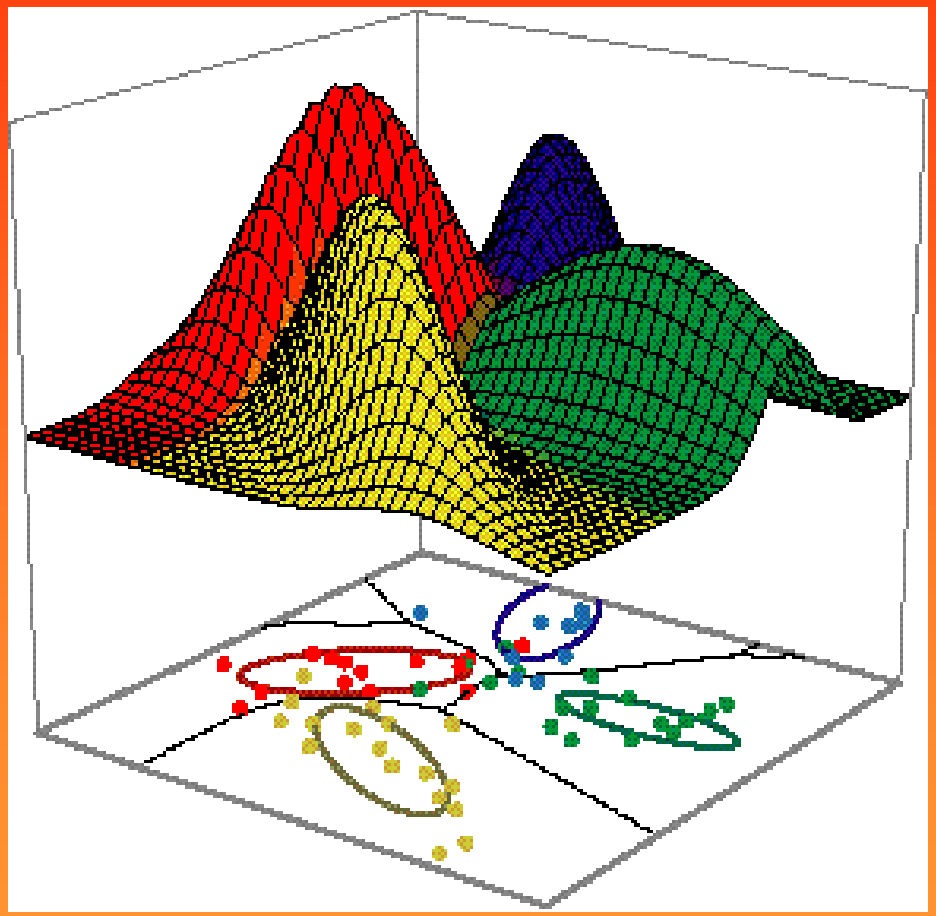


Pattern Classification

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): Nieparametryczne metody rozpoznawania (Sections 4.1-4.3)

- Wprowadzenie
- Estymacja funkcji gęstości
- Estymator Parzena

Wprowadzenie

- Wszystkie sparametryzowane funkcje gęstości są jednomodalne (mają jedno maksimum), natomiast wiele problemów spotykanych w praktyce wymaga opisu w postaci wielomodalnej funkcji gęstości

- Metody nieparametryczne mogą być stosowane zarówno przy ustalonej klasie rozkładu jak i przy nieznanej postaci funkcji gęstości

- Wyróżniamy dwa typy nieparametrycznych metod:
  - Estymacja rozkładów $P(x \mid \omega_j)$
  - Bezpośrednia estymacja prawdopodobieństw a'posteriori

Estymacja funkcji gęstości

Podstawy działania metody:

- Prawdopodobieństwo, że wektor x leży w pewnym obszarze $\mathcal{R}$ jest równe:

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

- *P* jest wygładzonym (uśrednionym) odpowiednikiem funkcji gęstości p(x) dla próby o liczebności n;

  Prawdopodobieństwo, że k punktów przypadnie na dany obszar $\mathcal{R}$ jest równe:

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

a wartość oczekiwana k jest równa: E(k) = nP (3)

Estymator ML dla $P = \theta$

$\underset{\theta}{Max}(P_k \mid \theta)$ jest dany wyrażeniem $\hat{\theta} = \frac{k}{n} \cong P$

Zatem stosunek k/n jest zgodnym estymatorem prawdopodobieństwa P oraz dla funkcji gęstości p.

Jeżeli *p(x)* jest ciągła, obszar *R* jest tak mały, że p nie zmienia się znacząco wewnątrz niego, to:

$$\int_R p(x')dx' \cong p(x)V \qquad (4)$$

gdzie x' to punkt w *R* a V objętością obszaru *R*.

Łącząc równania (1), (3) i (4) otrzymujemy:

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

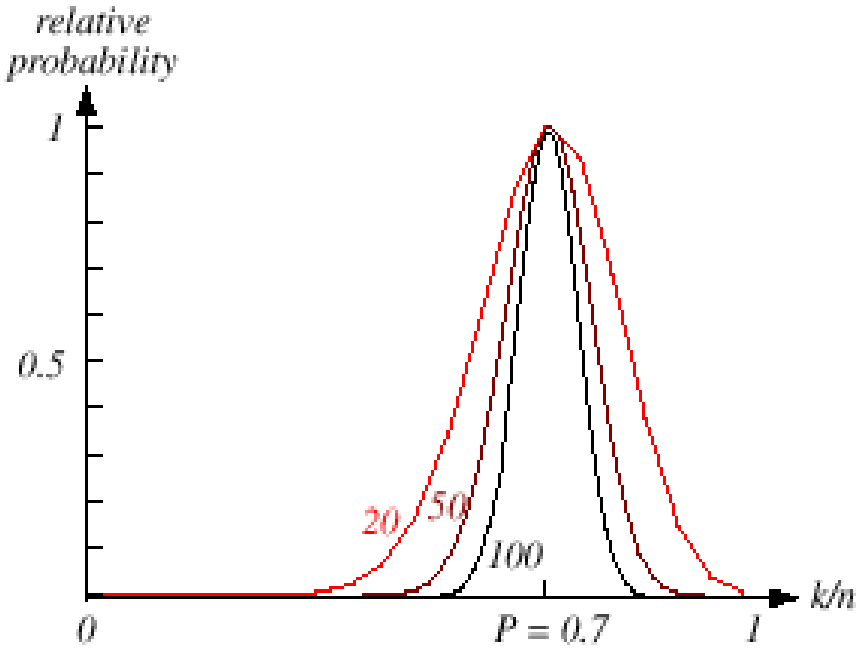


FIGURE 4.1. The relative probability an estimate given by Eq. 4 will yield a particular value for the probability density, here where the true probability was chosen to be 0.7. Each curve is labeled by the total number of patterns n sampled, and is scaled to give the same maximum (at the true probability). 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.

Estymacja funkcji gęstości

- Uzasadnienie równania (4)

$$\int_R p(x')dx' \cong p(x)V \qquad (4)$$

Zakładamy, że *p(x)* jest ciągła i obszar *R* jest tak mały, że p nie zmienia się znacząco wewnątrz niego.

Jeżeli *p(x)* = const., to można je wyciągać przed całkę.

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

gdzie $\mu(R)$ jest: polem powierzchni w przestrzeni euklidesowej R^2

objętością w przestrzeni euklidesowej R^3

hiperobjętością w przestrzeni euklidesowej R^n

Ponieważ $p(x) \cong p(x')$ = const., więc w przestrzeni euklidesowej R^3:

$$\int_R p(x')dx' \cong p(x)V$$

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

- Warunek zbieżności

Stosunek *k/(nV)* jest uśrednioną wartością *p(x).*

p(x) można wyznaczyć tylko wtedy, gdy V zmierza do zera.

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

Ale to jest przypadek, gdy obszar $\mathcal{R}$ nie zawiera żadnych próbek! Estymator jest wówczas rozbieżny.

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

- The volume V needs to approach 0 anyway if we want to use this estimation
  - V nie może być zbyt małe ponieważ liczba próbek jest zawsze ograniczona
  - Nie da się całkowicie zredukować wariancji wielkości k/n
  - Teoretycznie, jeśli dostępna jest nieskończona liczba próbek, nie ma tego problemu

    W celu estymacji funkcji gęstości x, konstruujemy sekwencję obszarów

    $R_1, R_2, \ldots$ zawierających x: pierwszy obszar zawiera jedną próbkę, kolejny dwie, itd.

