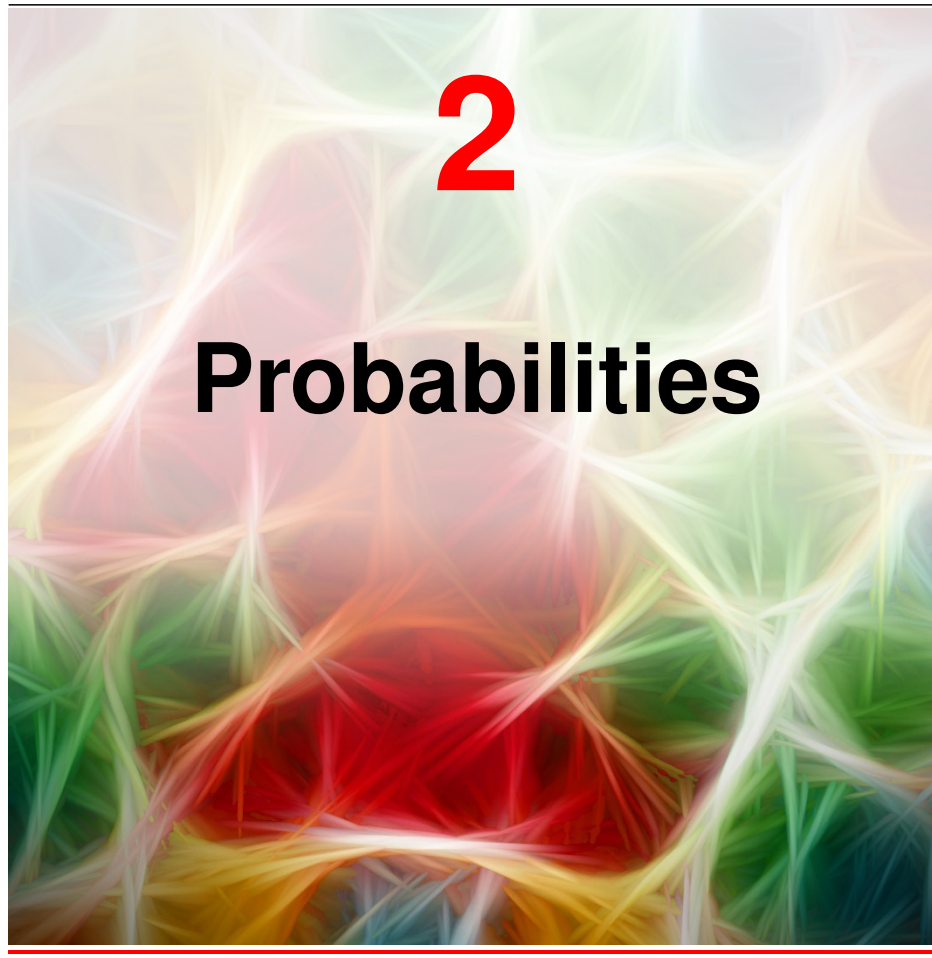

2 Probabilities

In almost every application of machine learning we have to deal with uncertainty. For example, a system that classifies images of skin lesions as benign or malignant can never in practice achieve perfect accuracy. We can distinguish between two kinds of uncertainty. The first is *epistemic uncertainty* (derived from the Greek word episteme meaning knowledge), sometimes called *systematic uncertainty*. It arises because we only get to see data sets of finite size. As we observe more data, for instance more examples of benign and malignant skin lesion images, we are better able to predict the class of a new example. However, even with an infinitely large data set, we would still not be able to achieve perfect accuracy due to the second kind of uncertainty known as *aleatoric uncertainty*, also called *intrinsic* or *stochastic* uncertainty, or sometimes simply called *noise*. Generally speaking, the noise arises because we are able to observe only partial information about the world, and therefore, one way to reduce this source of uncertainty is to gather different kinds of data. This is illustrated

C. M. Bishop, H. Bishop, *Deep Learning*, https://doi.org/10.1007/978-3-031-45468-4_2

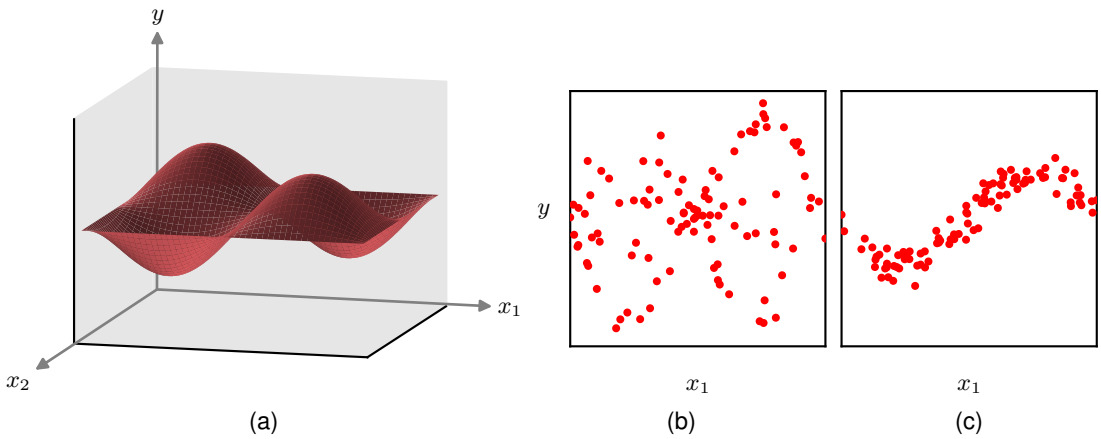

Figure 2.1 An extension of the simple sine curve regression problem to two dimensions. (a) A plot of the function $y(x_1, x_2) = \sin(2\pi x_1)\sin(2\pi x_2)$. Data is generated by selecting values for x_1 and x_2, computing the corresponding value of $y(x_1, x_2)$, and then adding Gaussian noise. (b) Plot of 100 data points in which x_2 is unobserved showing high levels of noise. (c) Plot of 100 data points in which x_2 is fixed to the value $x_2 = \frac{\pi}{2}$, simulating the effect of being able to measure x_2 as well as x_1, showing much lower levels of noise.

Section 1.2

using an extension of the sine curve example to two dimensions in Figure 2.1.

As a practical example of this, a biopsy sample of the skin lesion is much more informative than the image alone and might greatly improve the accuracy with which we can determine if a new lesion is malignant. Given both the image and the biopsy data, the intrinsic uncertainty might be very small, and by collecting a large training data set, we may be able to reduce the systematic uncertainty to a low level and thereby make predictions of the class of the lesion with high accuracy.

Both kinds of uncertainty can be handled using the framework of *probability theory*, which provides a consistent paradigm for the quantification and manipulation of uncertainty and therefore forms one of the central foundations for machine learning. We will see that probabilities are governed by two simple formulae known as the *sum rule* and the *product rule*. When coupled with *decision theory*, these rules allow us, at least in principle, to make optimal predictions given all the information available to us, even though that information may be incomplete or ambiguous.

Section 2.1
Section 5.2

The concept of probability is often introduced in terms of frequencies of repeatable events. Consider, for example, the bent coin shown in Figure 2.2, and suppose that the shape of the coin is such that if it is flipped a large number of times, it lands concave side up 60% of the time, and therefore lands convex side up 40% of the time. We say that the *probability* of landing concave side up is 60% or 0.6. Strictly, the probability is defined in the limit of an infinite number of 'trials' or coin flips in this case. Because the coin must land either concave side up or convex side up, these probabilities add to 100% or 1.0. This definition of probability in terms of the frequency of repeatable events is the basis for the *frequentist* view of statistics.

Now suppose that, although we know that the probability that the coin will land concave side up is 0.6, we are not allowed to look at the coin itself and we do not

Figure 2.2 Probability can be viewed either as a frequency associated with a repeatable event or as a quantification of uncertainty. A bent coin can be used to illustrate the difference, as discussed in the text.

know which side is heads and which is tails. If asked to take a bet on whether the coin will land heads or tails when flipped, then symmetry suggests that our bet should be based on the assumption that the probability of seeing heads is 0.5, and indeed a more careful analysis shows that, in the absence of any additional information, this is indeed the rational choice. Here we are using probabilities in a more general sense than simply the frequency of events. Whether the convex side of the coin is heads or tails is not itself a repeatable event, it is simply unknown. The use of probability as a quantification of uncertainty is the *Bayesian* perspective and is more general in that it includes frequentist probability as a special case. We can learn about which side of the coin is heads if we are given results from a sequence of coin flips by making use of Bayesian reasoning. The more results we observe, the lower our uncertainty as to which side of the coin is which.

Section 2.6

Exercise 2.40

Having introduced the concept of probability informally, we turn now to a more detailed exploration of probabilities and discuss how to use them quantitatively. Concepts developed in the remainder of this chapter will form a core foundation for many of the topics discussed throughout the book.

2.1. The Rules of Probability

In this section we will derive two simple rules that govern the behaviour of probabilities. However, in spite of their apparent simplicity, these rules will prove to be very powerful and widely applicable. We will motivate the rules of probability by first introducing a simple example.

2.1.1 A medical screening example

Consider the problem of screening a population in order to provide early detection of cancer, and let us suppose that 1% of the population actually have cancer. Ideally our test for cancer would give a positive result for anyone who has cancer and a negative result for anyone who does not. However, tests are not perfect, so we will suppose that when the test is given to people who are free of cancer, 3% of them will test positive. These are known as *false positives*. Similarly, when the test is given to people who do have cancer, 10% of them will test negative. These are called *false negatives*. The various error rates are illustrated in Figure 2.3.

Given this information, we might ask the following questions: (1) 'If we screen the population, what is the probability that someone will test positive?', (2) 'If some-

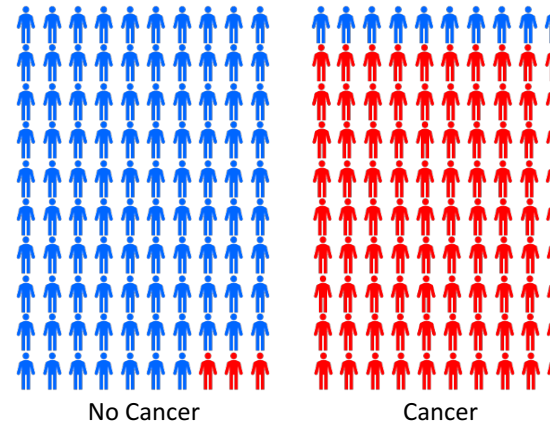

Figure 2.3 Illustration of the accuracy of a cancer test. Out of every hundred people taking the test who do not have cancer, shown on the left, on average three will test positive. For those who have cancer, shown on the right, out of every hundred people taking the test, on average 90 will test positive.

one receives a positive test result, what is the probability that they actually have cancer?'. We could answer such questions by working through the cancer screening case in detail. Instead, however, we will pause our discussion of this specific example and first derive the general rules of probability, known as the *sum rule of probability* and the *product rule*. We will then illustrate the use of these rules by answering our two questions.

2.1.2 The sum and product rules

To derive the rules of probability, consider the slightly more general example shown in Figure 2.4 involving two variables X and Y. In our cancer example, X could represent the presence or absence of cancer, and Y could be a variable denoting the outcome of the test. Because the values of these variables can vary from one person to another in a way that is generally unknown, they are called *random variables* or *stochastic variables*. We will suppose that X can take any of the values x_i where $i = 1, \ldots, L$ and that Y can take the values y_j where $j = 1, \ldots, M$. Consider a total of N trials in which we sample both of the variables X and Y, and let the number of such trials in which $X = x_i$ and $Y = y_j$ be n_{ij}. Also, let the number of trials in which X takes the value x_i (irrespective of the value that Y takes) be denoted by c_i, and similarly let the number of trials in which Y takes the value y_j be denoted by r_j.

