Università di Verona

A.Y. 2021-22

Machine Learning & Artificial Intelligence

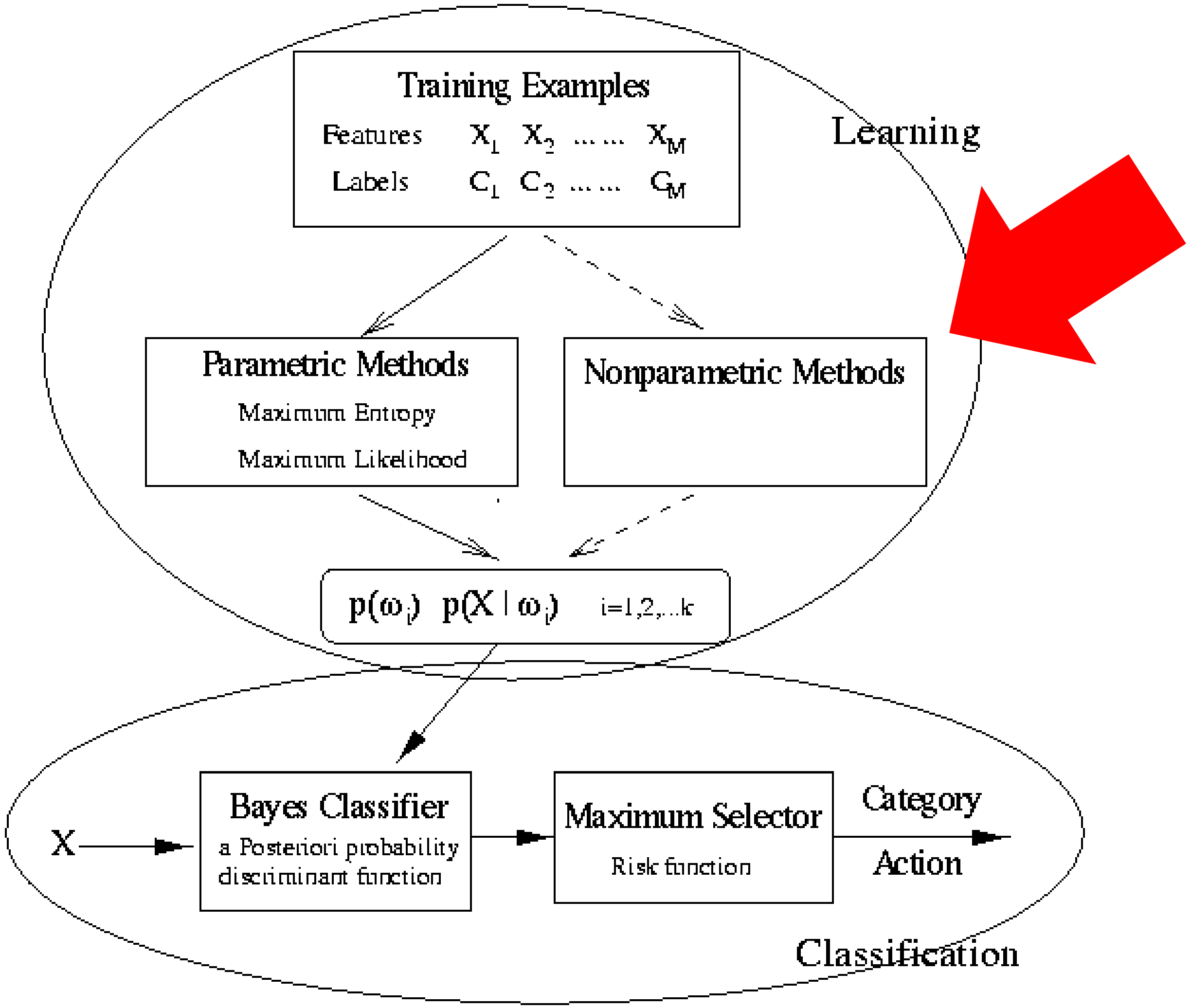
Non parametric techniques

Vittorio Murino

Summary

- Introduction
- Potential functions
- Conditional probability density estimation
- Parzen Windows
- Prototype Method
- k-NN – k Nearest-Neighbor Method

Overview


Training Examples
Features x_1 x_2 ...... x_M
Labels C_1 C_2 ...... C_M
Learning
Parametric Methods
Maximum Entropy
Maximum Likelihood
Nonparametric Methods
$p(\omega_i)$ $p(X \mid \omega_i)$ i=1,2,...k
X
Bayes Classifier
a Posteriori probability
discriminant function
Maximum Selector
Risk function
Category
Action
Classification

Introduction

- The Bayesian classifier minimizes the probability of error.
- The problem with this classifier is that it is based on prior and conditional probabilities: if these are unknown they must be estimated from the data.
- There are three types of methods for estimating these densities:
    - parametric methods: the shape of the density is assumed, its parameters are estimated from the data (e.g., gaussian)
    - non-parametric methods: no assumption on density form, fully estimated by data (e.g. KNN)
    - semi-parametric methods: mixed technique, it is assumed that there is a rather large family of density functions (e.g. neural networks)

- In parametric methods it is assumed that the form of probability densities is known, but this assumption cannot be made in many recognition problems.

- In particular, most parametric methods assume that probability densities are unimodal (i.e., they have a single maximum), but in reality many problems involve the use of multimodal densities.

- In this part we will illustrate some non-parametric methods, which allow to estimate the probability density functions starting directly from the samples.

- In particular, there is the problem of estimating the quantity

$$p(\mathbf{x}|\omega_i) \equiv \widehat{p}_i(\mathbf{x})$$

- This problem, if solved, allows you to use the Bayesian classifier (theoretical optimal)

- These methods have the common characteristic of estimating the required functions by means of a simpler set of functions, usually associated with each sample

Potential functions

- Basic idea: to establish an analogy between samples, thought as points in an appropriate space, and the concept of electric charge.
- By placing an electric charge at each point associated with a sample, it can be assumed that by appropriately defining a potential associated with the charge, the resulting electrostatic potential defines a discriminating function for the recognition problem considered.
- This problem formulation implies the ability to approximate a global *discriminant* function (referred to the entire space) by means of a set of potential functions.

- Under this perspective, the problem is analogous to finding an expression for the conditional probability starting from the samples and a set of potential functions associated with them.
- On the other hand, the two problems (definition of linear discriminating function and conditional probability) are directly related.
- In this case, however, unlike the Gaussian case for example, it is assumed not to know the shape of the probability density function.
- Therefore, we are talking about *non-parametric methods*, as it becomes necessary to estimate the probability density directly from the samples.

- Let $\gamma(\mathbf{x},\mathbf{y}_\mathrm{j})$ be a potential function for a generic sample $\mathbf{y}_\mathrm{j}$ belonging to class i.
- Then we can write:

$$\widehat{p}_i(\mathbf{x}) = \frac{1}{N_i}\sum_{j=1}^{N_i}\gamma\left(\mathbf{x},\mathbf{y}_j\right) \qquad \text{where } N_i = \#\text{ samples class } i$$

- To build a good approximation we must impose some constraints on the form of γ.

  1) $\gamma(\mathbf{x},\mathbf{y}) \geq 0$

  2) arg $\max_{\mathbf{x}} \gamma(\mathbf{x},\mathbf{y}_\mathrm{k}) = \mathbf{y}_\mathrm{k}$**,** that is, $\gamma(\mathbf{x},\mathbf{y}_\mathrm{k})$ is max for $\mathbf{x}=\mathbf{y}_\mathrm{k}$**;**

  3) $\gamma(\mathbf{x},\mathbf{y}_1) \cong \gamma(\mathbf{x},\mathbf{y}_2)$, se $|\mathbf{y}_2-\mathbf{y}_1| < \varepsilon$, that is, if the two vectors of the samples are "close enough" – this constraint serves to ensure that p does not vary abruptly or may have discontinuities

  4) $\gamma(\mathbf{x},\mathbf{y})$ continuous.

5) $\int_{-\infty}^{+\infty} \gamma(\mathbf{x}, \mathbf{y}_k) d\mathbf{x} = 1$ normalisation condition.

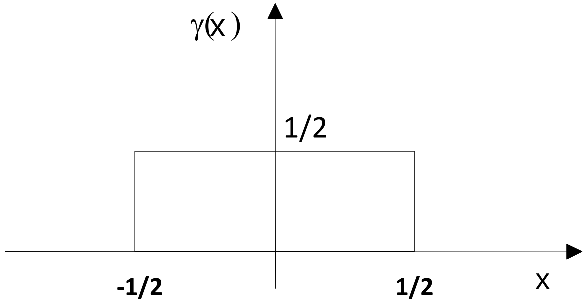
6) $\gamma(\mathbf{x},\mathbf{y}_{\mathrm{k}}) \cong 0$, if $\mathbf{x}$ is far away from $\mathbf{y}_{\mathrm{k}}$.

