

# 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

# Chapter 4 (Part 1): Non-Parametric Classification (Sections 4.1-4.3)

- Introduction
- Density Estimation
- Parzen Windows

# Introduction

- All Parametric densities are unimodal (have a single local maximum), whereas many practical problems involve multimodal densities
- Nonparametric procedures can be used with arbitrary distributions and without the assumption that the forms of the underlying densities are known
- There are two types of nonparametric methods:
  - Estimating  $P(x | w_j)$
  - Bypass probability and go directly to a-posteriori probability estimation

# Density Estimation

- Basic idea:
- Probability that a vector  $x$  will fall in region  $R$  is:

$$P = \int_{\mathfrak{R}} p(x') dx' \quad (1)$$

- $P$  is a smoothed (or averaged) version of the density function  $p(x)$  if we have a sample of size  $n$ ; therefore, the probability that  $k$  points fall in  $R$  is then:

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

and the expected value for  $k$  is:

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

ML estimation of  $P = \theta$

$$\text{Max}_{\theta}(P_k / \theta) \text{ is reached for } \hat{\theta} = \frac{k}{n} \cong P$$

Therefore, the ratio  $k/n$  is a good estimate for the probability  $P$  and hence for the density function  $p$ .

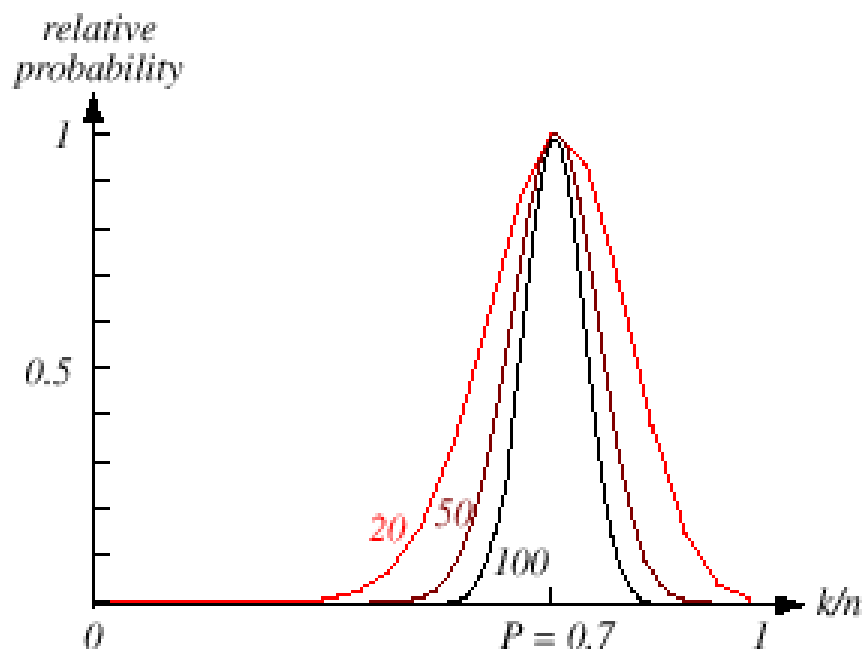
$p(x)$  is continuous and that the region  $\mathcal{R}$  is so small that  $p$  does not vary significantly within it, we can write:

$$\int_{\mathcal{R}} p(x') dx' \cong p(x) V \quad (4)$$

where  $x$  is a point within  $\mathcal{R}$  and  $V$  the volume enclosed by  $\mathcal{R}$ .

Combining equation (1) , (3) and (4) yields:

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



**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.

## Density Estimation (cont.)

- Justification of equation (4)

$$\int_{\mathcal{R}} p(x') dx' \cong p(x) V \quad (4)$$

We assume that  $p(x)$  is continuous and that region  $\mathcal{R}$  is so small that  $p$  does not vary significantly within  $\mathcal{R}$ . Since  $p(x) = \text{constant}$ , it is not a part of the sum.

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

Where:  $\mu(\mathcal{R})$  is: a surface in the Euclidean space  $R^2$

a volume in the Euclidean space  $R^3$

a hypervolume in the Euclidean space  $R^n$

Since  $p(x) \cong p(x') = \text{constant}$ , therefore in the Euclidean space  $R^3$ :

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

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



- Condition for convergence

The fraction  $k/(nV)$  is a space averaged value of  $p(x)$ .  
 $p(x)$  is obtained only if  $V$  approaches zero.

$$\lim_{V \rightarrow 0, k=0} p(x) = 0 \text{ (if } n = \text{fixed)}$$

This is the case where no samples are included in  $\mathcal{R}$ :  
 it is an uninteresting case!

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

In this case, the estimate diverges: it is an  
 uninteresting case!

- The volume  $V$  needs to approach 0 anyway if we want to use this estimation
  - Practically,  $V$  cannot be allowed to become small since the number of samples is always limited
  - One will have to accept a certain amount of variance in the ratio  $k/n$
  - Theoretically, if an unlimited number of samples is available, we can circumvent this difficulty

To estimate the density of  $x$ , we form a sequence of regions  $\mathcal{R}_1, \mathcal{R}_2, \dots$  containing  $x$ : the first region contains one sample, the second two samples and so on.