The probability that X will take the value x_i and Y will take the value y_j is written $p(X = x_i, Y = y_j)$ and is called the *joint* probability of $X = x_i$ and $Y = y_j$. It is given by the number of points falling in the cell i,j as a fraction of the total number of points, and hence

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}. \tag{2.1}$$

Here we are implicitly considering the limit $N \to \infty$. Similarly, the probability that X takes the value x_i irrespective of the value of Y is written as $p(X = x_i)$ and is

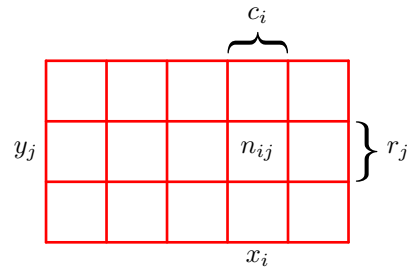

Figure 2.4 We can derive the sum and product rules of probability by considering a random variable X, which takes the values $\{x_i\}$ where $i = 1, \ldots, L$, and a second random variable Y, which takes the values $\{y_j\}$ where $j = 1, \ldots, M$. In this illustration, we have $L = 5$ and $M = 3$. If we consider the total number N of instances of these variables, then we denote the number of instances where $X = x_i$ and $Y = y_j$ by n_{ij}, which is the number of instances in the corresponding cell of the array. The number of instances in column i, corresponding to $X = x_i$, is denoted by c_i, and the number of instances in row j, corresponding to $Y = y_j$, is denoted by r_j.

given by the fraction of the total number of points that fall in column i, so that

$$p(X = x_i) = \frac{c_i}{N}. \tag{2.2}$$

Since $\sum_i c_i = N$, we see that

$$\sum_{i=1}^{L} p(X = x_i) = 1 \tag{2.3}$$

and, hence, the probabilities sum to one as required. Because the number of instances in column i in Figure 2.4 is just the sum of the number of instances in each cell of that column, we have $c_i = \sum_j n_{ij}$ and therefore, from (2.1) and (2.2), we have

$$p(X = x_i) = \sum_{j=1}^{M} p(X = x_i, Y = y_j), \tag{2.4}$$

which is the *sum rule* of probability. Note that $p(X = x_i)$ is sometimes called the *marginal* probability and is obtained by marginalizing, or summing out, the other variables (in this case Y).

If we consider only those instances for which $X = x_i$, then the fraction of such instances for which $Y = y_j$ is written $p(Y = y_j|X = x_i)$ and is called the *conditional* probability of $Y = y_j$ given $X = x_i$. It is obtained by finding the fraction of those points in column i that fall in cell i,j and, hence, is given by

$$p(Y = y_j|X = x_i) = \frac{n_{ij}}{c_i}. \tag{2.5}$$

Summing both sides over j and using $\sum_j n_{ij} = c_i$, we obtain

$$\sum_{j=1}^{M} p(Y = y_j|X = x_i) = 1 \tag{2.6}$$

showing that the conditional probabilities are correctly normalized. From (2.1), (2.2), and (2.5), we can then derive the following relationship:

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i} \cdot \frac{c_i}{N} = p(Y = y_j|X = x_i)p(X = x_i), \tag{2.7}$$

which is the *product rule* of probability.

So far, we have been quite careful to make a distinction between a random variable, such as X, and the values that the random variable can take, for example x_i. Thus, the probability that X takes the value x_i is denoted $p(X = x_i)$. Although this helps to avoid ambiguity, it leads to a rather cumbersome notation, and in many cases there will be no need for such pedantry. Instead, we may simply write $p(X)$ to denote a distribution over the random variable X, or $p(x_i)$ to denote the distribution evaluated for the particular value x_i, provided that the interpretation is clear from the context.

With this more compact notation, we can write the two fundamental rules of probability theory in the following form:

sum rule $$p(X) = \sum_Y p(X, Y) \tag{2.8}$$

product rule $$p(X, Y) = p(Y|X)p(X). \tag{2.9}$$

Here $p(X, Y)$ is a joint probability and is verbalized as 'the probability of X *and* Y'. Similarly, the quantity $p(Y|X)$ is a conditional probability and is verbalized as 'the probability of Y *given* X'. Finally, the quantity $p(X)$ is a marginal probability and is simply 'the probability of X'. These two simple rules form the basis for all of the probabilistic machinery that we will use throughout this book.

2.1.3 Bayes' theorem

From the product rule, together with the symmetry property $p(X, Y) = p(Y, X)$, we immediately obtain the following relationship between conditional probabilities:

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}, \tag{2.10}$$

which is called *Bayes' theorem* and which plays an important role in machine learning. Note how Bayes' theorem relates the conditional distribution $p(Y|X)$ on the left-hand side of the equation, to the 'reversed' conditional distribution $p(X|Y)$ on the right-hand side. Using the sum rule, the denominator in Bayes' theorem can be expressed in terms of the quantities appearing in the numerator:

$$p(X) = \sum_Y p(X|Y)p(Y). \tag{2.11}$$

Thus, we can view the denominator in Bayes' theorem as being the normalization constant required to ensure that the sum over the conditional probability distribution on the left-hand side of (2.10) over all values of Y equals one.

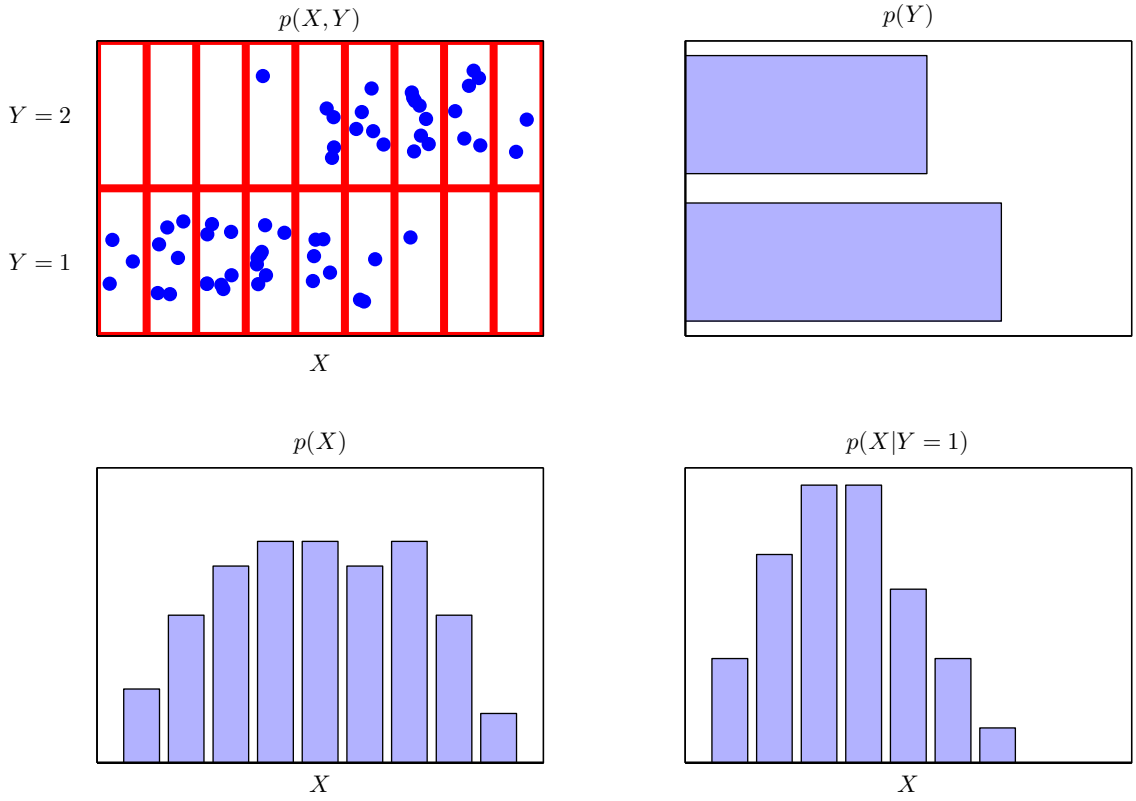

Figure 2.5 An illustration of a distribution over two variables, X, which takes nine possible values, and Y, which takes two possible values. The top left figure shows a sample of 60 points drawn from a joint probability distribution over these variables. The remaining figures show histogram estimates of the marginal distributions $p(X)$ and $p(Y)$, as well as the conditional distribution $p(X|Y = 1)$ corresponding to the bottom row in the top left figure.

Section 3.5.1

In Figure 2.5, we show a simple example involving a joint distribution over two variables to illustrate the concept of marginal and conditional distributions. Here a finite sample of $N = 60$ data points has been drawn from the joint distribution and is shown in the top left. In the top right is a histogram of the fractions of data points having each of the two values of Y. From the definition of probability, these fractions would equal the corresponding probabilities $p(Y)$ in the limit when the sample size $N \to \infty$. We can view the histogram as a simple way to model a probability distribution given only a finite number of points drawn from that distribution. The remaining two plots in Figure 2.5 show the corresponding histogram estimates of $p(X)$ and $p(X|Y = 1)$.

2.1.4 Medical screening revisited

Let us now return to our cancer screening example and apply the sum and product rules of probability to answer our two questions. For clarity, when working through this example, we will once again be explicit about distinguishing between the random variables and their instantiations. We will denote the presence or absence of cancer by the variable C, which can take two values: $C = 0$ corresponds to 'no cancer' and $C = 1$ corresponds to 'cancer'. We have assumed that one person in a hundred in the population has cancer, and so we have

$$p(C = 1) = 1/100 \tag{2.12}$$
$$p(C = 0) = 99/100, \tag{2.13}$$

respectively. Note that these satisfy $p(C = 0) + p(C = 1) = 1$.

Now let us introduce a second random variable T representing the outcome of a screening test, where $T = 1$ denotes a positive result, indicative of cancer, and $T = 0$ a negative result, indicative of the absence of cancer. As illustrated in Figure 2.3, we know that for those who have cancer the probability of a positive test result is 90%, while for those who do not have cancer the probability of a positive test result is 3%. We can therefore write out all four conditional probabilities:

$$p(T = 1|C = 1) = 90/100 \tag{2.14}$$
$$p(T = 0|C = 1) = 10/100 \tag{2.15}$$
$$p(T = 1|C = 0) = 3/100 \tag{2.16}$$
$$p(T = 0|C = 0) = 97/100. \tag{2.17}$$

Again, note that these probabilities are normalized so that

$$p(T = 1|C = 1) + p(T = 0|C = 1) = 1 \tag{2.18}$$

and similarly

$$p(T = 1|C = 0) + p(T = 0|C = 0) = 1. \tag{2.19}$$

We can now use the sum and product rules of probability to answer our first question and evaluate the overall probability that someone who is tested at random will have a positive test result:

$$\begin{aligned} p(T = 1) &= p(T = 1|C = 0)p(C = 0) + p(T = 1|C = 1)p(C = 1) \\ &= \frac{3}{100} \times \frac{99}{100} + \frac{90}{100} \times \frac{1}{100} = \frac{387}{10,000} = 0.0387. \end{aligned} \tag{2.20}$$

We see that if a person is tested at random there is a roughly 4% chance that the test will be positive even though there is a 1% chance that they actually have cancer. From this it follows, using the sum rule, that $p(T = 0) = 1 - 387/10,000 = 9613/10,000 = 0.9613$ and, hence, there is a roughly 96% chance that the do not have cancer.

Now consider our second question, which is the one that is of particular interest to a person being screened: if a test is positive, what is the probability that the person

has cancer? This requires that we evaluate the probability of cancer conditional on the outcome of the test, whereas the probabilities in (2.14) to (2.17) give the probability distribution over the test outcome conditioned on whether the person has cancer. We can solve the problem of reversing the conditional probability by using Bayes' theorem (2.10) to give

$$p(C=1|T=1) = \frac{p(T=1|C=1)p(C=1)}{p(T=1)} \tag{2.21}$$

$$= \frac{90}{100} \times \frac{1}{100} \times \frac{10,000}{387} = \frac{90}{387} \simeq 0.23 \tag{2.22}$$

so that if a person is tested at random and the test is positive, there is a 23% probability that they actually have cancer. From the sum rule, it then follows that $p(C = 0|T = 1) = 1 - 90/387 = 297/387 \simeq 0.77$, which is a 77% chance that they do not have cancer.

2.1.5 Prior and posterior probabilities

We can use the cancer screening example to provide an important interpretation of Bayes' theorem as follows. If we had been asked whether someone is likely to have cancer, before they have received a test, then the most complete information we have available is provided by the probability $p(C)$. We call this the *prior probability* because it is the probability available *before* we observe the result of the test. Once we are told that this person has received a positive test, we can then use Bayes' theorem to compute the probability $p(C|T)$, which we will call the *posterior probability* because it is the probability obtained *after* we have observed the test result T.

In this example, the prior probability of having cancer is 1%. However, once we have observed that the test result is positive, we find that the posterior probability of cancer is now 23%, which is a substantially higher probability of cancer, as we would intuitively expect. We note, however, that a person with a positive test still has only a 23% change of actually having cancer, even though the test appears, from Figure 2.3 to be reasonably 'accurate'. This conclusion seems counter-intuitive to many people. The reason has to do with the low prior probability of having cancer. Although the test provides strong evidence of cancer, this has to be combined with the prior probability using Bayes' theorem to arrive at the correct posterior probability.

Exercise 2.1

2.1.6 Independent variables

Finally, if the joint distribution of two variables factorizes into the product of the marginals, so that $p(X, Y) = p(X)p(Y)$, then X and Y are said to be *independent*. An example of independent events would be the successive flips of a coin. From the product rule, we see that $p(Y|X) = p(Y)$, and so the conditional distribution of Y given X is indeed independent of the value of X. In our cancer screening example, if the probability of a positive test is independent of whether the person has cancer, then $p(T|C) = p(T)$, which means that from Bayes' theorem (2.10) we have $p(C|T) = p(C)$, and therefore probability of cancer is not changed by observing the test outcome. Of course, such a test would be useless because the outcome of the test tells us nothing about whether the person has cancer.

