Chapter 4

Nonparametric Decision Making

4.1 Introduction

Chapter 3 assumed that the type of the density function, such as the normal or Poisson, was known for each class. Only the parameters of the densities, such as their means or variances, had to be estimated from the data before using them to estimate probabilities of class membership or to make classification decisions. This situation is referred to as parametric decision making.

In most real problems, even the types of the density functions of interest are unknown. Looking at histograms, scatterplots or tables of the data, or the application of statistical procedures may suggest that a particular type of class density may be used, or they may indicate that the data are not well fit by any of the standard types of densities or distributions. In this case, **nonparametric techniques** are needed. In this chapter, techniques for fitting an arbitrary density to a set of samples will be described. These are the histogram technique, discussed in the next section, and the kernel or window technique, discussed in Section 4.3. Other techniques, which classify samples without the explicit use of density functions, will also be described. These include several nearest neighbor techniques and some methods for obtaining discriminant functions directly from the data.

Another technique, which we will not discuss in detail, involves forming the sample cumulative distribution function, fitting a curve to it, and then taking the derivative to get a function, which after normalization to unit area is used as an estimated density function.

4.2 Histograms

One of the easiest ways of obtaining an approximate density function $\hat{p}(x)$ from sampled data if no parametric form is assumed for the underlying density is to form a **histogram** of the data such as the three shown in Figure 4.1. To form a histogram, the range of the feature variable x is divided into a finite number of adjacent **intervals** that include all of the data. These intervals are also called **cells** or **bins**. The number or fraction of samples falling within each interval is then plotted as a function of x as a bar graph. If a sample falls directly on a boundary between intervals, by convention, it is put into the interval to its right. The density is assumed to be constant within each interval of x. To use the histogram as an estimate of the true underlying continuous density function, the area under the histogram must equal one. The area under the density in each interval j is equal to the fraction of the total number N of the samples that fell into that interval, n_j/N, so the height of the density equals this area divided by the width of the interval: $\hat{p}_j = n_j/(Nw_j)$. When the approximate density function has been determined, decisions can be made using Bayes' theorem as in Chapter 3. When the feature x is discrete, its range can be divided into intervals and the same technique can be used to fit the distribution by a density function. If there are not too many possible values of x_i, the fraction of the samples that have each value of x_i can be used as an estimate of the discrete distribution $P(x_i)$. The sum of these $\hat{P}(x_i)$ will equal one.

Choosing the number and location of the histogram intervals is an art; no definitive theoretical guide for this choice is available. Figures 4.1b, 4.1c, and 4.1d show possible choices for histograms to describe 50 random numbers that were chosen from the normal density shown in Figure 4.1a. If a small number of wide intervals is used such as in Figure 4.1c, the number of samples falling within each interval will be relatively large, so the height of the rectangle and thus the area within the interval can be estimated quite accurately. However, the resulting approximate density will be flat over large regions and any fine structure (narrow fluctuations) in the true distribution will tend to be lost. Using a relatively large number of histogram intervals can preserve the fine re of the true density, but when too many intervals are used as in Figure 4.1 nce in their heights decreases. At first glance, it may appear to show ne structure in the data; however, most of the apparent structure ew samples, and thus cannot be very significant. People tend to he data even when the "structure" is due to random fluctuation, st "overfitting" the data, which degrades performance.

e, if the number of intervals were several times the number ervals would contain no samples, and most of the others one sample each. In this case, the histogram reduces to gles, nearly one for each sample point. The histogram of its teeth missing. This would not produce a useful