    Niech V_n będzie objętością obszaru of R_n, k_n liczbą próbek w obszarze R_n a $p_n(x)$ n-tym estymatorem $p(x)$:

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

Następujący zestaw warunków koniecznych musi być spełniony aby $p_n(x)$ zmierzał w granicy do $p(x)$:

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

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

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

Są dwa podejścia do konstrukcji sekwencji obszarów spełniających te warunk

(a) Zmniejszanie początkowego obszaru: $V_n = 1/\sqrt{n}$

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

To podejście nosi nazwę "estymacji przy użyciu okien Parzena"

(b) Ustal k_n jako funkcję n: $k_n = \sqrt{n}$; objętość V_n zwiększa się dotąd, aż n-ty obszar obejmie k_n próbek leżących w sąsiedztwie x. To podejście nosi nazwę "metody k_n-najbliższych sąsiadów".

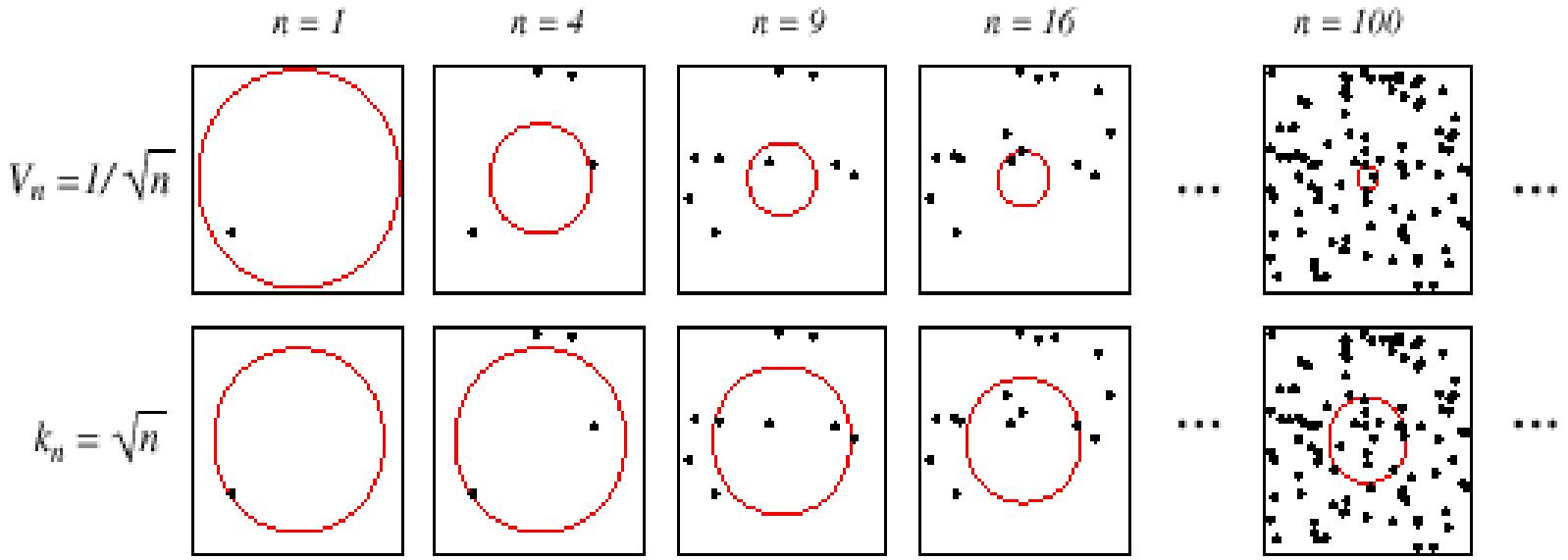


FIGURE 4.2. There are two leading methods for estimating the density at a point, here at the center of each square. The one shown in the top row is to start with a large volume centered on the test point and shrink it according to a function such as $V_n = 1/\sqrt{n}$. The other method, shown in the 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.

Estymator Parzena

- Zakładamy, że R_n jest d-wymiarowym hiperszaścianem

$$V_n = h_n^d \quad (\mathrm{h_n}: \text{ długość krawędzi } R_n)$$

Niech $\varphi(\mathrm{u})$ będzie funkcją okna:

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

- $\varphi((x-x_i)/h_n)$ przyjmuje wartość 1, jeżeli x_i leży wewnątrz hiperześcianu o objętości V_n i środku w x, a w przeciwnym przypadku przyjmuje wartość 0.

- Liczba próbek zawartych wewnątrz hipersześcianu:

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

Wstawiając wyrażenie na k_n do równania (7), otrzymujemy estymator:

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

$P_n(x)$ estymuje $p(x)$ jako średnią z funkcji zmiennej x i próbek (x_i) $(i = 1,\ldots,n)$. Funkcje φ mogą być dowolne!

Przykład

- Szczególny przypadek *p(x)* → *N(0,1)*

Niech $\varphi(u) = \frac{1}{\sqrt{2\pi}} \exp\left[-u^2/2\right]$ *i* $h_n = \frac{h_1}{\sqrt{n}} \quad (n > 1)$

(h_1: znany parametr)

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

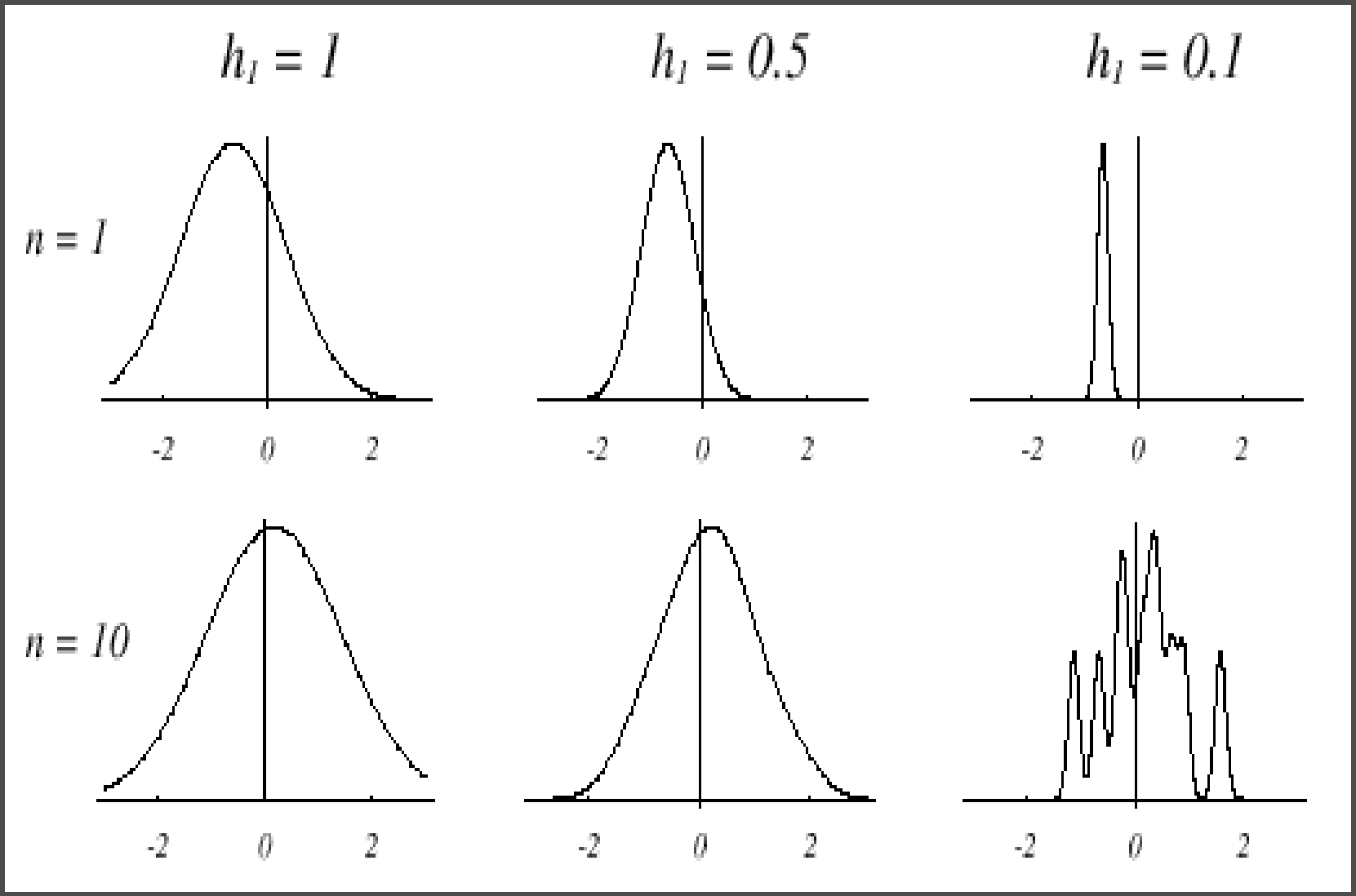
jest średnią rozkładów normalnych z wartościami oczekiwanymi równymi x_i.