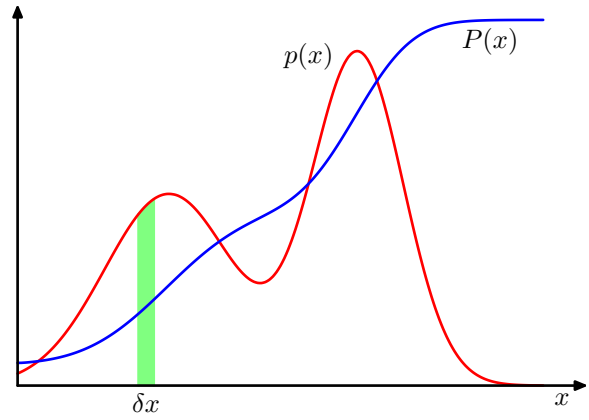

Figure 2.6 The concept of probability for discrete variables can be extended to that of a probability density $p(x)$ over a continuous variable x and is such that the probability of x lying in the interval $(x, x + \delta x)$ is given by $p(x)\delta x$ for $\delta x \to 0$. The probability density can be expressed as the derivative of a cumulative distribution function $P(x)$.

2.2. Probability Densities

As well as considering probabilities defined over discrete sets of values, we also wish to consider probabilities with respect to continuous variables. For instance, we might wish to predict what dose of drug to give to a patient. Since there will be uncertainty in this prediction, we want to quantify this uncertainty and again we can make use of probabilities. However, we cannot simply apply the concepts of probability discussed so far directly, since the probability of observing a specific value for a continuous variable, to infinite precision, will effectively be zero. Instead, we need to introduce the concept of a *probability density*. Here we will limit ourselves to a relatively informal discussion.

We define the probability density $p(x)$ over a continuous variable x to be such that the probability of x falling in the interval $(x, x + \delta x)$ is given by $p(x)\delta x$ for $\delta x \to 0$. This is illustrated in Figure 2.6. The probability that x will lie in an interval (a, b) is then given by

$$p(x \in (a, b)) = \int_a^b p(x)\,\mathrm{d}x. \tag{2.23}$$

Because probabilities are non-negative, and because the value of x must lie somewhere on the real axis, the probability density $p(x)$ must satisfy the two conditions

$$p(x) \geqslant 0 \tag{2.24}$$

$$\int_{-\infty}^{\infty} p(x)\,\mathrm{d}x = 1. \tag{2.25}$$

The probability that x lies in the interval $(-\infty, z)$ is given by the *cumulative distribution function* defined by

$$P(z) = \int_{-\infty}^{z} p(x)\,\mathrm{d}x, \tag{2.26}$$

which satisfies $P'(x) = p(x)$, as shown in Figure 2.6.

If we have several continuous variables $x_1, \ldots, x_D$, denoted collectively by the vector $\mathbf{x}$, then we can define a joint probability density $p(\mathbf{x}) = p(x_1, \ldots, x_D)$ such that the probability of $\mathbf{x}$ falling in an infinitesimal volume $\delta\mathbf{x}$ containing the point $\mathbf{x}$ is given by $p(\mathbf{x})\delta\mathbf{x}$. This multivariate probability density must satisfy

$$p(\mathbf{x}) \geqslant 0 \tag{2.27}$$

$$\int p(\mathbf{x})\,\mathrm{d}\mathbf{x} = 1 \tag{2.28}$$

in which the integral is taken over the whole of $\mathbf{x}$ space. More generally, we can also consider joint probability distributions over a combination of discrete and continuous variables.

The sum and product rules of probability, as well as Bayes' theorem, also apply to probability densities as well as to combinations of discrete and continuous variables. If $\mathbf{x}$ and $\mathbf{y}$ are two real variables, then the sum and product rules take the form

$$\textbf{sum rule} \qquad p(\mathbf{x}) = \int p(\mathbf{x}, \mathbf{y})\,\mathrm{d}\mathbf{y} \tag{2.29}$$

$$\textbf{product rule} \qquad p(\mathbf{x}, \mathbf{y}) = p(\mathbf{y}|\mathbf{x})p(\mathbf{x}). \tag{2.30}$$

Similarly, Bayes' theorem can be written in the form

$$p(\mathbf{y}|\mathbf{x}) = \frac{p(\mathbf{x}|\mathbf{y})p(\mathbf{y})}{p(\mathbf{x})} \tag{2.31}$$

where the denominator is given by

$$p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{y})p(\mathbf{y})\,\mathrm{d}\mathbf{y}. \tag{2.32}$$

A formal justification of the sum and product rules for continuous variables requires a branch of mathematics called *measure theory* (Feller, 1966) and lies outside the scope of this book. Its validity can be seen informally, however, by dividing each real variable into intervals of width Δ and considering the discrete probability distribution over these intervals. Taking the limit $\Delta \to 0$ then turns sums into integrals and gives the desired result.

2.2.1 Example distributions

There are many forms of probability density that are in widespread use and that are important both in their own right and as building blocks for more complex probabilistic models. The simplest form would be one in which $p(x)$ is a constant, independent of x, but this cannot be normalized because the integral in (2.28) will be divergent. Distributions that cannot be normalized are called *improper*. We can, however, have the uniform distribution that is constant over a finite region, say (c, d), and zero elsewhere, in which case (2.28) implies

$$p(x) = 1/(d - c), \qquad x \in (c, d). \tag{2.33}$$

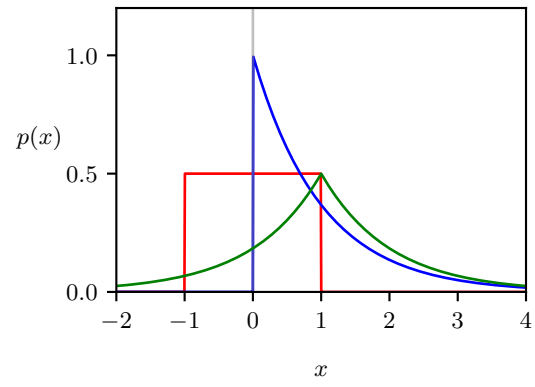

Figure 2.7 Plots of a uniform distribution over the range $(-1, 1)$, shown in red, the exponential distribution with $\lambda = 1$, shown in blue, and a Laplace distribution with $\mu = 1$ and $\gamma = 1$, shown in green.

Another simple form of density is the *exponential distribution* given by

$$p(x|\lambda) = \lambda \exp(-\lambda x), \qquad x \geqslant 0. \tag{2.34}$$

A variant of the exponential distribution, known as the *Laplace distribution*, allows the peak to be moved to a location μ and is given by

$$p(x|\mu, \gamma) = \frac{1}{2\gamma} \exp\left(-\frac{|x - \mu|}{\gamma}\right). \tag{2.35}$$

The constant, exponential, and Laplace distributions are illustrated in Figure 2.7.

Another important distribution is the *Dirac delta function*, which is written

$$p(x|\mu) = \delta(x - \mu). \tag{2.36}$$

This is defined to be zero everywhere except at $x = \mu$ and to have the property of integrating to unity according to (2.28). Informally, we can think of this as an infinitely narrow and infinitely tall spike located at $x = \mu$ with the property of having unit area. Finally, if we have a finite set of observations of x given by $\mathcal{D} = \{x_1, \ldots, x_N\}$ then we can use the delta function to construct the *empirical distribution* given by

$$p(x|\mathcal{D}) = \frac{1}{N} \sum_{n=1}^{N} \delta(x - x_n), \tag{2.37}$$

which consists of a Dirac delta function centred on each of the data points. The probability density defined by (2.37) integrates to one as required.

Exercise 2.6

2.2.2 Expectations and covariances

One of the most important operations involving probabilities is that of finding weighted averages of functions. The weighted average of some function $f(x)$ under a probability distribution $p(x)$ is called the *expectation* of $f(x)$ and will be denoted by $\mathbb{E}[f]$. For a discrete distribution, it is given by summing over all possible values of x in the form

$$\mathbb{E}[f] = \sum_{x} p(x) f(x) \tag{2.38}$$

where the average is weighted by the relative probabilities of the different values of x. For continuous variables, expectations are expressed in terms of an integration with respect to the corresponding probability density:

$$\mathbb{E}[f] = \int p(x)f(x)\,\mathrm{d}x. \tag{2.39}$$

In either case, if we are given a finite number N of points drawn from the probability distribution or probability density, then the expectation can be approximated as a finite sum over these points:

Exercise 2.7

$$\mathbb{E}[f] \simeq \frac{1}{N}\sum_{n=1}^{N} f(x_n). \tag{2.40}$$

The approximation in (2.40) becomes exact in the limit $N \to \infty$.

Sometimes we will be considering expectations of functions of several variables, in which case we can use a subscript to indicate which variable is being averaged over, so that for instance

$$\mathbb{E}_x[f(x, y)] \tag{2.41}$$

denotes the average of the function $f(x, y)$ with respect to the distribution of x. Note that $\mathbb{E}_x[f(x, y)]$ will be a function of y.

We can also consider a *conditional expectation* with respect to a conditional distribution, so that

$$\mathbb{E}_x[f|y] = \sum_x p(x|y)f(x), \tag{2.42}$$

which is also a function of y. For continuous variables, the conditional expectation takes the form

$$\mathbb{E}_x[f|y] = \int p(x|y)f(x)\,\mathrm{d}x. \tag{2.43}$$

The *variance* of $f(x)$ is defined by

$$\mathrm{var}[f] = \mathbb{E}\left[(f(x) - \mathbb{E}[f(x)])^2\right] \tag{2.44}$$

and provides a measure of how much $f(x)$ varies around its mean value $\mathbb{E}[f(x)]$. Expanding out the square, we see that the variance can also be written in terms of the expectations of $f(x)$ and $f(x)^2$:

Exercise 2.8

$$\mathrm{var}[f] = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2. \tag{2.45}$$

In particular, we can consider the variance of the variable x itself, which is given by

$$\mathrm{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2. \tag{2.46}$$

For two random variables x and y, the *covariance* measures the extent to which the two variables vary together and is defined by

$$\begin{aligned} \mathrm{cov}[x, y] &= \mathbb{E}_{x,y}\left[\{x - \mathbb{E}[x]\}\{y - \mathbb{E}[y]\}\right] \\ &= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y]. \end{aligned} \tag{2.47}$$

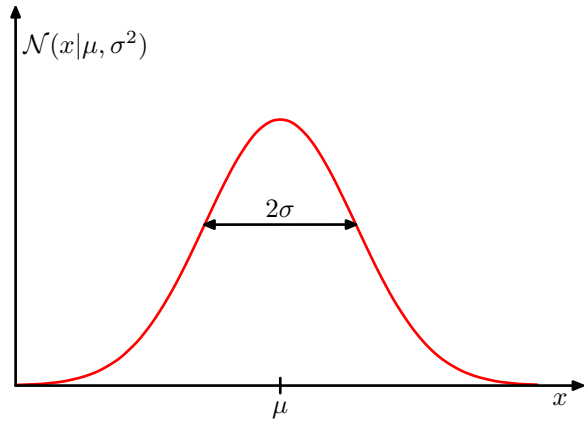

Figure 2.8 Plot of a Gaussian distribution for a single continuous variable x showing the mean μ and the standard deviation σ.

Exercise 2.9

If x and y are independent, then their covariance equals zero.

For two vectors $\mathbf{x}$ and $\mathbf{y}$, their covariance is a matrix given by

$$\begin{aligned}\mathrm{cov}[\mathbf{x},\mathbf{y}] &= \mathbb{E}_{\mathbf{x},\mathbf{y}}\left[\{\mathbf{x}-\mathbb{E}[\mathbf{x}]\}\{\mathbf{y}^{\mathrm{T}}-\mathbb{E}[\mathbf{y}^{\mathrm{T}}]\}\right] \\ &= \mathbb{E}_{\mathbf{x},\mathbf{y}}[\mathbf{x}\mathbf{y}^{\mathrm{T}}]-\mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{y}^{\mathrm{T}}]. \end{aligned} \tag{2.48}$$

If we consider the covariance of the components of a vector $\mathbf{x}$ with each other, then we use a slightly simpler notation $\mathrm{cov}[\mathbf{x}] \equiv \mathrm{cov}[\mathbf{x},\mathbf{x}]$.

2.3. The Gaussian Distribution

One of the most important probability distributions for continuous variables is called the *normal* or *Gaussian* distribution, and we will make extensive use of this distribution throughout the rest of the book. For a single real-valued variable x, the Gaussian distribution is defined by

$$\mathcal{N}\left(x|\mu,\sigma^2\right) = \frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}, \tag{2.49}$$

which represents a probability density over x governed by two parameters: μ, called the *mean*, and σ^2, called the *variance*. The square root of the variance, given by σ, is called the *standard deviation*, and the reciprocal of the variance, written as $\beta = 1/\sigma^2$, is called the *precision*. We will see the motivation for this terminology shortly. Figure 2.8 shows a plot of the Gaussian distribution. Although the form of the Gaussian distribution might seem arbitrary, we will see later that it arises naturally from the concept of maximum entropy and from the perspective of the central limit theorem.

