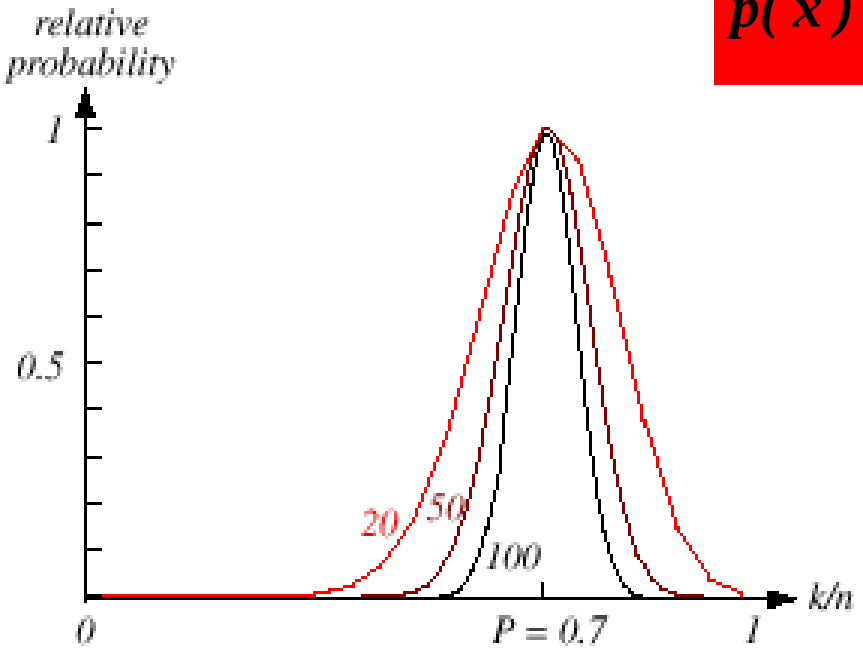


FIGURE 4.1. The relative probability an estimate given by Eq. 4 will yield a particular value for the probability density, here where the true probability was chosen to be 0.7. Each curve is labeled by the total number of patterns n sampled, and is scaled to give the same maximum (at the true probability). The form of each curve is binomial, as given by Eq. 2. For large n, such binomials peak strongly at the true probability. In the limit $n \rightarrow \infty$, the curve approaches a delta function, and we are guaranteed that our estimate will give the true probability. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Estimacion de densidades

- Justificamos un poco mas (4)

$$\int_{\mathfrak{R}} p(x')dx' \cong p(x)V \qquad (4)$$

Asumimos que *p(x)* es continua y que la region R es tan pequeña que p no varia significativamente en R. Como *p(x)* = constante, no es parte de la suma.

$$\int_{\Re} p(x')dx' = p(x')\int_{\Re} dx' = p(x')\int_{\Re} 1_{\Re}(x)dx' = p(x')\mu(\Re)$$

Donde: $\mu(R)$ es: una superficie en el espacio euclideo R^2

un volumen en el espacio euclideo R^3

un hipervolumen en el espacio euclideo R^n

Como $p(x) \cong p(x')$ = constante, en R^3:

$$\int_{\Re} p(x')dx' \cong p(x).V$$

$$p(x) \cong \frac{k}{nV}$$

Condiciones de convergencia

- La fraccion *k/(nV)* es un valor promedio de *p(x)* sobre el espacio.
- *p(x)* se obtiene exacta solo si V se acerca a cero .
- Si n es un numero fijo, puede no haber muestras en ***R***: por lo cual

$$\lim_{V \to 0,\, k=0} p(x) = 0 \quad (\text{si n} = \text{fijo})$$

- Si alguna o mas muestras coinciden con x, el estimador diverge

$$\lim_{V \to 0,\, k \neq 0} p(x) = \infty$$

- Desde un punto de vista practico, siempre hay una cantidad de muestras finita.
- El volumen V debe reducirse para que se pueda usar esta estimacion, pero no completamente porque puede quedarse sin puntos.
- se debe aceptar una variabilidad en el radio k/n y un promedio en la densidad p(x)

- Teoricamente, se puede dejar el numero de muestras crecer a infinito.