Rezultaty obliczeń numerycznych:

Dla $n = 1$ i $h_1 = 1$

$$\mathrm{p}_1(x) = \varphi(x - x_1) = \frac{1}{\sqrt{2\pi}} e^{-1/2}(x - x_1)^2 \rightarrow N(x_1, 1)$$

Dla $n = 10$ i $h = 0.1$,
wpływ pojedynczych próbek można łatwo ocenić


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

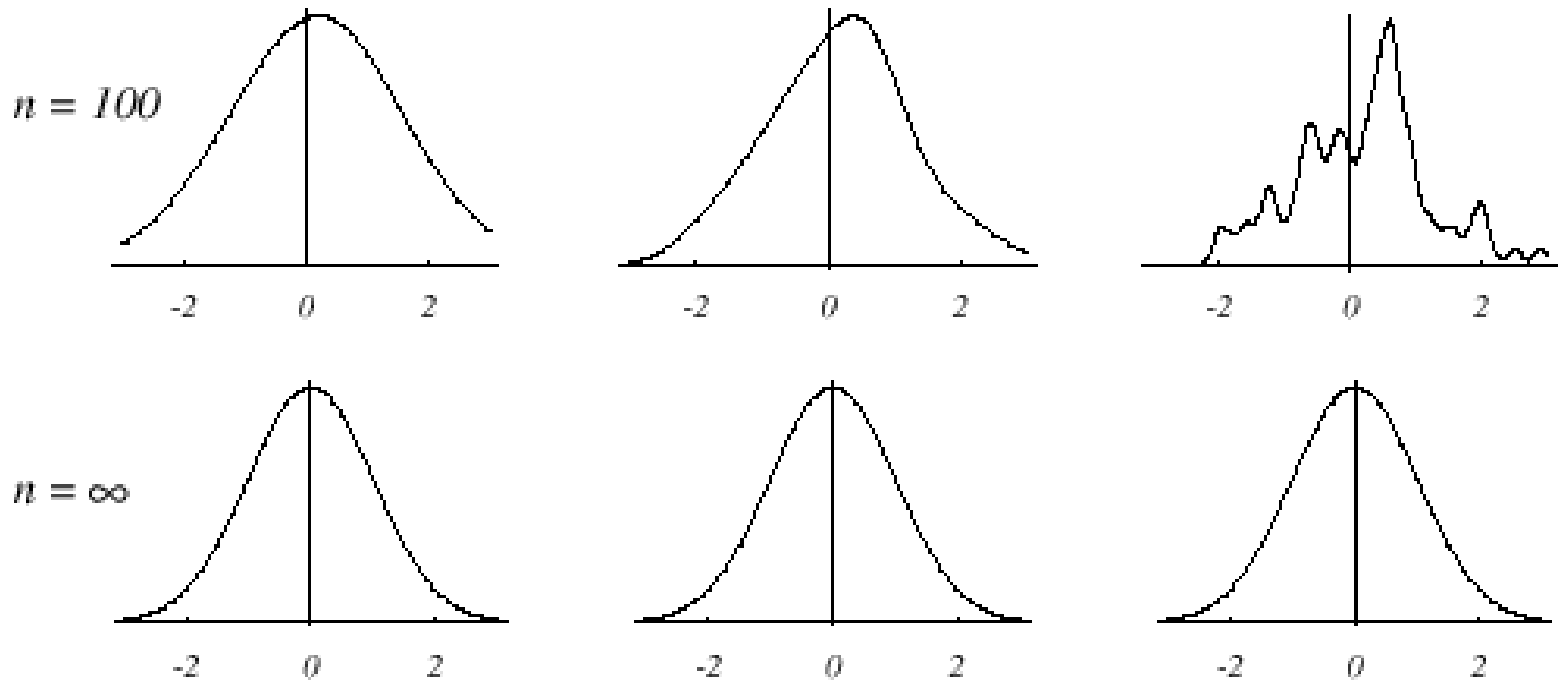
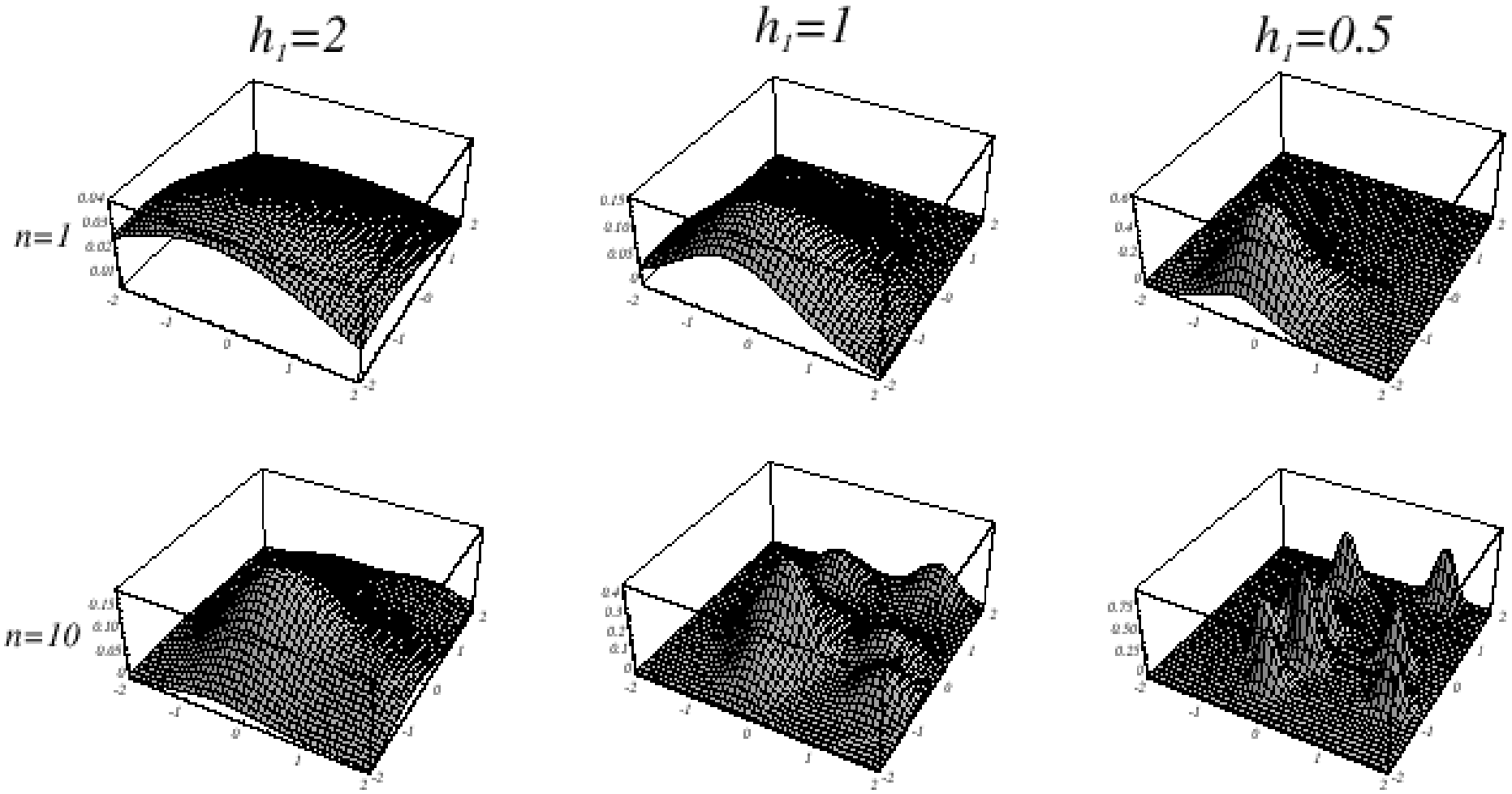


FIGURE 4.5. Parzen-window estimates of a univariate normal density using different window widths and numbers of samples. The vertical axes have been scaled to best show the structure in each graph. Note particularly that the $n = \infty$ estimates are the same (and match the true density function), regardless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

Analogiczne wyniki dla dwóch wymiarów:



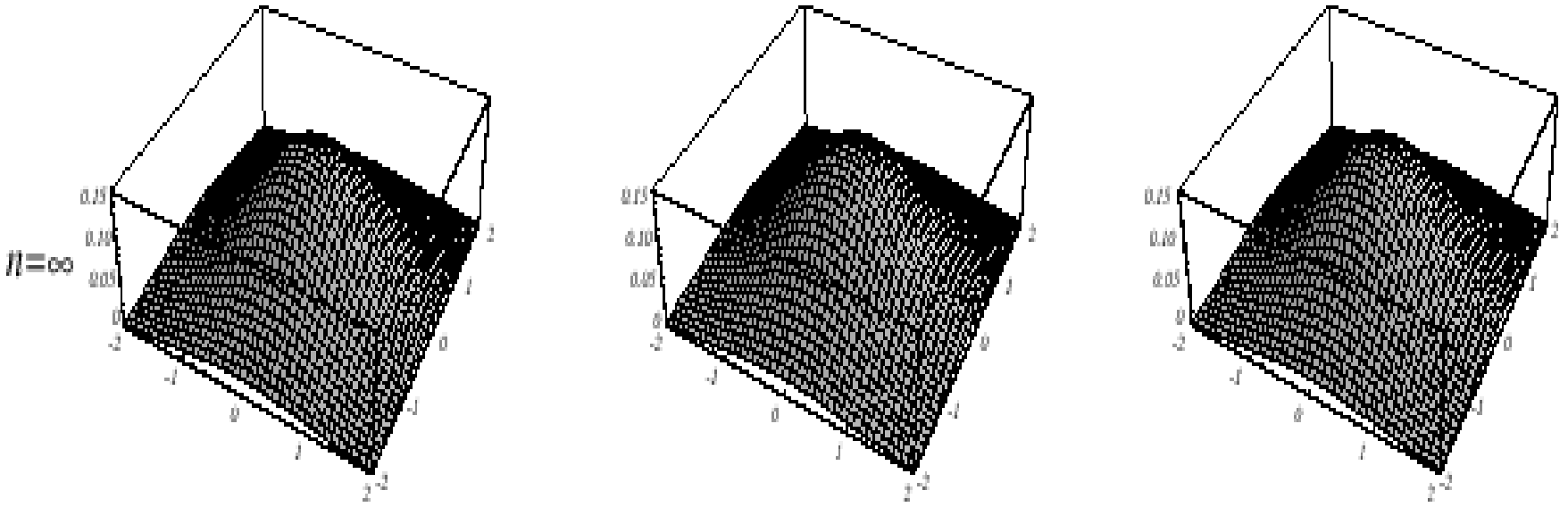
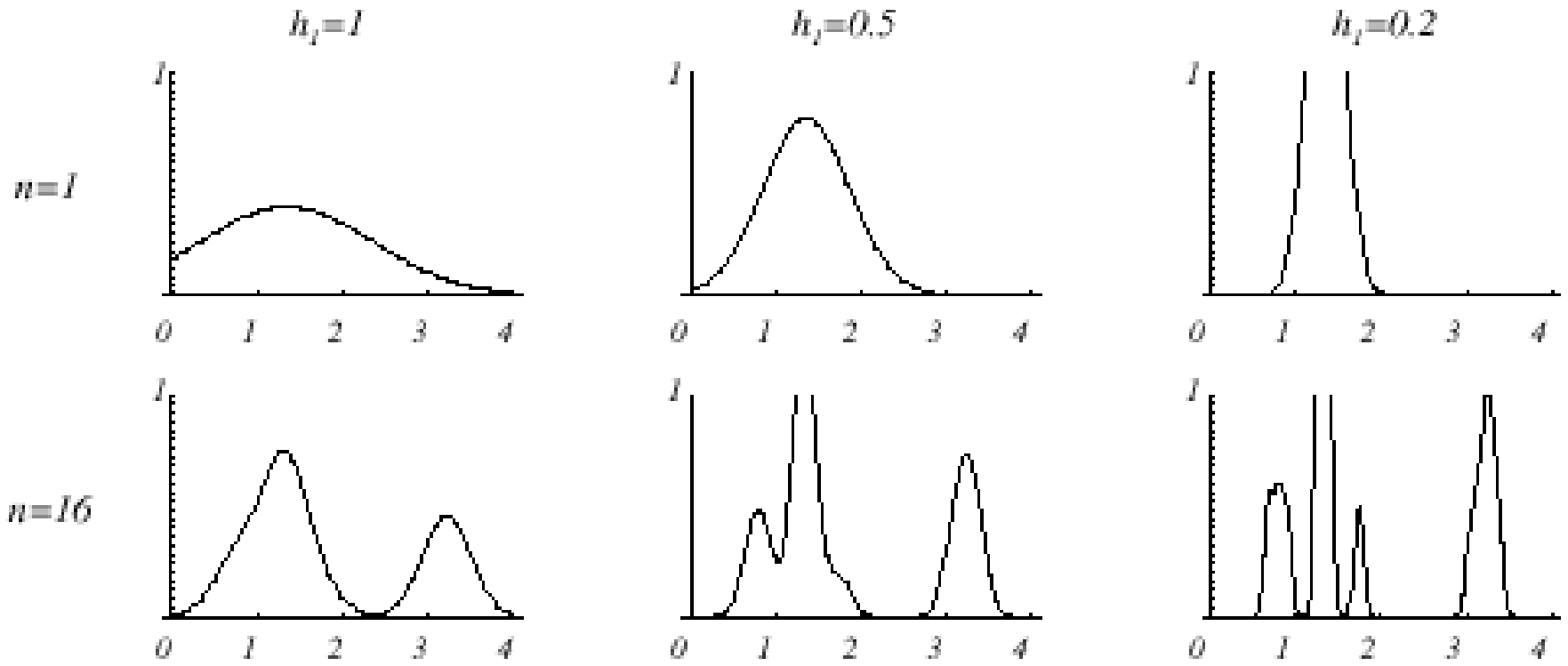


FIGURE 4.6. Parzen-window estimates of a bivariate normal density using different window widths and numbers of samples. The vertical axes have been scaled to best show the structure in each graph. Note particularly that the $n = \infty$ estimates are the same (and match the true distribution), regardless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

- Przypadek $p(x) = \lambda_1.U(a,b) + \lambda_2.T(c,d)$ (nieznany rozkład) (mieszanina rozkładów jednostajnego i trójkątnego)



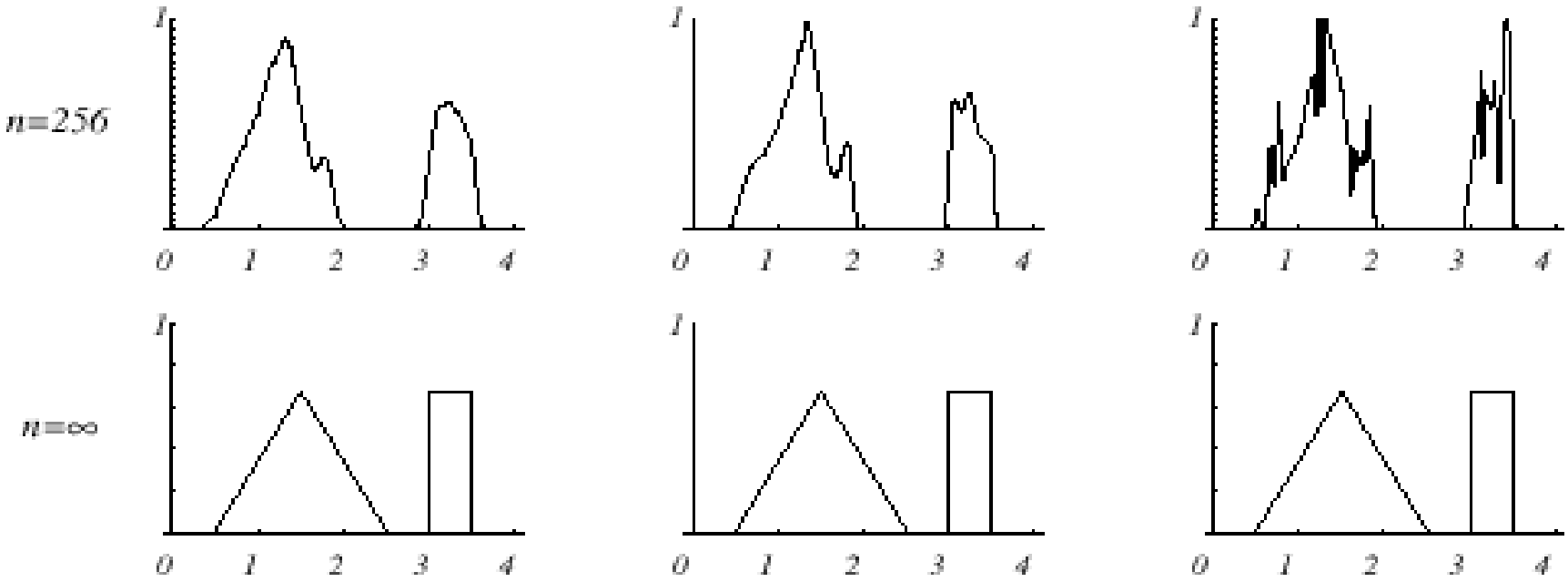


FIGURE 4.7. Parzen-window estimates of a bimodal distribution using different window widths and numbers of samples. Note particularly that the $n = \infty$ estimates are the same (and match the true distribution), regardless of window width. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern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