Section 2.5.4
Section 3.2

From (2.49) we see that the Gaussian distribution satisfies

$$\mathcal{N}(x|\mu,\sigma^2) > 0. \tag{2.50}$$

Exercise 2.12 Also, it is straightforward to show that the Gaussian is normalized, so that

$$\int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right)\,\mathrm{d}x = 1. \tag{2.51}$$

Thus, (2.49) satisfies the two requirements for a valid probability density.

2.3.1 Mean and variance

Exercise 2.13 We can readily find expectations of functions of x under the Gaussian distribution. In particular, the average value of x is given by

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right) x\,\mathrm{d}x = \mu. \tag{2.52}$$

Because the parameter μ represents the average value of x under the distribution, it is referred to as the mean. The integral in (2.52) is known as the *first-order moment* of the distribution because it is the expectation of x raised to the power one. We can similarly evaluate the second-order moment given by

$$\mathbb{E}[x^2] = \int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right) x^2\,\mathrm{d}x = \mu^2 + \sigma^2. \tag{2.53}$$

From (2.52) and (2.53), it follows that the variance of x is given by

$$\mathrm{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \sigma^2 \tag{2.54}$$

and hence σ^2 is referred to as the variance parameter. The maximum of a distribution *Exercise 2.14* is known as its mode. For a Gaussian, the mode coincides with the mean.

2.3.2 Likelihood function

Suppose that we have a data set of observations represented as a row vector $\mathbf{x} = (x_1, \ldots, x_N)$, representing N observations of the scalar variable x. Note that we are using the typeface $\mathbf{x}$ to distinguish this from a single observation of a D-dimensional vector-valued variable, which we represent by a column vector $\mathbf{x} = (x_1, \ldots, x_D)^{\mathrm{T}}$. We will suppose that the observations are drawn independently from a Gaussian distribution whose mean μ and variance σ^2 are unknown, and we would like to determine these parameters from the data set. The problem of estimating a distribution, given a finite set of observations, is known as *density estimation*. It should be emphasized that the problem of density estimation is fundamentally ill-posed, because there are infinitely many probability distributions that could have given rise to the observed finite data set. Indeed, any distribution $p(\mathbf{x})$ that is non-zero at each of the data points $\mathbf{x}_1, \ldots, \mathbf{x}_N$ is a potential candidate. Here we constrain the space of distributions to be Gaussians, which leads to a well-defined solution.

Data points that are drawn independently from the same distribution are said to be *independent and identically distributed*, which is often abbreviated to i.i.d. or IID. We have seen that the joint probability of two independent events is given by the product of the marginal probabilities for each event separately. Because our data

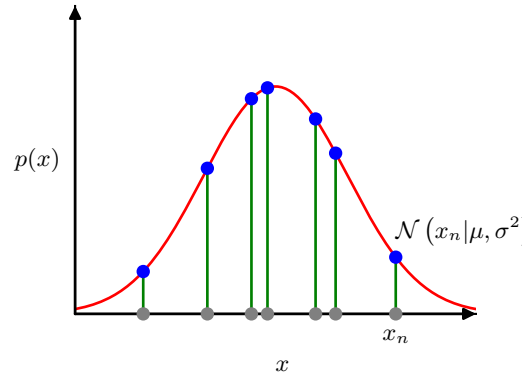

Figure 2.9 Illustration of the likelihood function for the Gaussian distribution shown by the red curve. Here the grey points denote a data set of values $\{x_n\}$, and the likelihood function (2.55) is given by the product of the corresponding values of $p(x)$ denoted by the blue points. Maximizing the likelihood involves adjusting the mean and variance of the Gaussian so as to maximize this product.

set $\mathbf{x}$ is i.i.d., we can therefore write the probability of the data set, given μ and σ^2, in the form

$$p(\mathbf{x}|\mu,\sigma^2) = \prod_{n=1}^{N} \mathcal{N}\left(x_n|\mu,\sigma^2\right). \tag{2.55}$$

When viewed as a function of μ and σ^2, this is called the *likelihood function* for the Gaussian and is interpreted diagrammatically in Figure 2.9.

One common approach for determining the parameters in a probability distribution using an observed data set, known as *maximum likelihood*, is to find the parameter values that maximize the likelihood function. This might appear to be a strange criterion because, from our foregoing discussion of probability theory, it would seem more natural to maximize the probability of the parameters given the data, not the probability of the data given the parameters. In fact, these two criteria are related.

Section 2.6.2

To start with, however, we will determine values for the unknown parameters μ and σ^2 in the Gaussian by maximizing the likelihood function (2.55). In practice, it is more convenient to maximize the log of the likelihood function. Because the logarithm is a monotonically increasing function of its argument, maximizing the log of a function is equivalent to maximizing the function itself. Taking the log not only simplifies the subsequent mathematical analysis, but it also helps numerically because the product of a large number of small probabilities can easily underflow the numerical precision of the computer, and this is resolved by computing the sum of the log probabilities instead. From (2.49) and (2.55), the log likelihood function can be written in the form

$$\ln p\left(\mathbf{x}|\mu,\sigma^2\right) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi). \tag{2.56}$$

Maximizing (2.56) with respect to μ, we obtain the maximum likelihood solution given by

Exercise 2.15

$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n, \tag{2.57}$$

which is the *sample mean*, i.e., the mean of the observed values $\{x_n\}$. Similarly, maximizing (2.56) with respect to σ^2, we obtain the maximum likelihood solution for the variance in the form

$$\sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{\mathrm{ML}})^2, \tag{2.58}$$

which is the *sample variance* measured with respect to the sample mean μ_{ML}. Note that we are performing a joint maximization of (2.56) with respect to μ and σ^2, but for a Gaussian distribution, the solution for μ decouples from that for σ^2 so that we can first evaluate (2.57) and then subsequently use this result to evaluate (2.58).

2.3.3 Bias of maximum likelihood

The technique of maximum likelihood is widely used in deep learning and forms the foundation for most machine learning algorithms. However, it has some limitations, which we can illustrate using a univariate Gaussian.

We first note that the maximum likelihood solutions μ_{ML} and σ^2_{ML} are functions of the data set values $x_1, \ldots, x_N$. Suppose that each of these values has been generated independently from a Gaussian distribution whose true parameters are μ and σ^2. Now consider the expectations of μ_{ML} and σ^2_{ML} with respect to these data set values. It is straightforward to show that

Exercise 2.16

$$\mathbb{E}[\mu_{\mathrm{ML}}] = \mu \tag{2.59}$$

$$\mathbb{E}[\sigma^2_{\mathrm{ML}}] = \left(\frac{N-1}{N}\right)\sigma^2. \tag{2.60}$$

We see that, when averaged over data sets of a given size, the maximum likelihood solution for the mean will equal the true mean. However, the maximum likelihood estimate of the variance will underestimate the true variance by a factor $(N-1)/N$. This is an example of a phenomenon called *bias* in which the estimator of a random quantity is systematically different from the true value. The intuition behind this result is given by Figure 2.10.

Note that bias arises because the variance is measured relative to the maximum likelihood estimate of the mean, which itself is tuned to the data. Suppose instead we had access to the true mean μ and we used this to determine the variance using the estimator

$$\widehat{\sigma}^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu)^2. \tag{2.61}$$

Exercise 2.17

Then we find that

$$\mathbb{E}\left[\widehat{\sigma}^2\right] = \sigma^2, \tag{2.62}$$

which is unbiased. Of course, we do not have access to the true mean but only to the observed data values. From the result (2.60) it follows that for a Gaussian distribution, the following estimate for the variance parameter is unbiased:

$$\widetilde{\sigma}^2 = \frac{N}{N-1}\sigma^2_{\mathrm{ML}} = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \mu_{\mathrm{ML}})^2. \tag{2.63}$$

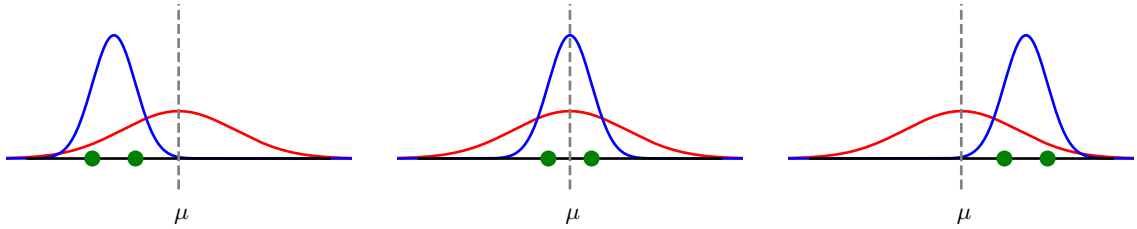

Figure 2.10 Illustration of how bias arises when using maximum likelihood to determine the mean and variance of a Gaussian. The red curves show the true Gaussian distribution from which data is generated, and the three blue curves show the Gaussian distributions obtained by fitting to three data sets, each consisting of two data points shown in green, using the maximum likelihood results (2.57) and (2.58). Averaged across the three data sets, the mean is correct, but the variance is systematically underestimated because it is measured relative to the sample mean and not relative to the true mean.

Correcting for the bias of maximum likelihood in complex models such as neural networks is not so easy, however.

Note that the bias of the maximum likelihood solution becomes less significant as the number N of data points increases. In the limit $N \to \infty$ the maximum likelihood solution for the variance equals the true variance of the distribution that generated the data. In the case of the Gaussian, for anything other than small N, this bias will not prove to be a serious problem. However, throughout this book we will be interested in complex models with many parameters, for which the bias problems associated with maximum likelihood will be much more severe. In fact, the issue of bias in maximum likelihood is closely related to the problem of *over-fitting*.

Section 2.6.3

2.3.4 Linear regression

Section 1.2

We have seen how the problem of linear regression can be expressed in terms of error minimization. Here we return to this example and view it from a probabilistic perspective, thereby gaining some insights into error functions and regularization.

The goal in the regression problem is to be able to make predictions for the target variable t given some new value of the input variable x by using a set of training data comprising N input values $\mathbf{x} = (x_1, \ldots, x_N)$ and their corresponding target values $\mathbf{t} = (t_1, \ldots, t_N)$. We can express our uncertainty over the value of the target variable using a probability distribution. For this purpose, we will assume that, given the value of x, the corresponding value of t has a Gaussian distribution with a mean equal to the value $y(x, \mathbf{w})$ of the polynomial curve given by (1.1), where $\mathbf{w}$ are the polynomial coefficients, and a variance σ^2. Thus, we have

$$p(t|x, \mathbf{w}, \sigma^2) = \mathcal{N}\left(t|y(x, \mathbf{w}), \sigma^2\right). \tag{2.64}$$

This is illustrated schematically in Figure 2.11.

We now use the training data $\{\mathbf{x}, \mathbf{t}\}$ to determine the values of the unknown parameters $\mathbf{w}$ and σ^2 by maximum likelihood. If the data is assumed to be drawn

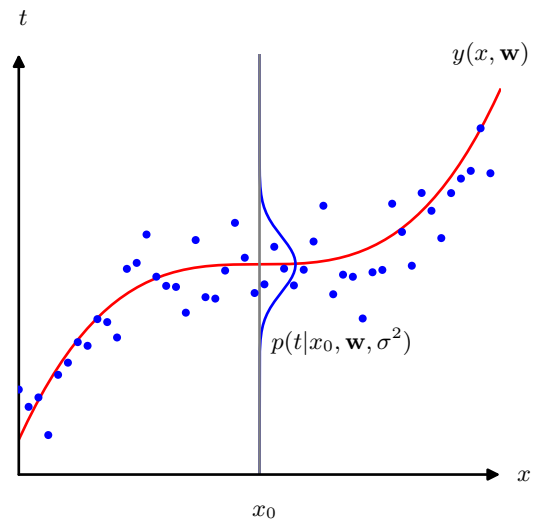

Figure 2.11 Schematic illustration of a Gaussian conditional distribution for t given x, defined by (2.64), in which the mean is given by the polynomial function $y(x, \mathbf{w})$, and the variance is given by the parameter σ^2.

independently from the distribution (2.64), then the likelihood function is given by

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}\left(t_n|y(x_n, \mathbf{w}), \sigma^2\right). \tag{2.65}$$

As we did for the simple Gaussian distribution earlier, it is convenient to maximize the logarithm of the likelihood function. Substituting for the Gaussian distribution, given by (2.49), we obtain the log likelihood function in the form

$$\ln p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \sigma^2) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi). \tag{2.66}$$

Consider first the evaluation of the maximum likelihood solution for the polynomial coefficients, which will be denoted by $\mathbf{w}_{\mathrm{ML}}$. These are determined by maximizing (2.66) with respect to $\mathbf{w}$. For this purpose, we can omit the last two terms on the right-hand side of (2.66) because they do not depend on $\mathbf{w}$. Also, note that scaling the log likelihood by a positive constant coefficient does not alter the location of the maximum with respect to $\mathbf{w}$, and so we can replace the coefficient $1/2\sigma^2$ with $1/2$. Finally, instead of maximizing the log likelihood, we can equivalently minimize the negative log likelihood. We therefore see that maximizing the likelihood is equivalent, so far as determining $\mathbf{w}$ is concerned, to minimizing the *sum-of-squares error function* defined by

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2. \tag{2.67}$$

Thus, the sum-of-squares error function has arisen as a consequence of maximizing the likelihood under the assumption of a Gaussian noise distribution.