- There are several possible forms of γ which can take these constraints into account, referring to the one-dimensional case.

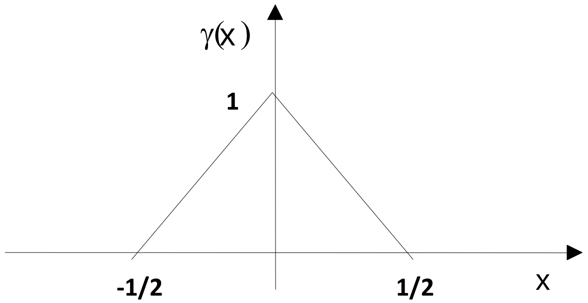
- In particular, it will refer to a $\gamma(\mathbf{z},\mathbf{y})$ function depending on only one variable as argument, $\mathbf{x}$, which is represented by the Euclidean norm of the difference between the two vectors $\mathbf{z}$ and $\mathbf{y}$, i.e., $\mathbf{x}=|\mathbf{z}-\mathbf{y}|$.

Potential function examples

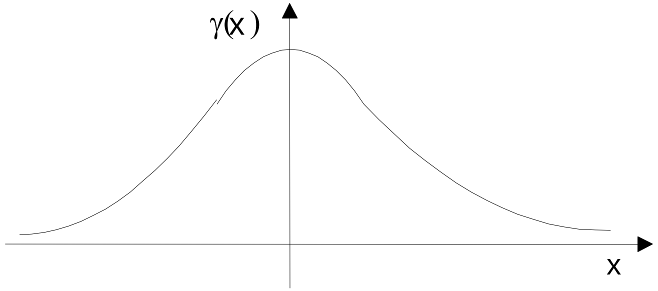
1) $$\gamma(\mathbf{x}) = \begin{cases} 0{,}5 & |\mathbf{x}| \le 1 \\ 0 & |\mathbf{x}| > 1 \end{cases}$$ Rectangle



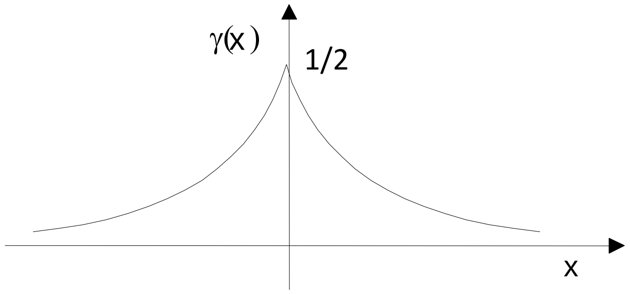
2) $$\gamma(\mathbf{x}) = \begin{cases} 1 - |\mathbf{x}| & |\mathbf{x}| \le 1 \\ 0 & |\mathbf{x}| > 1 \end{cases}$$ Triangle



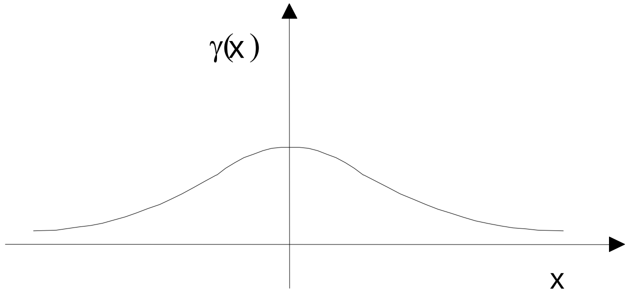
3) $$\gamma(\mathbf{x}) = (2\pi)^{-\frac{1}{2}} e^{-\left(\frac{\mathbf{x}^2}{2}\right)}$$ Gaussian



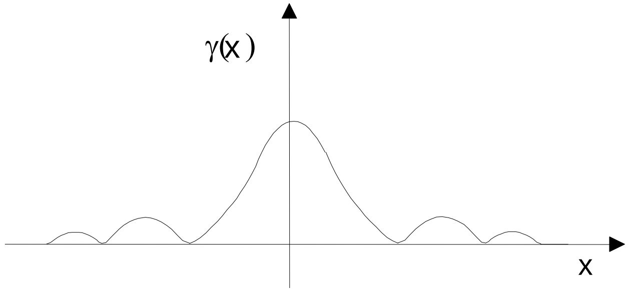
4) $$\gamma(\mathbf{x}) = \frac{1}{2} e^{-|\mathbf{x}|}$$ Decreasing Exponential



5) $$\gamma(\mathbf{x}) = \left[\pi\left(1+\mathbf{x}^2\right)\right]^{-1}$$ Distribution of Cauchy



6) $$\gamma(\mathbf{x}) = (2\pi)^{-1}\left(\frac{\sin\left(\frac{\mathbf{x}}{2}\right)}{\frac{\mathbf{x}}{2}}\right)^2$$ (Squared) Sinc function (sin x/x)^2



- In the last case, γ is not monotonic but it dampens with periodic trend.
- In n-dimensional space, $\mathbf{x}$ obviously becomes a vector.
- Problems:

  1) Selection of potential functions γ.

  2) Degree of overlap of γ's.

Conditional density estimation

Basic idea:

Problem: $p(\mathbf{x})$ estimation

- The probability that a vector $\mathbf{x}$ is in a region $\mathscr{R}$ is:

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

- P is a *smoothed* (or averaged) version of the density $p(\mathbf{x})$, and we can estimate the *smoothed* value of p by estimating the probability P.
- Consider a set of samples (i.i.d.) of cardinality n extracted according to $p(\mathbf{x})$: the probability that k points out of n are in $\mathscr{R}$ is given by the binomial law:

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

And the expected value for k is:

$$\mathrm{E}[k] = nP \qquad (3)$$

- The estimate ML of P $(= \theta)$

  $\max_{\theta}(P_k \mid \theta)$ is given by $\hat{\theta} = \dfrac{k}{n} \cong P$

- Therefore, the ratio k/n will be a good estimate for the probability P and hence for the smoothed density function p.

- If we assume that $p(\mathbf{x})$ is continuous and that the region $\mathcal{R}$ is so small that p does not vary appreciably within it (so that it can be approximated by a constant), we can write:

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

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

Combining the equations (1) , (3) e (4) we obtain: $p(x) \cong \frac{k/n}{V}$ (5)

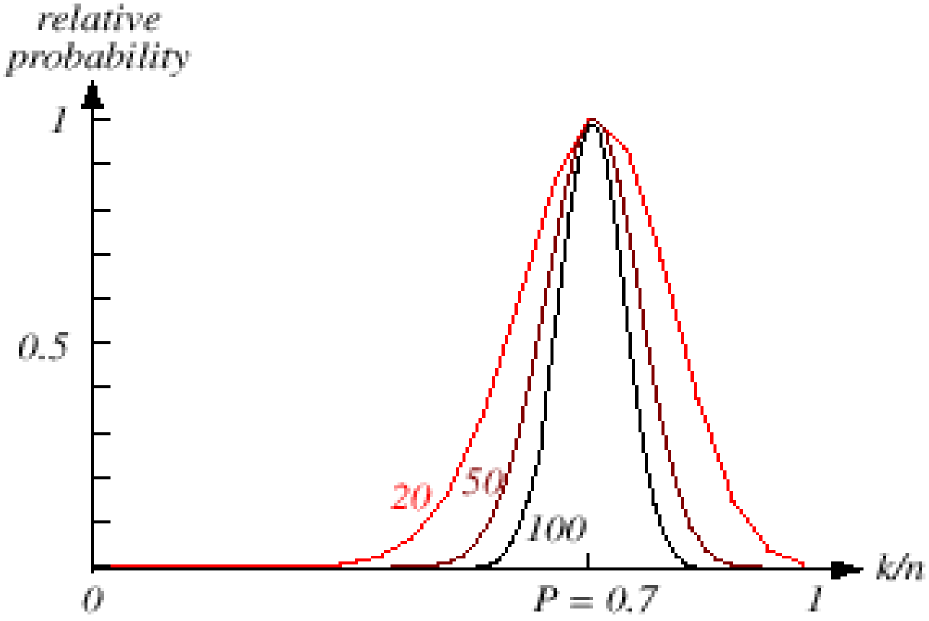


FIGURE 4.1. The relative probability an estimate given by Eq. 4 will yield a particular value for the probability density, here where the true probability was chosen to be 0.7. Each curve is labeled by the total number of patterns *n* sampled, and is scaled to give the same maximum (at the true probability).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

Conditions for the convergence

The fraction $k/(nV)$ is a value averaged in the region of $p(\mathbf{x})$.
True $p(\mathbf{x})$ is only obtained if V becomes small arbitrarily to zero.

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

This is the case where there are no samples included in $\mathcal{R}$

→ not interesting!

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

In this case, the estimate diverges → not interesting!

- It is necessary that V approaches to 0 in each case if we want to use the estimation
  - In practice, V cannot become small at will as the number of samples is always limited and a certain variance in the k/n ratio and a certain approximation (mediated value) of $p(\mathbf{x})$ must be accepted
  - Theoretically, if the number of samples is unlimited, we can overcome the problem.

