An Introduction to Probabilistic Graphical Models

Michael I. Jordan
University of California, Berkeley

June 30, 2003

Chapter 5

Statistical Concepts

It is useful to attempt to distinguish the activities of the probability theorist and the statistician. Our perspective in the previous chapters has been mainly that of the former—we have built graphical models involving sets of random variables and shown how to compute the probabilities of certain events associated with these random variables. Given a particular choice of graphical model, consisting of a graph and a set of local conditional probabilities or potentials, we have seen how to infer the probabilities of various events of interest, such as the marginal or conditional probability that a particular random variable takes on a particular value.

Statistics is in a certain sense the inverse of probability theory. In a statistical setting the random variables in our domain have been *observed* and are therefore no longer unknown, rather it is the model that is unknown. We wish to infer the model from the data rather than the data from the model.

The problem of "inferring the model from the data" is a deep one, raising fundamental questions regarding the nature of knowledge, reasoning, learning, and scientific inquiry. In statistics, the study of these fundamental questions has often come down to a distinction between two major schools of thought—the *Bayesian* and the *frequentist*. In the following section we briefly outline the key distinctions between these two schools. It is worth noting that our discussion here will be incomplete and that we will be returning to these distinctions at various junctures in the book as our development of graphical models begins to bring the distinctions into clearer relief. But an equally important point to make is that many of the problems—particularly the computational problems—faced in these frameworks are closely related, even identical. A great deal of important work can be done within the graphical models formalism that is equally useful to Bayesian and frequentist statistics.

Beyond our discussion of foundational issues, we will also introduce several classes of statistical problems in this chapter, in particular the core problems of *density estimation*, *regression* and *classification*. As in earlier chapters our goal is to present enough in the way of concrete details to make the discussion understandable, but to emphasize broad themes that will serve as landmarks for our more detailed presentation in later chapters.

5.1 Bayesian and frequentist statistics

Bayesian statistics is in essence an attempt to deny any fundamental distinction between probability theory and statistics. Probability theory itself provides the capability for inverting relationships between uncertain quantities—this is the essence of *Bayes rule*—and Bayesian statistics represents an attempt to treat all statistical inference as probabilistic inference.

Let us consider a problem in which we have already decided upon the model *structure* for a given problem domain—for example, we have chosen a particular graphical model including a particular pattern of connectivity—but we have not yet chosen the values of the model *parameters*—the numerical values of the local conditional probabilities or potentials. We wish to choose these parameter values on the basis of observed data. (In general we might also want to choose the model structure on the basis of observed data, but let us postpone that problem—see Section 5.3).

For every choice of parameter values we obtain a different numerical specification for the joint distribution of the random variables X. We will henceforth write this probability distribution as $p(x \,|\, \theta)$ to reflect this dependence. Putting on our hats as probability theorists, we view the model $p(x \,|\, \theta)$ as a conditional probability distribution; intuitively it is an assignment of probability mass to unknown values of X, given a fixed value of θ. Thus, θ is known and X is unknown. As statisticians, however, we view X as known—we have observed its realization x—and θ as unknown. We thus in some sense need to invert the relationship between x and θ. The Bayesian point of view implements this notion of "inversion" using Bayes rule:

$$p(\theta \,|\, x) = \frac{p(x \,|\, \theta)p(\theta)}{p(x)}. \tag{5.1}$$

The assumptions allowing us to write this equation are noteworthy. First, in order to interpret the left-hand side of the equation we must view θ as a random variable. This is characteristic of the Bayesian approach—all unknown quantities are treated as random variables. Second, we view the data x as a quantity to be conditioned on—our inference is conditional on the event $\{X = x\}$. Third, in order to calculate $p(\theta \,|\, x)$ we see (from the right-hand side of Eq. (5.1)) that we must have in hand the probability distribution $p(\theta)$—the *prior probability* of the parameters. Given that we are viewing θ as a random variable, it is formally reasonable to assign a (marginal) probability to it, but one needs to think about what such a prior probability means in terms of the problem we are studying. Finally, note that Bayes rule yields a distribution over θ—the *posterior probability* of θ given x, not a single estimate of θ. If we wish to obtain a single value, we must (and will) invoke additional principles, but it is worth noting at the outset that the Bayesian approach tends to resist collapsing distributions to points.

The frequentist approach wishes to avoid the use of prior probabilities in statistics, and thus avoids the use of Bayes rule for the purpose of assigning probabilities to parameters. The goal of frequentist methodology is to develop an "objective" statistical theory, in which two statisticians employing the methodology must necessarily draw the same conclusions from a particular set of data.

Consider in particular a coin-tossing experiment, where $X \in \{0, 1\}$ is a binary variable representing the outcome of the coin toss, and $\theta \in (0, 1)$ is a real-valued parameter denoting the probability of heads. Thus the model is the Bernoulli distribution, $p(x \,|\, \theta) = \theta^x(1-\theta)^{1-x}$. Approaching the problem from a Bayesian perspective requires us to assign a prior probability to θ before observing the outcome of the coin toss. Two different Bayesian statisticians may assign different priors to θ and thus obtain different conclusions from the experiment. The frequentist statistician wishes to avoid such "subjectivity." From another point of view, a frequentist may claim that θ is a fixed property of the coin, and that it makes no sense to assign probability to it. A Bayesian may agree with the former statement, but would argue that $p(\theta)$ need not represent anything about the physics of the situation, but rather represents the *statistician's uncertainty* about the value of θ. Tossing the coin reduces the statistician's uncertainty, and changes the prior probability into the posterior probability $p(\theta \mid x)$. Bayesian statistics views the posterior probability and the prior probability alike as (possibly) subjective.

There are situations in which frequentist statistics and Bayesian statistics agree that parameters can be endowed with probability distributions. Suppose that we consider a factory that makes coins in batches, where each batch is characterized by a smelting process that affects the fairness of the resulting coins. A coin from a given batch has a different probability of heads than a coin from a different batch, and ranging over batches we obtain a distribution on the probability of heads θ. A frequentist is in general happy to assign prior probabilities to parameters, as long as those probabilities refer to objective frequencies of observing values of the parameters in repeated experiments.

From the point of view of frequentist statistics, there is no single preferred methodology for inverting the relationship between parameters and data. Rather, the basic idea is to consider various *estimators* of θ, where an estimator is some function of the observed data x (we will discuss a particular example below). One establishes various general criteria for evaluating the quality of various estimators, and chooses the estimator that is "best" according to these criteria. (Examples of such criteria include the *bias* and *variance* of estimators; these criteria will be discussed in Chapter 26). An important feature of this evaluation process is that it generally requires that the data x be viewed as the result of a random experiment that can be repeated and in which other possible values of x could have been obtained. This is of course consistent with the general frequentist philosophy, in which probabilities correspond to objective frequencies.

There is one particular estimator that is widely used in frequentist statistics, namely the *maximum likelihood* estimator. This estimator is popular for a number of reasons, in particular because it often yields "natural estimators" (e.g., sample proportions and sample means) in simple settings and also because of its favorable asymptotic properties.

To understand the maximum likelihood estimator, we must understand the notion of "likelihood" from which it derives. Recall that the probability model $p(x \mid \theta)$ has the intuitive interpretation of assigning probability to X for each fixed value of θ. In the Bayesian approach this intuition is formalized by treating $p(x \mid \theta)$ as a conditional probability distribution. In the frequentist approach, however, such a formal interpretation is suspect, because it suggests that θ is a random variable that can be conditioned on. The frequentist instead treats the model $p(x \mid \theta)$ as a family of probability distributions indexed by θ, with no implication that we are conditioning on θ.[1] Moreover, to implement a notion of "inversion" between x and θ, we simply change our point

[1] To acknowledge this interpretation, frequentist treatments often adopt the notation $p_\theta(x)$ in place of $p(x \mid \theta)$. We will stick with $p(x \mid \theta)$, hoping that the frequentist-minded reader will forgive us this abuse of notation. It will

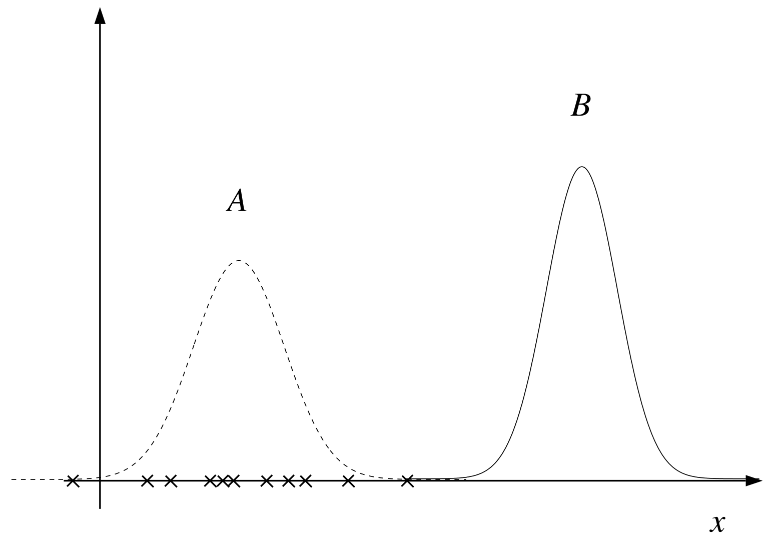

Figure 5.1: A univariate density estimation problem. (See Section 5.2.1 for a discussion of density estimation). The data $\{x_1, x_2, \ldots, x_N\}$ are given as X's along the abscissa. The parameter vector θ is the mean μ and variance σ^2 of a Gaussian density. Two candidate densities, involving different values of θ, are shown in the figure. Density A assigns higher probability to the observed data than density B, and thus would be preferred according to the principle of maximum likelihood.

of view—we treat $p(x \mid \theta)$ as a function of θ for fixed x. When interpreted in this way, $p(x \mid \theta)$ is referred to as the *likelihood function* and it provides the basis for maximum likelihood estimation.

As suggested in Figure 5.1, the likelihood function can be used to evaluate particular choices of θ. In particular, if for a given value of θ we find that the observed value of x is assigned low probability, then this is perhaps a poor choice of θ. A value of θ that assigns higher probability to x is preferred. Ranging over all possible choices of θ, we pick that value of θ that assigns maximal probability to x, and treat this value as an estimate of the true θ:

$$\hat{\theta}_{ML} = \mathrm{argmax}_\theta \; p(x \mid \theta). \tag{5.2}$$

Thus the maximum likelihood estimate is that value of θ that maximizes the likelihood function.

Regardless of whether one agrees that this justification of the maximum likelihood estimate is a natural one, it is certainly true that we have an estimator—a function of x—and we can evaluate the properties of this estimator under various frequentist criteria. It turns out that maximum likelihood is a good estimator under a variety of measures of quality, particularly in settings of large sample sizes when asymptotic analyses are meaningful (indeed, maximum likelihood estimates can be shown to be "optimal" in such settings). In other settings, particularly in cases of small sample sizes, maximum likelihood plays an important role as the starting point for the development of more complex estimators.

simplify our presentation throughout the rest of the book, liberating us from having to make distinctions between Bayesian and frequentist interpretations where none are needed or implied.

Another appealing feature of likelihood-based estimation is that it provides a link between Bayesian methods and frequentist methods. In particular, note that the distribution $p(x \mid \theta)$ appears in our basic Bayesian equation Eq. (5.1). Note moreover that Bayesian statisticians refer to this probability as a "likelihood" as do frequentist statisticians, even though the interpretation is different. Symbolically, we can interpret Eq. (5.1) as follows:

$$\textit{posterior} \; \propto \; \textit{likelihood} \; \times \textit{prior}, \tag{5.3}$$

where we see that in the Bayesian approach the likelihood can be viewed as a data-dependent operator that transforms between the prior probability and the posterior probability. At a bare minimum, Bayesian approaches and likelihood-based frequentist approaches have in common the need to calculate the likelihood for various values of θ. This is not a trivial fact—indeed a major focus of this book is the set of complex statistical models in which the computation of the likelihood is itself a daunting computational task. In working out effective computational procedures to deal with such models we are contributing to both Bayesian and frequentist statistics.

Let us explore this connection between Bayesian and frequentist approaches a bit further. Suppose in particular that we force the Bayesian to choose a particular value of θ; that is, to collapse the posterior distribution $p(\theta \mid x)$ to a point estimate. Various possibilities present themselves; in particular one could choose the mean of the posterior distribution or perhaps the mode. The mean of the posterior is often referred to as a *Bayes estimate*:

$$\hat{\theta}_{Bayes} = \int \theta \; p(\theta \mid x) d\theta, \tag{5.4}$$

and it is possible and worthwhile to study the frequentist properties of Bayes estimates. The mode of the posterior is often referred to as the *maximum a posteriori (MAP)* estimate:

$$\begin{aligned} \hat{\theta}_{MAP} &= \operatorname{argmax}_\theta \; p(\theta \mid x) & (5.5)\\ &= \operatorname{argmax}_\theta \; p(x \mid \theta) p(\theta), & (5.6) \end{aligned}$$

where in the second equation we have utilized the fact that the factor $p(x)$ in the denominator of Bayes rule is independent of θ. In a setting in which the prior probability is taken to be uniform on θ, the MAP estimate reduces to the maximum likelihood estimate. When the prior is not taken to be uniform, one can still view Eq. (5.6) as the maximization of a *penalized likelihood*. To see this, note that one generally works with logarithms when maximizing over probability distributions (the fact that the logarithm is a monotonic function implies that it does not alter the optimizing value). Thus one has:

$$\hat{\theta}_{MAP} = \operatorname{argmax}_\theta \; \{\log p(x \mid \theta) + \log p(\theta)\}, \tag{5.7}$$

as an alternative expression for the MAP estimate. Here the "penalty" is the additive term $\log p(\theta)$. Penalized log likelihoods are widely used in frequentist statistics to improve on maximum likelihood estimates in small sample settings (as we will see in Chapter 26).