Przykład rozpoznawania

Klasyfikator oparty o estymator Parzena

- Estymujemy funkcje gęstości dla każdej klasy i przypisujemy nowe próbki do klasy odpowiadającej maksimum a’posteriori

- Postać obszaru decyzyjnego klasyfikatora opartego o estymator Parzena zależy od wyboru funkcji okna, co przedstawiono na następnym rysunku.

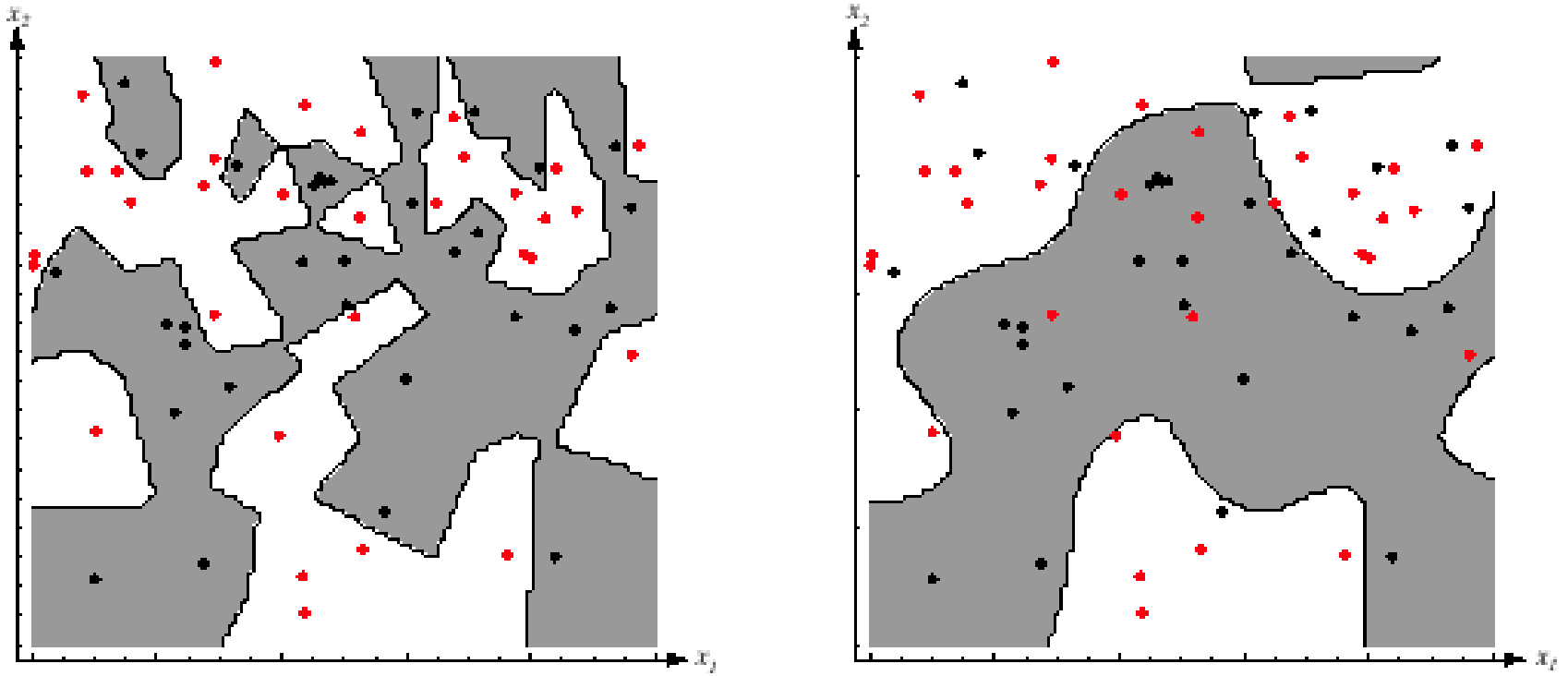


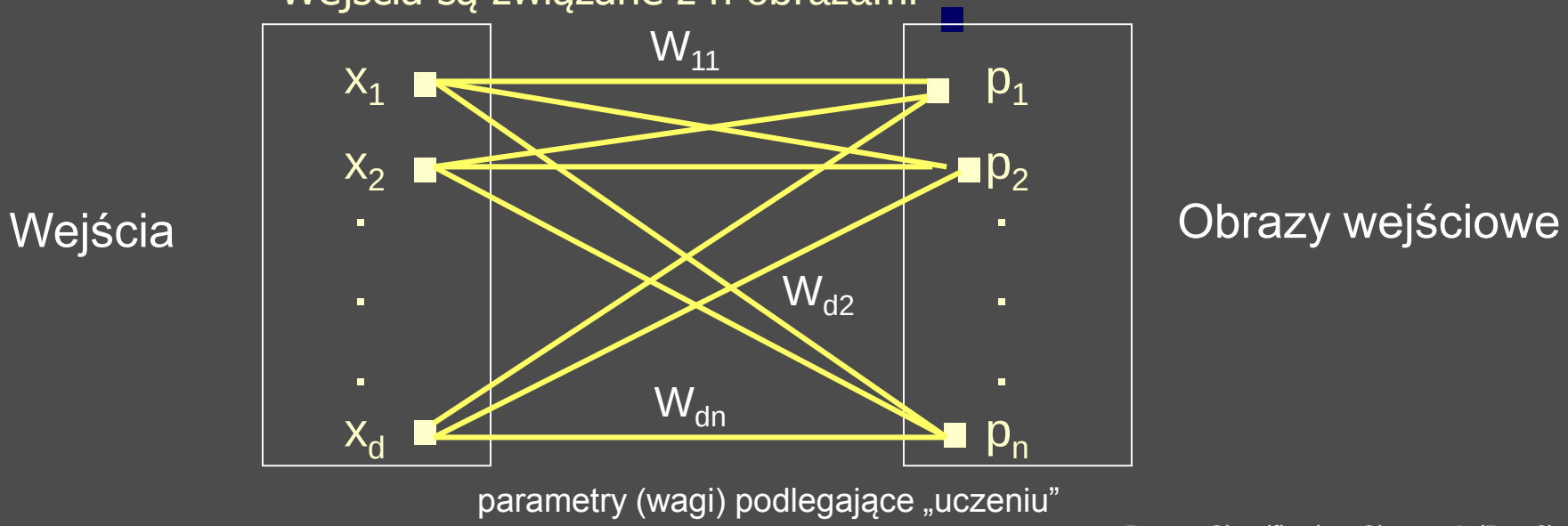
FIGURE 4.8. The decision boundaries in a two-dimensional Parzen-window dichotomizer depend on the window width h. At the left a small h leads to boundaries that are more complicated than for large h on same data set, shown at the right. Apparently, for these data a small h would be appropriate for the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