Exercise 2.18

We can also use maximum likelihood to determine the variance parameter σ^2. Maximizing (2.66) with respect to σ^2 gives

$$\sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}_{\mathrm{ML}}) - t_n\}^2 . \tag{2.68}$$

Note that we can first determine the parameter vector $\mathbf{w}_{\mathrm{ML}}$ governing the mean, and subsequently use this to find the variance σ^2_{ML} as was the case for the simple Gaussian distribution.

Having determined the parameters $\mathbf{w}$ and σ^2, we can now make predictions for new values of x. Because we now have a probabilistic model, these are expressed in terms of the *predictive distribution* that gives the probability distribution over t, rather than simply a point estimate, and is obtained by substituting the maximum likelihood parameters into (2.64) to give

$$p(t|x, \mathbf{w}_{\mathrm{ML}}, \sigma^2_{\mathrm{ML}}) = \mathcal{N}\left(t|y(x, \mathbf{w}_{\mathrm{ML}}), \sigma^2_{\mathrm{ML}}\right) . \tag{2.69}$$

2.4. Transformation of Densities

Chapter 18

We turn now to a discussion of how a probability density transforms under a nonlinear change of variable. This property will play a crucial role when we discuss a class of generative models called *normalizing flows*. It also highlights that a probability density has a different behaviour than a simple function under such transformations.

Consider a single variable x and suppose we make a change of variables $x = g(y)$, then a function $f(x)$ becomes a new function $\widetilde{f}(y)$ defined by

$$\widetilde{f}(y) = f(g(y)). \tag{2.70}$$

Now consider a probability density $p_x(x)$, and again change variables using $x = g(y)$, giving rise to a density $p_y(y)$ with respect to the new variable y, where the suffixes denote that $p_x(x)$ and $p_y(y)$ are different densities. Observations falling in the range $(x, x + \delta x)$ will, for small values of δx, be transformed into the range $(y, y + \delta y)$, where $x = g(y)$, and $p_x(x)\delta x \simeq p_y(y)\delta y$. Hence, if we take the limit $\delta x \to 0$, we obtain

$$\begin{aligned} p_y(y) &= p_x(x)\left|\frac{\mathrm{d}x}{\mathrm{d}y}\right| \\ &= p_x(g(y))\left|\frac{\mathrm{d}g}{\mathrm{d}y}\right| . \end{aligned} \tag{2.71}$$

Here the modulus $|\cdot|$ arises because the derivative $\mathrm{d}y/\,\mathrm{d}x$ could be negative, whereas the density is scaled by the ratio of lengths, which is always positive.

This procedure for transforming densities can be very powerful. Any density $p(y)$ can be obtained from a fixed density $q(x)$ that is everywhere non-zero by making a nonlinear change of variable $y = f(x)$ in which $f(x)$ is a monotonic function so that $0 \leqslant f'(x) < \infty$.

Exercise 2.19

One consequence of the transformation property (2.71) is that the concept of the maximum of a probability density is dependent on the choice of variable. Suppose $f(x)$ has a mode (i.e., a maximum) at $\widehat{x}$ so that $f'(\widehat{x}) = 0$. The corresponding mode of $\widetilde{f}(y)$ will occur for a value $\widehat{y}$ obtained by differentiating both sides of (2.70) with respect to y:

$$\widetilde{f}'(\widehat{y}) = f'(g(\widehat{y}))g'(\widehat{y}) = 0. \tag{2.72}$$

Assuming $g'(\widehat{y}) \neq 0$ at the mode, then $f'(g(\widehat{y})) = 0$. However, we know that $f'(\widehat{x}) = 0$, and so we see that the locations of the mode expressed in terms of each of the variables x and y are related by $\widehat{x} = g(\widehat{y})$, as one would expect. Thus, finding a mode with respect to the variable x is equivalent to first transforming to the variable y, then finding a mode with respect to y, and then transforming back to x.

Now consider the behaviour of a probability density $p_x(x)$ under the change of variables $x = g(y)$, where the density with respect to the new variable is $p_y(y)$ and is given by (2.71). To deal with the modulus in (2.71) we can write $g'(y) = s|g'(y)|$ where $s \in \{-1, +1\}$. Then (2.71) can be written as

$$p_y(y) = p_x(g(y))sg'(y)$$

where we have used $1/s = s$. Differentiating both sides with respect to y then gives

$$p'_y(y) = sp'_x(g(y))\{g'(y)\}^2 + sp_x(g(y))g''(y). \tag{2.73}$$

Due to the presence of the second term on the right-hand side of (2.73), the relationship $\widehat{x} = g(\widehat{y})$ no longer holds. Thus, the value of x obtained by maximizing $p_x(x)$ will not be the value obtained by transforming to $p_y(y)$ then maximizing with respect to y and then transforming back to x. This causes modes of densities to be dependent on the choice of variables. However, for a linear transformation, the second term on the right-hand side of (2.73) vanishes, and so in this case the location of the maximum transforms according to $\widehat{x} = g(\widehat{y})$.

This effect can be illustrated with a simple example, as shown in Figure 2.12. We begin by considering a Gaussian distribution $p_x(x)$ over x shown by the red curve in Figure 2.12. Next we draw a sample of $N = 50{,}000$ points from this distribution and plot a histogram of their values, which as expected agrees with the distribution $p_x(x)$. Now consider a nonlinear change of variables from x to y given by

$$x = g(y) = \ln(y) - \ln(1 - y) + 5. \tag{2.74}$$

The inverse of this function is given by

$$y = g^{-1}(x) = \frac{1}{1 + \exp(-x + 5)}, \tag{2.75}$$

which is a *logistic sigmoid* function and is shown in Figure 2.12 by the blue curve.

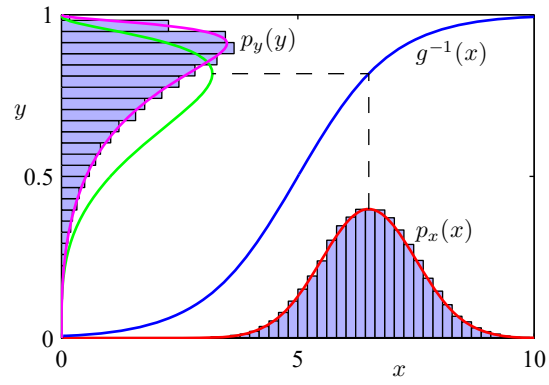

Figure 2.12 Example of the transformation of the mode of a density under a nonlinear change of variables, illustrating the different behaviour compared to a simple function.

If we simply transform $p_x(x)$ as a function of x we obtain the green curve $p_x(g(y))$ shown in Figure 2.12, and we see that the mode of the density $p_x(x)$ is transformed via the sigmoid function to the mode of this curve. However, the density over y transforms instead according to (2.71) and is shown by the magenta curve on the left side of the diagram. Note that this has its mode shifted relative to the mode of the green curve.

To confirm this result, we take our sample of $50{,}000$ values of x, evaluate the corresponding values of y using (2.75), and then plot a histogram of their values. We see that this histogram matches the magenta curve in Figure 2.12 and not the green curve.

2.4.1 Multivariate distributions

We can extend the result (2.71) to densities defined over multiple variables. Consider a density $p(\mathbf{x})$ over a D-dimensional variable $\mathbf{x} = (x_1, \ldots, x_D)^{\mathrm{T}}$, and suppose we transform to a new variable $\mathbf{y} = (y_1, \ldots, y_D)^{\mathrm{T}}$ where $\mathbf{x} = \mathbf{g}(\mathbf{y})$. Here we will limit ourselves to the case where $\mathbf{x}$ and $\mathbf{y}$ have the same dimensionality. The transformed density is then given by the generalization of (2.71) in the form

$$p_{\mathbf{y}}(\mathbf{y}) = p_{\mathbf{x}}(\mathbf{x})\,|\det \mathbf{J}| \tag{2.76}$$

where $\mathbf{J}$ is the *Jacobian matrix* whose elements are given by the partial derivatives $J_{ij} = \partial g_i / \partial y_j$, so that

$$\mathbf{J} = \begin{bmatrix} \dfrac{\partial g_1}{\partial y_1} & \cdots & \dfrac{\partial g_1}{\partial y_D} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial g_D}{\partial y_1} & \cdots & \dfrac{\partial g_D}{\partial y_D} \end{bmatrix}. \tag{2.77}$$

Intuitively, we can view the change of variables as expanding some regions of space and contracting others, with an infinitesimal region $\Delta\mathbf{x}$ around a point $\mathbf{x}$ being transformed to a region $\Delta\mathbf{y}$ around the point $\mathbf{y} = \mathbf{g}(\mathbf{x})$. The absolute value of the determinant of the Jacobian represents the ratio of these volumes and is the same factor

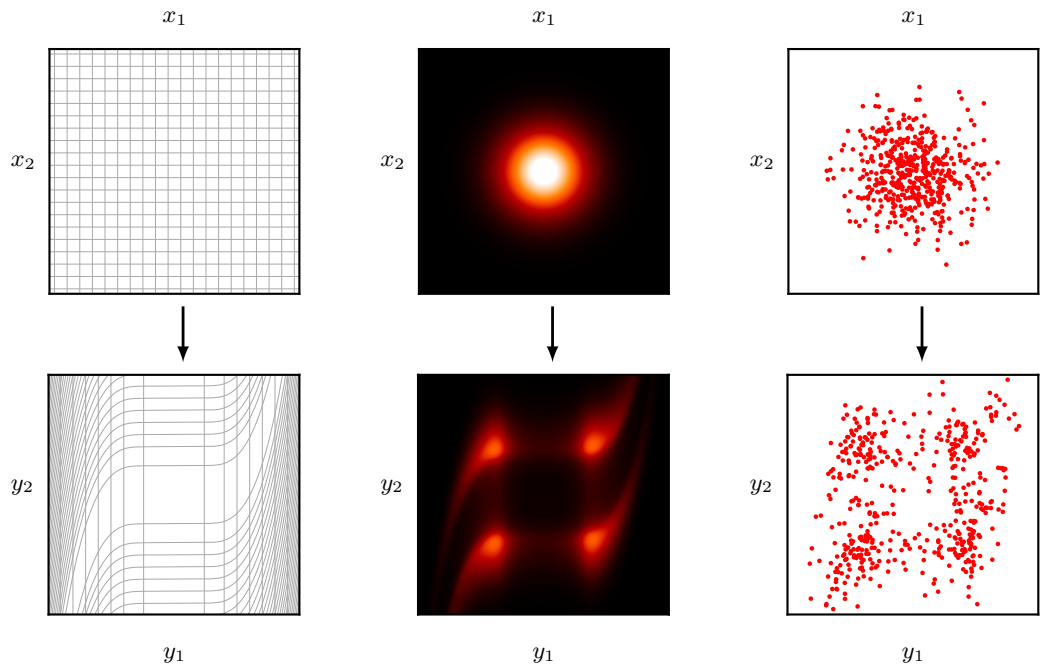

Figure 2.13 Illustration of the effect of a change of variables on a probability distribution in two dimensions. The left column shows the transforming of the variables whereas the middle and right columns show the corresponding effects on a Gaussian distribution and on samples from that distribution, respectively.

that arises when changing variables within an integral. The formula (2.77) follows from the fact that the probability mass in region $\Delta\mathbf{x}$ is the same as the probability mass in $\Delta\mathbf{y}$. Once again, we take the modulus to ensure that the density is non-negative.

We can illustrate this by applying a change of variables to a Gaussian distribution in two dimensions, as shown in the top row in Figure 2.13. Here the transformation from $\mathbf{x}$ to $\mathbf{y}$ is given by

Exercise 2.20

$$y_1 = x_1 + \tanh(5x_1) \tag{2.78}$$

$$y_2 = x_2 + \tanh(5x_2) + \frac{x_1^3}{3}. \tag{2.79}$$

Also shown on the bottom row are samples from a Gaussian distribution in $\mathbf{x}$-space along with the corresponding transformed samples in $\mathbf{y}$-space.

2.5. Information Theory

Probability theory forms the basis for another important framework called *information theory*, which quantifies the information present in a data set and which plays an important role in machine learning. Here we give a brief introduction to some of the key elements of information theory that we will need later in the book, including the important concept of entropy in its various forms. For a more comprehensive introduction to information theory, with connections to machine learning, see MacKay (2003).