Infinitas muestras

- Para estimar la densidad en x, se forma una sucesion de regiones R_j que contienen a x, y una sucesion de muestras de talla j, para ser usadas con R_j
- Sea V_n el volumen de R_n , k_n el numero de las n muestras que cae en R_n y $p_n(x)$ el nth estimador de $p(x)$:

$$p_n(x) = (k_n/n)/V_n \qquad (7)$$

- Se necesitan las siguientes hipotesis para segurar convergencia de p_n *a* p

$$1)\ \lim_{n\to\infty} V_n = 0$$

$$2)\ \lim_{n\to\infty} k_n = \infty$$

$$3)\ \lim_{n\to\infty} k_n / n = 0$$

$$1)\ \lim_{n\to\infty} V_n = 0$$

$$2)\ \lim_{n\to\infty} k_n = \infty$$

$$3)\ \lim_{n\to\infty} k_n / n = 0$$

- La primera condicion asegura que el radio P/V converge a p(x)
- La segunda condicion asegura que la frecuencia de observacion converge en probabilidad a P
- La tercera condicion asegura que si bien hay infinitas muestras en la ventana, solo son una parte pequeña del total de muestras

Estudiaremos dos formas diferentes de formar las regiones tales que

$$p_n(x) \underset{n\to\infty}{\longrightarrow} p(x) \quad (1)$$

(a) Reducir una region inicial mediante una funcion, como $V_n = 1/\sqrt{n}$, de tal forma que el k_n resultante se comporte bien y se cumpla (1)

Este método se llama el ““ metodo de estimacion con ventana de Parzen”

(b) Especificar k_n como una funcion de n, como $k_n = \sqrt{n}$; el volumen V_n crece hasta que engloba k_n vecinos de x.

Este metodo se llama el “metodo de los k_n-vecinos mas cercanos ”

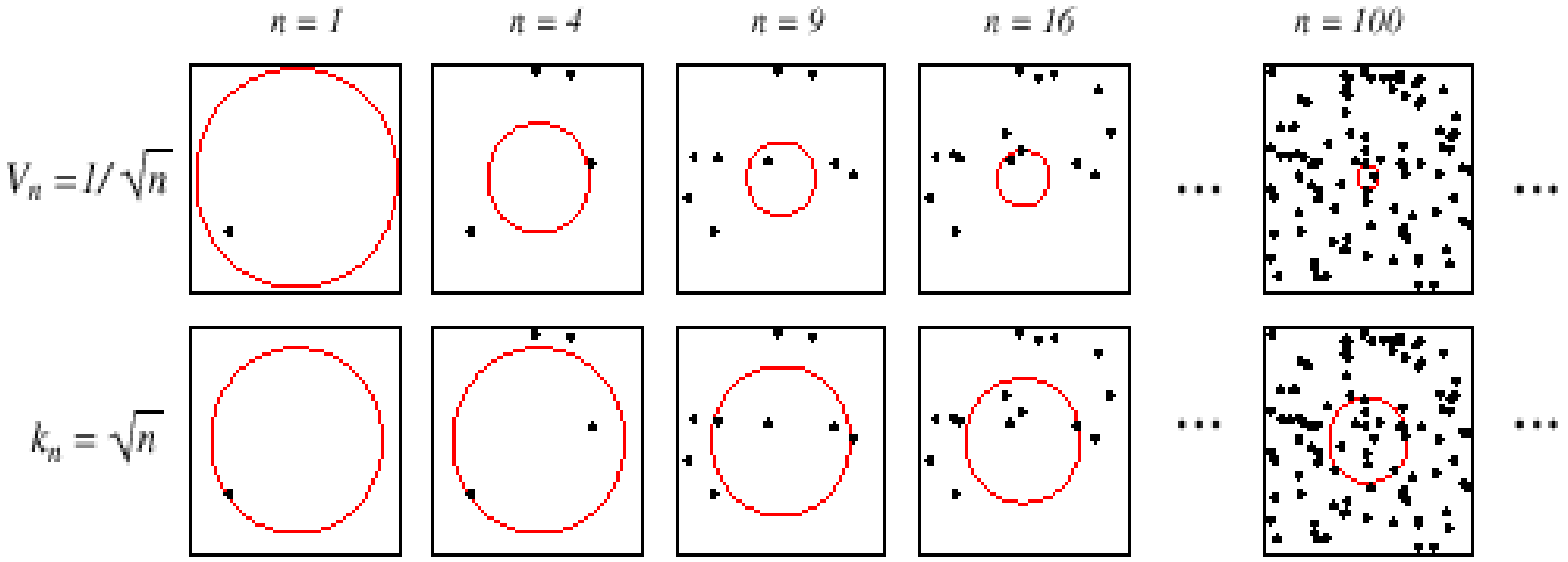


FIGURE 4.2. There are two leading methods for estimating the density at a point, here at the center of each square. The one shown in the top row is to start with a large volume centered on the test point and shrink it according to a function such as $V_n = 1/\sqrt{n}$. The other method, shown in the bottom row, is to decrease the volume in a data-dependent way, for instance letting the volume enclose some number $k_n = \sqrt{n}$ of sample points. The sequences in both cases represent random variables that generally converge and allow the true density at the test point to be calculated. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Ventana de Parzen