It is important to emphasize, however, that MAP estimation involves a rather un-Bayesian use of the Bayesian formalism, and it would be wrong to understand the distinction between Bayesian and frequentist statistics as merely a matter of how to interpret a penalized log likelihood. To clarify,

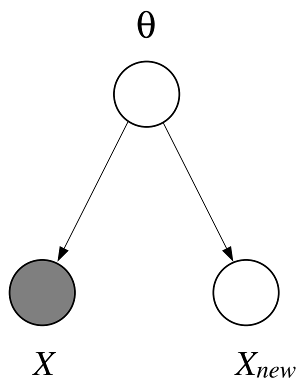

Figure 5.2: A graphical representation of the problem of prediction from a Bayesian point of view.

let us consider a somewhat broader problem in which the difference between MAP estimation and a fuller Bayesian approach is more salient. Let us consider the problem of *prediction*, where we are not interested in the value of θ per se, but are interested in using a model based on θ to predict future values of the random variable X. Let us suppose in particular that we have two random variables, X and X_{new}, which are characterized by the same distribution, and that we wish to use an observation of X to make a prediction regarding likely values of X_{new}. For simplicity, let us assume that X and X_{new} are independent; more precisely, we assume that they are conditionally independent given θ. We write:

$$\begin{aligned} p(x_{new} \mid x) &= \int p(x_{new}, \theta \mid x) d\theta && (5.8) \\ &= \int p(x_{new} \mid \theta, x) p(\theta \mid x) d\theta && (5.9) \\ &= \int p(x_{new} \mid \theta) p(\theta \mid x) d\theta. && (5.10) \end{aligned}$$

From the latter equation we see that the Bayesian prediction is based on combining the predictions across all values of θ, with the posterior distribution serving as a "weighting function." That is, interpreting the conditional probability $p(x_{new} \mid \theta)$ as the prediction of X_{new} given θ, we weight this prediction by the posterior probability $p(\theta \mid x)$, and integrate over all such weighted predictions. Note in particular that this calculation requires the entire posterior probability, not merely its value at a single point.

Within a frequentist approach, we are not allowed to treat θ as a random variable, and thus we do not attribute meaning to the integral in Eq. (5.10). Rather, we would consider various "estimates" of x_{new}; a natural choice might be the "plug-in estimate" $p(x_{new} \mid \hat{\theta}_{ML})$. Here we see that the difference between the frequentist approach and the Bayesian approach has become more significant; in the latter case we have to perform an integral in order to obtain a prediction. We can relate the two approaches if we approximate the posterior distribution by collapsing it to a delta function at $\hat{\theta}_{MAP}$, in which case the integral in Eq. (5.10) reduces to the plug-in estimate

$p(x_{new} \mid \hat{\theta}_{MAP})$. But in general this collapse would not satisfy the Bayesian (who views the integral as providing a better predictor than any predictor based on a point estimate) nor the frequentist (who wants to be free to consider a wider class of estimates than the plug-in estimate).

As a final note, consider the graphical model shown in Figure 5.2. This model captures the Bayesian point of view on the prediction problem that we have just discussed. The parameter θ is depicted as a node in the model; this is of course consistent with the Bayesian approach of treating parameters as random variables. Moreover, the conditional independence of X and X_{new} given θ is reflected as a Markov property in the graph. Finally, as we invite the reader to verify in Exercise **??**, applying the elimination algorithm to the graph yields exactly the calculation in Eq. (5.10). This is a reflection of a general fact—graphical models provide a nice way to visualize and organize Bayesian calculations. We will return to this point in later chapters. But let us emphasize here that this linkage, appealing as it is, does not reflect any special affinity between graphical models and Bayesian methods, but rather is a reflection of the more general link between Bayesian methods and probabilistic inference.

5.2 Statistical problems

Let us now descend from the somewhat ethereal considerations of statistical foundations to a rather more concrete consideration of problems in statistical estimation. In this section we will discuss three major classes of statistical problems—*density estimation*, *regression*, and *classification*. Not all statistical problems fall into one of these three classes, nor is it always possible to unambiguously characterize a given problem in terms of these classes, but there are certain core aspects of these three problem categories that are worth isolating and studying in a purified form.

We have two main goals in this section. The first is to introduce the graphical approach to representing statistical modeling problems, in particular emphasizing how the graphical representation helps makes modeling assumptions explicit. Second, we wish to begin to work with specific probability distributions, in particular the Gaussian and multinomial distributions. We will use this introductory section to illustrate some of the calculations that arise when using these distributions.

5.2.1 Density estimation

Suppose that we have in hand a set of observations on a random variable X—in general a vector-valued random variable—and we wish to use these observations to induce a probability density (probability mass function for discrete variables) for X. This problem—which we refer to generically as the problem of density estimation—is a very general statistical problem. Obtaining a model of the density of X allows us to assess whether a particular observation of X is "typical," an assessment that is required in many practical problems including *fault detection*, *outlier detection* and *clustering*. Density estimation also underlies many *dimensionality reduction* algorithms, where a joint density is projected onto a subspace or manifold, hopefully reducing the dimensionality of a data set while retaining its salient features. A related application is *compression*, where Shannon's fundamental relationship between code length and the negative logarithm of the density can be used to design a source code. Finally, noting that a joint density on X can be used to infer conditional densities among components of X, we can also use density estimates to solve problems in prediction.

To delimit the scope of the problem somewhat, note that in regression and classification the focus is on the relationship between a pair of variables, X and Y. That is, regression and classification problems differ from density estimation in that their focus is on a conditional density, $p(y \mid x)$, with the marginal $p(x)$ and the corresponding joint density of less interest, and perhaps not modeled at all. We develop methods that are specific to conditional densities in Sections 5.2.2 and 5.2.3.

Density estimation arises in many ways in the setting of graphical models. In particular we may be interested in inferring the density of a parentless node in a directed graphical model, or the density of a set of nodes in a larger model (in which case the density of interest is a marginal density), or the joint density of all of the nodes of our model.

Let us begin with an example. Our example will be one of the most classical of all statistical problems—that of estimating the mean and variance of a univariate Gaussian distribution.

Univariate Gaussian density estimation

Let us assume that X is a univariate random variable with a Gaussian distribution, that is:

$$p(x \mid \theta) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \right\}, \tag{5.11}$$

where μ and σ^2 are the mean and variance, respectively, and $\theta \triangleq (\mu, \sigma^2)$.[2] We wish to estimate θ based on observations of X. Here we are assuming that we know the parametric form of the density of X, and what is unknown are the numerical values of the parameters (cf. Figure 5.1). Plugging estimates of the parameters back into Eq. (5.11) provides an estimate of the density function.

Clearly a single observation of X provides no information about the variance and relatively poor information about the mean. Thus we need to consider multiple observations. What do we mean by "multiple observations"? Let us interpret this to mean that we have a *set of random variables*, $\{X_1, X_2, \ldots, X_N\}$, and that these random variables are *identically distributed*. Thus each of the variables X_n is characterized by a Gaussian distribution $p(x_n \mid \theta)$, with the same θ for each X_n.

In graphical model terms, we have a model with N nodes, one for each random variable. Which graphical model should we use? What connectivity pattern should we use? Let us suppose that the variables are not only identically distributed but that they are also *independent*. Thus we have the graphical model shown in Figure 5.3. It should be emphasized that these assumptions are by no means necessary; they are simply one possible set of assumptions, corresponding to a particular choice of graphical model. (We will be seeing significantly more complex graphical models on N Gaussian nodes; see, e.g., the Kalman filter in Chapter 15).

The nodes in Figure 5.3 are shaded, reflecting the fact that they are *observed data*. In general, "data" are designated by the shading of nodes in our models. In the context of the Bayesian approach to estimation, this use of shading is the same convention as we used in Chapter 2—in the Bayesian approach we *condition on the data* in order to compute probabilities for the parameters. In the context of frequentist approaches, where we no longer view ourselves as conditioning on the

[2]We will often denote this density as $\mathcal{N}(\mu, \sigma^2)$.

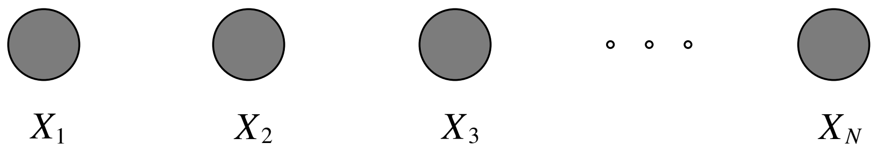

Figure 5.3: A graphical model representing the density estimation problem under an IID sampling model. The assumption that the data are sampled independently is reflected by the absence of links between the nodes. Each node is characterized by the same density.

data, we simply treat shading as a diagrammatic convention to indicate which nodes correspond to the observed data.

Letting X refer to the set of random variables $(X_1, X_2, \ldots, X_N)$, and letting x refer to the observations $(x_1, x_2, \ldots, x_N)$, we write the joint probability $p(x \mid \theta)$ as the product of local probabilities, one for each node in Figure 5.3:

$$\begin{aligned} p(x \mid \theta) &= \prod_{n=1}^{N} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x_n - \mu)^2\right\} && (5.12) \\ &= \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2\right\}, && (5.13) \end{aligned}$$

or alternatively, given that this particular graph can be interpreted as either a directed graph or an undirected graph, we can view this joint probability as a product of potential functions on the cliques of the graph (which are singleton nodes in this case).

Let us proceed to calculating parameter estimates. In particular let us calculate the maximum likelihood estimates of μ and σ^2. To do so we must maximize the likelihood $p(x \mid \theta)$ with respect to θ. We find it more convenient to maximize the logarithm of the likelihood, which, given that the logarithm is a monotonic function, will not change the results. Thus, let us define the *log likelihood*, denoted $l(\theta; x)$, as:

$$l(\theta; x) = \log p(x \mid \theta), \tag{5.14}$$

where we have reordered the variables on the left-hand side to emphasize that θ is to be viewed as the variable and x is to be viewed as a fixed constant. We now take the derivative of the log likelihood with respect to μ:

$$\begin{aligned} \frac{\partial l(\theta; x)}{\partial \mu} &= \frac{\partial}{\partial \mu}\left(-\frac{N}{2}\log(2\pi) - \frac{N}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2\right) && (5.15) \\ &= \frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n - \mu). && (5.16) \end{aligned}$$

Setting equal to zero and solving, we obtain:

$$\hat{\mu}_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n. \tag{5.17}$$

Thus we see that the maximum likelihood estimate of the mean of a Gaussian distribution is the sample mean.

Similarly let us take the derivative of the log likelihood with respect to σ^2:

$$\begin{aligned}
\frac{\partial l(\theta; x)}{\partial \sigma^2} &= \frac{\partial}{\partial \sigma^2}\left(-\frac{N}{2}\log(2\pi) - \frac{N}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2\right) && (5.18)\\
&= -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{n=1}^{N}(x_n - \mu)^2. && (5.19)
\end{aligned}$$

Setting equal to zero and solving, we obtain:

$$\hat{\sigma}^2_{ML} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \hat{\mu}_{ML})^2, \tag{5.20}$$

and we see that the maximum likelihood estimate of the variance is the sample variance. (Note that we are finding the joint estimates of μ and σ^2 by setting both partial derivatives equal to zero and solving simultaneously; this explains the presence of $\hat{\mu}_{ML}$ in the equation for $\hat{\sigma}^2_{ML}$).

Bayesian univariate Gaussian density estimation

In the Bayesian approach to density estimation the goal is to form a posterior density $p(\theta \mid x)$. Let us consider a simple version of this problem in which we take the variance σ^2 to be a known constant and restrict our attention to the mean μ. Thus we wish to obtain the posterior density $p(\mu \mid x)$, based on the prior density $p(\mu)$ and the Gaussian likelihood $p(x \mid \mu)$.

What prior distribution should we take for μ? This is a modeling decision, as was the decision to utilize a Gaussian for the probability of the data x in the first place. As we will see, it is mathematically convenient to take $p(\mu)$ to also be a Gaussian distribution. We will make this assumption in this section, but let us emphasize at the outset that mathematical convenience should not, and need not, dictate all of our modeling decisions. Indeed, a major thrust of this book is the development of methods for treating complex models, pushing back the frontier of what is "mathematically convenient" and, in the Bayesian setting, permitting a wide and expressive range of prior distributions.

If we take $p(\mu)$ to be a Gaussian distribution, then we face another problem: what should we take as the mean and variance of this distribution? To be consistent with the general Bayesian philosophy, we should treat these parameters as random variables and endow them with a prior distribution. This is indeed the approach of *hierarchical Bayesian modeling*, where we endow parameters with distributions characterized by "hyperparameters," which themselves can in turn

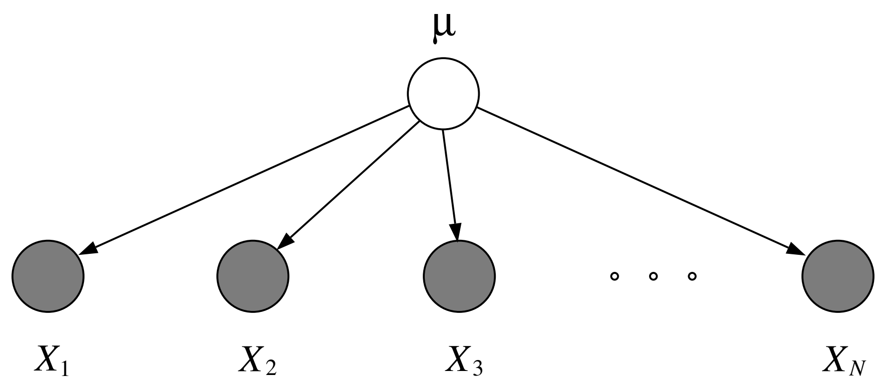

Figure 5.4: The graphical model for the Bayesian density estimation problem.

be endowed with distributions. While an infinite regress looms, in practice it is rare to take the hierarchical Bayesian approach to more than two or three levels, largely because there of diminishing returns—additional levels make little difference to the marginal probability of the data and thus to the expressiveness of our model.