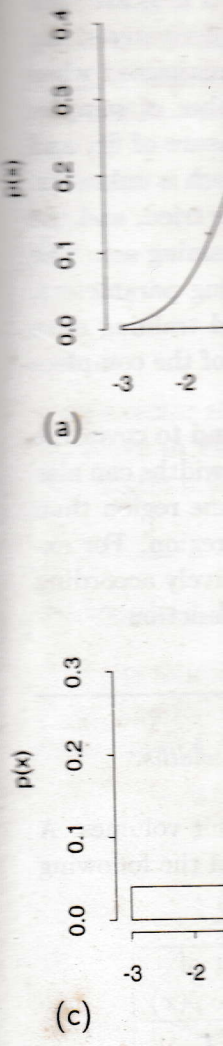

Figure 4.1: (a
(b) A histogr
A histogram
histogram of

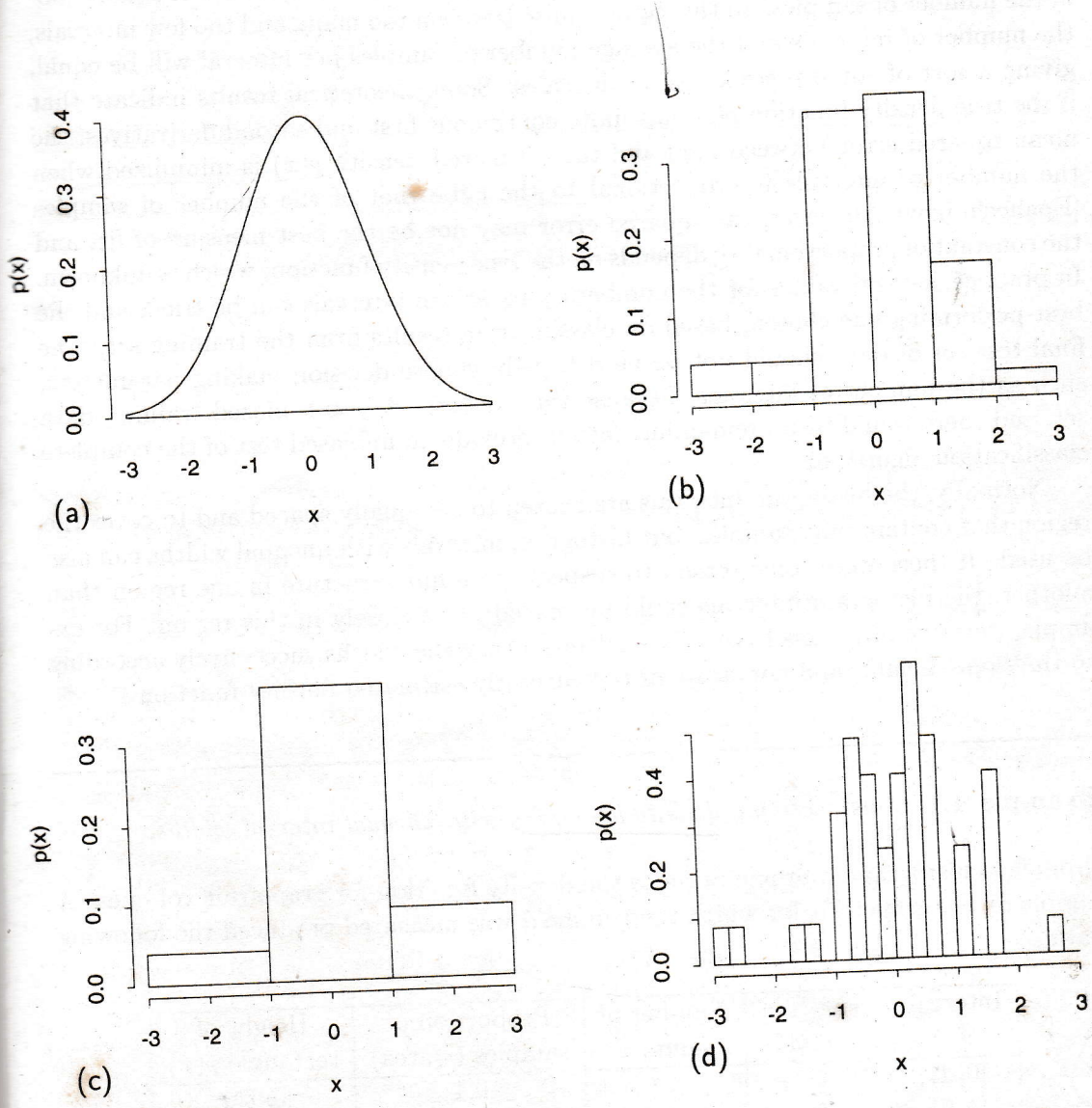

Figure 4.1: (a) The true normal density from which 50 random numbers were chosen. (b) A histogram of 50 normally distributed random numbers with six intervals. (c) A histogram of 50 normally distributed random numbers with three intervals. (d) A histogram of 50 normally distributed random numbers with 24 intervals.

estimate of the probability density function, but would look more like a representation of the actual data itself.

One rule of thumb is to choose the number of intervals to be equal to the square root of the number of samples. In this compromise between too many and too few intervals, the number of intervals and the average number of samples per interval will be equal, giving a sort of equal precision to both scales. Some theoretical results indicate that if the true density function $p(x)$ has finite continuous first and second derivatives, the mean squared error between $p(x)$ and the estimated density $\hat{p}(x)$ is minimized when the number of intervals is proportional to the cube root of the number of samples [Epanechnikov]. However, the squared error may not be the best measure of fit, and the constant of proportionality depends on the true density function, which is unknown. In practice, several values for the number of histogram intervals can be tried, and the best-performing one chosen, based on classification results from the training set. The final test set of data should not be used to help choose decision making parameters, such as the number of intervals, because this converts it into a biased training data set, and there would be no remaining data to provide an unbiased test of the complete classification algorithm.