- To estimate the density of $\mathbf{x}$, we form a sequence of regions:
  - $\mathcal{R}_1, \mathcal{R}_2, \ldots$ containing $\mathbf{x}$: the first region with one sample, the second with two, 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 $p_n(\mathbf{x})$ the n-th estimate for $p(\mathbf{x})$, then:

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

- $p_n(\mathbf{x})$ converges to $p(\mathbf{x})$ if three conditions occur:

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

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

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

- There are two common ways of obtaining sequences of regions that satisfy these conditions

(a) Take an initial region and shrink it by specifying the volume V_n as some function of n, such as $V_n = 1/\sqrt{n}$ and verifying that:

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

We obtain **Parzen-window method**

(b) Define k_n as a certain function of n, e.g., $k_n = \sqrt{n}$; and increase the volume V_n until it encloses k_n neighbors of $\mathbf{x}$.

We obtain **k-NN, k nearest neighbor**

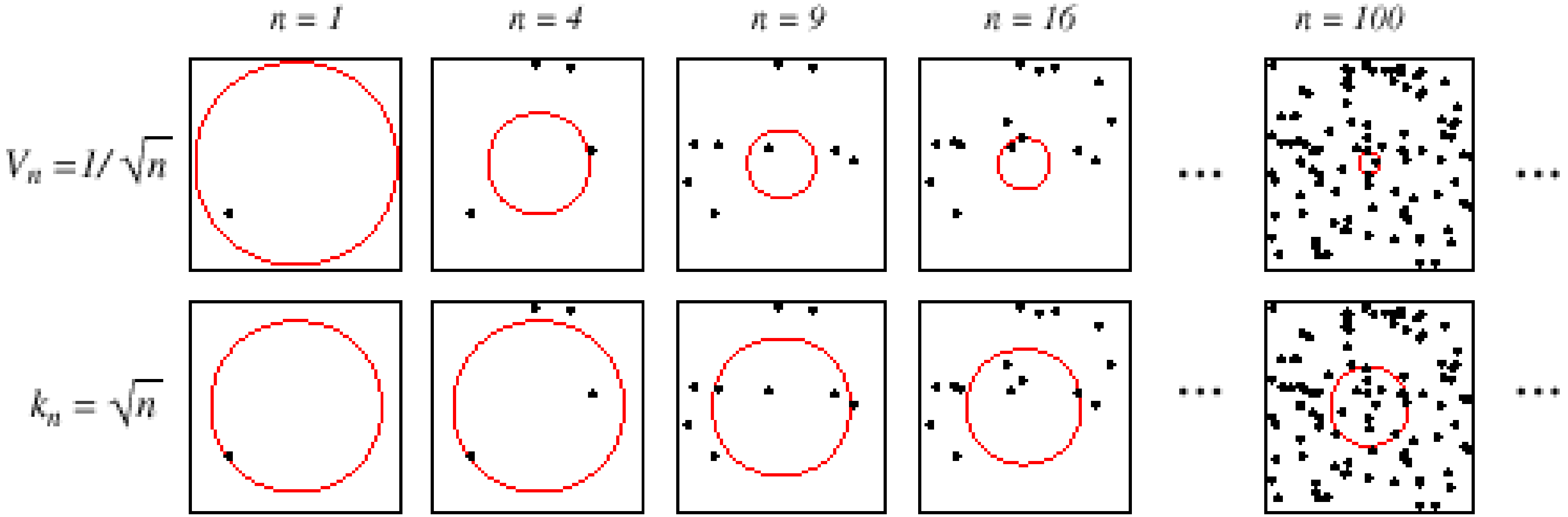
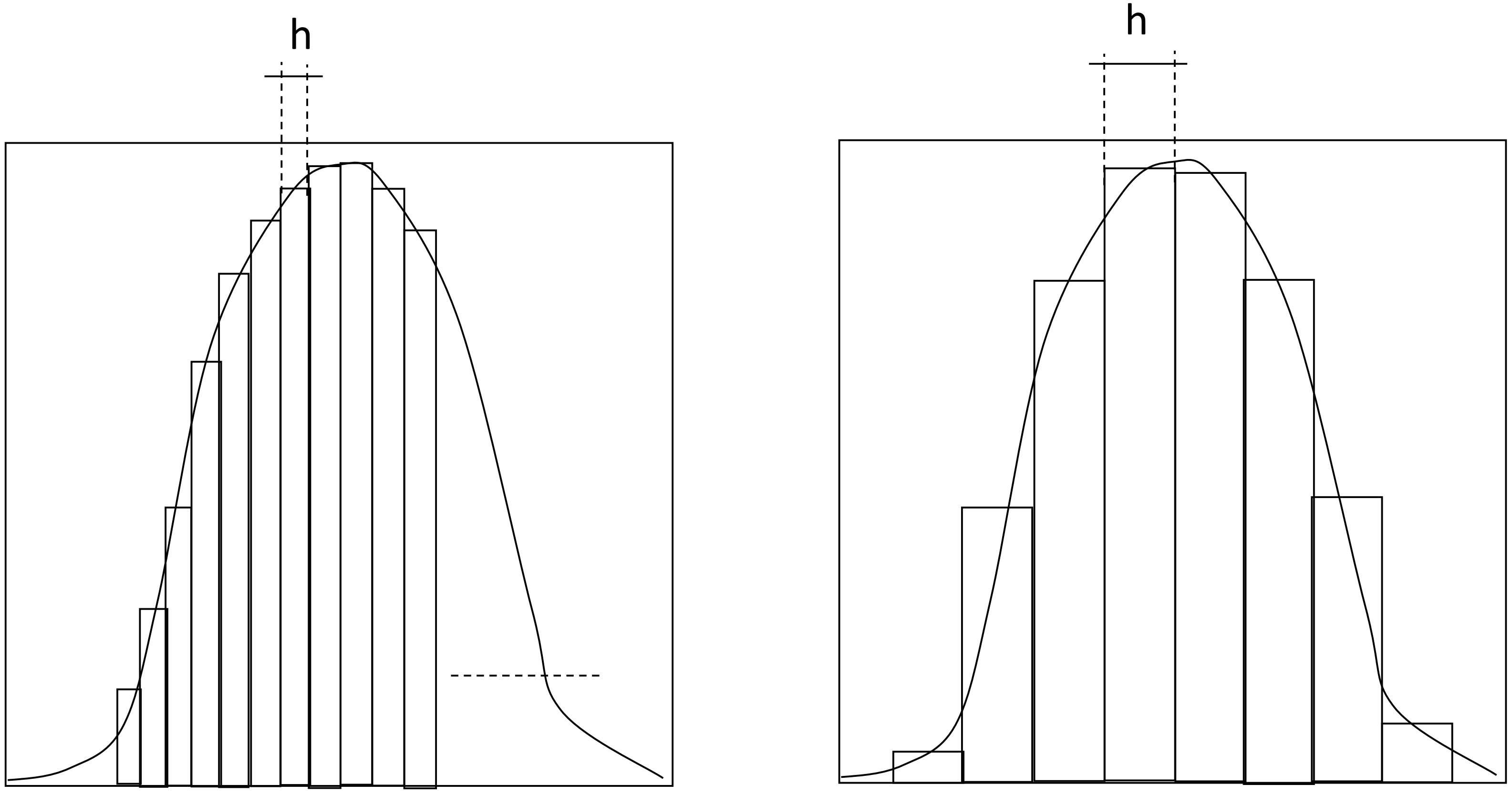


FIGURE 4.2. There are two leading methods for estimating the density at a point, here at the center of each square. The one shown in the top row is to start with a large volume centered on the test point and shrink it according to a function such as $V_n = 1/\sqrt{n}$. The other method, shown in the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

- In essence, viewed more intuitively, if we consider a one-dimensional case and have a set of x points taken from a $p(x)$ distribution, the easiest way to approximate it is by means of a histogram:



The probability that a sample x is in a certain bin can be estimated for each bin, so if given n samples and k (k_n) of these ones are in a bin, the relative probability can be estimated from the frequency ratio $P \cong k/n$

... that converges to the true P if $n \to \infty$ and, assuming the value of the pdf constant over the bin, this can be approximated as

$$\widehat{p}(x) \equiv \widehat{p}(\widehat{x}) \approx \frac{1}{h}\frac{k_n}{n}, \qquad |x - \widehat{x}| \leq \frac{h}{2}$$

where $\widehat{x}$ is the average value of the bin while assuming $p(x)$ continuous e h sufficiently small.

Parzen Windows

- If the region $\mathscr{R}$ is a small hypercube centred on $\mathbf{x}$ and we want to obtain the number k of samples falling in it, we can define the following window function:

$$\gamma(\mathbf{u}) = \begin{cases} 1, & |u_i| < 1/2 \\ 0, & altrimenti \end{cases} \qquad i = 1,.., D$$

  that defines a unit hypercube centered in the origin.