Let us take the mean of $p(\mu)$ to be a fixed constant μ_0 and take the variance to be a fixed constant τ^2, while recognizing that in general we might endow these parameters with distributions.

The graphical model characterizing our problem is shown in Figure 5.4. The graph has been augmented with a node for the unknown mean μ. Note that there is a single such node and that its children are the data $\{X_n\}$. Thus this graph provides more information than the graph of Figure 5.3; in particular the independence assumption is elaborated—the data are assumed to be *conditionally independent given the parameters.*

The likelihood is identical in form to the frequentist likelihood in Eq. (5.13). To obtain the posterior we therefore need only multiply by the prior:

$$p(\mu) = \frac{1}{(2\pi\tau^2)^{1/2}} \exp\left\{-\frac{1}{2\tau^2}(\mu - \mu_0)^2\right\} \tag{5.21}$$

to obtain the joint probability:

$$p(x, \mu) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2\right\} \frac{1}{(2\pi\tau^2)^{1/2}} \exp\left\{-\frac{1}{2\tau^2}(\mu - \mu_0)^2\right\}, \tag{5.22}$$

which when normalized yields the posterior $p(\mu \,|\, x)$. Multiplying the two exponentials together yields an exponent which is quadratic in the variable μ; thus, normalization involves "completing the square." Appendix A presents the algebra (and in Chapter 13 we present a general matrix-based approach to completing the square—an operation that crops up often when working with Gaussian random variables). The result takes the following form:

$$p(\mu \,|\, x) = \frac{1}{(2\pi\tilde{\sigma}^2)^{1/2}} \exp\left\{-\frac{1}{2\tilde{\sigma}^2}(\mu - \tilde{\mu})^2\right\}, \tag{5.23}$$

where

$$\tilde{\mu} = \frac{N/\sigma^2}{N/\sigma^2 + 1/\tau^2}\,\bar{x} + \frac{1/\tau^2}{N/\sigma^2 + 1/\tau^2}\,\mu_0, \tag{5.24}$$

where $\bar{x}$ is the sample mean, and where

$$\tilde{\sigma}^2 = \left(\frac{N}{\sigma^2} + \frac{1}{\tau^2}\right)^{-1}. \tag{5.25}$$

We see that the posterior probability is a Gaussian, with mean $\tilde{\mu}$ and variance $\tilde{\sigma}^2$.

Both the posterior variance and the posterior mean have an intuitive interpretation. Note first that σ^2/N is the variance of a sum of N independent random variables with variance σ^2, thus σ^2/N is the variance associated with the data. Eq. (5.25) says that we add the inverse of this variance to the inverse of the prior variance to obtain the inverse of the posterior variance. Thus, inverse variances add. From Eq. (5.24) we see that the posterior mean is obtained as a linear combination of the sample mean and the prior mean. The weights in this combination can be interpreted as the fraction of the posterior variance accounted for by the variance from the data term and the prior variance respectively. These weights sum to one; thus, the combination in Eq. (5.24) is a *convex combination.*

As the number of data points N becomes large, the weight associated with $\bar{x}$ goes to one and the weight associated with μ_0 approaches zero. Thus in the limit of large data sets, the Bayes estimate of μ approaches the maximum likelihood estimate of μ.

Plates

Let us take a quick detour to discuss a notational device that we will find useful. Graphical models representing independent, identically distributed (IID) sampling have a repetitive structure that can be captured with a formal device known as a *plate*. Plates allow repeated motifs to be represented in a simple way. In particular, the simple IID model shown in Figure 5.5(a) can be represented more succinctly using the plate shown in Figure 5.5(b).

For the Bayesian model in Figure 5.6(a) we obtain the representation in Figure 5.6(b). Note that the parameter μ appears *outside* the plate; this captures the fact that there is a single parameter value that is shared among the distributions for each of the X_n.

Formally, a plate is simply a graphical model "macro." That is, to interpret Figure 5.5(b) or Figure 5.6(b) we copy the graphical object in the plate N times, where the number N is recorded in the lower right-hand corner of the box, and apply the usual graphical model semantics to the result.

Density estimation for discrete data

Let us now consider the case in which the variables X_n are discrete variables, each taking on one of a finite number of possible values. We wish to study the density estimation problem in this setting, recalling that "probability density" means "probability mass function" in the discrete case.

As before, we will make the assumption that the data are IID, thus the modeling problem is represented by the plate shown in Figure 5.5(b). Each of the variables X_n can take on one of M

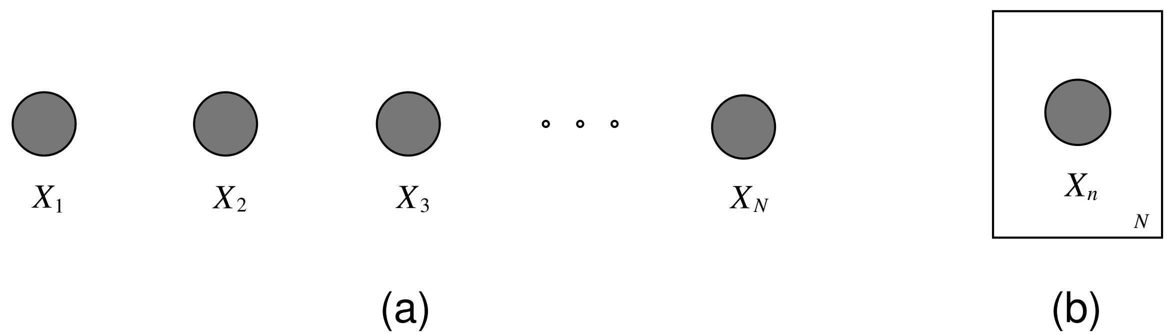

Figure 5.5: Repeated graphical motifs can be represented using plates. The IID sampling model for density estimation shown in (a) is represented using a plate in (b). The plate is interpreted by copying the graphical object within the box N times; thus the graph in (b) is a shorthand for the graph in (a).

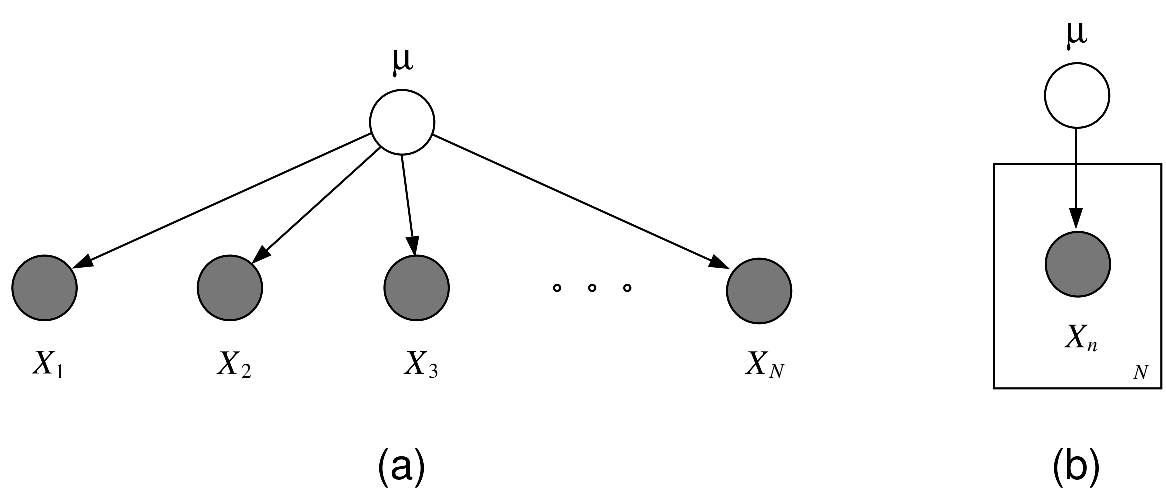

Figure 5.6: The Bayesian density estimation model shown in (a) is represented using a plate in (b). Again, the graph in (b) is to be interpreted as a shorthand for the graph in (a).

values. To represent this set of M values we will find it convenient to use a vector representation. In particular, let the range of X_n be the set of binary M-component vectors with one component equal to one and the other components equal to zero. Thus for a variable X_n taking on three values, we have:

$$X_n \in \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \right\}. \tag{5.26}$$

We use superscripts to refer to the components of these vectors, thus X_n^k refers to the kth component of the variable X_n. We have $X_n^k = 1$ if and only if the variable X_n takes on its kth value. Note that $\sum_k X_n^k = 1$ by definition.

Using this representation, we can write the probability distribution for X_n in a convenient general form. In particular, letting θ_k represent the probability that X_n takes on its kth value, i.e., $\theta_k \triangleq p(x_n^k = 1)$, we have:

$$p(x_n \,|\, \theta) = \theta_1^{x_n^1} \theta_2^{x_n^2} \cdots \theta_M^{x_n^M}. \tag{5.27}$$

This is the *multinomial* probability distribution, $\mathrm{Mult}(1, \theta)$, with parameter vector $\theta = (\theta_1, \theta_2, \ldots \theta_M)$. To calculate the probability of the observation x, we take the product over the individual multinomial probabilities:

$$\begin{aligned} p(x \,|\, \theta) &= \prod_{n=1}^{N} \theta_1^{x_n^1} \theta_2^{x_n^2} \cdots \theta_M^{x_n^M} && (5.28) \\ &= \theta_1^{\sum_{n=1}^N x_n^1} \theta_2^{\sum_{n=1}^N x_n^2} \cdots \theta_M^{\sum_{n=1}^N x_n^M}, && (5.29) \end{aligned}$$

where the exponent $\sum_{n=1}^N x_n^k$ is the count of the number of times the kth value of the multinomial variable is observed across the N observations.

To calculate the maximum likelihood estimates of the multinomial parameters we take the logarithm of Eq. (5.29) to obtain the log likelihood:

$$l(\theta; x) = \sum_{n=1}^{N} \sum_{k=1}^{M} x_n^k \log \theta_k, \tag{5.30}$$

and it is this expression that we must maximize with respect to θ.

This is a constrained optimization problem for which we use Lagrange multipliers. Thus we form the Lagrangian:

$$\tilde{l}(\theta; x) = \sum_{n=1}^{N} \sum_{k=1}^{M} x_n^k \log \theta_k + \lambda(1 - \sum_{k=1}^{M} \theta_k), \tag{5.31}$$

take derivatives with respect to θ_k:

$$\frac{\partial \tilde{l}(\theta; x)}{\partial \theta_k} = \frac{\sum_{n=1}^N x_n^k}{\theta_k} - \lambda \tag{5.32}$$

and set equal to zero:

$$\frac{\sum_{n=1}^{N} x_n^k}{\hat{\theta}_{k,ML}} = \lambda. \tag{5.33}$$

Multiplying through by $\hat{\theta}_{k,ML}$ and summing over k yields:

$$\begin{aligned}
\lambda &= \sum_{k=1}^{M}\sum_{n=1}^{N} x_n^k & (5.34)\\
&= \sum_{n=1}^{N}\sum_{k=1}^{M} x_n^k & (5.35)\\
&= N. & (5.36)
\end{aligned}$$

Finally, substituting Eq. (5.36) back into Eq. (5.33) we obtain:

$$\hat{\theta}_{k,ML} = \frac{1}{N}\sum_{n=1}^{N} x_n^k. \tag{5.37}$$

Noting again that $\sum_{n=1}^{N} x_n^k$ is the count of the number of times that the kth value is observed, we see that the maximum likelihood estimate of θ_k is a sample proportion.

Bayesian density estimation for discrete data

In this section we discuss a Bayesian approach to density estimation for discrete data. As in the Gaussian setting, we specify a prior using a parameterized distribution and show how to compute the corresponding posterior.

An appealing feature of the solution to the Gaussian problem was that the prior and the posterior have the same distribution—both are Gaussian distributions. Among other virtues, this implies that Eq. (5.24) and Eq. (5.25) can be used *recursively*—the posterior based on earlier observations can serve as the prior for additional observations. At each step the posterior distribution remains in the Gaussian family.

To achieve a similar closure property in the discrete problem we must find a prior distribution which when multiplied by the multinomial distribution yields a posterior distribution in the same family. Clearly, this can be achieved by a prior distribution of the form:

$$p(\theta) = C(\alpha)\theta_1^{\alpha_1-1}\theta_2^{\alpha_2-1}\cdots\theta_M^{\alpha_M-1}, \tag{5.38}$$

for $\sum_i \theta_i = 1$, where $\alpha = (\alpha_1, \ldots, \alpha_M)$ are hyperparameters and $C(\alpha)$ is a normalizing constant.[3] This distribution, known as the *Dirichlet distribution*, has the same functional form as the multinomial, but the θ_i are random variables in the Dirichlet distribution and parameters in the multinomial distribution. The constant $C(\alpha)$ is obtained via a bit of calculus (see Appendix B):

$$C(\alpha) = \frac{\Gamma(\sum_{i=1}^{M}\alpha_i)}{\prod_{i=1}^{M}\Gamma(\alpha_i)}, \tag{5.39}$$

[3]The negative one in the exponent is a convention; we could redefine the α_i to remove it.

where $\Gamma(\cdot)$ is the gamma function. In the rest of this section we will not bother with calculating the normalization; once we have a distribution in the Dirichlet form we can substitute into Eq. (5.39) to find the normalization factor.