Let  $V_n$  be the volume of  $\mathcal{R}_n$ ,  $k_n$  the number of samples falling in  $\mathcal{R}_n$  and  $p_n(x)$  be the  $n^{\text{th}}$  estimate for  $p(x)$ :

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

Three necessary conditions should apply if we want  $p_n(x)$  to converge to  $p(x)$

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

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

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

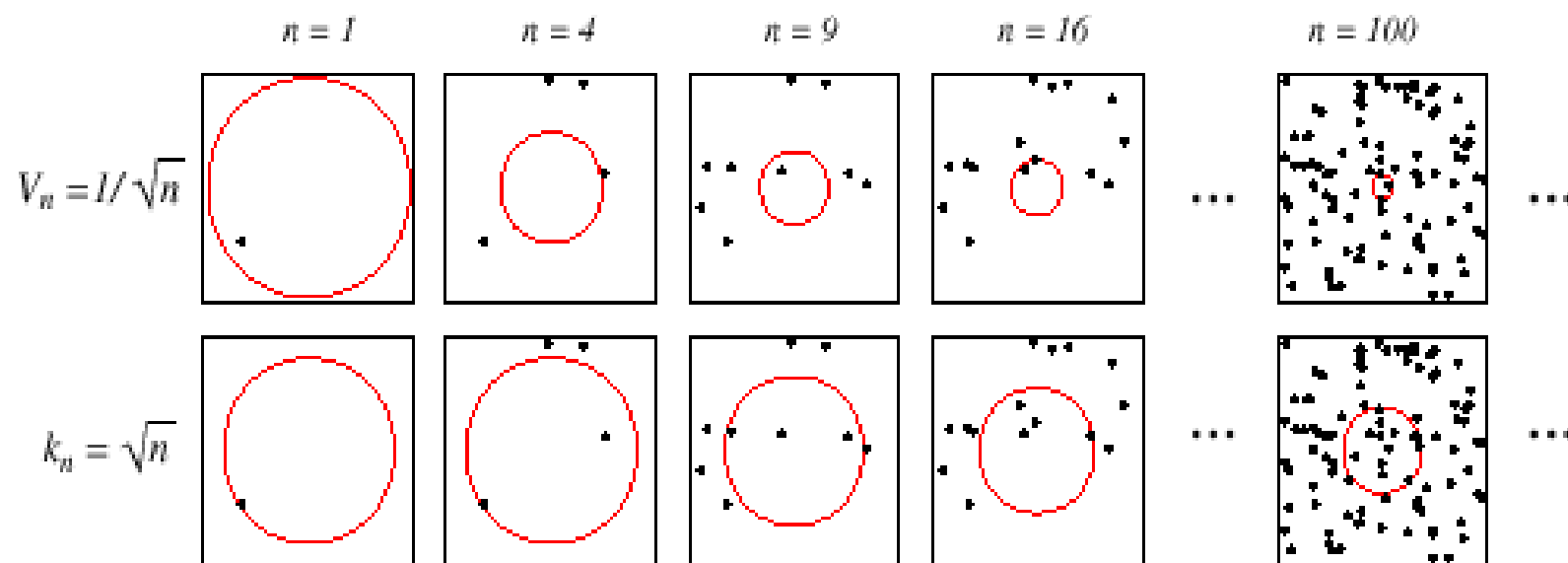
There are two different ways of obtaining sequences of regions that satisfy these conditions:

(a) Shrink an initial region where  $V_n = 1/\sqrt{n}$  and show that

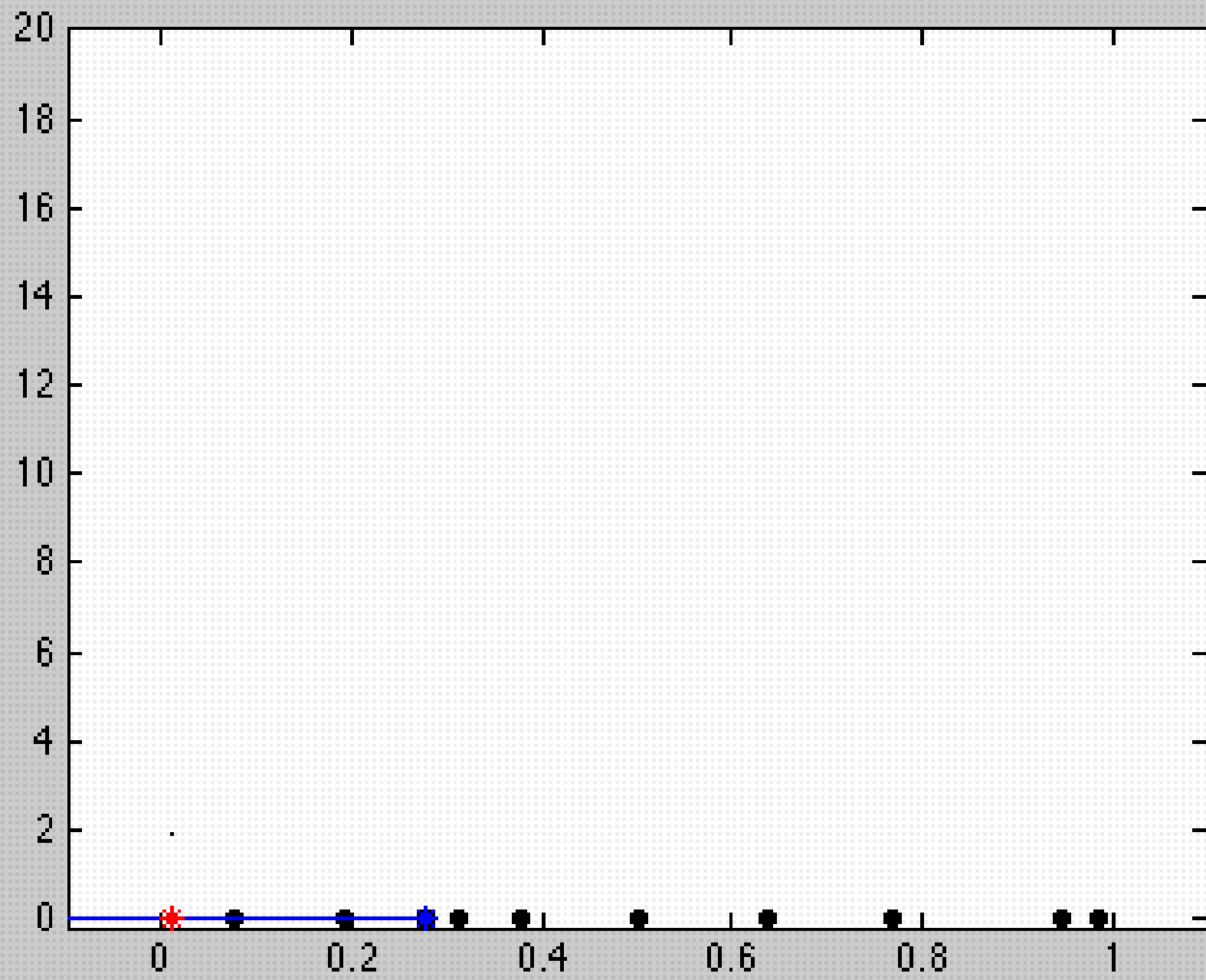
$$p_n(x) \xrightarrow{n \rightarrow \infty} p(x)$$

This is called “the Parzen-window estimation method”

(b) Specify  $k_n$  as some function of  $n$ , such as  $k_n = \sqrt{n}$ ; the volume  $V_n$  is grown until it encloses  $k_n$  neighbors of  $x$ . This is called “the  $k_n$ -nearest neighbor estimation method”



**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.



# Parzen Windows

- Parzen-window approach to estimate densities assume that the region  $\mathcal{R}_n$  is a d-dimensional hypercube

$V_n = h_n^d$  ( $h_n$  : length of the edge of  $\mathcal{R}_n$  )

Let  $\varphi(u)$  be the following window function :

$$\varphi(u) = \begin{cases} 1 & |u_j| \leq \frac{1}{2} \quad j = 1, \dots, d \\ 0 & \text{otherwise} \end{cases}$$

- $\varphi((x-x_i)/h_n)$  is equal to unity if  $x_i$  falls within the hypercube of volume  $V_n$  centered at  $x$  and equal to zero otherwise.

- The number of samples in this hypercube is:

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

By substituting  $k_n$  in equation (7), we obtain the following estimate:

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

$P_n(x)$  estimates  $p(x)$  as an average of functions of  $x$  and the samples  $(x_i)$  ( $i = 1, \dots, n$ ). These functions  $\varphi$  can be general!

- Illustration

- The behavior of the Parzen-window method

- Case where  $p(x) \rightarrow N(0, 1)$

Let  $\varphi(u) = (1/\sqrt{2\pi}) \exp(-u^2/2)$  and  $h_n = h_1/\sqrt{n}$  ( $n > 1$ )

( $h_1$ : known parameter)

Thus:

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

is an average of normal densities centered at the samples  $x_i$ .