2.5.1 Entropy

We begin by considering a discrete random variable x and we ask how much information is received when we observe a specific value for this variable. The amount of information can be viewed as the 'degree of surprise' on learning the value of x. If we are told that a highly improbable event has just occurred, we will have received more information than if we were told that some very likely event has just occurred, and if we knew that the event was certain to happen, we would receive no information. Our measure of information content will therefore depend on the probability distribution $p(x)$, and so we look for a quantity $h(x)$ that is a monotonic function of the probability $p(x)$ and that expresses the information content. The form of $h(\cdot)$ can be found by noting that if we have two events x and y that are unrelated, then the information gained from observing both of them should be the sum of the information gained from each of them separately, so that $h(x, y) = h(x) + h(y)$. Two unrelated events are statistically independent and so $p(x, y) = p(x)p(y)$. From these two relationships, it is easily shown that $h(x)$ must be given by the logarithm of $p(x)$ and so we have

Exercise 2.21

$$h(x) = -\log_2 p(x) \tag{2.80}$$

where the negative sign ensures that information is positive or zero. Note that low probability events x correspond to high information content. The choice of base for the logarithm is arbitrary, and for the moment we will adopt the convention prevalent in information theory of using logarithms to the base of 2. In this case, as we will see shortly, the units of $h(x)$ are bits ('binary digits').

Now suppose that a sender wishes to transmit the value of a random variable to a receiver. The average amount of information that they transmit in the process is obtained by taking the expectation of (2.80) with respect to the distribution $p(x)$ and is given by

$$\mathrm{H}[x] = -\sum_x p(x) \log_2 p(x). \tag{2.81}$$

This important quantity is called the *entropy* of the random variable x. Note that $\lim_{\epsilon \to 0} (\epsilon \ln \epsilon) = 0$ and so we will take $p(x) \ln p(x) = 0$ whenever we encounter a value for x such that $p(x) = 0$.

So far, we have given a rather heuristic motivation for the definition of information (2.80) and the corresponding entropy (2.81). We now show that these definitions

indeed possess useful properties. Consider a random variable x having eight possible states, each of which is equally likely. To communicate the value of x to a receiver, we would need to transmit a message of length 3 bits. Notice that the entropy of this variable is given by

$$\mathrm{H}[x] = -8 \times \frac{1}{8} \log_2 \frac{1}{8} = 3 \text{ bits}.$$

Now consider an example (Cover and Thomas, 1991) of a variable having eight possible states $\{a, b, c, d, e, f, g, h\}$ for which the respective probabilities are given by $(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64})$. The entropy in this case is given by

$$\mathrm{H}[x] = -\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{4} \log_2 \frac{1}{4} - \frac{1}{8} \log_2 \frac{1}{8} - \frac{1}{16} \log_2 \frac{1}{16} - \frac{4}{64} \log_2 \frac{1}{64} = 2 \text{ bits}.$$

We see that the nonuniform distribution has a smaller entropy than the uniform one, and we will gain some insight into this shortly when we discuss the interpretation of entropy in terms of disorder. For the moment, let us consider how we would transmit the identity of the variable's state to a receiver. We could do this, as before, using a 3-bit number. However, we can take advantage of the nonuniform distribution by using shorter codes for the more probable events, at the expense of longer codes for the less probable events, in the hope of getting a shorter average code length. This can be done by representing the states $\{a, b, c, d, e, f, g, h\}$ using, for instance, the following set of code strings: 0, 10, 110, 1110, 111100, 111101, 111110, and 111111. The average length of the code that has to be transmitted is then

$$\text{average code length} = \frac{1}{2} \times 1 + \frac{1}{4} \times 2 + \frac{1}{8} \times 3 + \frac{1}{16} \times 4 + 4 \times \frac{1}{64} \times 6 = 2 \text{ bits},$$

which again is the same as the entropy of the random variable. Note that shorter code strings cannot be used because it must be possible to disambiguate a concatenation of such strings into its component parts. For instance, 11001110 decodes uniquely into the state sequence c, a, d. This relation between entropy and shortest coding length is a general one. The *noiseless coding theorem* (Shannon, 1948) states that the entropy is a lower bound on the number of bits needed to transmit the state of a random variable.

From now on, we will switch to the use of natural logarithms in defining entropy, as this will provide a more convenient link with ideas elsewhere in this book. In this case, the entropy is measured in units of *nats* (from 'natural logarithm') instead of bits, which differ simply by a factor of $\ln 2$.

2.5.2 Physics perspective

We have introduced the concept of entropy in terms of the average amount of information needed to specify the state of a random variable. In fact, the concept of entropy has much earlier origins in physics where it was introduced in the context of equilibrium thermodynamics and later given a deeper interpretation as a measure of disorder through developments in statistical mechanics. We can understand this alternative view of entropy by considering a set of N identical objects that are to be divided amongst a set of bins, such that there are n_i objects in the ith bin. Consider

the number of different ways of allocating the objects to the bins. There are N ways to choose the first object, $(N-1)$ ways to choose the second object, and so on, leading to a total of $N!$ ways to allocate all N objects to the bins, where $N!$ (pronounced 'N factorial') denotes the product $N \times (N-1) \times \cdots \times 2 \times 1$. However, we do not wish to distinguish between rearrangements of objects within each bin. In the ith bin there are $n_i!$ ways of reordering the objects, and so the total number of ways of allocating the N objects to the bins is given by

$$W = \frac{N!}{\prod_i n_i!}, \tag{2.82}$$

which is called the *multiplicity*. The entropy is then defined as the logarithm of the multiplicity scaled by a constant factor $1/N$ so that

$$\mathrm{H} = \frac{1}{N} \ln W = \frac{1}{N} \ln N! - \frac{1}{N} \sum_i \ln n_i!. \tag{2.83}$$

We now consider the limit $N \to \infty$, in which the fractions n_i/N are held fixed, and apply Stirling's approximation:

$$\ln N! \simeq N \ln N - N, \tag{2.84}$$

which gives

$$\mathrm{H} = -\lim_{N\to\infty} \sum_i \left(\frac{n_i}{N}\right) \ln \left(\frac{n_i}{N}\right) = -\sum_i p_i \ln p_i \tag{2.85}$$

where we have used $\sum_i n_i = N$. Here $p_i = \lim_{N\to\infty}(n_i/N)$ is the probability of an object being assigned to the ith bin. In physics terminology, the specific allocation of objects into bins is called a *microstate*, and the overall distribution of occupation numbers, expressed through the ratios n_i/N, is called a *macrostate*. The multiplicity W, which expresses the number of microstates in a given macrostate, is also known as the *weight* of the macrostate.

We can interpret the bins as the states x_i of a discrete random variable X, where $p(X = x_i) = p_i$. The entropy of the random variable X is then

$$\mathrm{H}[p] = -\sum_i p(x_i) \ln p(x_i). \tag{2.86}$$

Distributions $p(x_i)$ that are sharply peaked around a few values will have a relatively low entropy, whereas those that are spread more evenly across many values will have higher entropy, as illustrated in Figure 2.14.

Because $0 \leqslant p_i \leqslant 1$, the entropy is non-negative, and it will equal its minimum value of 0 when one of the $p_i = 1$ and all other $p_{j\neq i} = 0$. The maximum entropy configuration can be found by maximizing H using a Lagrange multiplier to enforce the normalization constraint on the probabilities. Thus, we maximize

Appendix C

$$\widetilde{\mathrm{H}} = -\sum_i p(x_i) \ln p(x_i) + \lambda \left(\sum_i p(x_i) - 1 \right) \tag{2.87}$$

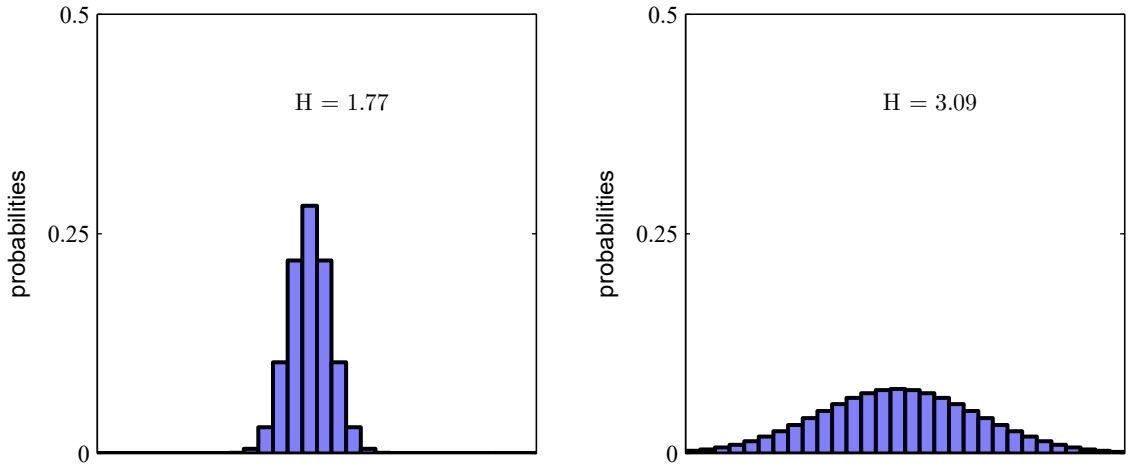

Figure 2.14 Histograms of two probability distributions over 30 bins illustrating the higher value of the entropy H for the broader distribution. The largest entropy would arise from a uniform distribution which would give $\mathrm{H} = -\ln(1/30) = 3.40$.

Exercise 2.22
Exercise 2.23

from which we find that all of the $p(x_i)$ are equal and are given by $p(x_i) = 1/M$ where M is the total number of states x_i. The corresponding value of the entropy is then $\mathrm{H} = \ln M$. This result can also be derived from Jensen's inequality (to be discussed shortly). To verify that the stationary point is indeed a maximum, we can evaluate the second derivative of the entropy, which gives

$$\frac{\partial \widetilde{\mathrm{H}}}{\partial p(x_i)\partial p(x_j)} = -I_{ij}\frac{1}{p_i} \tag{2.88}$$

where I_{ij} are the elements of the identity matrix. We see that these values are all negative and, hence, the stationary point is indeed a maximum.

2.5.3 Differential entropy

We can extend the definition of entropy to include distributions $p(x)$ over continuous variables x as follows. First divide x into bins of width Δ. Then, assuming that $p(x)$ is continuous, the *mean value theorem* (Weisstein, 1999) tells us that, for each such bin, there must exist a value x_i in the range $i\Delta \leqslant x_i \leqslant (i+1)\Delta$ such that

$$\int_{i\Delta}^{(i+1)\Delta} p(x)\,\mathrm{d}x = p(x_i)\Delta. \tag{2.89}$$

We can now quantize the continuous variable x by assigning any value x to the value x_i whenever x falls in the ith bin. The probability of observing the value x_i is then

$p(x_i)\Delta$. This gives a discrete distribution for which the entropy takes the form

$$\mathrm{H}_\Delta = -\sum_i p(x_i)\Delta \ln\left(p(x_i)\Delta\right) = -\sum_i p(x_i)\Delta \ln p(x_i) - \ln \Delta \tag{2.90}$$

where we have used $\sum_i p(x_i)\Delta = 1$, which follows from (2.89) and (2.25). We now omit the second term $-\ln\Delta$ on the right-hand side of (2.90), since it is independent of $p(x)$, and then consider the limit $\Delta \to 0$. The first term on the right-hand side of (2.90) will approach the integral of $p(x)\ln p(x)$ in this limit so that

$$\lim_{\Delta\to 0}\left\{-\sum_i p(x_i)\Delta \ln p(x_i)\right\} = -\int p(x)\ln p(x)\,\mathrm{d}x \tag{2.91}$$

where the quantity on the right-hand side is called the *differential entropy*. We see that the discrete and continuous forms of the entropy differ by a quantity $\ln\Delta$, which diverges in the limit $\Delta \to 0$. This reflects that specifying a continuous variable very precisely requires a large number of bits. For a density defined over multiple continuous variables, denoted collectively by the vector $\mathbf{x}$, the differential entropy is given by

$$\mathrm{H}[\mathbf{x}] = -\int p(\mathbf{x})\ln p(\mathbf{x})\,\mathrm{d}\mathbf{x}. \tag{2.92}$$

2.5.4 Maximum entropy

We saw for discrete distributions that the maximum entropy configuration corresponds to a uniform distribution of probabilities across the possible states of the variable. Let us now consider the corresponding result for a continuous variable. If this maximum is to be well defined, it will be necessary to constrain the first and second moments of $p(x)$ and to preserve the normalization constraint. We therefore maximize the differential entropy with the three constraints:

$$\int_{-\infty}^{\infty} p(x)\,\mathrm{d}x = 1 \tag{2.93}$$

$$\int_{-\infty}^{\infty} xp(x)\,\mathrm{d}x = \mu \tag{2.94}$$

$$\int_{-\infty}^{\infty} (x-\mu)^2 p(x)\,\mathrm{d}x = \sigma^2. \tag{2.95}$$

Appendix C The constrained maximization can be performed using Lagrange multipliers so that we maximize the following functional with respect to $p(x)$:

$$-\int_{-\infty}^{\infty} p(x)\ln p(x)\,\mathrm{d}x + \lambda_1\left(\int_{-\infty}^{\infty} p(x)\,\mathrm{d}x - 1\right) + \lambda_2\left(\int_{-\infty}^{\infty} xp(x)\,\mathrm{d}x - \mu\right) + \lambda_3\left(\int_{-\infty}^{\infty} (x-\mu)^2 p(x)\,\mathrm{d}x - \sigma^2\right). \tag{2.96}$$

Appendix B Using the calculus of variations, we set the derivative of this functional to zero giving

$$p(x) = \exp\left\{-1 + \lambda_1 + \lambda_2 x + \lambda_3 (x-\mu)^2\right\}. \tag{2.97}$$

Exercise 2.24 The Lagrange multipliers can be found by back-substitution of this result into the three constraint equations, leading finally to the result:

$$p(x) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \tag{2.98}$$

and so the distribution that maximizes the differential entropy is the Gaussian. Note that we did not constrain the distribution to be non-negative when we maximized the entropy. However, because the resulting distribution is indeed non-negative, we see with hindsight that such a constraint is not necessary.