- El estimador con ventana de asume que la region $\mathbf{R}_n$ es un hipercubo d-dimensional

$$V_n = h_n^d \quad (\mathrm{h_n} : \text{largo del lado de } \Re_n)$$

$$Sea\ \varphi(\mathrm{u})\ \text{la siguiente funcion de ventana :}$$

$$\varphi(\mathrm{u}) = \begin{cases} 1 & \left|\mathrm{u_j}\right| \le \frac{1}{2} \quad \mathrm{j} = 1,\ldots,\mathrm{d} \\ 0 & \text{en otro caso} \end{cases}$$

- $\varphi((x-x_i)/h_n)$ es igual a uno si x_i cae dentro del hipercubo con volumen V_n centrado en x, e igual a cero en otro caso.

El numero de muestras que cae en el hipercubo es:

$$k_n = \sum_{i=1}^{n} \varphi\left(\frac{x - x_i}{h_n}\right)$$

Sustituyendo k_n en la ecuacion (7), se obtiene el siguiente estimador

$$p_n(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{V_n}\varphi\left(\frac{x - x_i}{h_n}\right)$$

$P_n(x)$ estima *$p(x)$* como el promedio de funciones φ de *x, y de las muestras* (x_i) $(i = 1, \ldots, n)$. Esas funciones φ pueden ser generales

Ejemplo

- El comportamiento del metodo de la ventana de Parzen cuando p(x)=N(0,1)
- Sea $\varphi(u) = (1/\sqrt{(2\pi)}) \exp(-u^2/2)$ y $h_n = h_1/\sqrt{n}$ $(n>1)$ *con* h_1 *un parametro conocido*
- Entonces

$$p_n(x) = \frac{1}{n}\sum_{i=1}^{i=n}\frac{1}{h_n}\varphi\left(\frac{x - x_i}{h_n}\right)$$

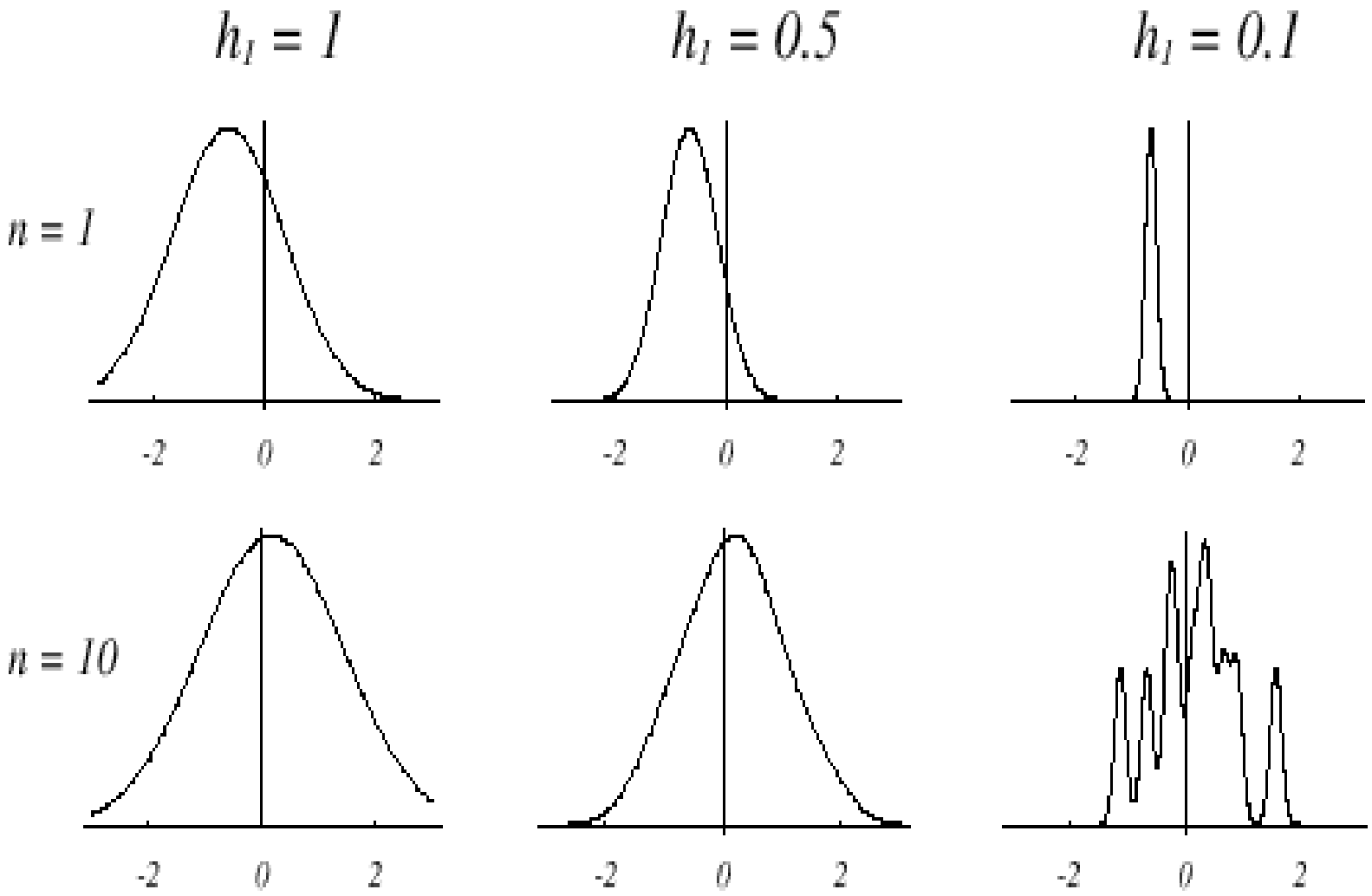
- es un promedio de densidades normales centradas en las muestras x_i.