- γ(**u**) is an example of *kernel* function, in this context called Parzen window.
- Therefore:

$$k = \sum_{j=1}^{N} \gamma\left(\frac{\mathbf{x} - \mathbf{y}_j}{h}\right)$$

- When we substitute this into (5) we obtain the estimate

$$p(x) = \frac{1}{N} \sum_{j=1}^{N} \frac{1}{h^D} \gamma\left(\frac{\mathbf{x} - \mathbf{y}_j}{h}\right), \quad \text{where} \quad V = h^D$$

- As in the case of the histogram, there are problems due to the presence of discontinuities between the hypercubes, but this can be remedied by using a **smoother** kernel function
- The Parzen windows’ method can be seen as an instance of the more general method of potential functions.
- Parzen windows help to estimate probability density as follows (one-dimensional case, for simplicity):

$$\widehat{p}_i(x) = \frac{1}{N_i} \sum_{j=1}^{N_i} \frac{1}{h} \gamma\left(\frac{x - y_j}{h} \right) \qquad \text{where } N_i = \#\text{ samples} \in \omega_i$$

where h is the Parzen window’s dimension and γ can be one of the potential functions seen above: h in practice regulates the overlap between the γ.

- The choice of h has a relevant effect on the found estimate:
  - if h is too big, the resulting estimate will be characterised by low resolution,
  - if h is too small, there will be of a large statistical variability (samples will just slightly interact).

- Given the previous equation, the larger N_i , the better the resulting estimate.
- So, we can also write:

$$\widehat{p}_i(x) = \lim_{N_i \to \infty} \left\{ \frac{1}{N_i} \sum_{j=1}^{N_i} \frac{1}{h} \gamma\left(\frac{x - y_j}{h} \right) \right\}$$

where y_j are the samples of the class ω_i.

Example

- Let γ be of Gaussian form, that is:

$$\gamma(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} \quad \text{and} \quad \hat{p}(x|\omega_k) = \frac{1}{N_k} \sum_{j=1}^{N_k} \frac{1}{h_k} \gamma\left(\frac{x - y_j}{h_k} \right)$$

- Suppose the input data used in the example is distributed as a Gaussian with N_k number of *training* samples for class ω_k.

- We try to see whether, using this data and the Parzen windows' method and the γ function introduced above, we are able to approximate the Gaussian probability density function that originated the data itself.

- The h_{k} value used in the estimate can be considered at first dependent on the type of γ function used and the number of samples.

- A practical rule is introduced to link these measures, whereby: $h_k = \dfrac{h_1}{N_k}$

- To see the effect of a different selection of parameters, we can try different values of h_1 and N_k, and see how the estimation of the conditional probability function varies with the number of samples and h.

- In particular, if we use values of h_1 too small (or large N_k), we can still distinguish the effects of individual samples in the estimation.

- Hence, we are in the case of large statistical variability, that is, too low interpolating effect.

- As h_k grows we have wider windows, that is, stronger overlap.

- To choose the value of h_1 to be used, the results in terms of $p(x)$ are evaluated:
  - an excessive presence of peaks in the probability density $\hat{p}(\mathbf{x})$ should be avoided, as well as a too high level of *smoothing* (i.e., almost constant curve or "very filtered").

- In the end, the following conditions are needed to apply this method:

  a) a large number of samples;

  b) verify the choice of γ and h;

  c) store all samples to have $\hat{p}(\mathbf{x})$.

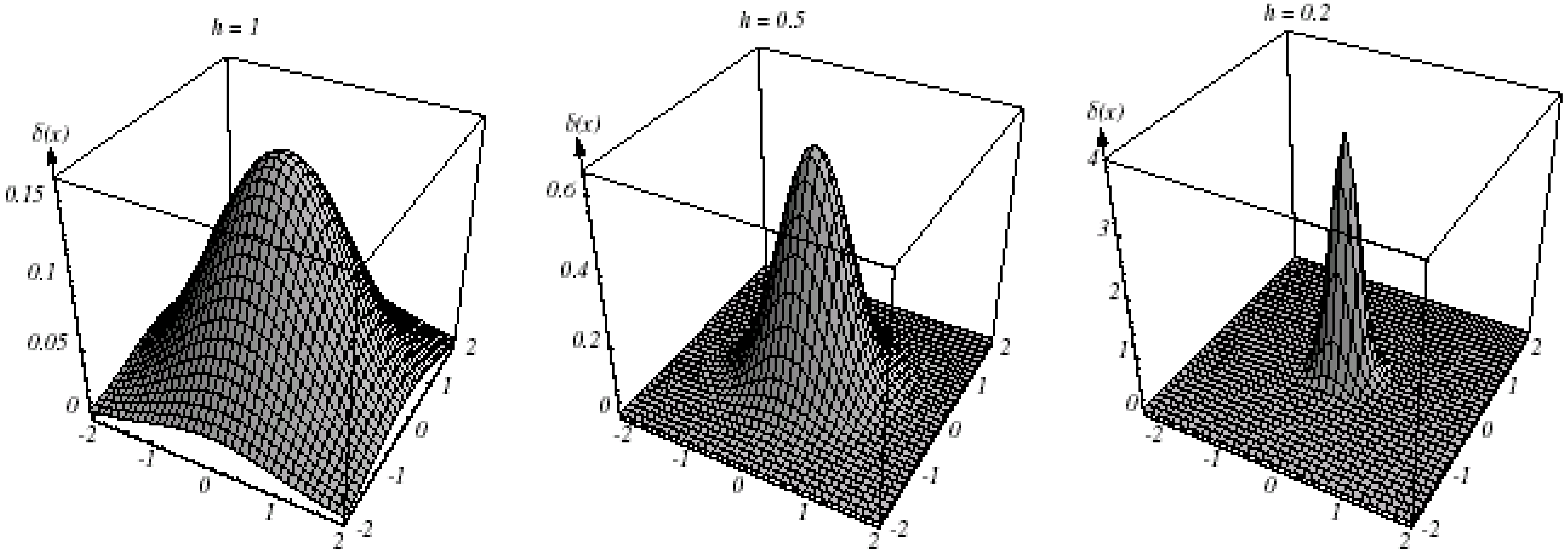


FIGURE 4.3. Examples of two-dimensional circularly symmetric normal Parzen windows for three different values of h. Note that because the $\delta(\mathbf{x})$ are normalized, different vertical scales must be used to show their structure. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

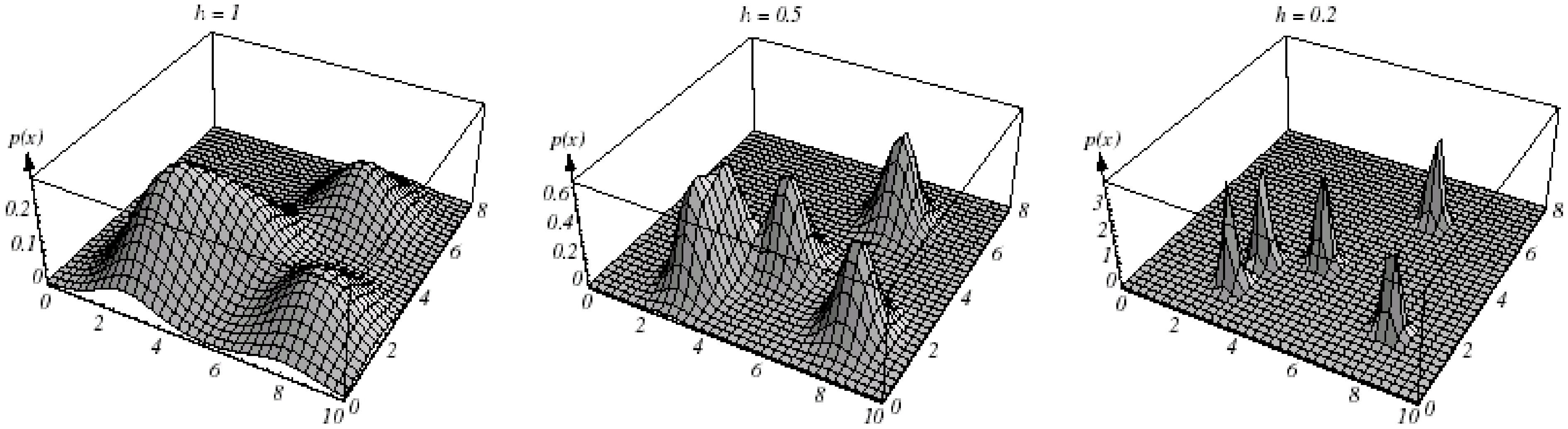


FIGURE 4.4. Three Parzen-window density estimates based on the same set of five samples, using the window functions in Fig. 4.3. As before, the vertical axes have been scaled to show the structure of each distribution. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Parzen windows with Gaussian kernel

- Parzen windows method in case of potential Gaussian function γ (also called Specht functions).
- In particular we write the conditional probability as:

$$\hat{p}_i(\mathbf{x}) = \frac{1}{\left(2\pi\sigma^2\right)^{\frac{n}{2}} N_i} \sum_{j=1}^{N_i} \exp\left\{ - \frac{\left(\mathbf{x}-\mathbf{y}_j\right)^{\mathrm{t}}\left(\mathbf{x}-\mathbf{y}_j\right)}{\left(2\sigma^2\right)} \right\} \qquad (1)$$

- σ is the so-called *smoothing* parameter.
  - for $\sigma = 0$, the probability becomes expressed by a sum of Dirac pulses centered on each $\mathbf{y}_j$;
  - per $\sigma = \infty$, $\hat{p}(\mathbf{x})$ becomes constant.