Normally, the histogram intervals are chosen to be equally spaced and to cover the region that contains any samples, but histogram intervals with unequal widths can also be used. If there were some reason to suspect more fine structure in one region than another, the histogram intervals could be spaced more closely in this region. For example, iterative programs have been written to vary the widths successively according to the slope, height, and curvature of the currently estimated density function.

Example 4.1 *Constructing a density histogram with unequal interval widths.*

Produce a histogram approximation to the density function for grapefruit volumes. A sample of 100 grapefruit for which the volume x was measured produced the following data:

Interval of x	Length of interval	Number of samples	Proportion of samples (=area)	Height of rectangle $\hat{p}(x)$
$[0, 4)$	4	10	0.1	0.025
$[4, 6)$	2	30	0.3	0.150
$[6, 7)$	1	30	0.3	0.300
$[7, 8)$	1	20	0.2	0.200
$[8, 10]$	2	10	0.1	0.050

The height of each rectangle is equal to the fraction of samples falling within its interval divided by the length of the interval (the base of the rectangle). For example, the height

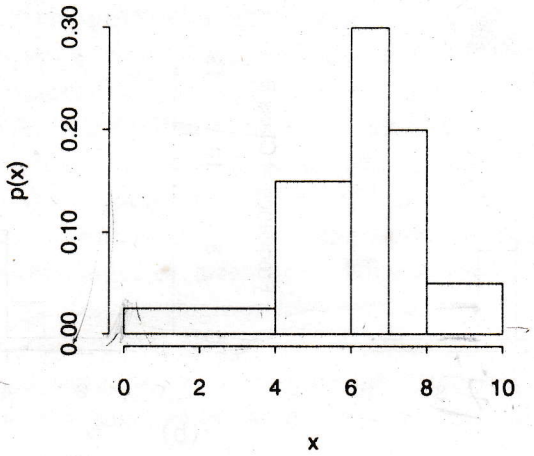

Figure 4.2: The histogram of Example 4.1.

of the rectangle for the interval from 0 to 4 is $0.1/(4-0) = 0.025$. The resulting density is shown in Figure 4.2. Because it is a density histogram, the sum of the areas of the rectangles must equal one.