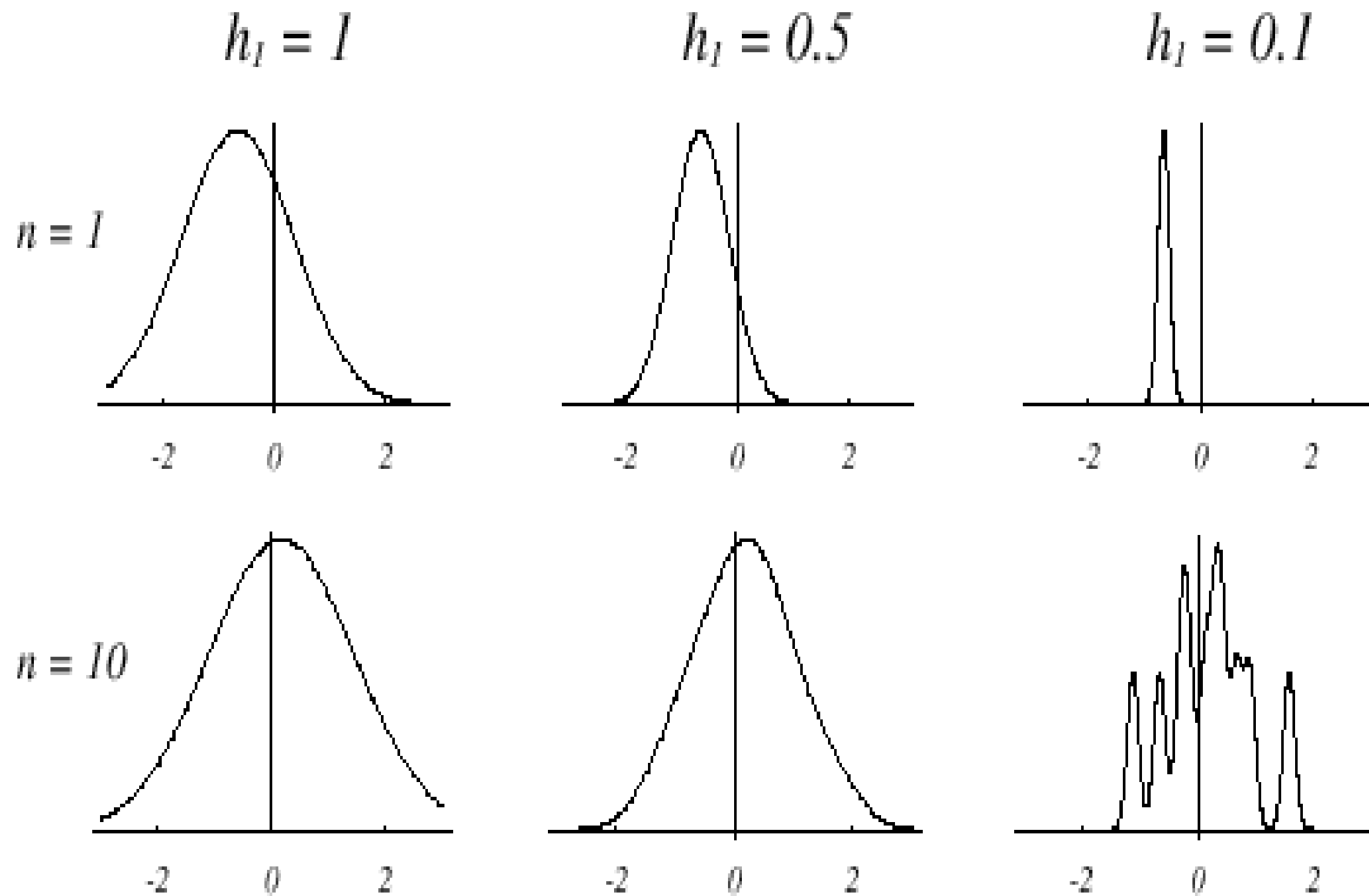
- Numerical results:

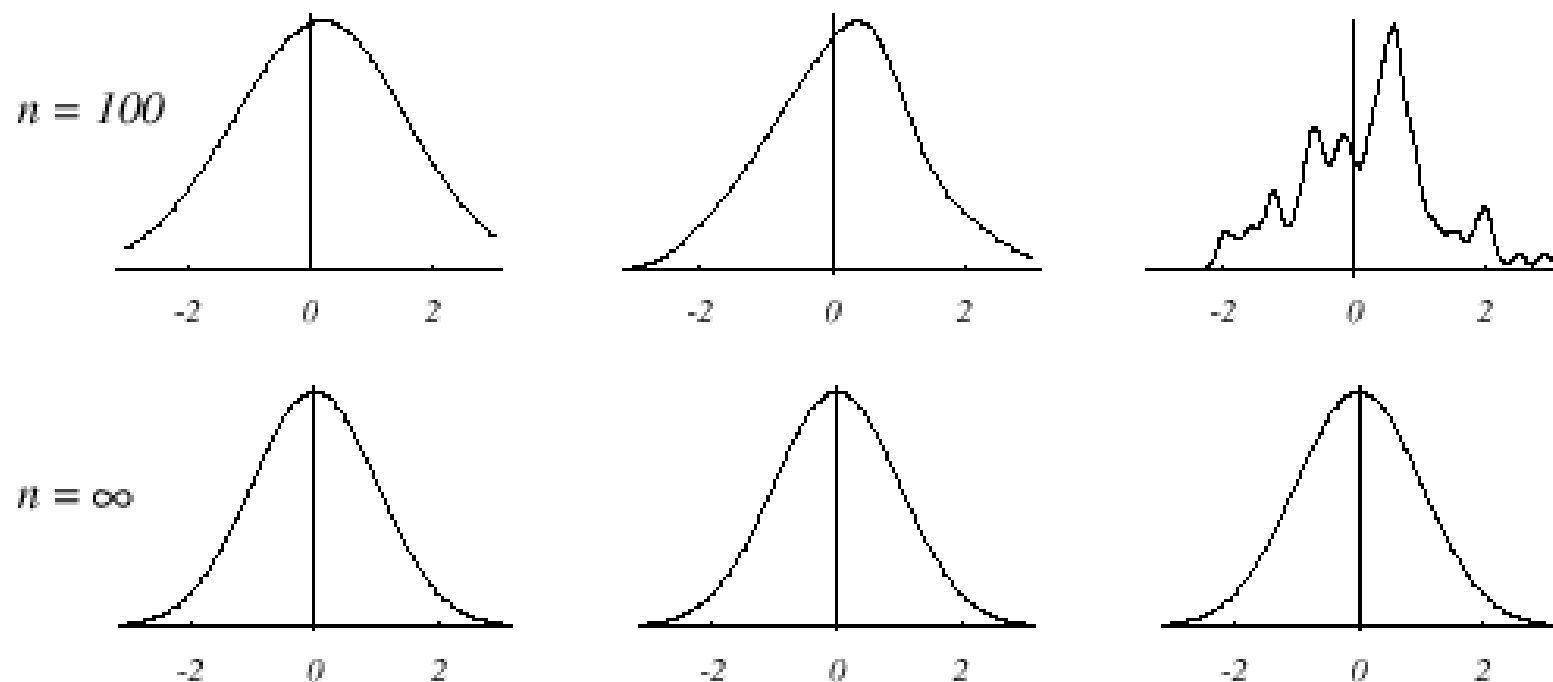
For  $n = 1$  and  $h_1=1$

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

$$p_n(x) = \frac{1}{n} \sum_{i=1, \dots, n} \frac{1}{h_n} \phi\left(\frac{x - x_i}{h_n}\right)$$

For  $n = 10$  and  $h = 0.1$ , the contributions of the individual samples are clearly observable !

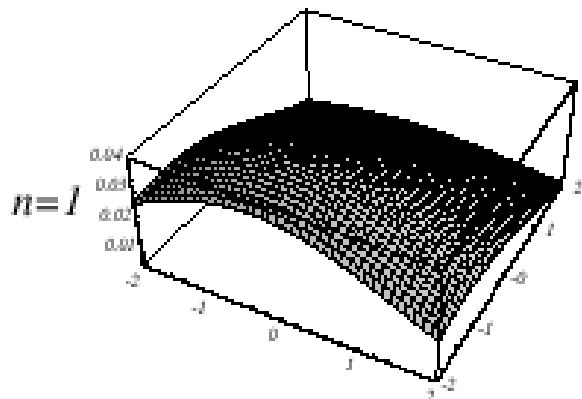




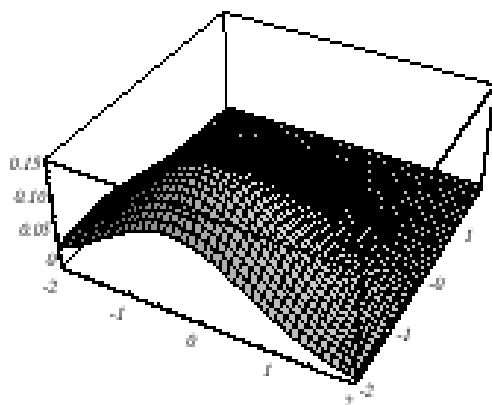
**FIGURE 4.5.** Parzen-window estimates of a univariate normal density using different window widths and numbers of samples. The vertical axes have been scaled to best show the structure in each graph. Note particularly that the  $n = \infty$  estimates are the same (and match the true density function),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

Analogous results are also obtained in two dimensions as illustrated:

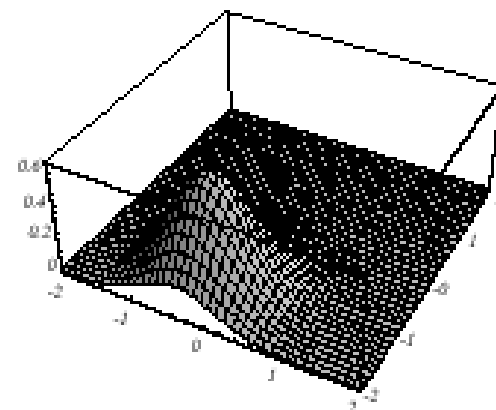
$h_1=2$



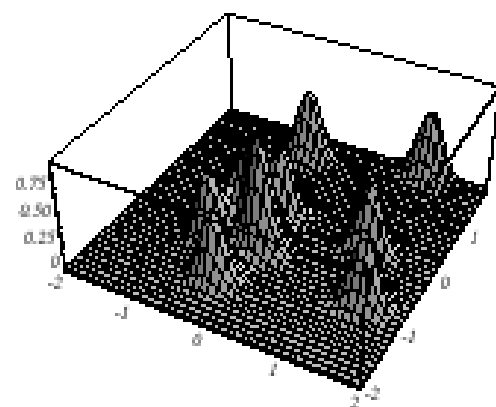
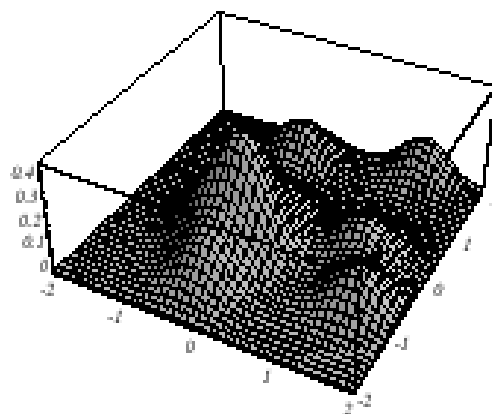
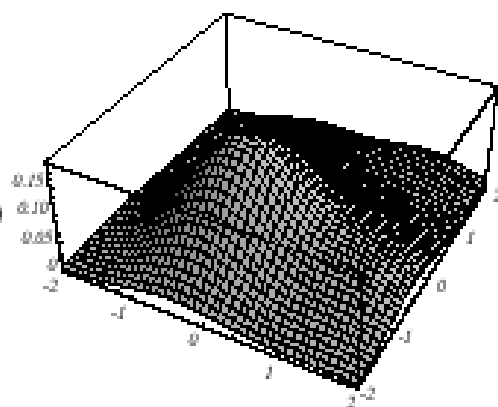
$h_1=1$