Exercise 2.25 If we evaluate the differential entropy of the Gaussian, we obtain

$$\mathrm{H}[x] = \frac{1}{2}\left\{1 + \ln(2\pi\sigma^2)\right\}. \tag{2.99}$$

Thus, we see again that the entropy increases as the distribution becomes broader, i.e., as σ^2 increases. This result also shows that the differential entropy, unlike the discrete entropy, can be negative, because $\mathrm{H}(x) < 0$ in (2.99) for $\sigma^2 < 1/(2\pi e)$.

2.5.5 Kullback–Leibler divergence

So far in this section, we have introduced a number of concepts from information theory, including the key notion of entropy. We now start to relate these ideas to machine learning. Consider some unknown distribution $p(\mathbf{x})$, and suppose that we have modelled this using an approximating distribution $q(\mathbf{x})$. If we use $q(\mathbf{x})$ to construct a coding scheme for transmitting values of $\mathbf{x}$ to a receiver, then the average *additional* amount of information (in nats) required to specify the value of $\mathbf{x}$ (assuming we choose an efficient coding scheme) as a result of using $q(\mathbf{x})$ instead of the true distribution $p(\mathbf{x})$ is given by

$$\begin{aligned}\mathrm{KL}(p\|q) &= -\int p(\mathbf{x}) \ln q(\mathbf{x})\,\mathrm{d}\mathbf{x} - \left(-\int p(\mathbf{x}) \ln p(\mathbf{x})\,\mathrm{d}\mathbf{x}\right) \\ &= -\int p(\mathbf{x}) \ln\left\{\frac{q(\mathbf{x})}{p(\mathbf{x})}\right\}\mathrm{d}\mathbf{x}. \end{aligned} \tag{2.100}$$

This is known as the *relative entropy* or *Kullback–Leibler divergence*, or *KL divergence* (Kullback and Leibler, 1951), between the distributions $p(\mathbf{x})$ and $q(\mathbf{x})$. Note that it is not a symmetrical quantity, that is to say $\mathrm{KL}(p\|q) \not\equiv \mathrm{KL}(q\|p)$.

We now show that the Kullback–Leibler divergence satisfies $\mathrm{KL}(p\|q) \geqslant 0$ with equality if, and only if, $p(\mathbf{x}) = q(\mathbf{x})$. To do this we first introduce the concept of *convex* functions. A function $f(x)$ is said to be convex if it has the property that every chord lies on or above the function, as shown in Figure 2.15.

Any value of x in the interval from $x = a$ to $x = b$ can be written in the form $\lambda a + (1-\lambda)b$ where $0 \leqslant \lambda \leqslant 1$. The corresponding point on the chord

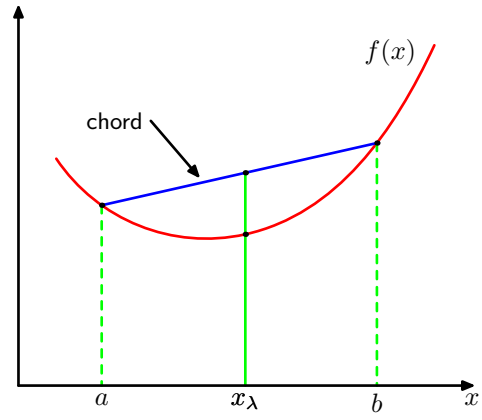

Figure 2.15 A convex function $f(x)$ is one for which every chord (shown in blue) lies on or above the function (shown in red).

is given by $\lambda f(a) + (1-\lambda)f(b)$, and the corresponding value of the function is $f(\lambda a + (1-\lambda)b)$. Convexity then implies

$$f(\lambda a + (1-\lambda)b) \leqslant \lambda f(a) + (1-\lambda)f(b). \tag{2.101}$$

Exercise 2.32 This is equivalent to the requirement that the second derivative of the function be everywhere positive. Examples of convex functions are $x \ln x$ (for $x > 0$) and x^2. A function is called *strictly convex* if the equality is satisfied only for $\lambda = 0$ and $\lambda = 1$. If a function has the opposite property, namely that every chord lies on or below the function, it is called *concave*, with a corresponding definition for *strictly concave*. If a function $f(x)$ is convex, then $-f(x)$ will be concave.

Exercise 2.33 Using the technique of proof by induction, we can show from (2.101) that a convex function $f(x)$ satisfies

$$f\left(\sum_{i=1}^{M} \lambda_i x_i\right) \leqslant \sum_{i=1}^{M} \lambda_i f(x_i) \tag{2.102}$$

where $\lambda_i \geqslant 0$ and $\sum_i \lambda_i = 1$, for any set of points $\{x_i\}$. The result (2.102) is known as *Jensen's inequality*. If we interpret the λ_i as the probability distribution over a discrete variable x taking the values $\{x_i\}$, then (2.102) can be written

$$f\left(\mathbb{E}[x]\right) \leqslant \mathbb{E}[f(x)] \tag{2.103}$$

where $\mathbb{E}[\cdot]$ denotes the expectation. For continuous variables, Jensen's inequality takes the form

$$f\left(\int \mathbf{x}p(\mathbf{x})\,\mathrm{d}\mathbf{x}\right) \leqslant \int f(\mathbf{x})p(\mathbf{x})\,\mathrm{d}\mathbf{x}. \tag{2.104}$$

We can apply Jensen's inequality in the form (2.104) to the Kullback–Leibler divergence (2.100) to give

$$\mathrm{KL}(p\|q) = -\int p(\mathbf{x}) \ln \left\{\frac{q(\mathbf{x})}{p(\mathbf{x})}\right\} \mathrm{d}\mathbf{x} \geqslant -\ln \int q(\mathbf{x})\,\mathrm{d}\mathbf{x} = 0 \tag{2.105}$$

where we have used $-\ln x$ is a convex function, together with the normalization condition $\int q(\mathbf{x})\,\mathrm{d}\mathbf{x} = 1$. In fact, $-\ln x$ is a strictly convex function, so the equality will hold if, and only if, $q(\mathbf{x}) = p(\mathbf{x})$ for all $\mathbf{x}$. Thus, we can interpret the Kullback–Leibler divergence as a measure of the dissimilarity of the two distributions $p(\mathbf{x})$ and $q(\mathbf{x})$.

We see that there is an intimate relationship between data compression and density estimation (i.e., the problem of modelling an unknown probability distribution) because the most efficient compression is achieved when we know the true distribution. If we use a distribution that is different from the true one, then we must necessarily have a less efficient coding, and on average the additional information that must be transmitted is (at least) equal to the Kullback–Leibler divergence between the two distributions.

Suppose that data is being generated from an unknown distribution $p(\mathbf{x})$ that we wish to model. We can try to approximate this distribution using some parametric distribution $q(\mathbf{x}|\boldsymbol{\theta})$, governed by a set of adjustable parameters $\boldsymbol{\theta}$. One way to determine $\boldsymbol{\theta}$ is to minimize the Kullback–Leibler divergence between $p(\mathbf{x})$ and $q(\mathbf{x}|\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. We cannot do this directly because we do not know $p(\mathbf{x})$. Suppose, however, that we have observed a finite set of training points $\mathbf{x}_n$, for $n = 1, \ldots, N$, drawn from $p(\mathbf{x})$. Then the expectation with respect to $p(\mathbf{x})$ can be approximated by a finite sum over these points, using (2.40), so that

$$\mathrm{KL}(p\|q) \simeq \frac{1}{N}\sum_{n=1}^{N}\Big\{-\ln q(\mathbf{x}_n|\boldsymbol{\theta}) + \ln p(\mathbf{x}_n)\Big\}. \tag{2.106}$$

The second term on the right-hand side of (2.106) is independent of $\boldsymbol{\theta}$, and the first term is the negative log likelihood function for $\boldsymbol{\theta}$ under the distribution $q(\mathbf{x}|\boldsymbol{\theta})$ evaluated using the training set. Thus, we see that minimizing this Kullback–Leibler divergence is equivalent to maximizing the log likelihood function.

Exercise 2.34

2.5.6 Conditional entropy

Now consider the joint distribution between two sets of variables $\mathbf{x}$ and $\mathbf{y}$ given by $p(\mathbf{x}, \mathbf{y})$ from which we draw pairs of values of $\mathbf{x}$ and $\mathbf{y}$. If a value of $\mathbf{x}$ is already known, then the additional information needed to specify the corresponding value of $\mathbf{y}$ is given by $-\ln p(\mathbf{y}|\mathbf{x})$. Thus the average additional information needed to specify $\mathbf{y}$ can be written as

$$\mathrm{H}[\mathbf{y}|\mathbf{x}] = -\iint p(\mathbf{y}, \mathbf{x}) \ln p(\mathbf{y}|\mathbf{x})\,\mathrm{d}\mathbf{y}\,\mathrm{d}\mathbf{x}, \tag{2.107}$$

which is called the *conditional entropy* of $\mathbf{y}$ given $\mathbf{x}$. It is easily seen, using the product rule, that the conditional entropy satisfies the relation:

Exercise 2.35

$$\mathrm{H}[\mathbf{x}, \mathbf{y}] = \mathrm{H}[\mathbf{y}|\mathbf{x}] + \mathrm{H}[\mathbf{x}] \tag{2.108}$$

where $\mathrm{H}[\mathbf{x}, \mathbf{y}]$ is the differential entropy of $p(\mathbf{x}, \mathbf{y})$ and $\mathrm{H}[\mathbf{x}]$ is the differential entropy of the marginal distribution $p(\mathbf{x})$. Thus, the information needed to describe $\mathbf{x}$ and $\mathbf{y}$ is given by the sum of the information needed to describe $\mathbf{x}$ alone plus the additional information required to specify $\mathbf{y}$ given $\mathbf{x}$.

2.5.7 Mutual information

When two variables $\mathbf{x}$ and $\mathbf{y}$ are independent, their joint distribution will factorize into the product of their marginals $p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{y})$. If the variables are not independent, we can gain some idea of whether they are 'close' to being independent by considering the Kullback–Leibler divergence between the joint distribution and the product of the marginals, given by

$$\begin{aligned} \mathrm{I}[\mathbf{x}, \mathbf{y}] &\equiv \mathrm{KL}(p(\mathbf{x}, \mathbf{y}) \| p(\mathbf{x})p(\mathbf{y})) \\ &= -\iint p(\mathbf{x}, \mathbf{y}) \ln \left(\frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x}, \mathbf{y})} \right) \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y}, \end{aligned} \tag{2.109}$$

which is called the *mutual information* between the variables $\mathbf{x}$ and $\mathbf{y}$. From the properties of the Kullback–Leibler divergence, we see that $I[\mathbf{x}, \mathbf{y}] \geqslant 0$ with equality if, and only if, $\mathbf{x}$ and $\mathbf{y}$ are independent. Using the sum and product rules of probability, we see that the mutual information is related to the conditional entropy through

Exercise 2.38

$$\mathrm{I}[\mathbf{x}, \mathbf{y}] = \mathrm{H}[\mathbf{x}] - \mathrm{H}[\mathbf{x}|\mathbf{y}] = \mathrm{H}[\mathbf{y}] - \mathrm{H}[\mathbf{y}|\mathbf{x}]. \tag{2.110}$$

Thus, the mutual information represents the reduction in the uncertainty about $\mathbf{x}$ by virtue of being told the value of $\mathbf{y}$ (or vice versa). From a Bayesian perspective, we can view $p(\mathbf{x})$ as the prior distribution for $\mathbf{x}$ and $p(\mathbf{x}|\mathbf{y})$ as the posterior distribution after we have observed new data $\mathbf{y}$. The mutual information therefore represents the reduction in uncertainty about $\mathbf{x}$ as a consequence of the new observation $\mathbf{y}$.