We now calculate the posterior probability:

$$p(\theta \mid x) \;\propto\; \theta_1^{\sum_{n=1}^N x_n^1} \theta_2^{\sum_{n=1}^N x_n^2} \cdots \theta_M^{\sum_{n=1}^N x_n^M} \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \cdots \theta_M^{\alpha_M - 1} \tag{5.40}$$

$$= \theta_1^{\sum_{n=1}^N x_n^1 + \alpha_1 - 1} \theta_2^{\sum_{n=1}^N x_n^2 + \alpha_2 - 1} \cdots \theta_M^{\sum_{n=1}^N x_n^M + \alpha_M - 1}. \tag{5.41}$$

This is a Dirichlet density, with parameters $\sum_{n=1}^N x_n^k + \alpha_k$. We see that to update the prior into a posterior we simply add the count $\sum_{n=1}^N x_n^k$ to the prior parameter α_k.

It is worthwhile to consider the special case of the multinomial distribution when $M = 2$. In this setting, X_n is best treated as a binary variable rather than a vector; thus: $x_n \in \{0, 1\}$. The multinomial distribution reduces to:

$$p(x_n \mid \theta) = \theta^{x_n}(1 - \theta)^{1 - x_n}; \tag{5.42}$$

the *Bernoulli distribution*. The parameter θ encodes the probability that X_n takes the value one.

In the case $M = 2$, the Dirichlet distribution specializes to the *beta distribution*:

$$p(\theta) = C(\alpha)\theta^{\alpha_1 - 1}(1 - \theta)^{\alpha_2 - 1}, \tag{5.43}$$

where $\alpha = (\alpha_1, \alpha_2)$ is the hyperparameter. The beta distribution has its support on the interval $[0, 1]$. Plots of the beta distribution are shown in Figure 5.7 for various values of α_1 and α_2. Note that the *uniform distribution* is the special case of the beta distribution when $\alpha_1 = 1$ and $\alpha_2 = 1$.

As the number of data points N becomes large, the sums $\sum_{n=1}^N x_n^k$ dominate the prior terms α_k in the posterior probability. In this limit, the posterior approaches the log likelihood in Eq. (5.30) and the Bayes estimate of θ approaches the maximum likelihood estimate of θ.

Mixture models

It is important to recognize that the Gaussian and multinomial densities are by no means the universally best choices of density model. Suppose, for example, if the data are continuous data restricted to the half-infinite interval $[0, \infty)$. The Gaussian, which assigns density to the entire real line, is unnatural here, and densities such as the gamma or lognormal, whose support is $[0, \infty)$, may be preferred. Similarly, the multinomial distribution treats discrete data as an unordered, finite set of values. In problems involving ordered sets, and/or infinite ranges, probability distributions such as the Poisson or geometric may be more appropriate. Maximum likelihood and Bayesian estimates are available for these distributions, and indeed there is a general family known as the *exponential* family—which includes all of the distributions listed above and many more—in which explicit formulas can be obtained. (We will discuss the exponential family in Chapter 8).

This larger family of distributions is still, however, restrictive. Consider the probability density shown in Figure 5.8. This density is bimodal and we are unable to represent it within the family of Gaussian, gamma or lognormal densities. Given a data set $\{x_n\}$ sampled from this density, we

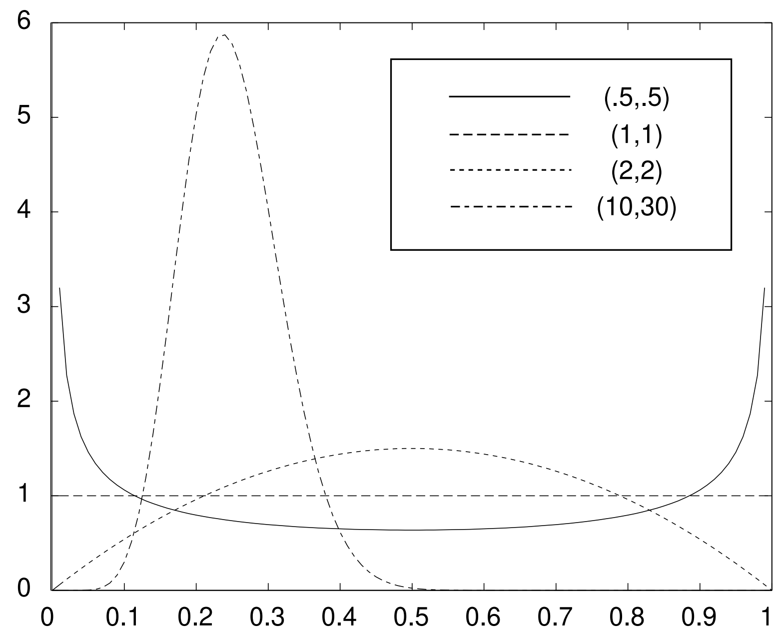

Figure 5.7: The beta(α_1, α_2) distribution for various values of the parameters α_1 and α_2.

can naively fit a Gaussian density, but the likelihood that we achieve will in general be significantly smaller than the likelihood of the data under the true density, and the resulting density estimate will bear little relationship to the truth.

Multimodal densities often reflect the presence of *subpopulations* or *clusters* in the population from which we are sampling. Thus, for example, we would expect the density of heights of trees in a forest to be multimodal, reflecting the different distributions of heights of different species. It may be that for a particular species the heights are unimodal and reasonably well modeled by a simple density, such as a density in the exponential family. If so, this suggests a "divide-and-conquer" strategy in which the overall density estimation is broken down into a set of smaller density estimation problems that we know how to handle. Let us proceed to develop such a strategy.

Let $f_k(x \mid \theta_k)$ be the density for the kth subpopulation, where θ_k is a parameter vector. We define a *mixture density* for a random variable X by taking the convex sum over the component densities $f_k(x \mid \theta_k)$:

$$p(x \mid \theta) = \sum_{k=1}^{K} \alpha_k f_k(x \mid \theta_k), \tag{5.44}$$

where the α_k are nonnegative constants that sum to one:

$$\sum_{k=1}^{K} \alpha_k = 1. \tag{5.45}$$

The densities $f_k(x \mid \theta_k)$ are referred to in this setting as *mixture components* and the parameters α_k

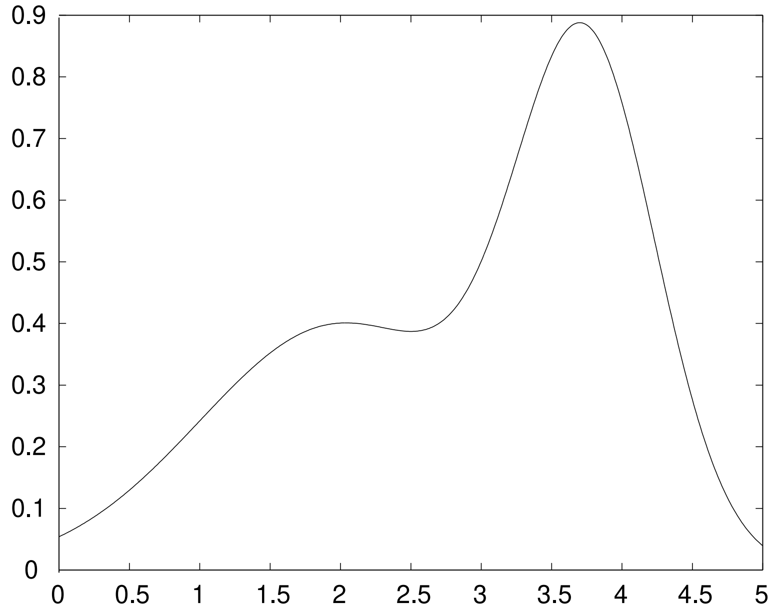

Figure 5.8: A bimodal probability density.

are referred to as *mixing proportions*. The parameter vector θ is the collection of all of the parameters, including the mixing proportions: $\theta \triangleq (\alpha_1, \ldots, \alpha_K, \theta_1, \ldots, \theta_K)$. That the function $p(x \mid \theta)$ that we have defined is in fact a density follows from the constraint that the mixing proportions sum to one.

The example shown in Figure 5.8 is a mixture density with $K = 2$:

$$p(x \mid \theta) = \alpha_1 \mathcal{N}(x \mid \mu_1, \sigma_1^2) + \alpha_2 \mathcal{N}(x \mid \mu_2, \sigma_2^2), \tag{5.46}$$

where the mixture components are Gaussian distributions with means μ_k and variances σ_k^2. Gaussian mixtures are a popular form of mixture model, particular in multivariate settings (see Chapter 10).

It is illuminating to express the mixture density in Eq. (5.44) in a way that makes explicit its interpretation in terms of subpopulations. Let us do this using the machinery of graphical models. As shown in Figure 5.9, we introduce a multinomial random variable Z into our model. We also introduce an edge from Z to X. Following the recipe from Chapter 2 we endow this graph with a joint probability distribution by assigning a marginal probability to Z and a conditional probability to X. Let α_k be the probability that Z takes on its kth value; thus, $\alpha_k \triangleq p(z^k = 1)$. Moreover, conditional on Z taking on its kth value, let the conditional probability of X, $p(x \mid z^k = 1)$, be given by $f_k(x \mid \theta_k)$. The joint probability is therefore given by:

$$\begin{aligned} p(x, z^k = 1 \mid \theta) &= p(x \mid z^k = 1, \theta) p(z^k = 1 \mid \theta) & (5.47) \\ &= \alpha_k f_k(x \mid \theta_k), & (5.48) \end{aligned}$$

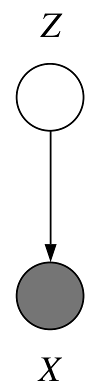

Figure 5.9: A mixture model represented as a graphical model. The latent variable Z is a multinomial node taking on one of K values.

where $\theta \triangleq (\alpha_1, \ldots, \alpha_K, \theta_1, \ldots, \theta_K)$. To obtain the marginal probability of X we sum over k:

$$p(x \mid \theta) = \sum_{k=1}^{K} p(x, z^k = 1 \mid \theta) \tag{5.49}$$

$$= \sum_{k=1}^{K} \alpha_k f_k(x \mid \theta_k), \tag{5.50}$$

which is the mixture model in Eq. (5.44).

This model gives us our first opportunity to invoke our discussion of probabilistic inference from Chapter 3. In particular, given an observation x, we can use Bayes rule to invert the arrow in Figure 5.9 and calculate the conditional probability of Z:

$$p(z^k = 1 \mid x, \theta) = \frac{p(x \mid z^k = 1, \theta_k) p(z^k = 1)}{\sum_j p(x \mid z^j = 1, \theta_j) p(z^j = 1)} \tag{5.51}$$

$$= \frac{\alpha_k f_k(x \mid \theta_k)}{\sum_j \alpha_j f_j(x \mid \theta_j)}. \tag{5.52}$$

This calculation allows us to use the mixture model to *classify* or *categorize* the observation x into one of the subpopulations or clusters that we assume to underly the model. In particular we might classify x into the class k that maximizes $p(z^k = 1 \mid x, \theta)$.

Let us turn to the problem of estimating the parameters of the mixture model from data. We again assume for simplicity a sampling model in which we have N IID observations $\{x_n; n = 1, \ldots, N\}$, while again noting that we will move beyond the IID setting in later chapters. The IID assumption corresponds to replicating our basic graphical model N times, yielding the plate shown in Figure 5.10. Note again that the variables Z_n are unshaded—they are *unobserved* or *latent* variables. We have introduced them into our model in order to make explicit the structural assumptions that lie behind the mixture density that we are using, but we need not assume that these variables are observed.

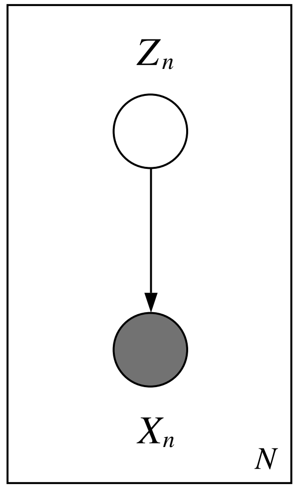

Figure 5.10: The mixture model under an IID sampling assumption.

The log likelihood is given by taking the logarithm of the joint probability associated with the model, which in the IID case becomes a sum of log probabilities. Again letting $x = (x_1, \ldots, x_N)$, we have:

$$l(\theta; x) = \sum_{n=1}^{N} \log \sum_{k=1}^{K} \alpha_k f_k(x_n \,|\, \theta_k). \tag{5.53}$$

To obtain maximum likelihood estimates we take derivatives with respect to θ and set to zero. The resulting equations are, however, nonlinear and do not admit a closed-form solution; solving these equations requires iterative methods. While any of a variety of numerical methods can be used, there is a particular iterative method—the *Expectation-Maximization (EM) algorithm*—that is natural not only for mixture models but also for more general graphical models. The EM algorithm involves an alternating pair of steps, the *E step* and the *M step*. The E step involves running an inference algorithm—for example the elimination algorithm that we discussed in Chapter 3—to essentially "fill in" the values of the unobserved nodes given the observed nodes. In the case of mixture models, this reduces to the invocation of Bayes rule in Eq. (5.52). The M step treats the "filled-in" graph as if all of the filled-in values had been observed, and updates the parameters to obtain improved values. In the mixture model setting this essentially reduces to finding separate density estimates for the separate subpopulations. We will present the EM algorithm formally in Chapter 11, and present its application to mixture models in Chapter 10.

