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 structure of the true density, but when too many intervals are used as in Figure 4.1d, the confidence in their heights decreases. At first glance, it may appear to show some interesting fine structure in the data; however, most of the apparent structure depends on only a few samples, and thus cannot be very significant. People tend to perceive structure in the data even when the "structure" is due to random fluctuation, so we must guard against "overfitting" the data, which degrades performance.

As an extreme example, if the number of intervals were several times the number of samples, most of the intervals would contain no samples, and most of the others would probably contain only one sample each. In this case, the histogram reduces to a number of very narrow rectangles, nearly one for each sample point. The histogram would resemble a comb with most of its teeth missing. This would not produce a useful

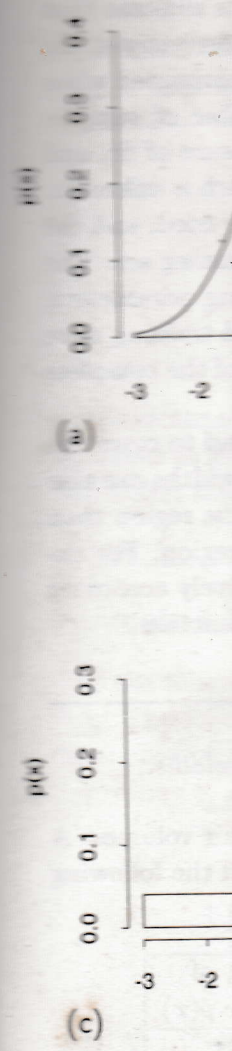

Figure 4.1: (a
(b) A histogr
A histogram
histogram of