- Since the smoothing parameter σ is considered uniform on all directions, it is essential to make a normalization on the *features* before estimating the probability, that is to change the *features* so that $\sigma_1 = \sigma_2 = ... = \sigma_n = \sigma$.
- It is also necessary to choose σ so that there is no excessive overlap between the potential functions centered on the different samples.
- For this purpose, it is suggested to estimate σ by taking the L points closest to the generic $\mathbf{y}_i$ sample and to calculate the average of the distances.

$$\sigma = \frac{1}{L}\sum_{j=1}^{L}\left\|\mathbf{y}_j - \mathbf{y}_i\right\| = \frac{1}{L}\sum_{j=1}^{L}\sqrt{\left(\mathbf{y}_j - \mathbf{y}_i\right)^2}$$

- Usually you choose heuristically $L \cong 0.05N$.
- The choice of L changes the degree of overlap between the functions.
- To find the discriminating function in the case of Specht functions, we can first approximate the Taylor series exponential around the mean value (zero):

$$e^{x} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots$$

- Substituting in (1) we have:

$$\hat{p}_i(\mathbf{x}) = \frac{1}{\left(2\pi\sigma^2\right)^{\frac{n}{2}} N_i} \mathrm{e}^{-\frac{\mathbf{x}^{\mathrm{t}}\mathbf{x}}{\left(2\sigma^2\right)}} \sum_{j=1}^{N_i} \mathrm{e}^{\frac{\mathbf{x}^{\mathrm{t}}\mathbf{y}_j}{\sigma^2}} \underbrace{\mathrm{e}^{-\frac{\left\|\mathbf{y}_j\right\|^2}{\left(2\sigma^2\right)}}}_{\text{noto}} \quad \cong$$

$$\cong \frac{1}{\left(2\pi\sigma^2\right)^{\frac{n}{2}} N_i} \mathrm{e}^{-\frac{\mathbf{x}^{\mathrm{T}}\mathbf{x}}{\left(2\sigma^2\right)}} \sum_{j=1}^{N_i} \mathrm{e}^{-c_j} \sum_{h=0}^{r} \left(\mathbf{x}^{\mathrm{t}}\mathbf{y}_j\right)^{h} \frac{1}{\sigma^{2h} h!}$$

- As we can see, the term c_j is known:

$$c_j = \frac{\left\| y_j \right\|^2}{2\sigma^2}$$

- The above expression applies for any class ω_i , therefore, we can write the discriminating function of class ω_k, using Bayes, such as:

$$g_k(\mathbf{x}) = \widehat{p}\left(x \middle| \omega_k\right)\widehat{p}(\omega_k) \cong \frac{\widehat{p}(\omega_k)}{N_k} \sum_{j=1}^{N_k} \mathrm{e}^{-c_j^k} \sum_{h=0}^{r} \left(\mathbf{x}^{\mathrm{t}} \mathbf{y}_j^k\right)^h \frac{1}{\sigma^{2h} h!}$$

- This expression can be obtained by simplifying the common terms of the type

$$\exp\left\{-\frac{\mathbf{x}^{\mathbf{t}}\mathbf{x}}{\left(2\sigma^2\right)}\right\}$$

- When σ is large enough, it's sufficient an approximation with r small, and the discriminating function becomes linear, at the limit for $\sigma \rightarrow \infty$ (Prototype Method).

- In the case of $\sigma \rightarrow 0$, an approximation with large r is required and the discriminating function becomes a sum of Dirac delta (as can be seen by making the limit for $\sigma \rightarrow 0$ in equation (1)).

- The method becomes substantially equivalent to the classification method of k Nearest Neighbors (k-points neighbors, k-NN) (see below).

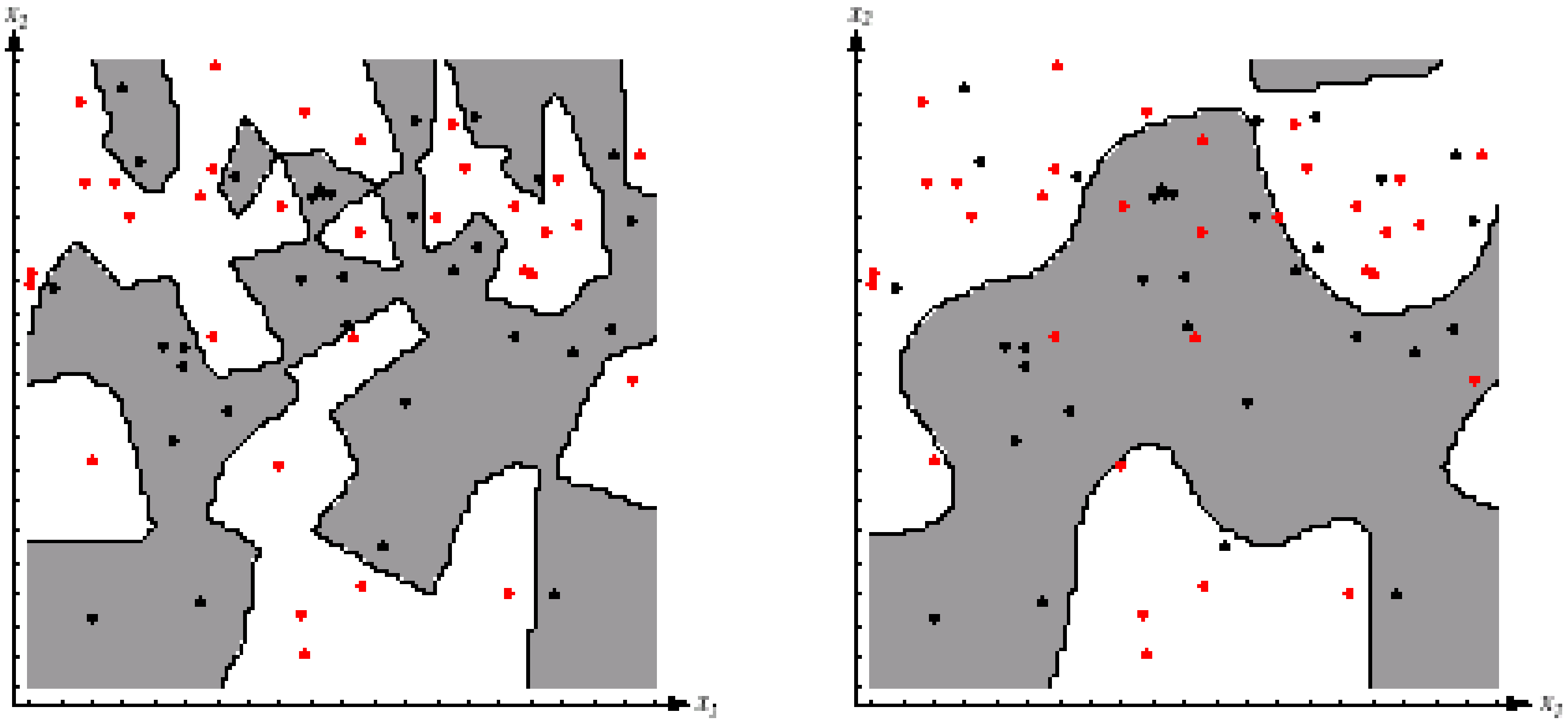


FIGURE 4.8. The decision boundaries in a two-dimensional Parzen-window dichotomizer depend on the window width h. At the left a small h leads to boundaries that are more complicated than for large h on same data set, shown at the right. Apparently, for these data a small h would be appropriate for the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