Example 4.2 *Classification of samples using histograms and Bayes' Theorem.*

Use the following data to classify a sample with $x = 7.5$, given that $P(A) = P(B) = 0.5$. The following data are the values of feature x for 60 randomly chosen samples from class A:

0.80	0.91	0.93	0.95	1.32	1.53	1.57	1.63	1.67	1.74
2.01	2.18	2.27	2.31	2.40	2.61	2.64	2.64	2.67	2.85
2.96	2.97	3.17	3.17	3.38	3.67	3.73	3.83	3.99	4.06
4.10	4.12	4.18	4.20	4.23	4.27	4.27	4.39	4.40	4.46
4.47	4.61	4.64	4.89	4.96	5.12	5.15	5.33	5.33	5.47
5.64	5.85	5.99	6.29	6.42	6.53	6.70	6.78	7.18	7.22

And the following measurements are 60 values of x for some random samples from class B:

3.54	3.88	4.24	4.30	4.30	4.70	4.75	4.97	5.21	5.42
5.60	5.77	5.87	5.94	5.95	6.04	6.05	6.15	6.19	6.21
6.33	6.41	6.43	6.49	6.52	6.58	6.60	6.63	6.65	6.75
6.90	6.92	7.03	7.08	7.18	7.29	7.33	7.41	7.41	7.46
7.61	7.67	7.68	7.68	7.78	7.96	8.03	8.12	8.20	8.22
8.33	8.36	8.44	8.45	8.49	8.75	8.76	9.14	9.20	9.86

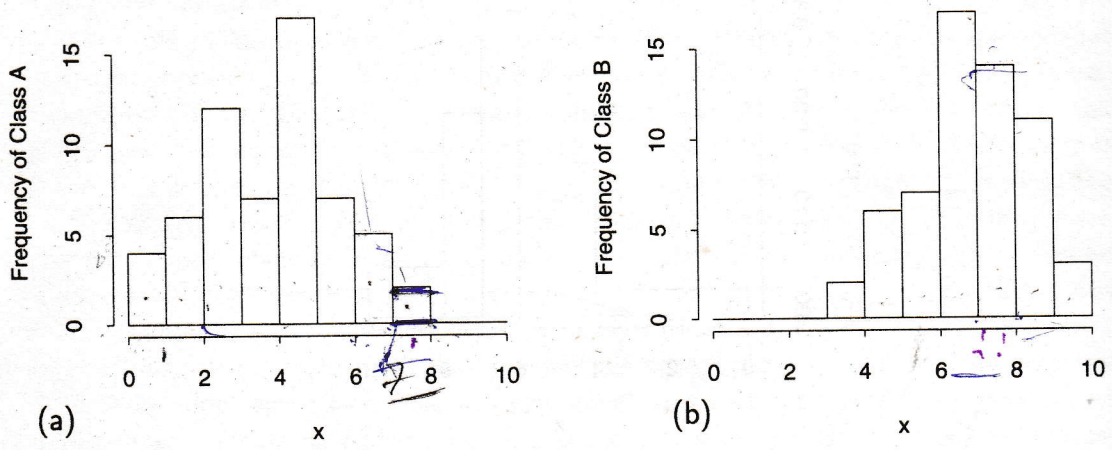

Figure 4.3: Histograms of the feature x for (a) class A and (b) class B.

Figure 4.3 shows histograms of the samples in each unit interval of x for classes A and B. To convert these to density functions, these numbers must be divided by the total number of samples (60) and the interval width (1).

To classify a sample with $x = 7.5$, compare the heights of the two histograms at 7.5. Because the class interval containing 7.5 is $[7, 8]$ for both classes A and B, $\hat{p}(7.5|A) = 2/[60(8-7)]$, and $\hat{p}(7.5|B) = 14/[60(8-7)]$. Using Bayes' Theorem,

$$P(A|7.5) = \frac{p(7.5|A)P(A)}{p(7.5|A)P(A) + p(7.5|B)P(B)}$$
$$= \frac{(2/60)(0.5)}{(2/60)(0.5) + (14/60)(0.5)} = 1/8 = 0.125$$

Also $P(B|7.5) = 1 - P(A|7.5) = 0.875$. Therefore, $P(A|7.5) < P(B|7.5)$, so the sample should be classified into class B.