Resultados numericos

- Para n=1 y $h_1=1$

$$p_1(x) = \varphi(x - x_1) = \frac{1}{\sqrt{2\pi}} e^{-1/2(x-x_1)^2} \rightarrow N(x_1, 1)$$

- Para $n = 10$ y $h = 0.1$, las contribuciones de lasmuestras pueden observarse claramente !


$h_1 = 1$
$h_1 = 0.5$
$h_1 = 0.1$
$n = 1$
$n = 10$
-2
0
2

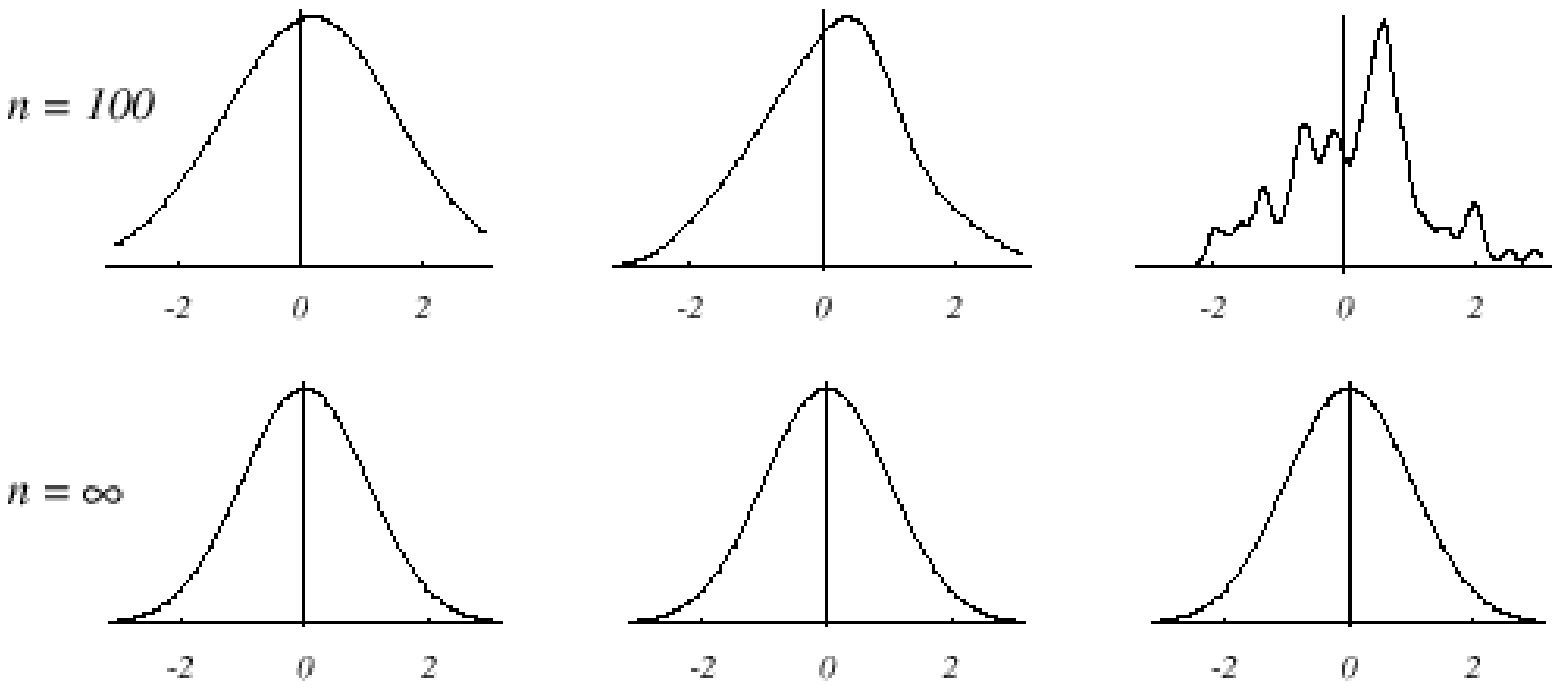
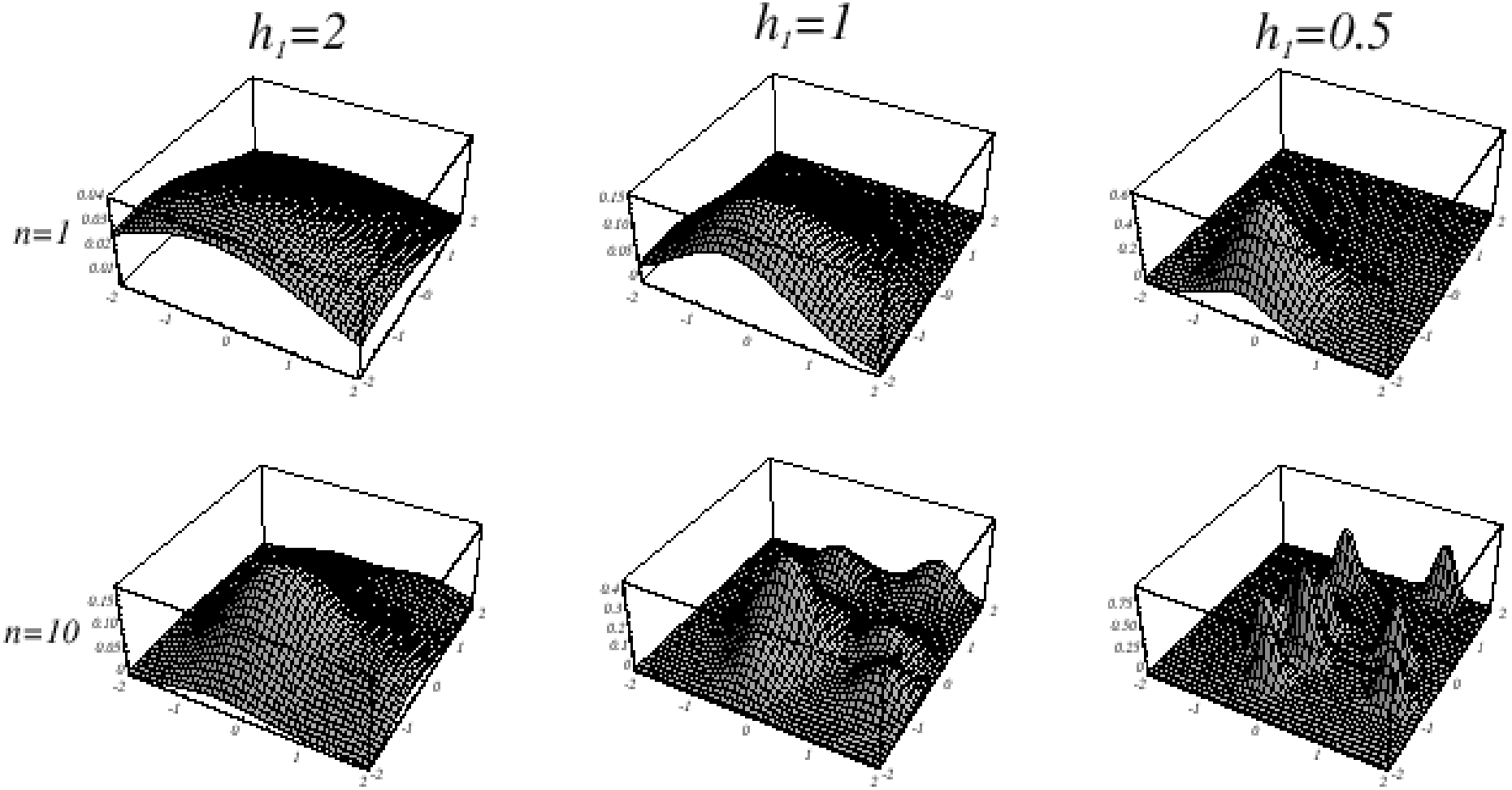