2.6. Bayesian Probabilities

When we considered the bent coin in Figure 2.2, we introduced the concept of probability in terms of the frequencies of random, repeatable events, such as the probability of the coin landing concave side up. We will refer to this as the *classical* or *frequentist* interpretation of probability. We also introduced the more general *Bayesian* view, in which probabilities provide a quantification of uncertainty. In this case, our uncertainty is whether the concave side of the coin is heads or tails.

The use of probability to represent uncertainty is not an ad hoc choice but is inevitable if we are to respect common sense while making rational and coherent inferences. For example, Cox (1946) showed that if numerical values are used to represent degrees of belief, then a simple set of axioms encoding common sense properties of such beliefs leads uniquely to a set of rules for manipulating degrees of belief that are equivalent to the sum and product rules of probability. It is therefore natural to refer to these quantities as (Bayesian) probabilities.

For the bent coin we assumed, in the absence of further information, that the probability of the concave side of the coin being heads is 0.5. Now suppose we are told the results of flipping the coin a few times. Intuitively, it seems that this should provide us with some information as to whether the concave side is heads. For instance, suppose we see many more flips that land tails than land heads. Given

that the coin is more likely to land concave side up, this provides evidence to suggest that the concave side is more likely to be tails. In fact, this intuition is correct, and furthermore, we can quantify this using the rules of probability. Bayes' theorem now acquires a new significance, because it allows us to convert the prior probability for the concave side being heads into a posterior probability by incorporating the data provided by the coin flips. Moreover, this process is iterative, meaning the posterior probability becomes the prior for incorporating data from further coin flips.

Exercise 2.40

One aspect of the Bayesian viewpoint is that the inclusion of prior knowledge arises naturally. Suppose, for instance, that a fair-looking coin is tossed three times and lands heads each time. The maximum likelihood estimate of the probability of landing heads would give 1, implying that all future tosses will land heads! By contrast, a Bayesian approach with any reasonable prior will lead to a less extreme conclusion.

Section 3.1.2

2.6.1 Model parameters

The Bayesian perspective provides valuable insights into several aspects of machine learning, and we can illustrate these using the sine curve regression example. Here we denote the training data set by $\mathcal{D}$. We have already seen in the context of linear regression that the parameters can be chosen using *maximum likelihood*, in which $\mathbf{w}$ is set to the value that maximizes the likelihood function $p(\mathcal{D}|\mathbf{w})$. This corresponds to choosing the value of $\mathbf{w}$ for which the probability of the observed data set is maximized. In the machine learning literature, the negative log of the likelihood function is called an *error function*. Because the negative logarithm is a monotonically decreasing function, maximizing the likelihood is equivalent to minimizing the error. This leads to a specific choice of parameter values, denoted $\mathbf{w}_{\mathrm{ML}}$, which are then used to make predictions for new data.

Section 1.2

We have seen that different choices of training data set, for example containing different numbers of data points, give rise to different solutions for $\mathbf{w}_{\mathrm{ML}}$. From a Bayesian perspective, we can also use the machinery of probability theory to describe this uncertainty in the model parameters. We can capture our assumptions about $\mathbf{w}$, *before* observing the data, in the form of a prior probability distribution $p(\mathbf{w})$. The effect of the observed data $\mathcal{D}$ is expressed through the likelihood function $p(\mathcal{D}|\mathbf{w})$, and Bayes' theorem now takes the form

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})}, \tag{2.111}$$

which allows us to evaluate the uncertainty in $\mathbf{w}$ *after* we have observed $\mathcal{D}$ in the form of the posterior probability $p(\mathbf{w}|\mathcal{D})$.

It is important to emphasize that the quantity $p(\mathcal{D}|\mathbf{w})$ is called the likelihood function when it is viewed as a function of the parameter vector $\mathbf{w}$, and it expresses how probable the observed data set is for different values of $\mathbf{w}$. Note that the likelihood $p(\mathcal{D}|\mathbf{w})$ is not a probability distribution over $\mathbf{w}$, and its integral with respect to $\mathbf{w}$ does not (necessarily) equal one.

Given this definition of likelihood, we can state Bayes' theorem in words:

$$\text{posterior} \propto \text{likelihood} \times \text{prior} \tag{2.112}$$

where all of these quantities are viewed as functions of $\mathbf{w}$. The denominator in (2.111) is the normalization constant, which ensures that the posterior distribution on the left-hand side is a valid probability density and integrates to one. Indeed, by integrating both sides of (2.111) with respect to $\mathbf{w}$, we can express the denominator in Bayes' theorem in terms of the prior distribution and the likelihood function:

$$p(\mathcal{D}) = \int p(\mathcal{D}|\mathbf{w})p(\mathbf{w})\,\mathrm{d}\mathbf{w}. \tag{2.113}$$

In both the Bayesian and frequentist paradigms, the likelihood function $p(\mathcal{D}|\mathbf{w})$ plays a central role. However, the manner in which it is used is fundamentally different in the two approaches. In a frequentist setting, $\mathbf{w}$ is considered to be a fixed parameter, whose value is determined by some form of 'estimator', and error bars on this estimate are determined (conceptually, at least) by considering the distribution of possible data sets $\mathcal{D}$. By contrast, from the Bayesian viewpoint there is only a single data set $\mathcal{D}$ (namely the one that is actually observed), and the uncertainty in the parameters is expressed through a probability distribution over $\mathbf{w}$.

2.6.2 Regularization

Section 1.2.5

We can use this Bayesian perspective to gain insight into the technique of regularization that was used in the sine curve regression example to reduce over-fitting. Instead of choosing the model parameters by maximizing the likelihood function with respect to $\mathbf{w}$, we can maximize the posterior probability (2.111). This technique is called the *maximum a posteriori* estimate, or simply *MAP* estimate. Equivalently, we can minimize the negative log of the posterior probability. Taking negative logs of both sides of (2.111), we have

$$-\ln p(\mathbf{w}|\mathcal{D}) = -\ln p(\mathcal{D}|\mathbf{w}) - \ln p(\mathbf{w}) + \ln p(\mathcal{D}). \tag{2.114}$$

The first term on the right-hand side of (2.114) is the usual log likelihood. The third term can be omitted since it does not depend on $\mathbf{w}$. The second term takes the form of a function of $\mathbf{w}$, which is added to the log likelihood, and we can recognize this as a form of regularization. To make this more explicit, suppose we choose the prior distribution $p(\mathbf{w})$ to be the product of independent zero-mean Gaussian distributions for each of the elements of $\mathbf{w}$ such that each has the same variance s^2 so that

$$p(\mathbf{w}|s) = \prod_{i=0}^{M} \mathcal{N}(w_i|0, s^2) = \prod_{i=0}^{M} \left(\frac{1}{2\pi s^2}\right)^{1/2} \exp\left\{-\frac{w_i^2}{2s^2}\right\}. \tag{2.115}$$

Substituting into (2.114), we obtain

$$-\ln p(\mathbf{w}|\mathcal{D}) = -\ln p(\mathcal{D}|\mathbf{w}) + \frac{1}{2s^2}\sum_{i=0}^{M} w_i^2 + \text{const}. \tag{2.116}$$

If we consider the particular case of the linear regression model whose log likelihood is given by (2.66), then we find that maximizing the posterior distribution is equivalent to minimizing the function

Exercise 2.41

$$E(\mathbf{w}) = \frac{1}{2\sigma^2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{1}{2s^2}\mathbf{w}^{\mathrm{T}}\mathbf{w}. \tag{2.117}$$

We see that this takes the form of the regularized sum-of-squares error function encountered earlier in the form (1.4).

2.6.3 Bayesian machine learning

The Bayesian perspective has allowed us to motivate the use of regularization and to derive a specific form for the regularization term. However, the use of Bayes' theorem alone does not constitute a truly Bayesian treatment of machine learning since it is still finding a single solution for $\mathbf{w}$ and does not therefore take account of uncertainty in the value of $\mathbf{w}$. Suppose we have a training data set $\mathcal{D}$ and our goal is to predict some target variable t given a new input value x. We are therefore interested in the distribution of t given both x and $\mathcal{D}$. From the sum and product rules of probability, we have

$$p(t|x, \mathcal{D}) = \int p(t|x, \mathbf{w})p(\mathbf{w}|\mathcal{D})\,\mathrm{d}\mathbf{w}. \tag{2.118}$$

We see that the prediction is obtained by taking a weighted average $p(t|x, \mathbf{w})$ over all possible values of $\mathbf{w}$ in which the weighting function is given by the posterior probability distribution $p(\mathbf{w}|\mathcal{D})$. The key difference that distinguishes Bayesian methods is this integration over the space of parameters. By contrast, conventional frequentist methods use point estimates for parameters obtained by optimizing a loss function such as a regularized sum-of-squares.

This fully Bayesian treatment of machine learning offers some powerful insights. For example, the problem of over-fitting, encountered earlier in the context of polynomial regression, is an example of a pathology arising from the use of maximum likelihood, and does not arise when we marginalize over parameters using the Bayesian approach. Similarly, we may have multiple potential models that we could use to solve a given problem, such as polynomials of different orders in the regression example. A maximum likelihood approach simply picks the model that gives the highest probability of the data, but this favours ever more complex models, leading to over-fitting. A fully Bayesian treatment involves averaging over all possible models, with the contribution of each model weighted by its posterior probability. Moreover, this probability is typically highest for models of intermediate complexity. Very simple models (such as polynomials of low order) have low probability as they are unable to fit the data well, whereas very complex models (such as polynomials of very high order) also have low probability because the Bayesian integration over parameters automatically and elegantly penalizes complexity. For a comprehensive overview of Bayesian methods applied to machine learning, including neural networks, see Bishop (2006).

Section 1.2

Section 9.6

Unfortunately, there is a major drawback with the Bayesian framework, and this is apparent in (2.118), which involves integrating over the space of parameters. Modern deep learning models can have millions or billions of parameters and even simple approximations to such integrals are typically infeasible. In fact, given a

limited compute budget and an ample source of training data, it will often be better to apply maximum likelihood techniques, generally augmented with one or more forms of regularization, to a large neural network rather than apply a Bayesian treatment to a much smaller model.

Exercises

2.1 ($\star$) In the cancer screening example, we used a prior probability of cancer of $p(C = 1) = 0.01$. In reality, the prevalence of cancer is generally very much lower. Consider a situation in which $p(C = 1) = 0.001$, and recompute the probability of having cancer given a positive test $p(C = 1|T = 1)$. Intuitively, the result can appear surprising to many people since the test seems to have high accuracy and yet a positive test still leads to a low probability of having cancer.

2.2 ($\star\star$) Deterministic numbers satisfy the property of *transitivity*, so that if $x > y$ and $y > z$ then it follows that $x > z$. When we go to random numbers, however, this property need no longer apply. Figure 2.16 shows a set of four cubical dice that have been arranged in a cyclic order. Show that each of the four dice has a 2/3 probability of rolling a higher number than the previous die in the cycle. Such dice are known as *non-transitive dice*, and the specific examples shown here are called *Efron dice*.

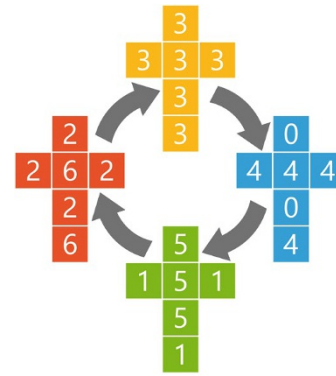

Figure 2.16 An example of non-transitive cubical dice, in which each die has been 'flattened' to reveal the numbers on each of the faces. The dice have been arranged in a cycle, such that each die has a 2/3 probability of rolling a higher number than the previous die in the cycle.

2.3 ($\star$) Consider a variable $\mathbf{y}$ given by the sum of two independent random variables $\mathbf{y} = \mathbf{u} + \mathbf{v}$ where $\mathbf{u} \sim p_{\mathbf{u}}(\mathbf{u})$ and $\mathbf{v} \sim p_{\mathbf{v}}(\mathbf{v})$. Show that the distribution $p_{\mathbf{y}}(\mathbf{y})$ is given by

$$p(\mathbf{y}) = \int p_{\mathbf{u}}(\mathbf{u}) p_{\mathbf{v}}(\mathbf{y} - \mathbf{u}) \, \mathrm{d}\mathbf{u}. \tag{2.119}$$

This is known as the *convolution* of $p_{\mathbf{u}}(\mathbf{u})$ and $p_{\mathbf{v}}(\mathbf{v})$.

2.4 ($\star\star$) Verify that the uniform distribution (2.33) is correctly normalized, and find expressions for its mean and variance.

2.5 ($\star\star$) Verify that the exponential distribution (2.34) and the Laplace distribution (2.35) are correctly normalized.