Histograms are not restricted to one-dimensional densities, but can be used in any number of dimensions. For example, $p(x, y)$ can be approximated by dividing both x and y into intervals, and determining the number of samples that fall within each rectangular histogram bin with dimensions Δx and Δy. The volume under the surface of this two-dimensional histogram is then normalized to equal one, to yield an estimate of the density function $p(x, y)$. The square root rule of thumb can be generalized to produce an **equal precision rule**. When there are two features, the number of Δx intervals, the number of Δy intervals, and the average number of samples per bin are each set equal to the cube root of the total number of samples. When there are n features, the $(n+1)$st root is used.

The histogram technique becomes impractical for spaces of high dimension. For example, even in a five-dimensional space in which each of 5 discrete features could have 10 possible values, or where each of 5 continuous features has been divided into 10 intervals, there would be $10^5 = 100,000$ histogram bins, and several times this number of samples would be required to obtain a reasonable estimate of $p(x_1, \ldots, x_5)$. Sample sets of this size are usually not available. If there were 10 features, each with 100 possible values, there would be $100^{10} = 10^{20}$ bins, which is obviously impractical. If there were 1,000 samples and 10 features, the equal precision rule would recommend $1,000^{1/11} = 1.874$ intervals for each feature, with an average of 1.874 samples per bin. Rounding this to two intervals (high or low) for each feature would give $2^{10} = 1,024$ bins, with an average of $1,000/1,024$ or only about one sample per bin. This may not produce a useful histogram. However, in some cases, many of the bins contain no data, so the average number of samples in the populated bins could be considerably larger.

4.3 Kernel and Window Estimators

The samples themselves can be thought of as a very rough approximation to the true density function, namely a set of spikes or delta functions, one at each sample value, each with a very small width and a very large height, such that the combined area of all the spikes is one. The area of each spike is the number of samples lying at that point divided by the total number of samples. Figure 4.4 shows an example of delta functions approximating the density of the samples located at $x = 1, 2, 4$. This approximation to a continuous density function is not useful in decision making. However, if the delta functions at each sample point are replaced by other functions called **kernels**—such as rectangles, triangles, or normal density functions, which have been scaled so that their combined area equals one—their sum produces a smoother, more satisfactory estimate. A triangular kernel with a base of 3 has been used to produce the estimated density function shown in Figure 4.5a for the same data. Rectangular, triangular, and normal kernels, each scaled to have a standard deviation of 1, are shown in Figure 4.6.

Example 4.3 *Using a triangular kernel.*

Consider the data set with one feature x and three samples at $x = 1$, 2, and 4. We have decided to use a triangular kernel with a base of three units. Since there are three samples and the total area of the approximated density function must be 1, the area of each triangle must be 1/3, so the height of each triangle must be 2/9. The three triangular kernels and the resulting estimated density function, which is their sum, are shown in Figure 4.5a.

$\frac{1}{3} = \frac{1}{2} \times 3 \times h$

$h = \frac{2}{9}$

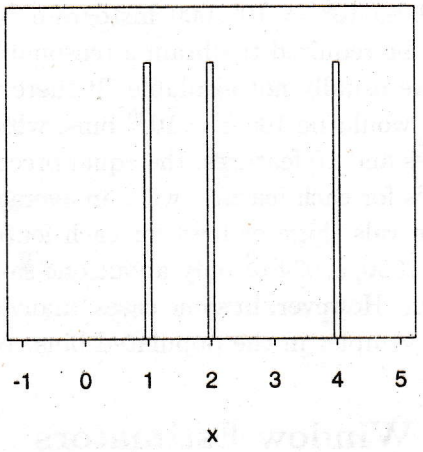

Figure 4.4: A density approximated by three delta functions. The height and width of each bar are $1/(n\delta)$ and δ, respectively.