FIGURE 4.5. Parzen-window estimates of a univariate normal density using different window widths and numbers of samples. The vertical axes have been scaled to best show the structure in each graph. Note particularly that the $n = \infty$ estimates are the same (and match the true density function), regardless of window width. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

En dos dimensiones se ven las mismas contribuciones de las muestras:



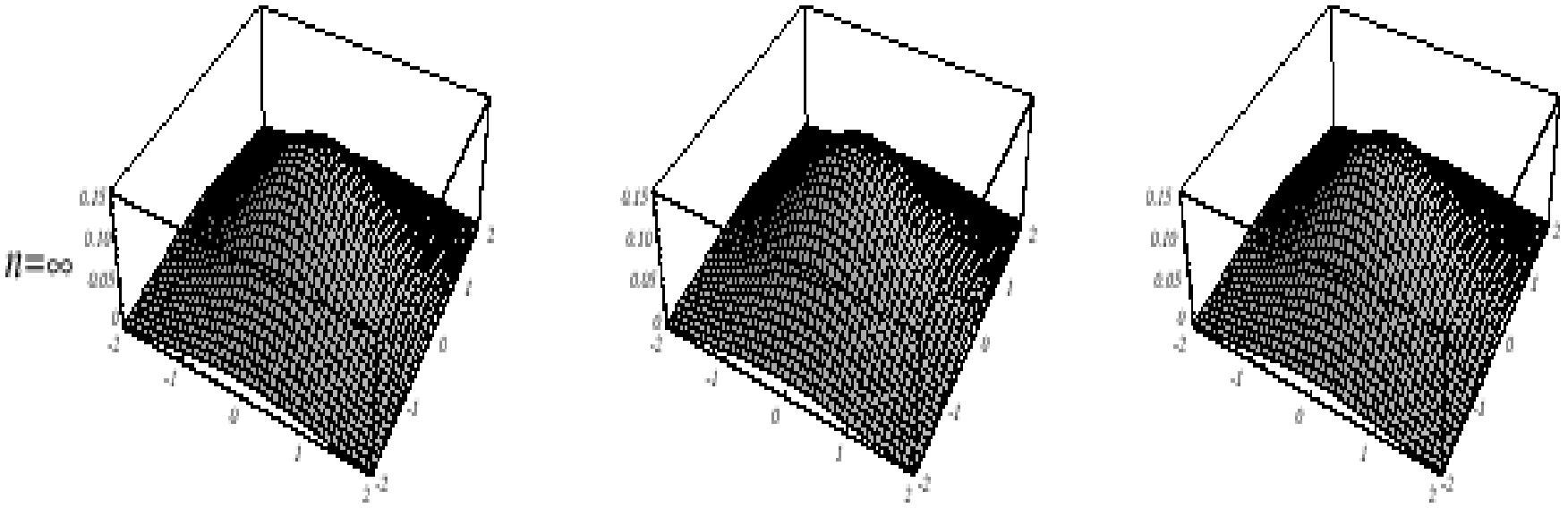
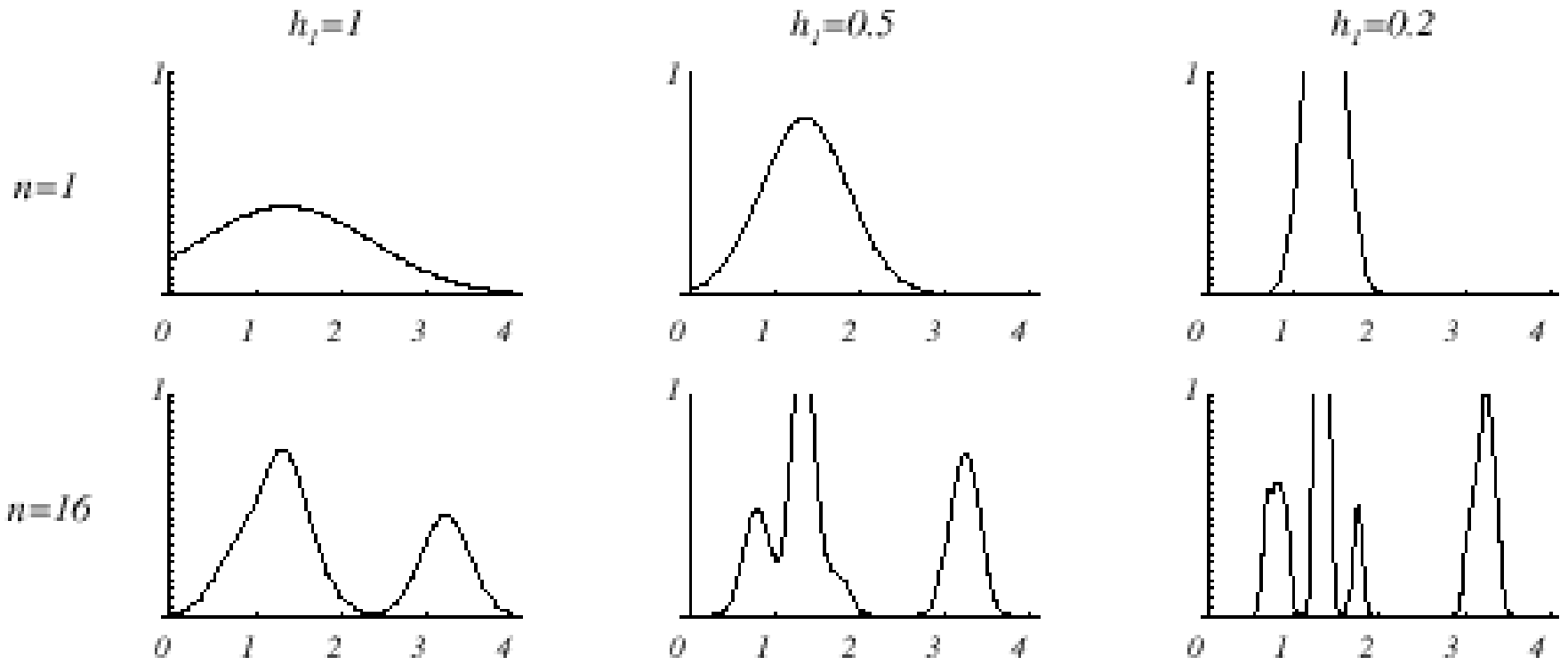