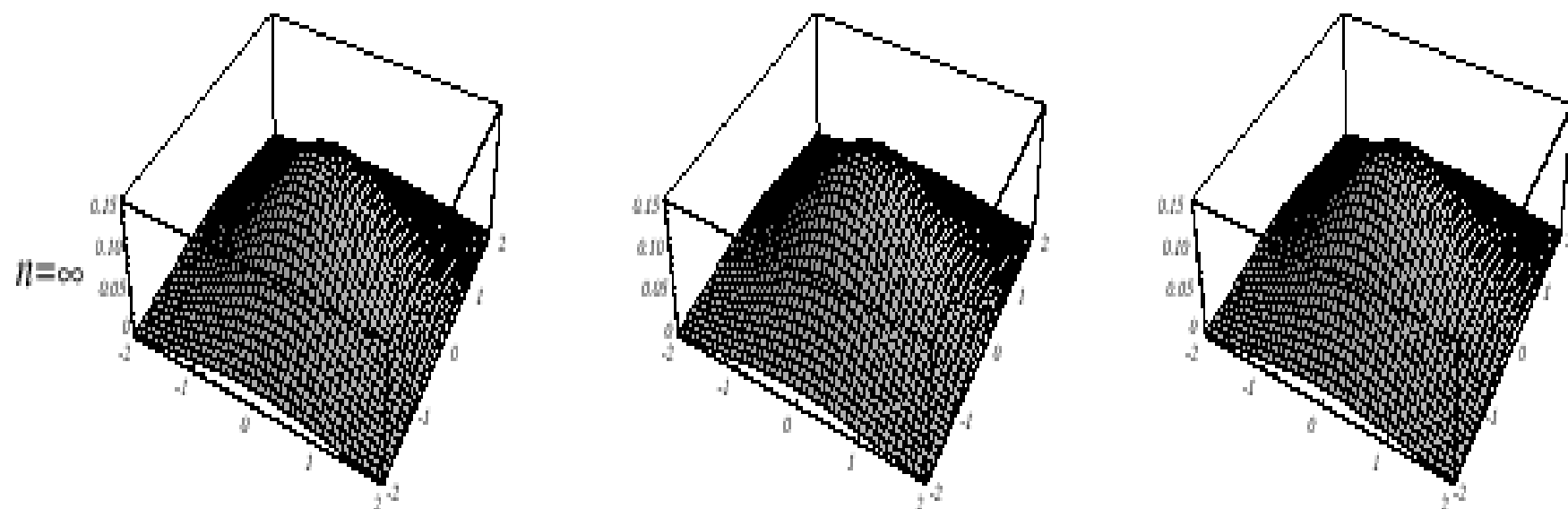


$h_1=0.5$



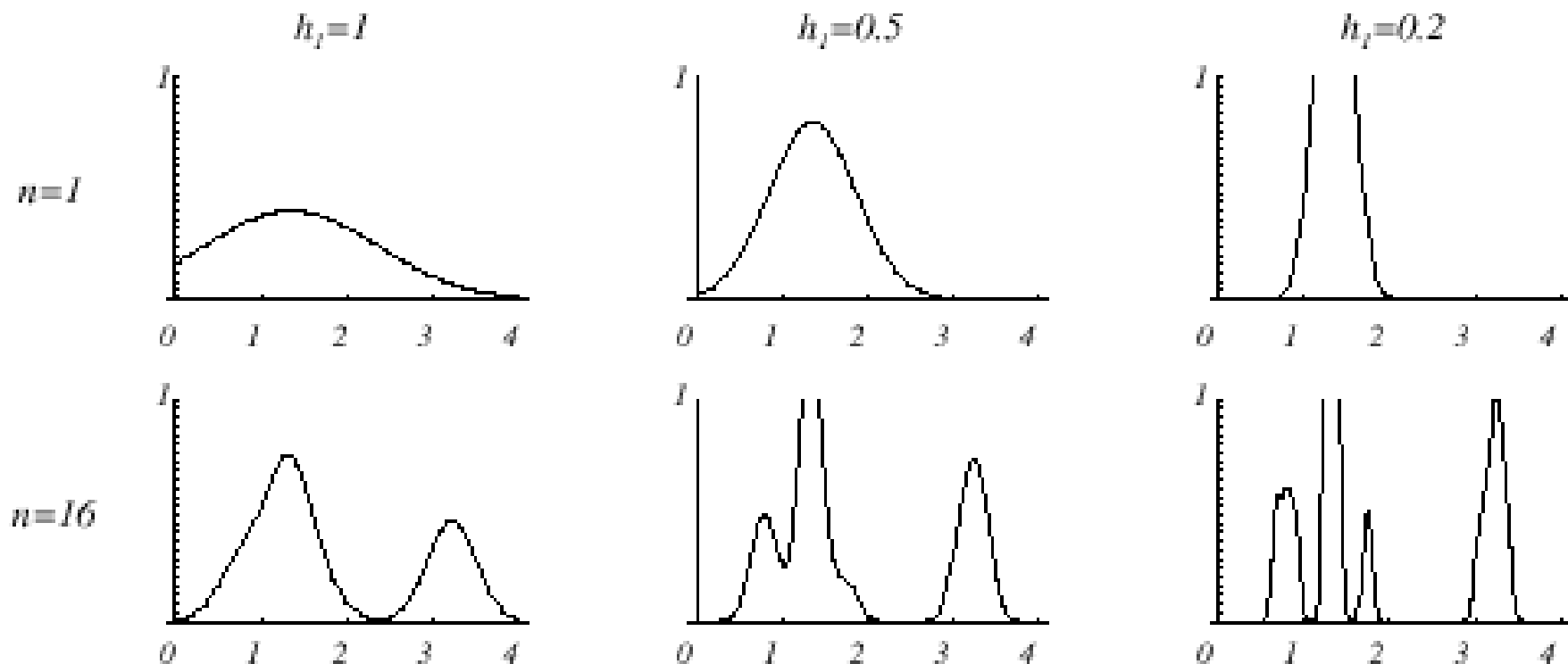
$n=10$

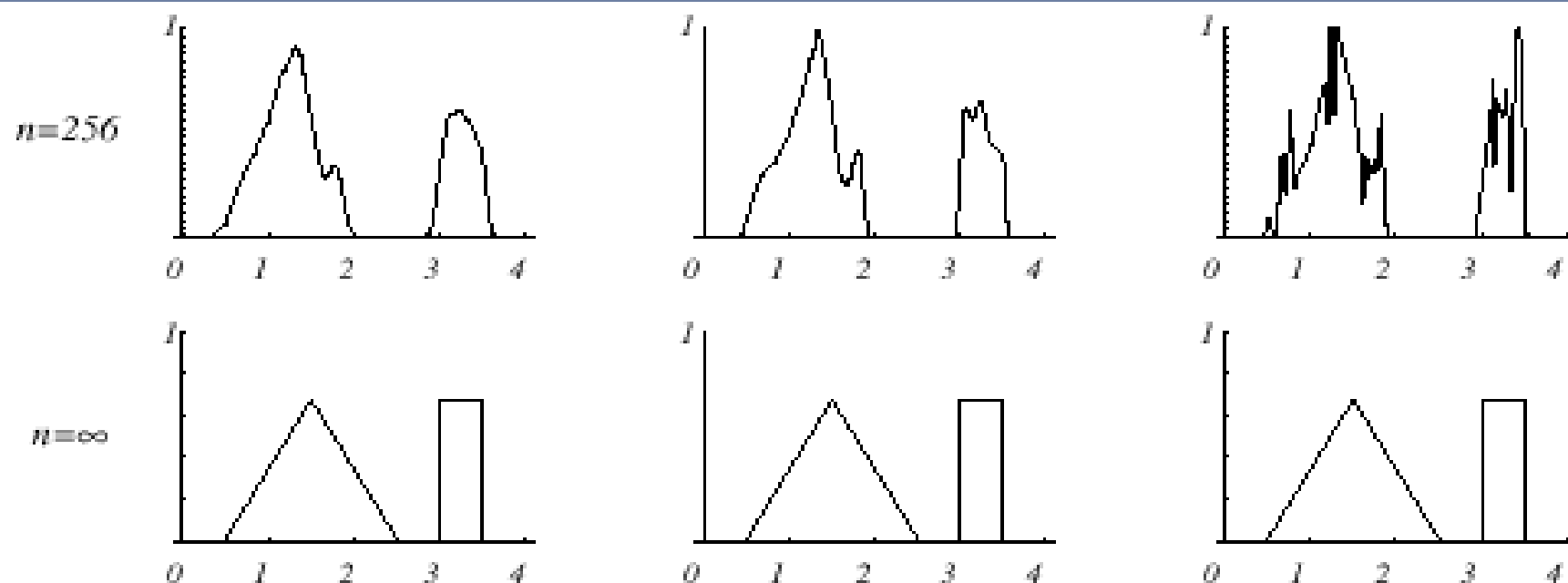




**FIGURE 4.6.** Parzen-window estimates of a bivariate normal density using different window widths and numbers of samples. The vertical axes have been scaled to best show the structure in each graph. Note particularly that the  $n = \infty$  estimates are the same (and match the true distribution),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

- Case where  $p(x) = \lambda_1 \cdot U(a,b) + \lambda_2 \cdot T(c,d)$   
(unknown density) (mixture of a uniform and a triangle density)





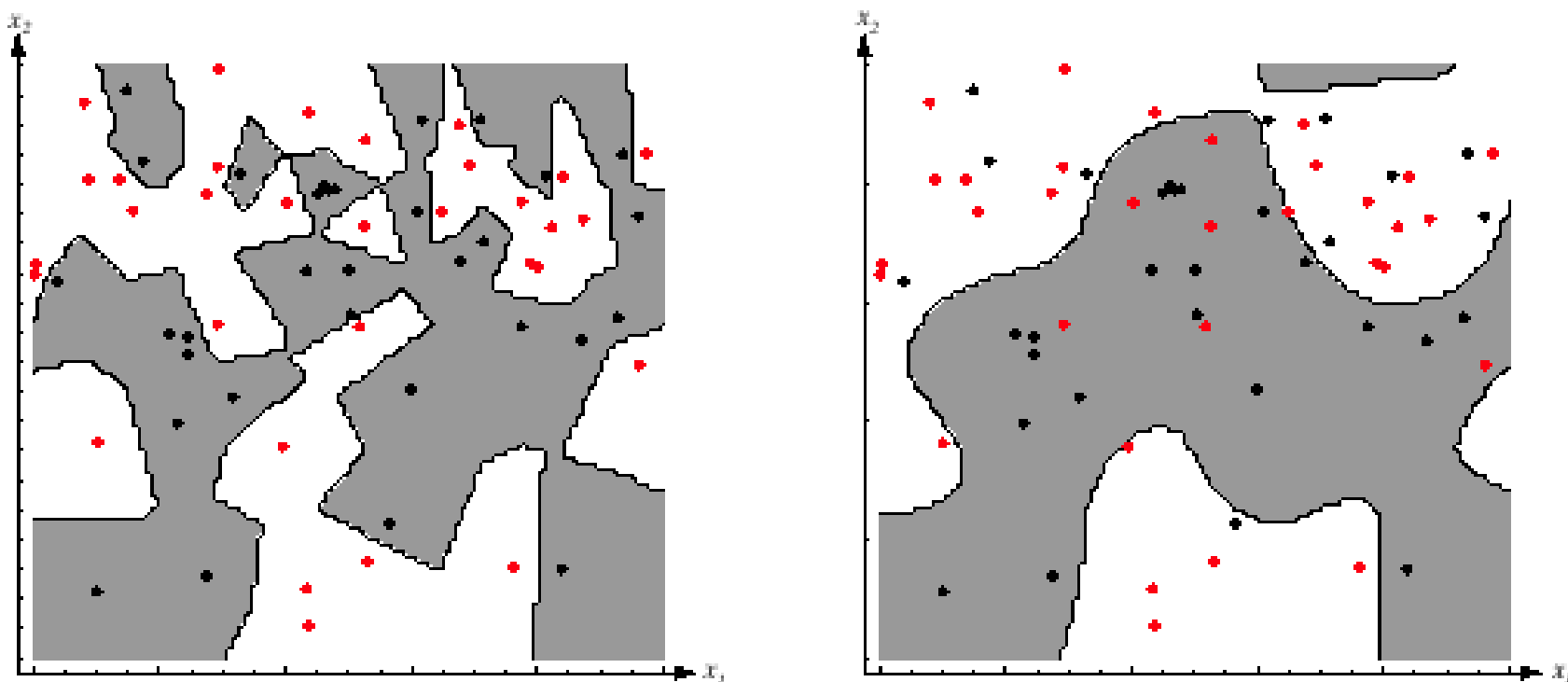
**FIGURE 4.7.** Parzen-window estimates of a bimodal distribution using different window widths and numbers of samples. Note particularly that the  $n = \infty$  estimates are the same (and match the true distribution), regardless of window width. From: Richard O. Duda, Peter E. Hart, and David G. Stork,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

- Classification example

In classifiers based on Parzen-window estimation:

- We estimate the densities for each category and classify a test point by the label corresponding to the maximum posterior
- The decision region for a Parzen-window classifier depends upon the choice of window function as illustrated in the following figure.





**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.