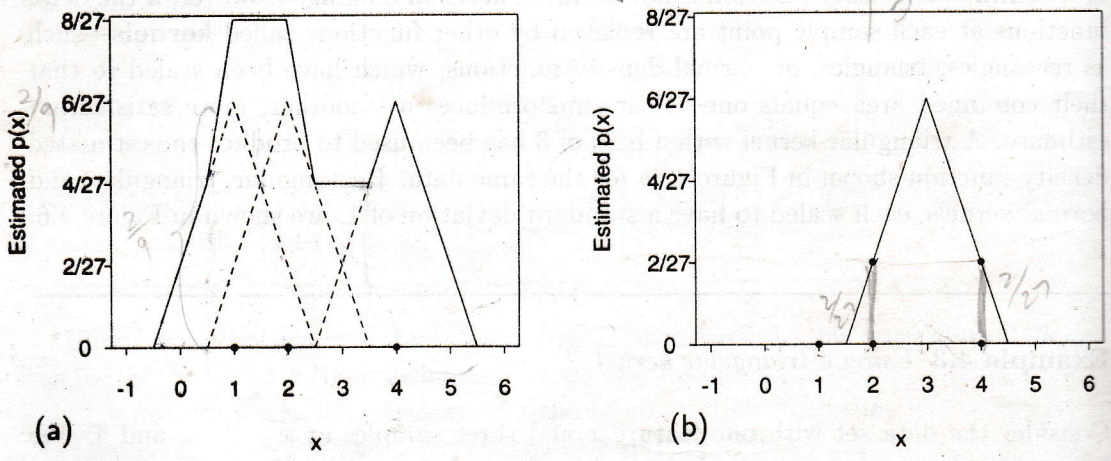

Figure 4.5: (a) Calculating $\hat{p}(x)$ with the kernel method. The kernels (dotted) and their sum (solid) are shown. (b) Calculating $\hat{p}(3)$ with the window method. $\hat{p}(3) = 2/27 + 2/27$.

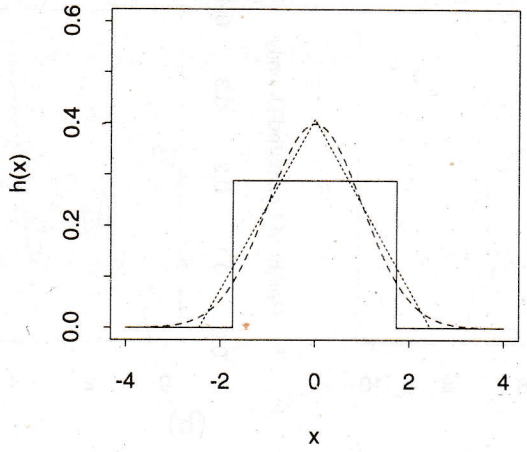

Figure 4.6: Three kernel functions, each with an area and a standard deviation of 1: rectangular (solid), triangular (dotted), normal (dashed).

To classify a sample at $x = x_0$, the entire densities $p(x|C_i)$ are not needed, only their values at $p(x_0|C_i)$. For example, let us assume that $p(x)$ at $x = 3$ is desired in Example 4.3. The result will, of course, be equal to the sum of the heights at $x = 3$ of all the kernel functions which have ranges which include the point $x = 3$. Another way of obtaining this result is to first reflect the kernel function about its center to produce a **window function**, center it at $x = 3$, and to sum its heights over each of the sample points included in this window. Symmetric kernel functions are usually used. They are unchanged by reflection to produce windows. The reason for the reflection of asymmetric kernels is that, for example, if the sample at point x_i affects mainly the estimated density at higher values of x rather than lower values, then the resulting density at x_i is mainly affected by samples at lower values of x. (In linear systems theory, if the kernel is taken to be an impulse response function, the use of a kernel would be equivalent to linear superposition and the use of a window would be equivalent to convolution. Convolution will be discussed further in Section 7.6.)