FIGURE 4.6. Parzen-window estimates of a bivariate normal density using different window widths and numbers of samples. The vertical axes have been scaled to best show the structure in each graph. Note particularly that the $n = \infty$ estimates are the same (and match the true distribution), regardless of window width. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Caso donde $p(x) = \lambda_1.U(a,b) + \lambda_2.T(c,d)$ (densidad desconocida es mezcla de una uniforme y una densidad triangulo



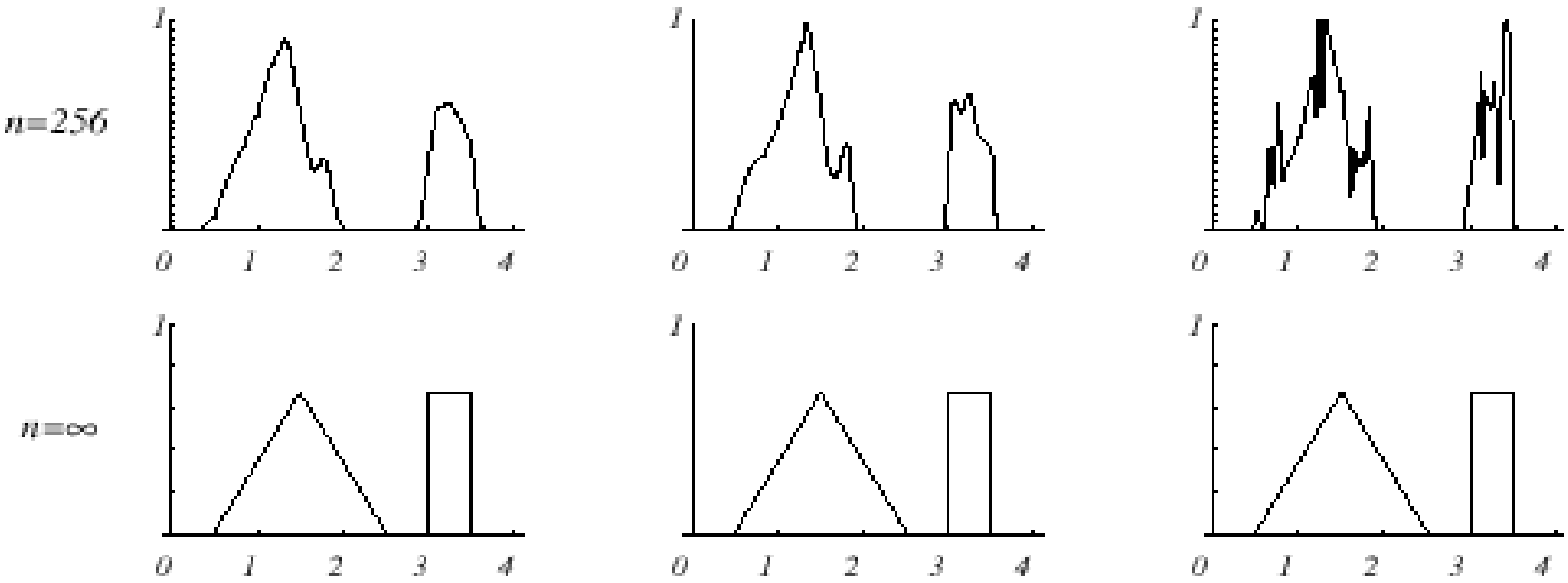


FIGURE 4.7. Parzen-window estimates of a bimodal distribution using different window widths and numbers of samples. Note particularly that the $n = \infty$ estimates are the same (and match the true distribution), regardless of window width. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Ejemplo de clasificacion con estimador de ventana de Parzen