Nonparametric density estimation

In many cases data may come from a complex mechanism about which we have little or no prior knowledge. The density underlying the data may not fall into one of the "standard" forms. The density may be multimodal, but we may have no reason to suspect underlying subpopulations and may have no reason to attribute any particular meaning to the modes. When we find ourselves in

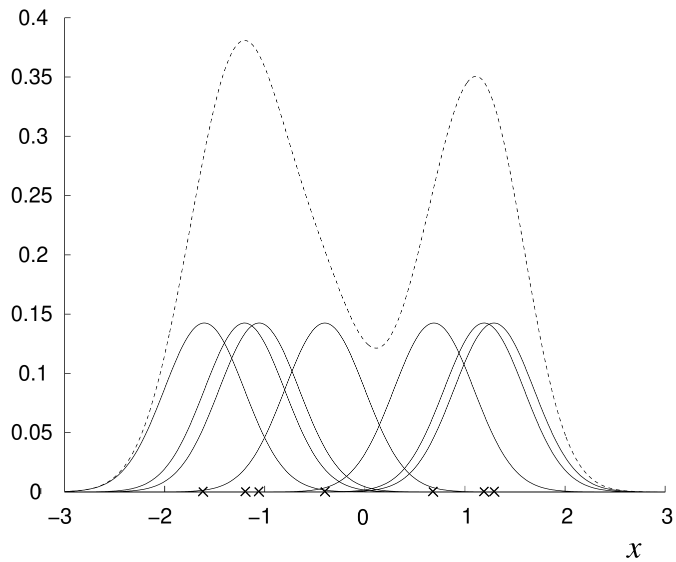

Figure 5.11: An example of kernel density estimation. The kernel functions are Gaussians centered at the data points x_n (shown as crosses on the abscissa). Each Gaussian has a standard deviation $\lambda = 0.35$. The Gaussians have been scaled by dividing by the number of data points ($N = 8$). The density estimate (shown as a dotted curve) is the sum of these scaled kernels.

such a situation—by no means uncommon—what do we do?

Nonparametric density estimation provides a general class of methods for dealing with such knowledge-poor cases. In this section we introduce this approach via a simple, intuitive nonparametric method known as a *kernel density estimator*. We return to a fuller discussion of nonparametric methods in Chapter 25.

The basic intuition behind kernel density estimation is that each data point x_n provides evidence for non-zero probability density at that point. A simple way to harness this intuition is to place an "atom" of mass at that point (see Figure 5.11). Moreover, making the assumption that the underlying probability density is smooth, we let the atoms have a non-zero "width." Superimposing N such atoms, one per data point, we obtain a density estimate.

More formally, let $k(x, x_n, \lambda)$ be a *kernel function*—a nonnegative function integrating to one (with respect to x). The argument x_n determines the location of the kernel function; kernels are generally symmetric about x_n. The parameter λ is a general "smoothing" parameter that determines the width of the kernel functions and thus the smoothness of the resulting density estimate. Superimposing N such kernel functions, and dividing by N, we obtain a probability density:

$$\hat{p}(x) = \frac{1}{N} \sum_{n=1}^{N} k(x, x_n, \lambda). \tag{5.54}$$

This density is the kernel density estimate of the underlying density $p(x)$.

A variety of different kernel functions are used in practice. Simple (e.g., piecewise polynomial) functions are often preferred, partly for computational reasons (calculating the density at a given point x requires N function evaluations). Gaussian functions are sometimes used, in which case x_n plays the role of the mean and λ plays the role of the standard deviation.

While the kernel function is often chosen a priori, the value of λ is generally chosen based on the data. This is a nontrivial estimation problem for which classical estimation methods are often of little help. In particular, it is important to understand that maximum likelihood is *not* appropriate for solving this problem. Suppose that we interpret the density in Eq. (5.54) as a likelihood function, with λ as the parameter. For most reasonable kernels, this "likelihood" increases monotonically as λ goes to zero, because the kernel assigns more probability density to the points x_n for smaller values of λ. Indeed, in the limit of $\lambda = 0$, the kernel generally approaches a delta function, giving infinite likelihood to the data. A sum of delta functions is obviously a poor density estimate.

We will discuss methods for choosing smoothing parameters in Chapter 25. As we will see, most practical methods involve some form of *cross-validation*, in which a fraction of the data are held out and used to evaluate various choices of λ. Both overly small and overly large values of λ will tend to assign small probability density to the held-out data, and this provides a rational approach to choosing λ.

The problem here is a general one, motivating a distinction between *parametric models* and *nonparametric models* and suggesting the need for distinct methods for their estimation. Understanding the distinction requires us to consider how a given model would change if the number of data points N were to increase. For parametric models the basic structure of the model remains fixed as N increases. In particular, for the Gaussian estimation problem treated in Section 5.2.1, the class of densities that are possible fits to the data remains the same whatever the value of N; for each N we obtain a Gaussian density with estimated parameters $\hat{\mu}$ and $\hat{\sigma}^2$. Increasing the number of data points increases the precision of these estimates, but it does not increase the class of densities that we are considering. In the nonparametric case, on the other hand, the class of densities increases as N increases. In particular, with $N + 1$ data points it is possible to obtain densities with $N + 1$ modes; this is not possible with N data points.

An alternative perspective is to view the locations of the kernels as "parameters"; the number of such "parameters" increases with the number of data points. In effect, we can view nonparametric models as parametric, but with an unbounded, data-dependent, number of parameters. Indeed, in an alternative language that is often used, parametric models are referred to as "finite-dimensional models," and nonparametric models are referred to as "infinite-dimensional models."

It is worthwhile to compare the kernel density estimator in Eq. (5.54) to the mixture model in Eq. (5.44). Consider in particular the case in which Gaussian mixture components are used in Eq. (5.44) and Gaussian kernel functions are used in Eq. (5.54). In this case the kernel estimator can be viewed as a mixture model in which the means are fixed to the data point locations, the variances are set to λ^2, and the mixing proportions are set to $1/N$. In what sense are the two different approaches to density estimation really different?

Again, the key difference between the two approaches is revealed when we let the number of data points N grow. The mixture model is generally viewed as a parametric model, in which case the number of mixture components, K, does not increase as the number of data points grows. This is consistent with our interpretation of a mixture model in terms of a set of K underlying subpopulations—if we believe that these subpopulations exist, then we do not vary K as N increases. In the kernel estimation approach, on the other hand, we have no commitment to underlying subpopulations, and we accord no special treatment to the number of kernels. As the number of data points grows, we allow the number of kernels to grow. Moreover we generally expect that λ will shrink as N grows to allow an increasingly close fit to the details of the true density.

There are several caveats to this discussion. First, in the mixture model setting, we may not know the number K of mixture components in practice and we may wish to estimate K from the data. This is a model selection problem (see Section 5.3). Solutions to model selection problems generally involve allowing K to increase as the number of data points increases, based on the fact that more data points are generally needed to provide more compelling evidence for multiple modes. Second, mixture models can also be used nonparametrically. In particular, a *mixture sieve* is a mixture model in which the number of components is allowed to grow with the number of data points. This differs from kernel density estimation in that the location of the mixture components are treated as free parameters rather than being fixed at the data points; moreover, each mixture component generally has its own (free) scale parameter. Also, the growth rate of the number of "parameters" in mixture sieves is slower than that of kernel density estimation (e.g., $\log N$ vs. N). As this discussion begins to suggest, however, it becomes difficult to enforce a clear boundary between parametric and nonparametric methods. A given approach can be treated in one way or the other, depending on a modeler's goals and assumptions.

There is a general tradeoff between flexibility and statistical efficiency that is relevant to this discussion. If the underlying "true" density is a Gaussian, then we probably want to estimate this density using a parametric approach, we can also use a kernel density estimate. The latter estimate will eventually converge to the true density, but it may require very many data points. A parametric estimator will converge more rapidly. Of course, if the true density is not a Gaussian, then the parametric estimate would still converge, but to the wrong density, whereas the nonparametric estimate would eventually converge to the true density. In sum, if we are willing to make more assumptions then we get faster convergence, but with the possibility of poor performance if reality does not match our assumptions. Nonparametric estimators allow us to get away with fewer assumptions, while requiring more data points for comparable levels of performance.

There is also a general point to be made with respect to the representation of densities in graphical models. As suggested in Figure 5.12, there are two ways to represent a multi-modal density as a graphical model. As shown in Figure 5.12(a), we can allow the class of densities $p(x)$ at node X to include multi-modal densities, such as mixtures or kernel density estimates. Alternatively, we can use the "structured" model depicted in Figure 5.12(b), where we obtain a mixture distribution for X_n by marginalizing over the latent variable Z_n. Although it may seem natural to reserve the latter representation for parametric modeling, in particular for the setting in which we attribute a "meaning" to the latent variable, such a step is in general unwarranted. The mixture sieve exemplifies a situation in which we may wish to use graphical machinery to represent the structure of a nonparametric model explicitly. In general, the choices of how to use and how to interpret graphical structure are modeling decisions. While we may wish to use graphical representations to express domain-specific structural knowledge, we may also be guided by other factors, including mathematical convenience and the availability of computational tools.

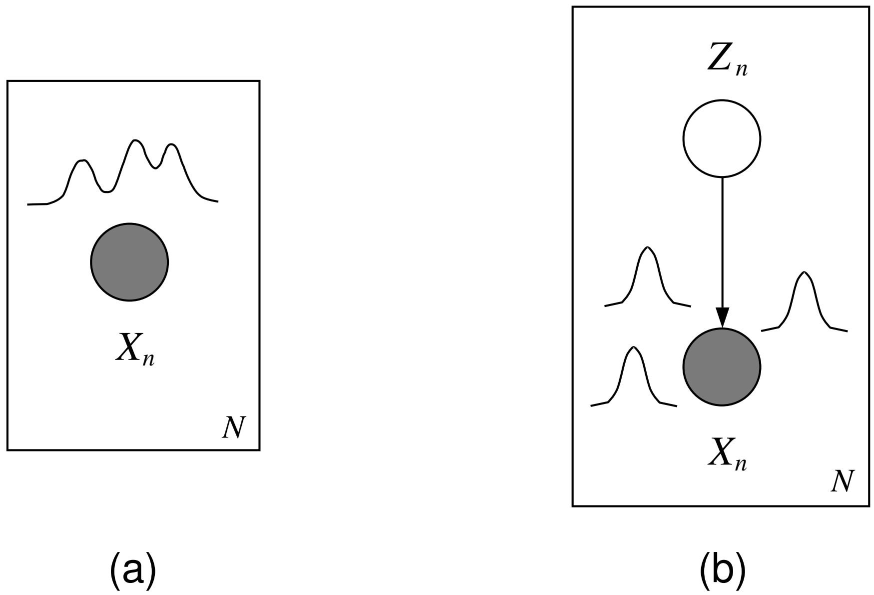

Figure 5.12: Two ways to represent a multi-modal density within the graphical model formalism. (a) The local probability model at each node is a mixture or a kernel density estimate. (b) A latent variable is used to represent mixture components explicitly; marginalizing over the latent variable yields a mixture model for the observable X_n.

There is nothing inappropriate about letting such factors be a guide, but in doing so we must be cautious about any interpretation or meaning that we attach to the model.

Summary of density estimation

Our goal in this section has not been to provide a full treatment of density estimation; indeed we have only scratched the surface of what is an extensive literature in statistics. We do hope, however, to have introduced a few key ideas—the calculation of maximum likelihood and Bayesian parameter estimates for Gaussian and multinomial densities, the use of mixture models to obtain a richer class of density models, and the distinction between parametric and nonparametric density estimation. Each of these ideas will be picked up and pursued in numerous contexts throughout the book.

5.2.2 Regression

In a *regression model* the goal is to model the dependence of a *response* or *output* variable Y on a *covariate* or *input* variable X. We capture this dependence via a conditional probability distribution $p(y \,|\, x)$. In graphical model terms, we have a two-node model in which X is the parent and Y is the child (see Figure 5.13).

One way to treat regression problems is to estimate the joint density of X and Y and to calculate

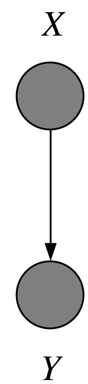

Figure 5.13: A regression model.

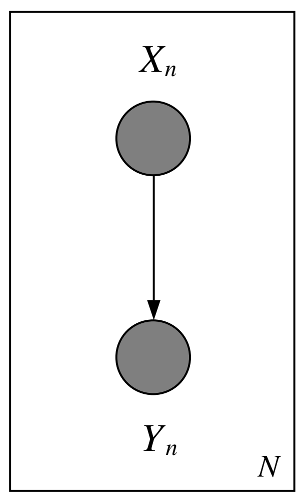

Figure 5.14: The IID regression model represented graphically.

the conditional $p(y \mid x)$ from the estimated joint. This approach forces us to model X, however, which may not be desired. Indeed, in many applications of regression, X is high-dimensional and hard to model. Moreover, the observations of X are often fixed by experimental design or another form of non-random process, and it is problematic to treat them via a simple sampling model, such as the IID model. In summary, it is necessary to develop methods appropriate to conditional densities.

Our discussion here will be brief, with a focus on basic representational issues.

We assume that we have a set of pairs of observed data, $\{(x_n, y_n); n = 1, \ldots, N\}$, where x_n is an observation of the input variable and y_n is a corresponding observation of the output variable. We again assume an independent, identical distributed (IID) sampling model for simplicity. The graphical representation of the IID regression model is shown as a plate in Figure 5.14.

Let us now consider some of the possible choices for the conditional probability model $p(y_n \mid x_n)$.