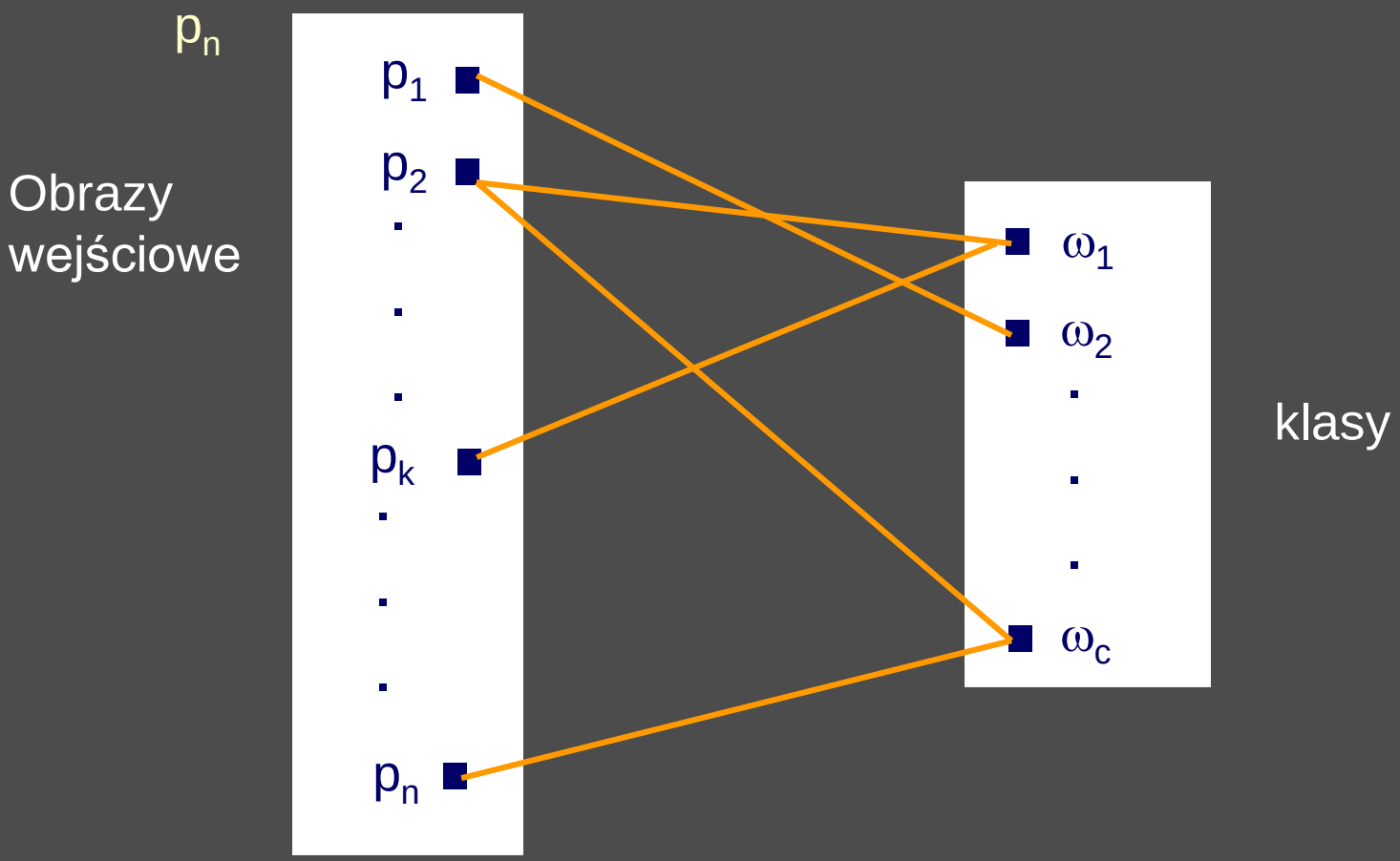
Chapter 4 (part 2): Nieparametryczne metody rozpoznawania (Sections 4.3-4.5)

- Estymator Parzena (ciąg dalszy)
- Estymator Kn –najbliższych sąsiadów
- Algorytm najbliższego sąsiada

Estymator parzena (ciąg dalszy)

- Estymator Parzena– implementacja w postaci Probabilistycznej Sieci Neuronowej
  - Wyznaczenie estymatora Parzena przy użyciu n próbek
    - d-wymiarowe obrazy (wektory cech), c klas
    - Wejścia są związane z n obrazami




p_n
Obrazy wejściowe
p_1
p_2
p_k
p_n
ω_1
ω_2
ω_c
klasy
Nieliniowe funkcje aktywacji

Algorytm uczenia sieci

1. Normalizacja do wartości 1 obrazów x z ciągu uczącego
2. Prezentacja pierwszego obrazu na wejście sieci
3. Ustal wartości parametrów połączeń pomiędzy wejściami sieci i pierwszym obrazem w taki sposób, że: $w_1 = x_1$
4. Utwórz pojedyncze połączenie pomiędzy pierwszym obrazem a klasą, która w ciągu uczącym jest związana z tym obrazem
5. Powtórz kroki 2-4 dla pozostałych obrazów z ciągu uczącego poprzez ustalenie parametrów (wag) tak, że: $w_k = x_k$ $(k = 1, 2, \ldots, n)$

Ostatecznie otrzymujemy następującą sieć:

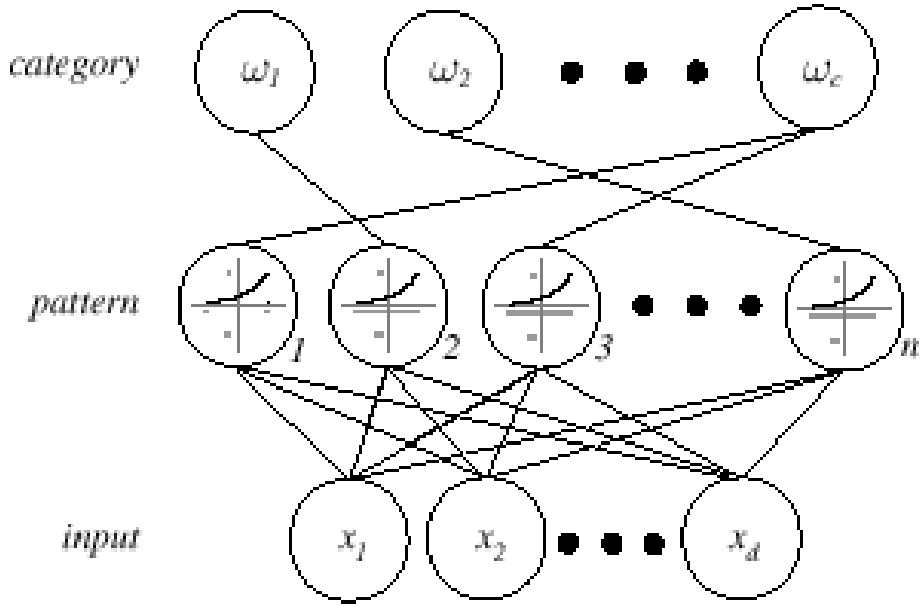


FIGURE 4.9. A probabilistic neural network (PNN) consists of d input units, n pattern units, and c category units. Each pattern unit forms the inner product of its weight vector and the normalized pattern vector $\mathbf{x}$ to form $z = \mathbf{w}^t\mathbf{x}$, and then it emits $\exp[(z-1)/\sigma^2]$. Each category unit sums such contributions from the pattern unit connected to it. This ensures that the activity in each of the category units represents the Parzen-window density estimate using a circularly symmetric Gaussian window of covariance $\sigma^2\mathbf{I}$, where $\mathbf{I}$ is the $d \times d$ identity matrix. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Algorytm testowania sieci

1. Normalizuj próbki x z ciągu testującego i podaj je na wejście sieci
2. Dla każdego obrazu wyznacz iloczyn skalarny:

$$net_k = w_k^T x$$

   a rezultat podaj na wejście funkcji aktywacji:

$$f(net_k) = \exp\left[\frac{net_k - 1}{\sigma^2}\right]$$

3. Wyznacz wartość wyjścia jako sumę wkładów od wszystkich obrazów połączonych z tym wyjściem

$$P_n(x \mid \omega_j) = \sum_{i=1}^{n} \varphi_i \propto P(\omega_j \mid x)$$

4. Jako numer klasy zwróć taką wartość *j* , dla której $P_n(x \mid \omega_j)$ *(j = 1, …, c)* jest największe.

- Estymator K_n – najbliższych sąsiadów

  - Cel: rozwiązanie problemu wyboru nieznanej „najlepszej” funkcji okna

    - Niech objętość hipersześcianu będzie funkcją danych uczących
    - Ustal położenie środka hipersześcianu nad x i zwiększaj jego objętość, aż w jego wnętrzu znajdzie się k_n próbek $(k_n = f(n))$
    - k_n noszą nazwę k_n najbliższych sąsiadów próbki x

  Wyróżniamy dwa przypadki:

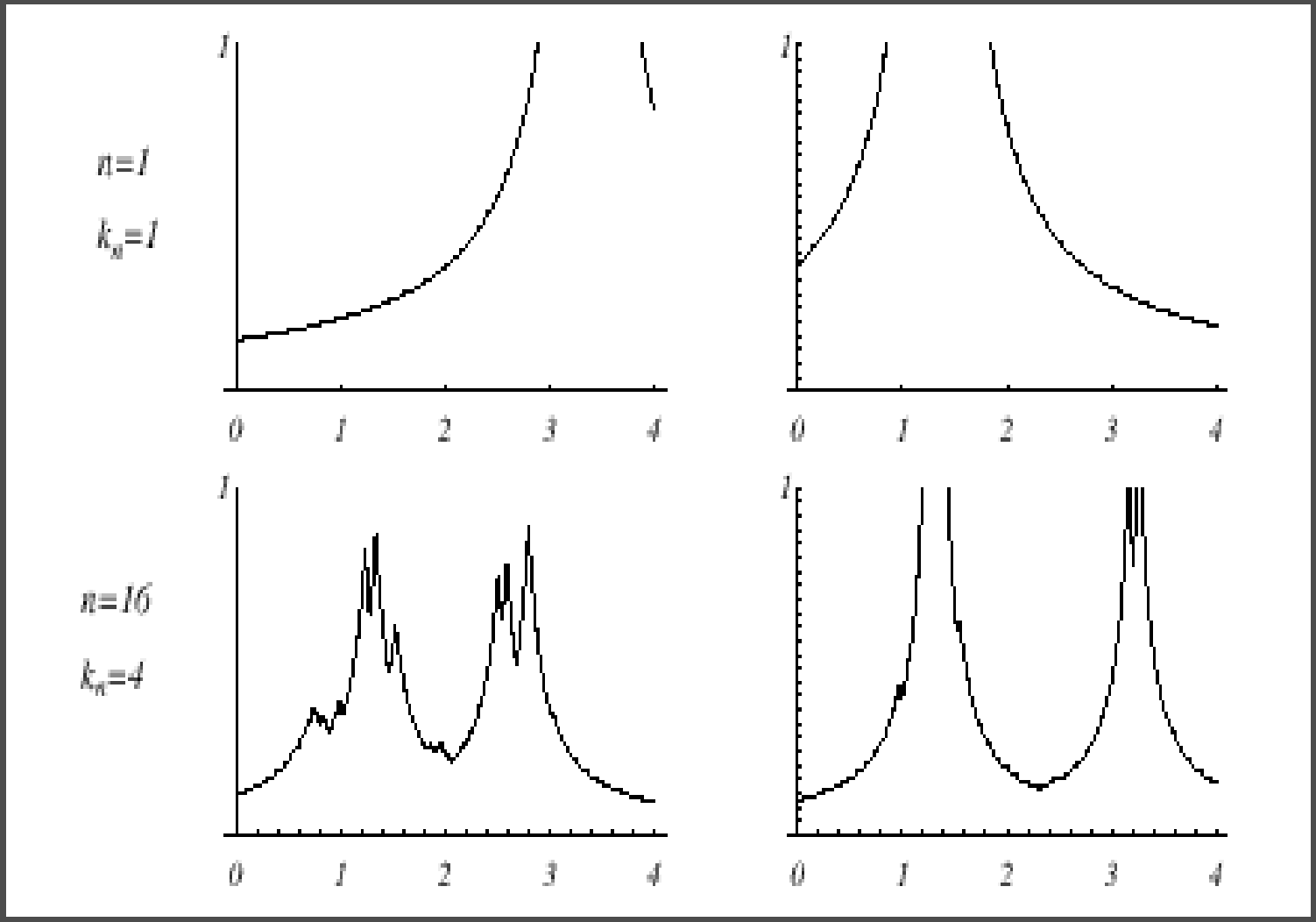
    - W otoczeniu x znajduje się dużo sąsiadów; objętość hipersześcianu jest mała, co skutkuje dobrą „rozdzielczością”
    - W otoczeniu x znajduje się niewiele sąsiadów; objętość hipersześcianu staje się duża a jej wzrost zatrzymuje się po osiągnięciu obszarów skupiających wiele próbek (sąsiadów)

  Możemy otrzymać rodzinę estymatorów, przyjmując $k_n=k_1/\sqrt{n}$ i wybierając różne wartości k_1

Przykład

Dla $k_n = \sqrt{n} = 1$; estymator ma postać:

$$P_n(x) = k_n / nV_n = 1 / V_1 = 1 / 2|x-x_1|$$


n=1
k_n=1
n=16
k_n=4
1
0
1
2
3
4

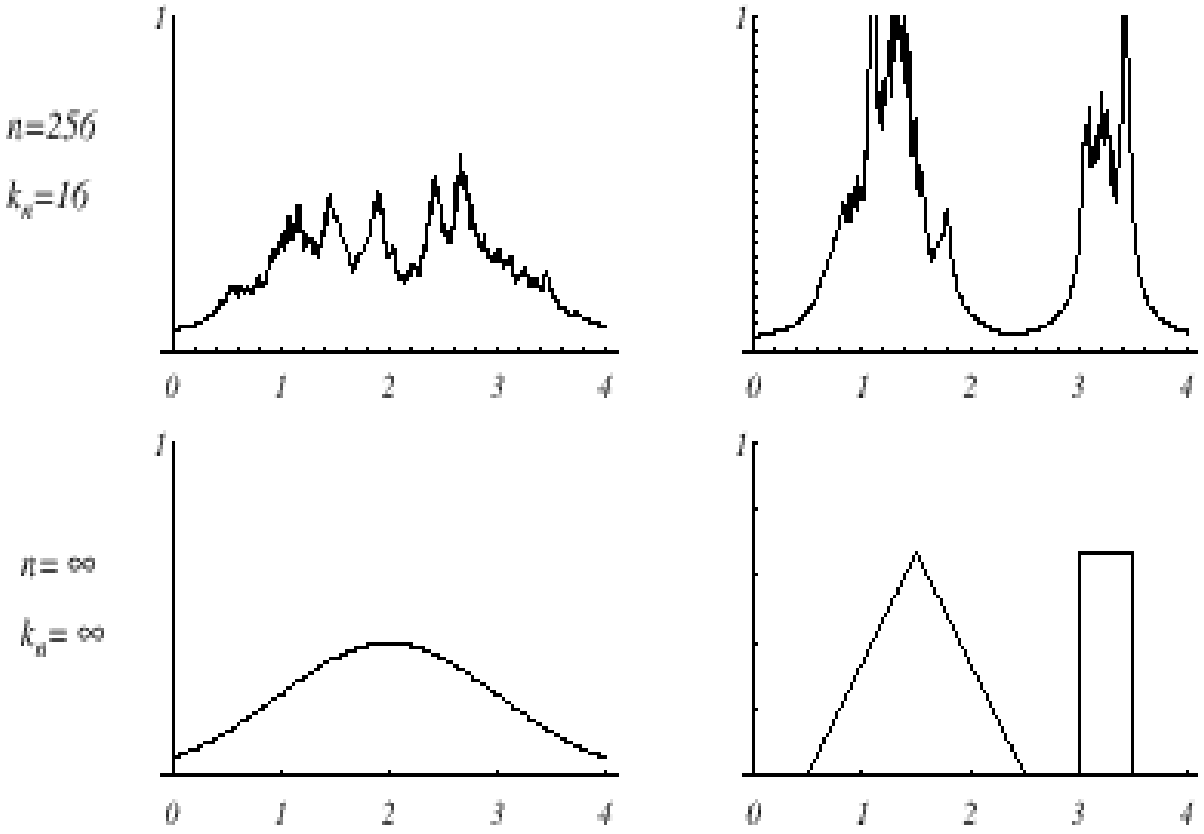


FIGURE 4.12. Several *k*-nearest-neighbor estimates of two unidimensional densities: a Gaussian and a bimodal distribution. Notice how the finite *n* estimates can be quite "spiky." From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Estymacja prawdopodobieństw a'posteriori

- Cel: estymacja $P(\omega_i \mid x)$ na podstawie ciągu uczącego n sklasyfikowanych próbek

- Konstruujemy hipersześcian o objętości V wokół x i zbieramy k próbek
- k_i próbek spośród k będzie pochodzić z klasy ω_i i wówczas:

$$p_n(x, \omega_i) = k_i / nV$$

Estymator $p_n(\omega_i \mid x)$ ma postać:

$$p_n(\omega_i \mid x) = \frac{p_n(x, \omega_i)}{\sum_{j=1}^{j=c} p_n(x, \omega_j)} = \frac{k_i}{k}$$

- k_i/k określa względną liczebność próbek z klasy ω_i wewnątrz hiperześcianu
- Aby zminimalizować liczbę błędów, wybierana jest klasa najczęściej reprezentowana wewnątrz hiperześcianu
- Algorytm działa najlepiej, jeżeli k jest duże i hiperześcian dostatecznie mały