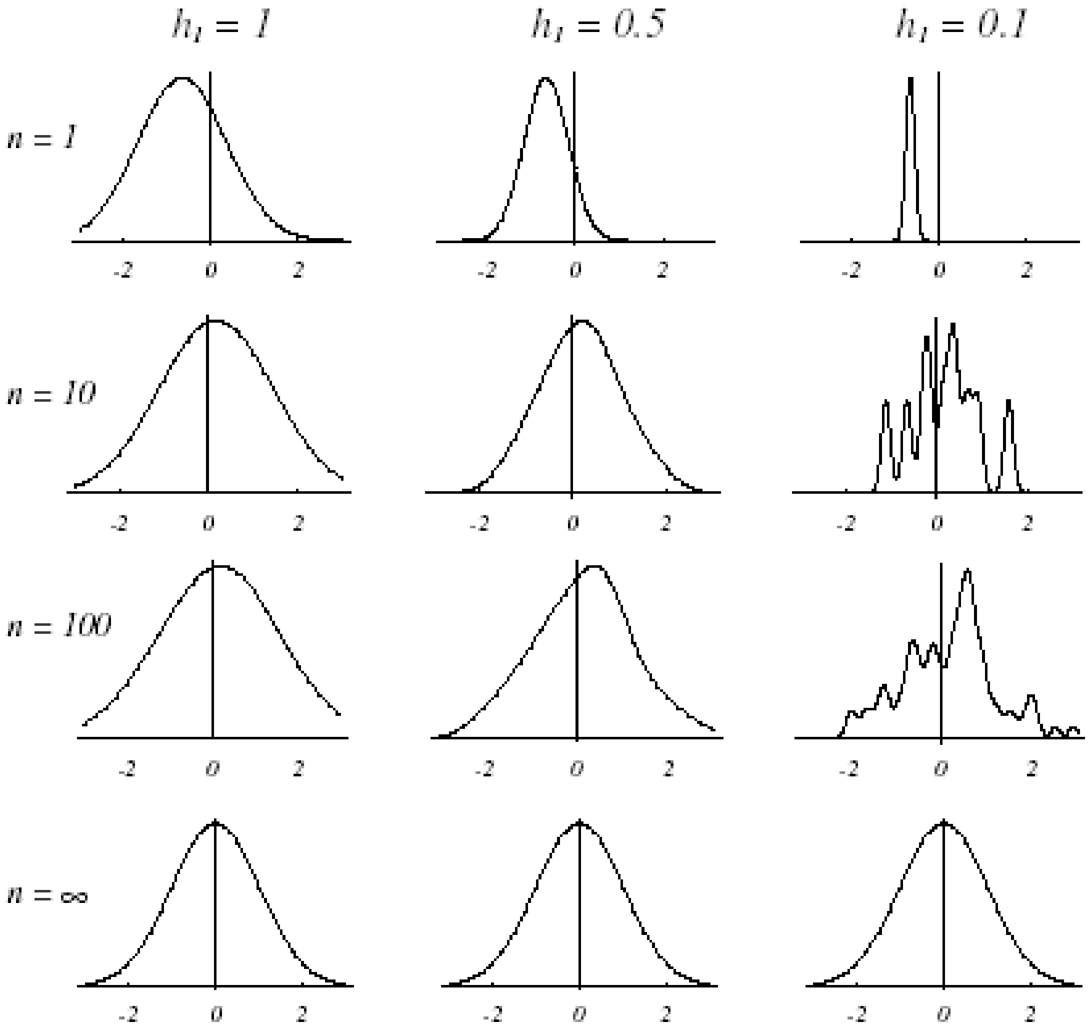
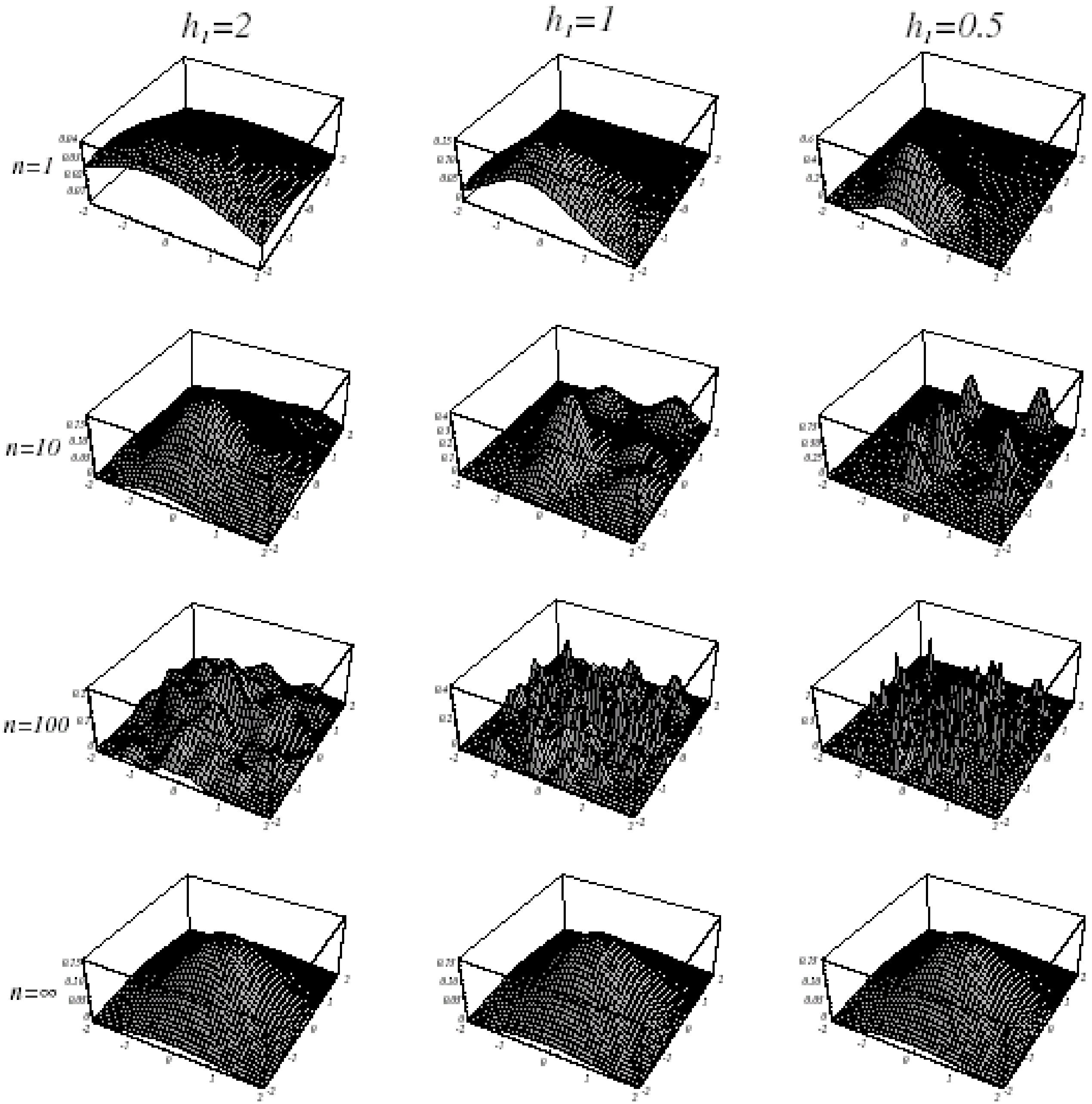
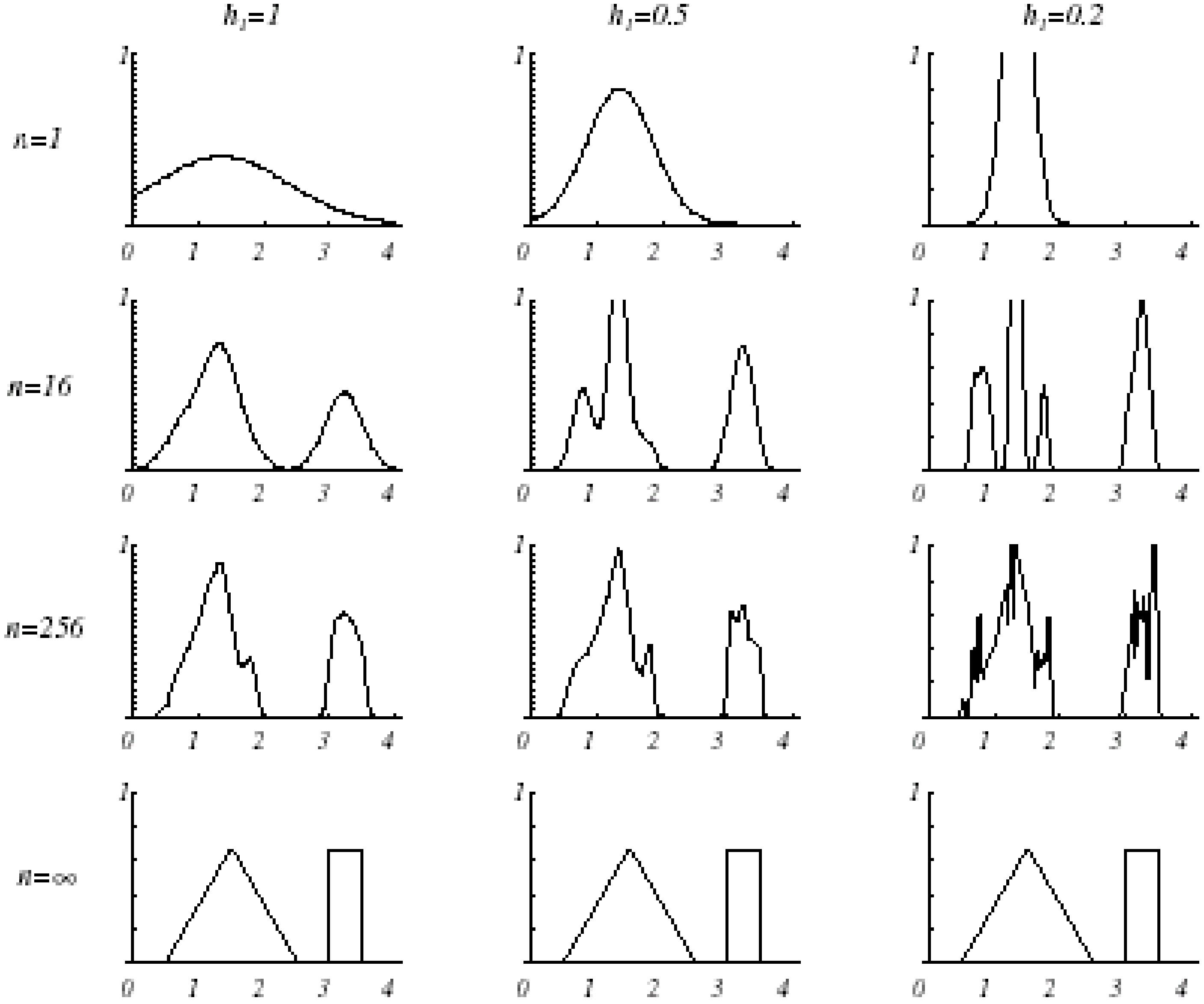