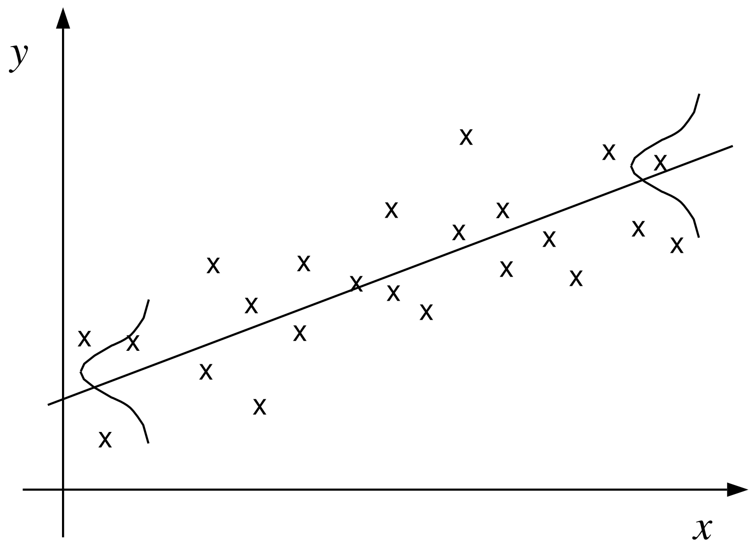

Figure 5.15: The linear regression model expresses the response variable Y in terms of the conditional mean function—the line in the figure—and input-independent random variation around the conditional mean.

As in the case of density estimation, we have a wide spectrum of possibilities, including parametric models, mixture models, and nonparametric models. We will discuss these models in detail in Chapters 6, 10, and 25, respectively, but let us sketch some of the possibilities here.

A *linear regression* model expresses Y_n as the sum of (1) a purely deterministic component that depends parametrically on x_n, and (2) a purely random component that is functionally independent of x_n:

$$Y_n = \beta^T x_n + \epsilon_n, \tag{5.55}$$

where β is a parameter vector and ϵ_n is a random variable having zero mean. Taking the conditional expectation of both sides of this equation yields $E[Y_n \,|\, x_n] = \beta^T x_n$. Thus the linear regression model expresses Y_n in terms of input-independent random variation ϵ_n around the conditional mean $\beta^T x_n$ (see Figure 5.15). The choice of the distribution of ϵ_n, which completes the specification of the model, is analogous to the choice of a density model in density estimation, and depends on the nature of Y_n. "Linear regression" generally refers to the case in which Y_n is real-valued and the distribution is taken to be $\mathcal{N}(0, \sigma^2)$. (In Chapter 8 we will be discussing "generalized linear models," which are regression models that are appropriate for other types of response variables). In the linear regression case, we have:

$$P(y_n \,|\, x_n, \theta) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{ -\frac{1}{2\sigma^2}(y_n - \beta^T x_n)^2 \right\}, \tag{5.56}$$

where for simplicity we have taken y_n to be univariate. The parameter vector θ includes β, which determines the conditional mean, and σ^2, which is the variance of ϵ_n and determines the scale of the variation around the conditional mean.

Linear regression is in fact broader than it may appear at first sight, in that the function $\beta^T x_n$ need only be linear in β and in particular may be nonlinear in x_n. Thus the model:

$$Y_n = \beta^T \phi(x_n) + \epsilon_n, \tag{5.57}$$

where $\phi(\cdot)$ is a vector-valued function of x_n, is a linear regression model. This model is a parametric model in that $\phi(\cdot)$ is fixed and our freedom in modeling the data comes only from the finite set of parameters β.

The problem of estimating the parameters of regression models is in principle no different from the corresponding estimation problem for density estimation. In the maximum likelihood approach, we form the log likelihood:

$$l(\theta; x) = \sum_{n=1}^{N} \log p(y_n \,|\, x_n, \theta), \tag{5.58}$$

take derivatives with respect to θ, set to zero and (attempt to) solve. We will discuss the issues that arise in carrying out this calculation in later chapters.

Conditional mixture models

Mixture models provide a way to move beyond the strictures of linear regression modeling. We can consider both a broader class of conditional mean functions as well as a broader class of density models for ϵ_n. Consider in particular the graphical model shown in Figure 5.16(a). We have introduced a multinomial latent variable Z_n that depends on the input X_n; moreover, the response Y_n depends on both X_n and Z_n. This graph corresponds to the following probabilistic model:

$$p(y_n \,|\, x_n, \theta) = \sum_{k=1}^{K} p(z_n^k = 1 \,|\, x_n, \theta) p(y_n \,|\, z_n^k = 1, x_n, \theta), \tag{5.59}$$

a *conditional mixture model.* Each mixture component $p(y_n \,|\, z_n^k = 1, x_n)$ corresponds to a different regression model, one for each value of k. The mixing proportions $p(z_n^k = 1 \,|\, x_n)$ "switch" among the regression models as a function of x_n. Thus, as suggested in Figure 5.16(a), the mixing proportions can be used to pick out regions of the input space where different regression functions are used. We can parameterize both the mixing proportions and the regression models and estimate both sets of parameters from data. This is a "divide-and-conquer" methodology in the regression domain. (We provide a fuller description of this model in Chapter 10).

The example in Figure 5.16(a) utilizes mixing proportions that are sharp, nearly binary functions of X_n, but it is also of interest to consider models in which these functions are smoother, allowing overlap in the component regression functions. Indeed, in the limiting case we obtain the model shown in Figure 5.16(b) in which the latent variable Z_n is independent of X_n. Here the presence of the latent variable serves only to induce multimodality in the conditional distribution $p(y_n \,|\, x_n)$. Much as in the case of density estimation, such a regression model may arise from a set of subpopulations, each characterized by a different "conditional mean."

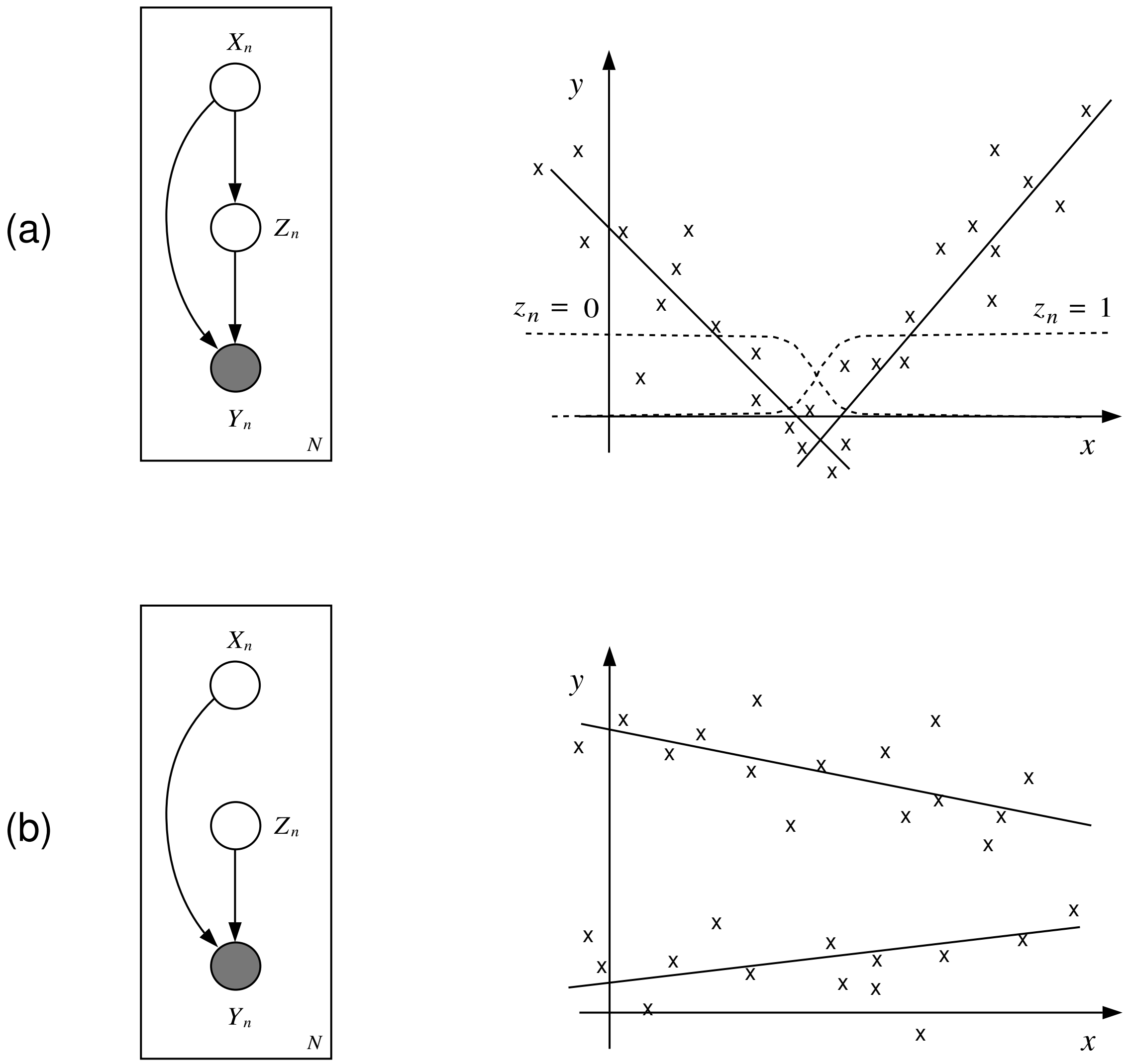

Figure 5.16: Two variants of conditional regression model. In (a), the latent variable Z_n is dependent on X_n. This corresponds to breaking up the input space into (partially overlapping) regions labeled by the values of Z_n. An example with binary Z_n is shown in the figure on the right, where the dashed line labeled by $z_n = 1$ is the probability $p(z_n = 1 \mid x_n)$, and the dashed line labeled by $z_n = 0$ is the probability $p(z_n = 0 \mid x_n)$. The two lines are the conditional means of the regressions, $p(y_n \mid z_n, x_n)$, for the two values of z_n, with the leftmost line corresponding to $z_n = 0$ and the rightmost line corresponding to $z_n = 1$. In (b), the latent variable Z_n is independent of X_n. This corresponds to total overlap of the regions corresponding to the values of Z_n and yields an input-independent mixture density for each value of x_n.

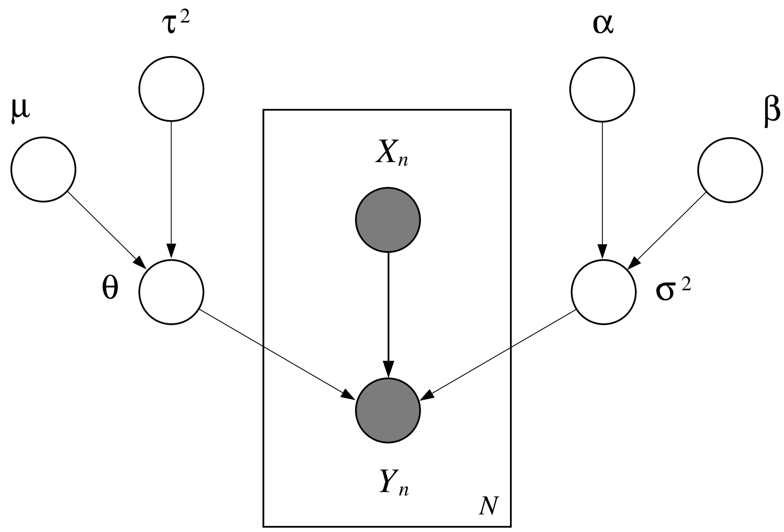

Figure 5.17: A Bayesian linear regression model. The parameter vector θ is endowed with a Gaussian prior, $\mathcal{N}(\mu, \tau^2)$. The variance σ^2 is endowed with an inverse gamma prior, $IG(\alpha, \beta)$.

The parameters of conditional mixture models can be estimated using the EM algorithm, as discussed in Chapter 10. Indeed, EM can be used quite generically for latent variable models such as those in Figure 5.16.

Nonparametric regression

Let us briefly consider the nonparametric approach to regression. While it is possible to use nonparametric methods to expand the repertoire of probability models for ϵ_n, a more common usage of nonparametric ideas involves allowing a wider class of conditional mean functions. The basic idea is to break up the input space into (possibly overlapping) regions, with one such region for each data point. Let us give an example from the class of methods known as *kernel regression*. As in kernel density estimation, let $k(x, x_n, \lambda)$ be a *kernel function* centered around the data point x_n. Denoting the conditional mean function as $f(x)$, we form an estimate as follows:

$$\hat{f}(x) = \frac{\sum_{n=1}^{N} k(x, x_n, \lambda) y_n}{\sum_{m=1}^{N} k(x, x_m, \lambda)} \tag{5.60}$$

That is, we estimate the conditional mean at x as the convex sum of the observed values y_n, where the weights in the sum are given by the normalized values of the kernel functions, one for each x_n, evaluated at x. Given that kernel functions are generally chosen to be "local," having most of their support near x_n, we see that the kernel regression estimate at x is a local average of the values y_n in the neighborhood of x.

We can once again forge a link between the mixture model approach and the nonparametric kernel regression approach. As we ask the reader to verify in Exercise **??**, taking the conditional

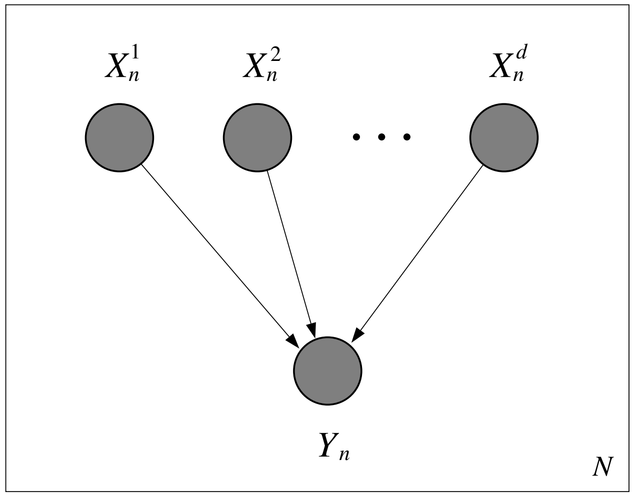

Figure 5.18: A graphical representation of the regression model in which the components of the input vector are treated as explicit nodes.

mean of Eq. (5.59) yields a weighted sum of conditional mean functions, one for each component k, where the weights are the mixing proportions $p(z_n^k = 1 \mid x_n)$. The kernel regression estimate in Eq. (5.60) can be viewed as an instance of this model, if we treat the normalized kernels $k(x, x_n, \lambda)/\sum_{m=1}^{N} k(x, x_m, \lambda)$ as mixing proportions, and the values y_n as (constant) conditional means. The same comments apply to this reduction as to the analogous reduction in the case of density estimation. In particular, as N increases, the number of components K in a parametric conditional mixture model generally remain fixed, whereas the number of kernels in the kernel regression model grow. We can, however, consider *conditional mixture sieves*, and obtain a nonparametric variant of a mixture model.