Algorytm najbliższego sąsiada

- $D_n = \{x_1, x_2, \ldots, x_n\}$ – zbiór uczący
- Niech $x' \in D_n$ będzie próbką uczącą leżącą najbliżej nowej próbki x; Klasyfikacja próbki x sprowadza się do przypisania jej tej samej klasy, z której pochodzi x'
- Algorytm najbliższego sąsiada prowadzi do błędu większego niż błąd popełniany przez optymalny klasyfikator Bayesa
- Jeżeli liczba próbek w ciągu uczącym jest bardzo duża (nieskończona), wówczas algorytm najbliższego sąsiada popełnia nie więcej niż dwukrotnie więcej błędów od optymalnego klasyfikatora Bayesa
- Jeżeli $n \to \infty$, zawsze można znaleźć x' takie, że:

$$P(\omega_i \mid x') \cong P(\omega_i \mid x)$$

Przykład:

$x = (0.68, 0.60)^T$

Obrazy z ciągu uczącego	Klasy	Estymowane prawdopodobieństwa a'posteriori
$(0.50, 0.30)$	ω_2 ω_3	0.25 $0.75 = P(\omega_m \mid x)$
$(0.70, 0.65)$	ω_5 ω_6	0.70 0.30

<u>Decyzja:</u> próbce x przypisujemy klasę ω_5

- Jeżeli $P(\omega_m \mid x) \cong 1$, to rezultaty algorytm najbliższego sąsiada są bardzo podobne do rezultatów optymalnego klasyfikatora Bayesa

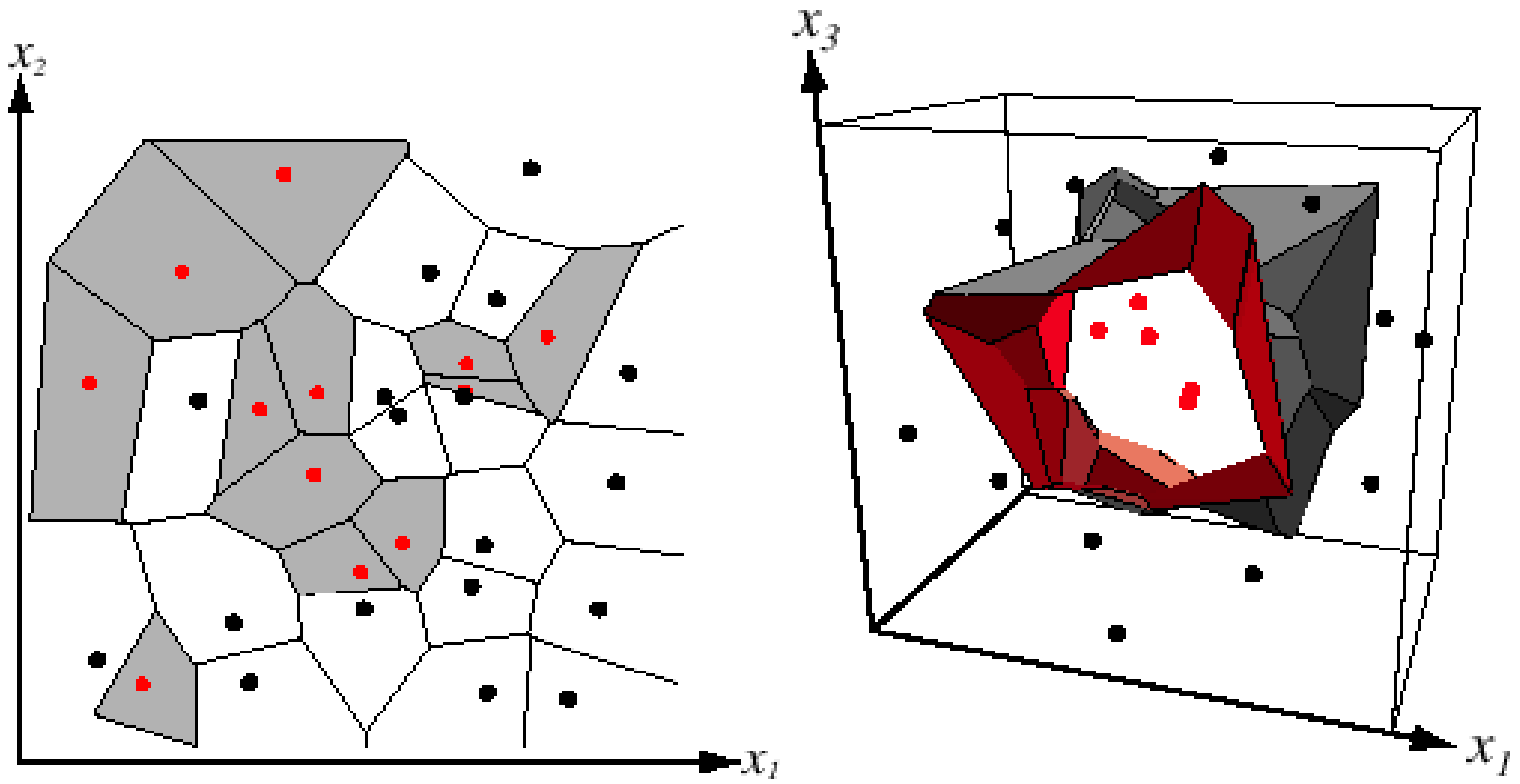


FIGURE 4.13. In two dimensions, the nearest-neighbor algorithm leads to a partitioning of the input space into Voronoi cells, each labeled by the category of the training point it contains. In three dimensions, the cells are three-dimensional, and the decision boundary resembles the surface of a crystal. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

- Reguła k – najbliższych sąsiadów
  - Cel: Klasyfikacja próbki x przez przypisanie jej klasy najczęściej reprezentowanej przez k najbliższych próbek z ciągu uczącego

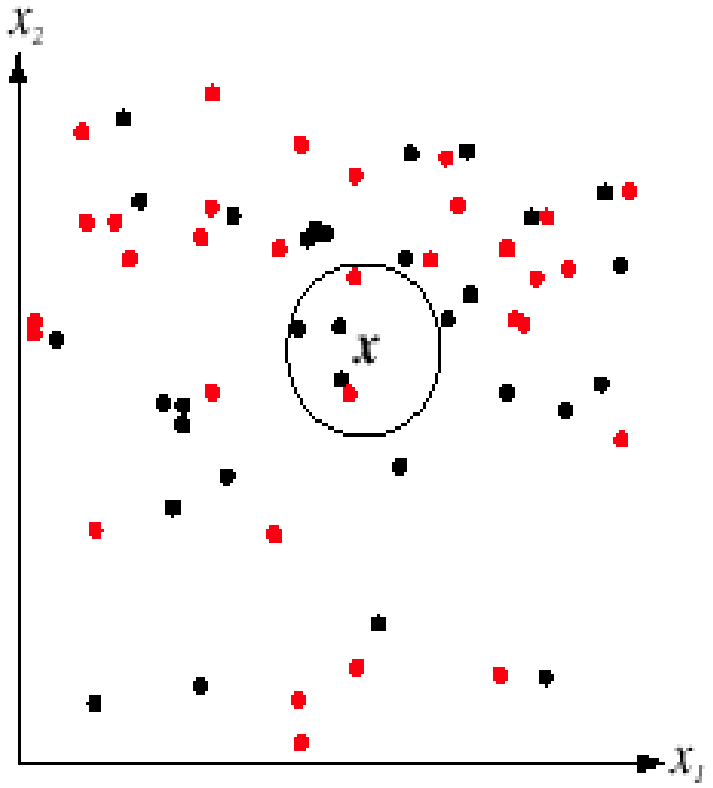


FIGURE 4.15. The k-nearest-neighbor query starts at the test point $\mathbf{x}$ and grows a spherical region until it encloses k training samples, and it labels the test point by a majority vote of these samples. In this $k = 5$ case, the test point $\mathbf{x}$ would be labeled the category of the black points. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Przykład:

$k = 3$ (liczba nieparzysta) i $x = (0.10, 0.25)^T$

Ciąg uczący	Klasy
(0.15, 0.35)	ω_1
(0.10, 0.28)	ω_2
(0.09, 0.30)	ω_5
(0.12, 0.20)	ω_2

Próbki z ciągu uczącego najbliższe próbce x wraz z ich klasami:

$\{(0.10, 0.28, \omega_2); (0.12, 0.20, \omega_2); (0.15, 0.35, \omega_1)\}$

ω_2 pojawia jest najczęściej reprezentowana i to ta klasa jest przypisana próbce x