FIGURE 4.5. Parzen-window estimates of a univariate normal density using different window widths and numbers of samples. The vertical axes have been scaled to best show the structure in each graph. Note particularly that the $n = \infty$ estimates are the same (and match the true density function), regardless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


$h_1=2$
$h_1=1$
$h_1=0.5$
$n=1$
$n=10$
$n=100$
$n=\infty$

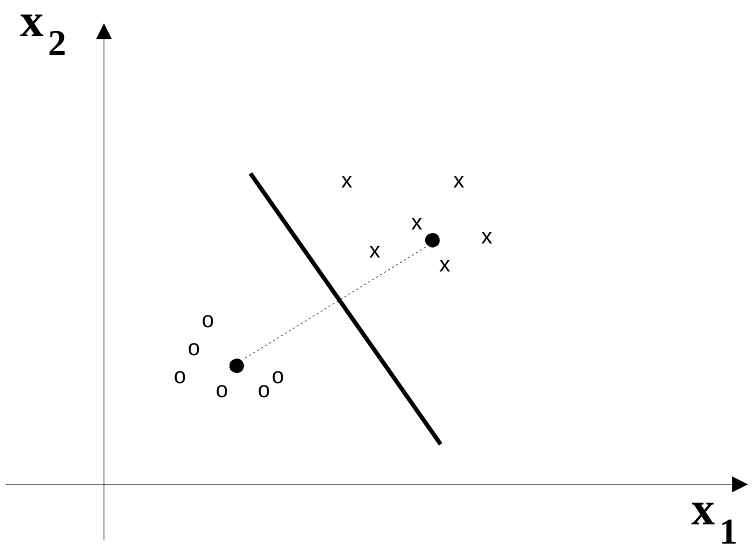


$h_1=1$
$h_1=0.5$
$h_1=0.2$
$n=1$
$n=16$
$n=256$
$n=\infty$
1
0
1
2
3
4

Method of prototypes (or archetypes)

- The method of prototypes is typically applied to find the decision regions of a classifier based on Minimum Distance, MDC. They generally belong to the class of the so-called simplified classification methods.
- The distances used can be several.
- Typically, it is supposed that the samples of a class tend to concentrate tightly around a pattern representative of the class itself.
- This situation is typical of cases where the variability of the pattern or the noise/disturbances in the observation phase are regular and therefore can be well modelled.
- Following this analogy, we can say that the prototyping method is fine when the patterns to be recognized are an alteration of a reality that we know deterministically.
- In this case, the minimum distance classifiers are extremely efficient.

- For example, we consider all the samples of the class as described by the centre of gravity, $\mathbf{m}$, of the class itself.



- Let $D(\mathbf{x},\mathbf{m}_i)$ be a metric in the feature space.
- The decision rule applied by an MDC is to choose the class

  ω_k if $D(\mathbf{x},\mathbf{m}_k) = \min D(\mathbf{x},\mathbf{m}_j)$, for $j = 1,..,M$ and $j \neq k$

- The method for determining discriminating functions according to the prototype method is divided into three steps.

    1) The centre of gravity $\mathbf{m}_i$ of the i-th class is determined.

    2) The straight line joining the barycentres of two classes is determined, characterized by the equation $r_{ij}(\mathbf{x})$.

    3) Calculate the hyperplane $g_{ij}^*(\mathbf{x})$ perpendicular to $r_{ij}(\mathbf{x})$.

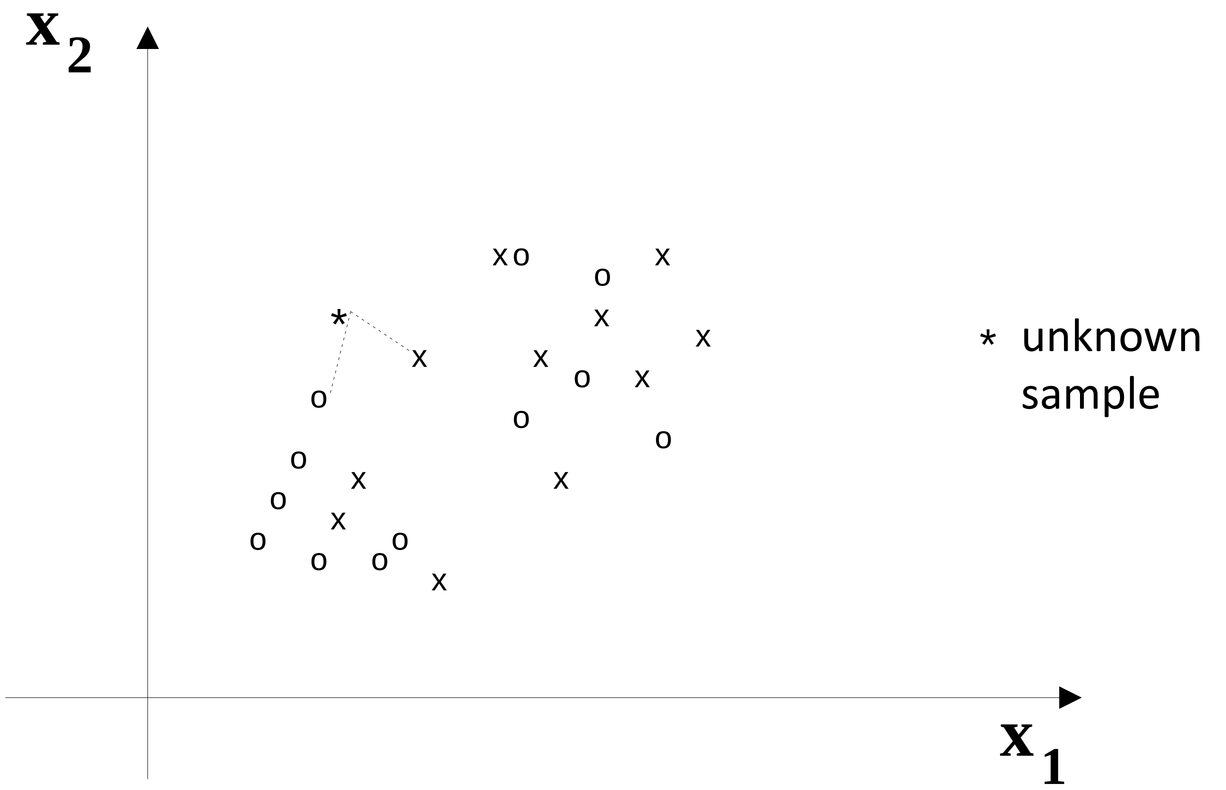
    4) The crossing point of this hyperplane in the feature space is determined on the basis of the different probabilities of the two classes. The more likely a class is or has a greater importance, the more the estimated discriminating function will be far from the barycentre of the class itself.

    5) Points 1 to 4 are repeated for all class pairs.