Figure 4.5b shows that the window function covers two of the samples. Its height over each sample is $2/27$ so that the estimate of $p(3)$ is $4/27$. This agrees with the value obtained in Figure 4.5a. Estimates of density functions in n dimensions, $p(\mathbf{x})$ where $\mathbf{x}$ is an n-dimensional vector, are obtained similarly as sums of n-dimensional kernels centered at the samples, normalized so that their total hypervolume or probability equals 1. To convert an n-dimensional kernel to an equivalent window, it must be reflected in each dimension.

The problem of choosing good widths or standard deviations for kernel or window functions is similar to the problem of choosing a good interval size for the histogram technique. If the width is too large, fine structure will be lost, but if the width is

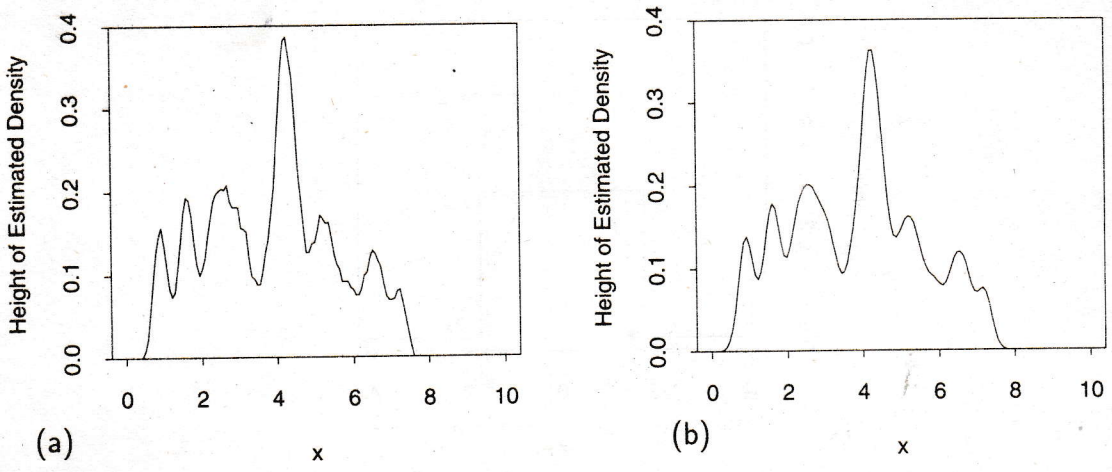

Figure 4.7: The estimated density functions $\hat{p}(x|A)$ for Example 4.2 using (a) a triangular kernel and (b) a normal kernel.

too small, the resulting approximation will not be sufficiently smooth, but will contain noisy fine structure due to the random locations of the samples. The width or the standard deviation of the window should be sufficiently large so that several samples will fall within this range, on the average. "Several" could again be taken to be the square root of the number of samples, or any other reasonable value. In practice, several different widths are usually tried, and the best is chosen intuitively or according to prior knowledge or according to its classification performance on the training data of interest. As the number of samples within the window approaches infinity and the window width itself approaches zero, the estimated density function will approach the true density, for any reasonably well-behaved window function and density.

Figure 4.7a shows the estimated density function for feature x for members of class A from Example 4.2 when using a triangular kernel with a base width of 2, and Figure 4.7b shows the estimated density when a normal kernel with $\sigma = 1$ is used. The triangular kernel produces a piecewise linear estimated density that is continuous and therefore smoother than a histogram, which is discontinuous at the interval endpoints, but it is rougher than the result using the normal kernel. This is because a triangular window and the resulting estimated density has discontinuities in its slope but the normal kernel is infinitely differentiable, that is, none of its derivatives contain any discontinuities. Window and kernel techniques are sometimes referred to as **Parzen estimation** [Young].