Bayesian approaches to regression

All of the models that we have considered in this section can be treated via Bayesian methods, where we endow the parameters (or entire conditional mean functions) with prior distributions. We then invoke Bayes rule to calculate posterior distributions. Figure 5.17 illustrates one such Bayesian regression model.

Remarks

Let us make one final remark regarding the graphical representation of regression models. Note that in this section we have treated the input variables X_n as single nodes, not availing ourselves of the opportunity to represent the components of these vector-valued variables as separate nodes (see Figure 5.18). This is consistent with our treatment of X_n as fixed variables to be conditioned on; representing the components as separate nodes would imply marginal independence between the components, an assumption that we may or may not wish to make. It is important to note,

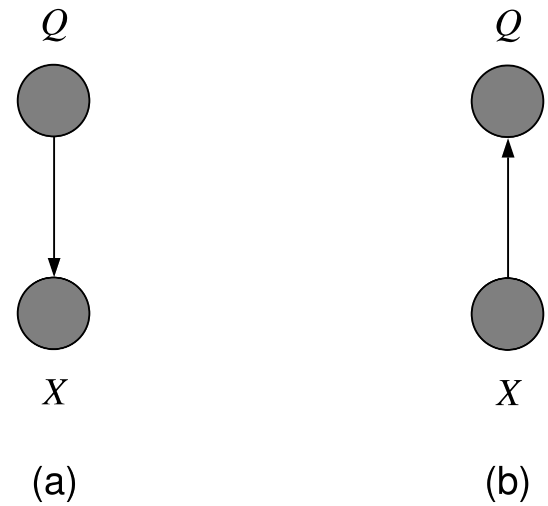

Figure 5.19: (a) The generative approach to classification represented as a graphical model. Fitting the model requires estimating the marginal probability $p(q)$ and the conditional probability $p(x \mid q)$. (b) The discriminative approach to classification represented as a graphical model. Fitting the model requires estimating the conditional probability $p(q \mid x)$.

however, that regression methods are agnostic regarding modeling assumptions about the conditioning variables. Regression methods form an estimate of $p(y \mid x)$ and this conditional density can be composed with an estimate of $p(x)$ to obtain an estimate of the joint. This allows us to use regression models as components of larger models. In particular, in the context of a graphical model in which a node A has multiple parents $B_1, B_2, \ldots, B_k$, we are free to use regression methods to represent $p(A \mid B_1, B_2, \ldots, B_k)$, regardless of the modeling assumptions made regarding the nodes B_i. Indeed each of the B_i may themselves be modeled in terms of regressions on variables further "upstream."

5.2.3 Classification

Classification problems are related to regression problems in that they involve pairs of variables. The distinguishing feature of classification problems is that the response variable ranges over a finite set, a seemingly minor issue that has important implications.

In classification we often refer to the covariate X as a *feature vector*, and the corresponding discrete response, which we denote by Q, as a *class label*. We typically view the feature vectors as descriptions of objects, and the goal is to label the objects, i.e., to classify the objects into one of a finite set of categories.

There are two basic approaches to classification problems, which can be interpreted graphically in terms of the direction of the edge between X and Q. The first approach, which we will refer to as *generative*, is based on the graphical model shown in Figure 5.19(a), in which there is an arrow from Q to X. This approach is closely related to density estimation—for each value of the discrete

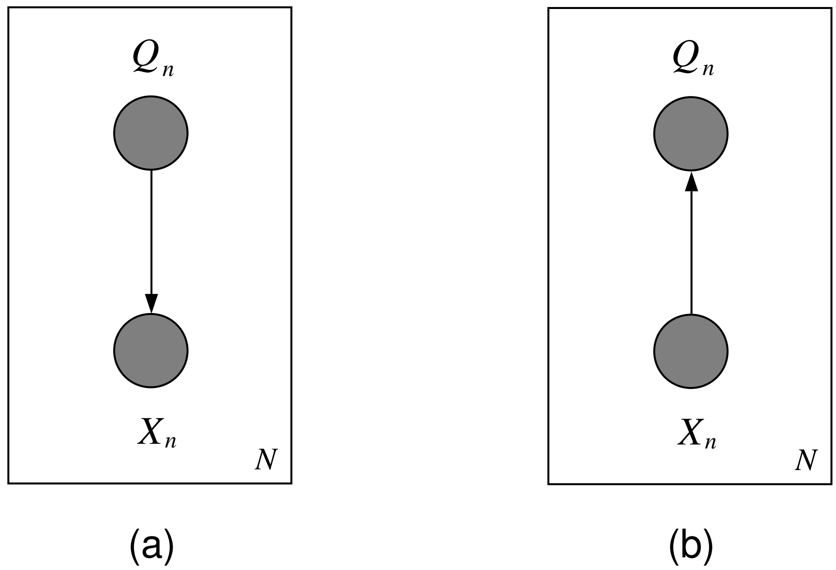

Figure 5.20: The IID classification models for the (a) generative approach and (b) discriminative approach.

variable Q we have a density, $p(x \mid q)$, which we refer to as a *class-conditional density.* We also require the marginal probability $p(q)$, which we refer to as the *prior probability* of the class Q (it is the probability of the class before a feature vector X is observed). This marginal probability is required if we are going to be able to "invert the arrow" and compute $p(q \mid x)$—the *posterior probability* of class Q.

The second approach to classification, which we refer to as *discriminative*, is closely related to regression. Here we represent the relationship between the feature vectors and the labels in terms of an arrow from X to Q (see Figure 5.19(b)). That is, we represent the relationship in terms of the conditional probability $p(q \mid x)$. When classifying an object we simply plug the corresponding feature vector x into the conditional probability and calculate $p(q \mid x)$. Performing this calculation, which tells us which class label has the highest probability, makes no reference to the marginal probability $p(x)$ and, as in regression, we may wish to abstain from incorporating such a marginal into the model.

As in regression, we have a set of data pairs $\{(x_n, q_n) : n = 1, \ldots, N\}$, assumed IID for simplicity. The representations of the classification problem as plates are shown in Figure 5.20.

Once again we postpone a general presentation of particular representations for the conditional probabilities in classification problems until later chapters. But let us briefly discuss a canonical example that will illustrate some typical representational choices, as well as illustrate some of the relationships between the generative and the discriminative approaches to classification. This example and several others will be developed in considerably greater detail in later chapters.

We specialize to two classes. Let us choose Gaussian class-conditional densities with equal covariance matrices for the two classes. An example of these densities (where we have assumed equal

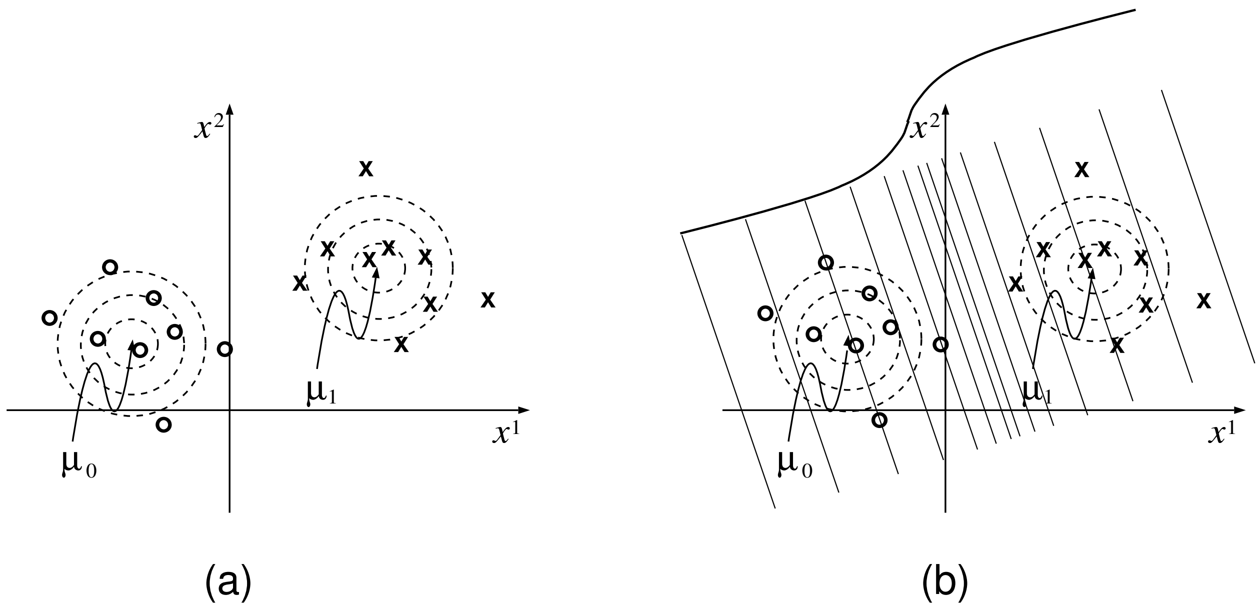

Figure 5.21: (a) Contour plots and samples from two Gaussian class-conditional densities for two-dimensional feature vectors $x_n = (x_n^1, x_n^2)$. The Gaussians have means μ_0 and μ_1 for class $q_n = 0$ and $q_n = 1$, respectively, and equal covariance matrices. (b) The solid lines are the contours of the posterior probability, $p(q_n = 1 \mid x_n)$. In the direction orthogonal to the linear contours, the posterior probability is a monotonically increasing function given by (Eq. (5.61)). This function is sketched at the top of the figure.

class priors) is shown in Figure 5.21(a). We use Bayes rule to compute the posterior probability that a given feature vector x_n belongs to class $q_n = 1$. Intuitively, we expect to obtain a ramp-like function which is zero in the vicinity of the class $q_n = 0$, increases to one-half in the region between the two classes, and approaches one in the vicinity of the class $q_n = 1$. This posterior probability function is shown in Figure 5.21(b), where indeed we see the ramp-like shape.

Analytically, as we show in Chapter 7, for Gaussian class-conditional densities the ramp-like posterior probability turns out to be the *logistic function*:

$$p(q_n = 1 \mid x_n) = \frac{1}{1 + e^{-\theta^T x_n}}, \tag{5.61}$$

where θ is a parameter vector that depends on the particular choices of means and covariances for the class-conditional densities, as well as the class priors. The inner product between θ and x_n is a projection operation that is responsible for the linear contours that we see in Figure 5.21(b).

Given these parametric forms for the class-conditional densities (the Gaussian densities) and the posterior probability (the logistic function), we must specify how to estimate the parameters based on the data. It is here that the generative and discriminative approaches begin to diverge. From the generative point of view, the problem is that of estimating the means and covariances of the Gaussian class-conditional densities, as well as the class priors. These are density estimation

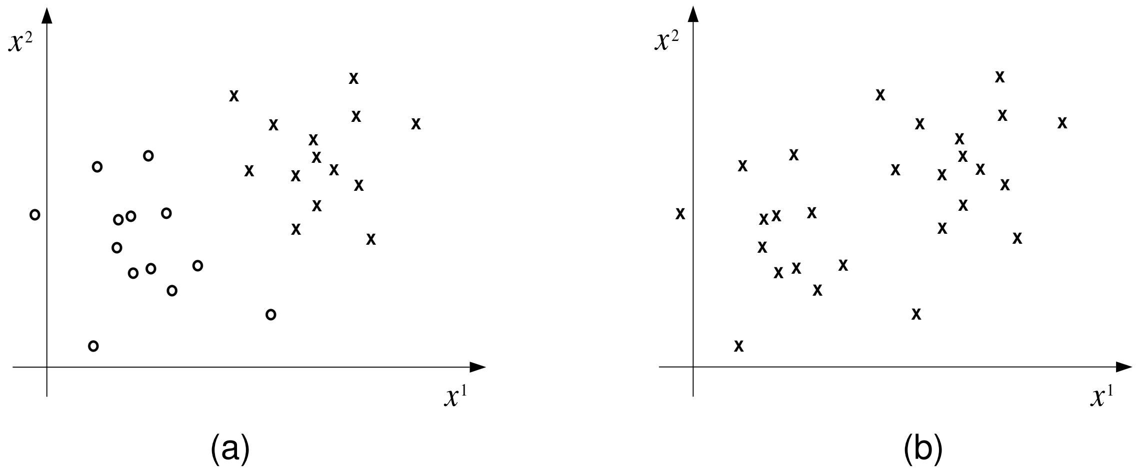

Figure 5.22: (a) A classification problem with the class $q_n = 0$ labeled with a "0" and the class $q_n = 1$ labeled with a "x". (b) The same feature vectors x_n as in (a), but with the labels erased.

problems, and the machinery of Section 5.2.1 is invoked to solve them. With these density estimates in hand, we derive an estimate of θ and thereby calculate an estimate of the posterior probability. Essentially, the goal is to model the classes, without any direct attempt to discriminate between the classes.