- In this way, $f_{ij}(\mathbf{x})$ discriminating functions are determined.
- This method identifies linear discriminating functions: the prototyping method is therefore a particular case of a linear classifier.
- **Advantages**
  - Simplicity
  - Low memory occupancy
- **Disadvantages**
  - Assumption that an articulated reality can be represented by only one sample (prototype or archetype of the class), as such sample may not even exist in physical reality.
- The lowest distance classifier classifies an observation on the basis of the nearest distance (that is, the most probable "match") between the observation and the prototypes of the classes.
- This approach is equivalent to the *template matching* method (correlation of a model with data).

k nearest-neighbors, k-NN

- The method of the *k nearest-neighbors* is derived from the prototypes' method, in the sense that it applies the same rule of the MDC classifier taking as reference not the center of gravity of all classes, but a set of variable points for a subset of classes among those found in the training phase.



- It can therefore be seen as a method applicable to the case where *each class is described by an enlarged set of prototypes.*
- Given an unknown point x^*, we consider $s_i \in \{s_1,\ldots,s_N\} = S$ the nearest point (NN) of the point x^* if

$$d(x^*, s_i) = \min_l d(x^*, s_l), \quad l = 1\ldots N$$

where $d(.)$ is a distance measure defined on the feature space

- If you interpret each point of set S as a prototype of a class, you can then see the equation above as a rule of classification 1-NN associating the sample x to the class j to which the point s_i belongs.

- Rule 1-NN can be generalized to define a k-NN rule, which consists of determining the k points belonging to the set S closest to the observation x.
- The classification rule k-NN associates observation x with class i having the largest number of elements among the nearest k.
- We call $\mathrm{U}(x)$ the set of k points closer to x.
- For example, with odd k and two classes, ω_1 and ω_2, the decision rule can choose the class with the most samples in $\mathrm{U}(x)$.
- The alternative classification rule to the majority rule can be more complicated.

- I decide to uniquely characterize the subsets of prototypes belonging to the different classes between k neirest neighbors, for example by calculating their center of gravity.
- I can then measure the distance of the point x from these points to decide the class.
- For $k = 1$ the set $\mathrm{U}(x)$ is given by the sample closest to the unknown sample.
- For large k, i.e. $k \cong N$, the method becomes equivalent to the prototypes' method, because the measured centre of gravity corresponds to that of the class and we can use an MDC classifier.

- To use this method we have to take into account the following considerations:
  - the metric must be "good", that is, discriminating
  - only a small amount of information about the feature space is used, that is, only k points of that space
  - it is necessary to store all samples
  - therefore, the method is more appropriate whenever the number of samples N is low
  - you must always apply the normalization of the features during training
  - the decision surfaces created are <u>non-linear</u>
  - a typical choice is $k \simeq \sqrt{N}$
  - the test set must be extensive and possibly with few errors.

How *k-NN* was born: pdf estimation

- The archetype method and its generalization to k-NN can be considered as a "simplified" classification method
- But it can also be used as a method of estimating $p(\mathbf{x})$, and then we can interpret it in its classical meaning as a traditional classification method
- Instead of fixing V as in the case of kernels, we fix the value of k and enlarge the radius of the sphere to include k samples
- Let’s see the a-priori probability:

given a class ω_j, the frequency of occurrence of samples N_j of class j is generally simply measured in relation to the total number of samples N, i.e.,

$$P(\omega_j) \equiv \widehat{p}(\omega_j) = \frac{N_j}{N}$$

Density estimation and Nearest-Neighbor rule

- 2 classes, ω_i and ω_j, containing N_i and N_j samples, $N = N_i + N_j$
- The local density estimation for ω_i is calculated as (and similarly for ω_j)

$$\hat{p}(x|\omega_i) = \frac{1}{V}\frac{k_i}{N_i}$$

i.e., ratio of k_i (k_j) points to the total N_i (N_j) belonging to the class ω_i (ω_j) contained in the volume V

- The Bayes rule says $p(x|\omega_i)P(\omega_i) > p(x|\omega_j)P(\omega_j)$, then

$$\hat{p}(x|\omega_i)\hat{p}(\omega_i) > \hat{p}(x|\omega_j)\hat{p}(\omega_j)$$

$$\Rightarrow \frac{1}{V}\frac{k_i}{N_i}\frac{N_i}{N} > \frac{1}{V}\frac{k_j}{N_j}\frac{N_j}{N} \quad \Rightarrow \quad k_i > k_j$$

- If there are undetermined cases (ties) there are some alternatives:
  - arbitrary choice
  - assign x to the class (among those "tie") that has the nearest average sample (calculated between the k_i samples)
  - assign x to the class that has the k_i samples with least distance from it
  - etc.

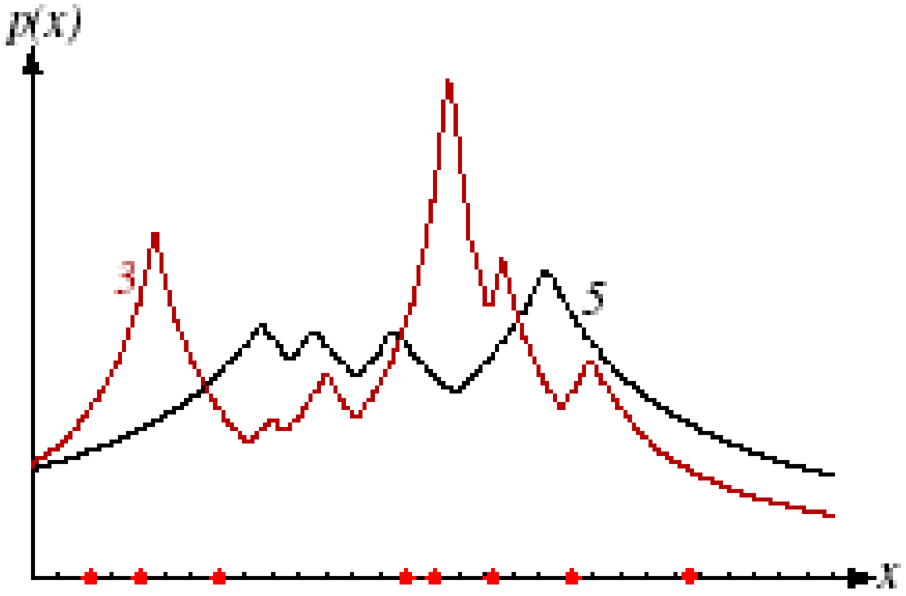


FIGURE 4.10. Eight points in one dimension and the *k*-nearest-neighbor density estimates, for $k = 3$ and 5. Note especially that the discontinuities in the slopes in the estimates generally lie *away* from the positions of the prototype points. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

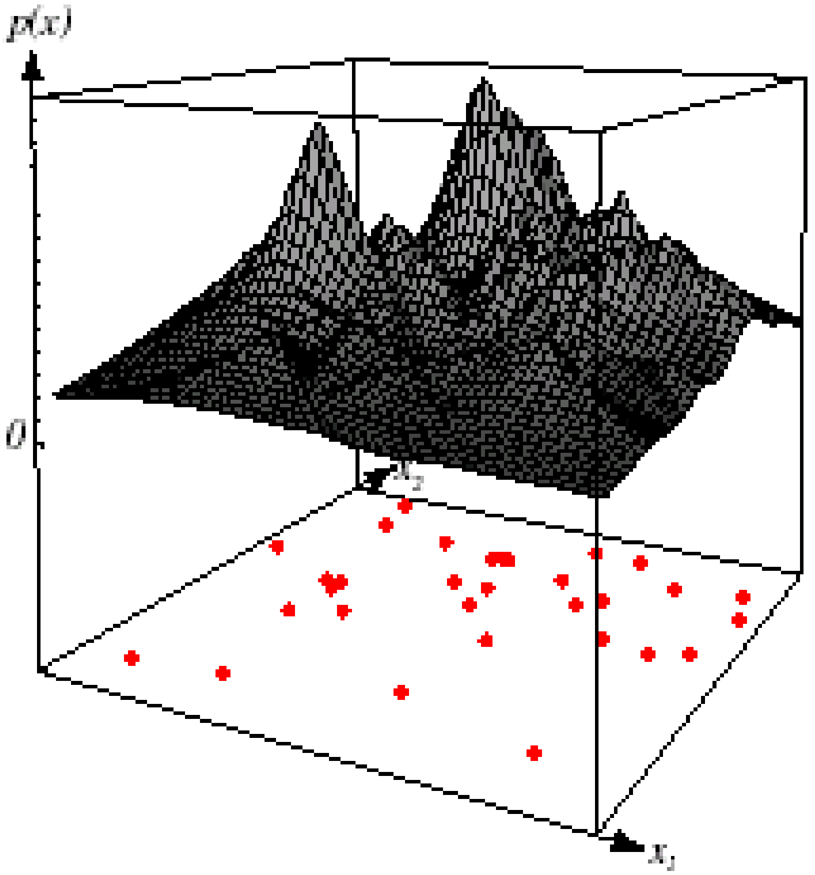


FIGURE 4.11. The k-nearest-neighbor estimate of a two-dimensional density for $k = 5$. Notice how such a finite n estimate can be quite "jagged," and notice that discontinuities in the slopes generally occur along lines away from the positions of the points themselves. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.