4.4 Nearest Neighbor Classification Techniques

The Single Nearest Neighbor Technique

The **single nearest neighbor** technique bypasses the problem of probability densities completely and simply classifies an unknown sample as belonging to the same class as the most **similar** or "nearest" sample point in the training set of data, which is often called a **reference set**. Nearest can be taken to mean the smallest **Euclidean distance** in n-dimensional feature space, which is the usual distance between two points $\mathbf{a} = (a_1, \ldots, a_n)$ and $\mathbf{b} = (b_1, \ldots, b_n)$, defined by

$$d_e(\mathbf{a}, \mathbf{b}) = \sqrt{\sum_{i=1}^{n} (b_i - a_i)^2} \tag{4.1}$$

where n is the number of features. This is an extension of the Pythagorean Theorem to n dimensions, and would be the distance measured by a ruler in one-, two-, or three-dimensional space. To save computing time, the square root would not actually be performed because the reference point with the smallest squared distance to the sample being classified also has the smallest distance to the sample.

Although Euclidean distance is probably the most commonly used distance function or measure of **dissimilarity** between feature vectors, it is not always the best metric. The fact that the distances in each dimension are squared before summation places great emphasis on those features for which the dissimilarity is large. A more moderate approach might be to use the sum of the **absolute differences** in each feature, rather than their squares, as the overall measure of dissimilarity. This would also save computing time. This distance metric would then be

$$d_{cb}(\mathbf{a}, \mathbf{b}) = \sum_{i=1}^{n} |b_i - a_i| \tag{4.2}$$

where n is the number of features. This sum of absolute distances in each dimension is sometimes called the **city block distance**, the **Manhattan metric**, or the **taxicab distance** because in the two-dimensional case it represents the distance traveled between two locations in a city if travel is restricted to lie along a rectangular grid of two-way streets. For example, the number of blocks north (or south) plus the number of blocks east (or west) would equal the total distance traveled.

A metric that would deemphasize single large feature differences and be more influenced by numerous small ones could be obtained by taking a sublinear function, such as the square root of the absolute values of the individual feature differences before summing. An extreme metric which considers only the most dissimilar pair of features is the **maximum distance** metric

$$d_m(\mathbf{a}, \mathbf{b}) = \max_{i=1}^{n} |b_i - a_i|. \tag{4.3}$$

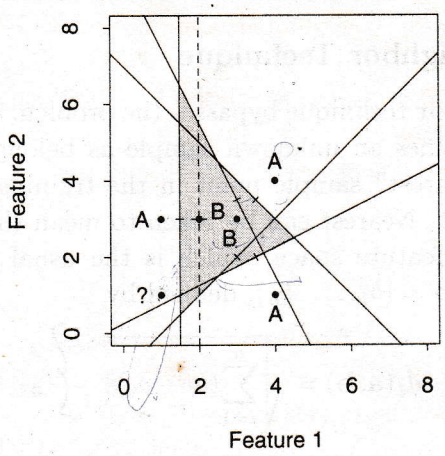

Figure 4.8: Plot for Example 4.4.

A generalization of the three distances (4.1), (4.2), and (4.3) is the **Minkowski distance** defined by

$$d_r(\mathbf{a}, \mathbf{b}) = \left[\sum_{i=1}^{n} |b_i - a_i|^r \right]^{1/r} \tag{4.4}$$

where r is an adjustable parameter (see Problem 4.18).

Example 4.4 *Nearest neighbor classification.*

Consider a feature space that contains three samples from class A and two samples from class B, as indicated in Figure 4.8. Suppose a sample of unknown class is located at $(1, 1)$. Using the Euclidean distance metric, the closest point of known classification is a sample from class A located at $(1, 3)$. The unknown sample would thus be classified as belonging to class A.

Although the result in Example 4.4 can be obtained by inspection, in general the computer must calculate the distance from the unknown sample to each sample in the reference set of data for which the classes are known, and choose the class of the point closest to the unknown sample. This could take a considerable amount of computer time if the reference set of data was large.

If Euclidean distance is used, the decision boundaries of the decision regions produced by the nearest neighbor technique are always **piecewise linear** because they consist of a number of line segments that are equidistant from a pair of samples of