In the discriminative approach, on the other hand, the logistic function is the central object of analysis. Indeed, in Chapter 7, we describe a regression-like method for estimating θ directly from data, without making reference to the means and covariances of an underlying generative model. Intuitively, this method can be viewed as an attempt to orient and position the ramp-like posterior probability in Figure 5.21(b) so as to assign a posterior probability that is near zero to the points x_n having label $q_n = 0$, and a posterior probability near one to the points x_n having label $q_n = 1$. Essentially, the goal is to discriminate between the classes, without any direct attempt to model the classes.

More generally, in a discriminative approach to classification we are not restricted to the logistic function, or to any other function that is derived from a generative model. Rather we can choose functions whose contours appear to provide a natural characterization of boundaries between classes. On the other hand, it may not always be apparent how to choose such functions, and in such cases we may prefer to take advantage of the generative approach, in which the boundaries arise implicitly via Bayes rule. In general, both the discriminative and the generative approaches are important tools to have in a modeling toolbox.

Mixture models revisited

Suppose we consider a classification problem in which none of the class labels are present. Is this a sensible problem to pose? What can one possibly learn from unlabeled data, particularly data that are completely unlabeled?

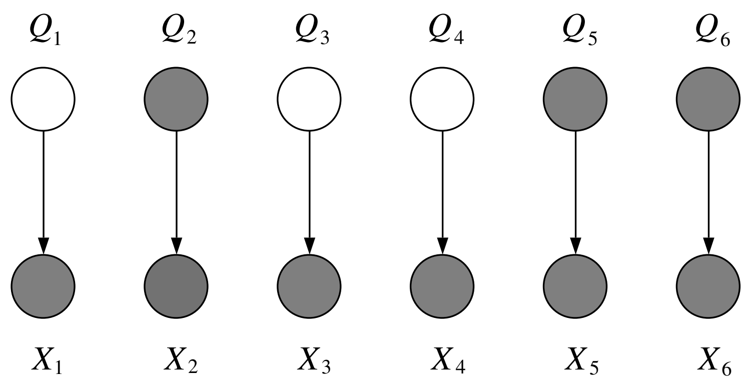

Figure 5.23: A model for partially labeled data in which the feature vectors x_2, x_5 and x_6 are labeled and the other feature vectors are unlabeled.

Consider Figure 5.22(a), where we have depicted a typical classification problem with two classes. Now consider Figure 5.22(b), where we have retained the feature vectors x_n, but erased the labels q_n. As this latter plot makes clear, although the labels are missing, there is still substantial statistical structure in the problem. Rather than solving a classification problem, we can solve a *clustering* problem, making explicit the fact that the data appear to fall into two clusters and assigning feature vectors to clusters.

In fact we have already solved this problem. The mixture model approach to density estimation discussed in Section 5.2.1 treats the density in terms of a set of underlying "subpopulations" labeled by a latent variable Z. The inferential calculation $p(z_n \mid x_n)$ given in Eq. (5.52) explicitly calculates the probability that the feature vector x_n belongs to each of the subpopulations.

The relationship between classification and mixture models is also clarified by comparing the "generative" graphical model in Figure 5.19(b) and the mixture model in Figure 5.9(a). These are the same graphical model—the only difference is the shading, corresponding to the assumption that the labels Q_n are observed in classification whereas the latent variables Z_n are unobserved in mixture modeling. In the setting of unlabeled data the generative classification model becomes identical to a mixture model.

In a more general setting we may have a "partially labeled" case in which the labels Q_n are observed for some data points and unobserved for other data points. This situation is represented graphically in Figure 5.23. We will be able to treat the problem of estimation in this case using the EM algorithm; indeed this "partially labeled" case requires no additional machinery beyond that already required for the mixture model.

It is common to refer to classification and regression models as "supervised learning" models and to refer to density estimation models as "unsupervised learning" models. In the comparison between mixture models and classification models just discussed, the distinction refers to the observation of the labels Q_n; one says that the labels in classification are provided by a "supervisor." While this terminology can be useful in making broad distinctions between models, it is our view that the terminology does not reflect a fundamental underlying distinction and we will tend to avoid its use in this book. It is our feeling that many models are neither entirely "supervised" nor entirely "unsupervised," and invoking the distinction often forces us to group together methods that have little in common as well as to separate methods that are closely related. We feel that a better way to understand relationships between models is to make them explicit as graphs. Models can then be compared in terms of graphical features such as which variables are considered latent and which observed, the directionalities of arcs that are used to represent conditional relationships, and the presence or absence of particular structural motifs.

Remarks

We have already indicated a relationship between mixture models and classification, but there are other roles for mixture models in the classification setting. In particular, we can use mixtures as class-conditional densities in the generative approach to classification, just as we used mixture models in the density estimation setting to extend the range of models that we considered. Also, in the context of the discriminative approach to classification, we can use conditional mixtures to represent the posterior probability $p(q \mid x)$, breaking this function into overlapping pieces, much as we did with the conditional mean in the case of regression.

Similarly, nonparametric methods have many roles to play in classification models. We can either extend the generative approach to allow nonparametric estimates of the class-conditional densities, or extend the discriminative approach to allow nonparametric estimates of the posterior probability.

Finally, there are once again Bayesian approaches in all of these cases. From a graphical point of view, these Bayesian approaches essentially involve making the parameters explicit as nodes, and using hyperparameters to express prior probability distributions on these nodes.

5.3 Model selection and model averaging

Thus far we have assumed that a specific model has been chosen in advance and we have focused on representing the model graphically and estimating its parameters. In some cases this assumption is reasonable—the model is determined by the problem and there is no need to consider data-driven approaches to choosing the model. More commonly, however, we wish to use the data to make informed choices regarding the model. We present a brief discussion of this problem—known as the *model selection* problem—in this section, anticipating our more detailed presentation in Chapter 26.

We consider a class $\mathcal{M}$ of possible models, letting $m \in \mathcal{M}$ denote a specific model in this family. We also augment our earlier notation to include explicit reference to the model; thus, $p(x \mid \theta, m)$ refers to the probability model for the random variable X, given a specific model and a specific choice of parameters for that model.[4] Also, in the Bayesian approach, $p(\theta \mid m)$ refers to the prior probability that we attach to the parameters θ, and $p(\theta \mid x, m)$ refers to the corresponding posterior. We wish to develop methods for choosing m based on the data x.

Let us begin with the Bayesian approach. Recall that unknowns are treated as random variables in the Bayesian approach; thus we introduce a random variable M to denote the model. The range

[4] For simplicity we use the same notation θ to represent the parameters in each of the models; in general we could allow the parameterization to vary with m.

of M is $\mathcal{M}$, and m denotes a realization of M. The goal of Bayesian analysis is to calculate the posterior probability of M, conditioning on the data x:

$$p(m \mid x) = \frac{p(x \mid m)p(m)}{p(x)}. \tag{5.62}$$

Note two important features of this equation. First, as in the case of parameter estimation, we require a prior probability; in particular, we need to specify the prior probability $p(m)$ of the model m. Second, note the absence of explicit mention of the parameter θ. The probabilities needed for Bayesian model selection are *marginal* probabilities.

Let us consider this latter issue in more detail. The calculation of the posterior probability in Eq. (5.62) requires the probability $p(x \mid m)$, a conditional probability that is referred to as the *marginal likelihood*. We compute the marginal likelihood from the likelihood by integrating over the parameters:

$$p(x \mid m) = \int p(x, \theta \mid m) d\theta \tag{5.63}$$

$$= \int p(x \mid \theta, m) p(\theta \mid m) d\theta, \tag{5.64}$$

where the prior probability $p(\theta \mid m)$ plays the role of a weighting function. Multiplying the marginal likelihood by the prior probability $p(m)$ yields the desired posterior $p(m \mid x)$, up to the normalization factor $p(x)$.

If we wish to use the posterior to *select* a model, then we must collapse the posterior to a point. As in the case of parameter estimation, various possibilities present themselves; in particular, a popular approach is to pick the model that maximizes the posterior probability. An advantage of this approach is that it obviates the need to calculate the normalization constant $p(x)$.

More generally, however, the Bayesian approach aims to use the entire posterior. To illustrate the use of the model posterior, let us again consider the problem of prediction. Taking X_{new} to be conditionally independent of X, given θ and m, we have:

$$p(x_{new} \mid x) = \int\int p(x_{new}, \theta, m \mid x) d\theta dm \tag{5.65}$$

$$= \int\int p(x_{new} \mid \theta, m) p(\theta, m \mid x) d\theta dm \tag{5.66}$$

$$= \int\int p(x_{new} \mid \theta, m) p(\theta \mid x, m) p(m \mid x) d\theta dm. \tag{5.67}$$

From this latter equation, we see that a full Bayesian approach to prediction requires two posterior probabilities: the model posterior $p(m \mid x)$ from Eq. (5.62) and the parameter posterior $p(\theta \mid x, m)$ from Eq. (5.1). These posteriors can be viewed as "weights" for the prediction $p(x_{new} \mid \theta, m)$; the total prediction can be viewed as a "weighted prediction." This approach to prediction is referred to as *model averaging.*

It should be acknowledged that it is a rare circumstance in which the integrals in Eq. (5.64) and Eq. (5.67) can be done exactly, and Bayesian model averaging and model selection generally involve making approximations. We will discuss some of these approximations in Chapter 26.

Frequentist approaches to model selection avoid the use of prior probabilities and Bayes rule. Rather, one considers various model selection *procedures*, and evaluates these procedures in terms of various frequentist criteria. For example, one could consider a scenario in which the true probability density is assumed to lie within the class $\mathcal{M}$, and ask that a model selection procedure pick the true model with high frequency. Alternatively, one could ask that the procedure select the "best" model in $\mathcal{M}$, where "best" is defined in terms of a measure such as the Kullback-Leibler divergence between a model and the true probability density.

It is important to understand that maximum likelihood itself cannot be used as a model selection procedure. Augmenting a model with additional parameters cannot decrease the likelihood, and thus maximum likelihood will prefer more complex models. More complex models may of course be better than simpler models, if they provide access to probability densities that are significantly closer to the true density, but at some point there are diminishing returns and more complex models principally provide access to additional poor models. The fact that we have to estimate parameters implies that with some probability we will select one of the poor models. Thus the "variance" introduced by the parameter estimation process can lead to poorer performance with a more complex model. Maximum likelihood is unable to address this "overfitting" phenomenon.

One approach to frequentist model selection is to "correct" maximum likelihood to account for the variance due to parameter estimation. The AIC method to be discussed in Chapter 26 exemplifies this approach. An alternative approach, also discussed in Chapter 26, is the *cross-validation* idea, in which the data are partitioned in subsets, with one subset used to fit parameters for various models, and another subset used to evaluate the resulting models.

5.4 Appendix A

In this section we calculate the posterior density of μ in the univariate Gaussian density estimation problem. Recall that the joint probability of x and μ is given by:

$$p(x, \mu) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2\right\} \frac{1}{(2\pi\tau^2)^{1/2}} \exp\{-\frac{1}{2\tau^2}(\mu - \mu_0)^2\}, \tag{5.68}$$

and the problem is to normalize this joint probability.

Let us focus on the terms involving μ, treating all other terms as "constants" and dropping them throughout. We have:

$$\begin{aligned}
p(x, \mu) &\propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}\left(x_n^2 - 2x_n\mu + \mu^2\right) - \frac{1}{2\tau^2}\left(\mu^2 - 2\mu_0\mu + \mu_0^2\right)\right\} &(5.69)\\
&= \exp\left\{-\frac{1}{2}\sum_{n=1}^{N}\left[\frac{1}{\sigma^2}\left(x_n^2 - 2x_n\mu + \mu^2\right) + \frac{1}{\tau^2}\left(\frac{\mu^2}{N} - 2\frac{\mu_0\mu}{N} + \frac{\mu_0^2}{N}\right)\right]\right\} &(5.70)\\
&= \exp\left\{-\frac{1}{2}\sum_{n=1}^{N}\left[\left(\frac{1}{\sigma^2} + \frac{1}{N\tau^2}\right)\mu^2 - 2\left(\frac{x_n}{\sigma^2} + \frac{\mu_0}{N\tau^2}\right)\mu + C\right]\right\} &(5.71)
\end{aligned}$$

$$\propto \exp\left\{-\frac{1}{2}\left[\left(\frac{N}{\sigma^2}+\frac{1}{\tau^2}\right)\mu^2 - 2\left(\frac{N\bar{x}}{\sigma^2}+\frac{\mu_0}{\tau^2}\right)\mu\right]\right\} \tag{5.72}$$

$$\propto \exp\left\{-\frac{1}{2}\left(\frac{N}{\sigma^2}+\frac{1}{\tau^2}\right)\left[\mu^2 - 2\left(\frac{N}{\sigma^2}+\frac{1}{\tau^2}\right)^{-1}\left(\frac{N\bar{x}}{\sigma^2}+\frac{\mu_0}{\tau^2}\right)\mu\right]\right\} \tag{5.73}$$

$$= \exp\left\{-\frac{1}{2\tilde{\sigma}^2}\left[\mu^2 - 2\tilde{\mu}\mu\right]\right\}, \tag{5.74}$$

where

$$\tilde{\sigma}^2 = \left(\frac{N}{\sigma^2}+\frac{1}{\tau^2}\right)^{-1} \tag{5.75}$$

and

$$\tilde{\mu} = \frac{N/\sigma^2}{N/\sigma^2+1/\tau^2}\ \bar{x} + \frac{1/\tau^2}{N/\sigma^2+1/\tau^2}\ \mu_0, \tag{5.76}$$

This identifies the posterior as a Gaussian distribution with mean $\tilde{\mu}$ and variance $\tilde{\sigma}^2$.

5.5 Historical remarks and bibliography