- S estiman densidades por cada categoria y se clasifican los puntos con la etiqueta correspondiente al maximo a posteriori

- La region de decision de una regla de clasificacion con ventana de Parzen depende de la eleccion de la ventana, como se ilustra en la siguiente figura.

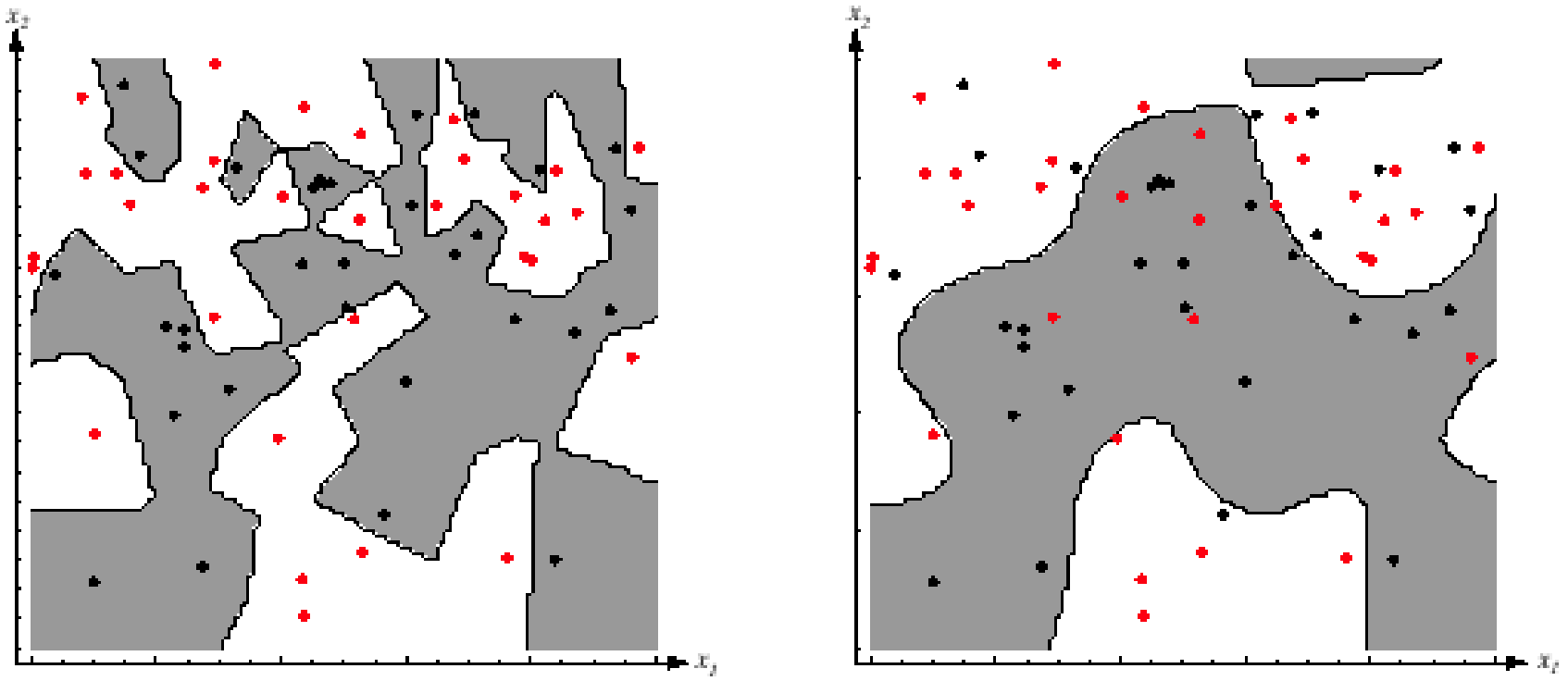


FIGURE 4.8. The decision boundaries in a two-dimensional Parzen-window dichotomizer depend on the window width h. At the left a small h leads to boundaries that are more complicated than for large h on same data set, shown at the right. Apparently, for these data a small h would be appropriate for the 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.