1

Orientation

1.1 Risky Inference

Inference is the process of arriving at a conclusion based on evidence or data. Sometimes it is possible to draw conclusions which are certain to be true. Here is a well-known example. You are in a strange land where each person either always lies or always tells the truth. One day you meet a couple, Hilda and Godiva. You ask Hilda whether at least one of them is a liar; Hilda says 'yes'. What kind of people are Hilda and Godiva? Hilda's answer is your evidence, and there is only one possible conclusion. Godiva is a liar and Hilda is a truth teller.

■ ***Exercise*** 1.1 To explain this inference, we can list all the possible combinations of liar (L) and truth teller (T). Writing Hilda's type first and Godiva's second, these are $\{TT, TL, LT, LL\}$. Why is TL the only combination consistent with the evidence?

Now suppose that liars and truth tellers occasionally act against type. That is, a truth teller might spontaneously lie once in a while, and a liar will speak the truth from time to time. What can you infer from Hilda's answer now? 'Godiva is a liar and Hilda is a truth teller' might still be a good guess, but (on our new supposition) *all* the hypotheses listed in Exercise 1.1 are consistent with the evidence. For example, Hilda and Godiva might both be natural truth tellers, but on the evening you met them, Hilda decided to lie. Whatever your conclusion, you can no longer be sure it is correct. You are dealing with a *risky inference* problem. To get a better grip on it, more information – some *observational data* – would be helpful. How often do liars and truth tellers behave against type? Are liars and truth tellers equally common in the population, or does one type predominate? Do liars tend to couple up with liars and truth tellers with truth tellers, or do opposites attract? If you knew the answers to those questions, you might be able to attach *probabilities* to the four possible conclusions. But you still won't know for sure which one is true.

■ ***Exercise*** 1.2 Suppose that people act against type 20% of the time, and that a priori (i.e. before Hilda speaks) you attach equal probability to each of the four combinations of liar and truth teller. Show that after Hilda speaks, the probability you should attach to the conclusion 'Godiva is a liar and Hilda is a truth teller' is $\frac{4}{7}$.

Risky inference problems are very common, and this is a book about how to tackle them. The methods we present will not always involve probabilistic reasoning, but they often will.

As our title suggests, we engage with two closely related disciplines: statistical modelling

https://doi.org/10.1017/9781009630696.002 Published online by Cambridge University Press

and machine learning. Both are centrally concerned with problems of risky inference, and they tackle these problems by learning from data. Historically the goal of statistical modelling was to infer probabilistic models from data, and to do so in a mathematically rigorous way. The goal of machine learning was to enable computers (machines) to learn from data without explicit programming. The boundary between the two has blurred. Statistical modelling now relies heavily on computational tools, and probabilistic methods are fundamental to machine learning, providing theoretical and practical tools for handling uncertainty and generalisation. Some differences persist. Machine learning researchers remain more willing than statisticians to use 'black box' models (e.g. large neural networks) whose inner workings are difficult for humans to interpret. In contrast, statistical modelling places greater emphasis on understanding data-generating processes, favouring models that are relatively simple and human-interpretable. This book does not dwell on disciplinary distinctions. With the partial exception of Chapter 13, our focus is on ideas and techniques that are central to both fields.

1.2 Supervised Learning

Here are some examples of risky inference tasks:

- given a series of clinical measurements from a single individual, including blood sugar levels, weight, height, blood pressure and so on, predict whether that person has diabetes;
- given a photograph of a plant, predict the plant's species;
- given an audio recording of a person speaking, predict the person's age.

Although these tasks look superficially dissimilar, they share a common structure. There is some variable, y, whose value you wish to predict. (In simple cases y will just be a number or a categorical label.) In order to make your prediction, you have some case-specific information, x. This might be one or more numbers or measurements, an image, an audio file, some text, or some combination of these things. We often refer to x as the *predictor*, and y as the *response*. The aim is to find a *prediction function*, $\hat{g}$, which makes a prediction, $\hat{y}$, of y, given x:

$$\hat{y} = \hat{g}(x).$$

We follow standard practice here and use hat symbols to indicate predictions or estimates. The prediction function gets a hat because we think of it as an estimate of an unknown ideal function g that makes the best possible predictions of y given x.

Now, suppose that a dataset consisting of many previous examples of (x, y) pairs is available. If $\hat{g}$ is a good prediction function, we would expect it to perform well on these past examples. That is, we would expect $\hat{g}(x)$ to lie close to y for most of the (x, y) pairs in the dataset. Seeking a function that has this property is a powerful and useful strategy, but we will see later that it must be applied cautiously. It is almost always possible to construct a 'wiggly' prediction function which matches previous examples very well.[1] Unfortunately, such functions can often be poor at predicting y for values of x which were not in the original dataset. This pitfall is called *overfitting* (see Figure 1.1), and we will need to find ways to

[1] One situation in which this is not possible is where the dataset contains many pairs of examples of the form (x, y) and (x, y'), with y and y' taking wildly dissimilar values.

https://doi.org/10.1017/9781009630696.002 Published online by Cambridge University Press

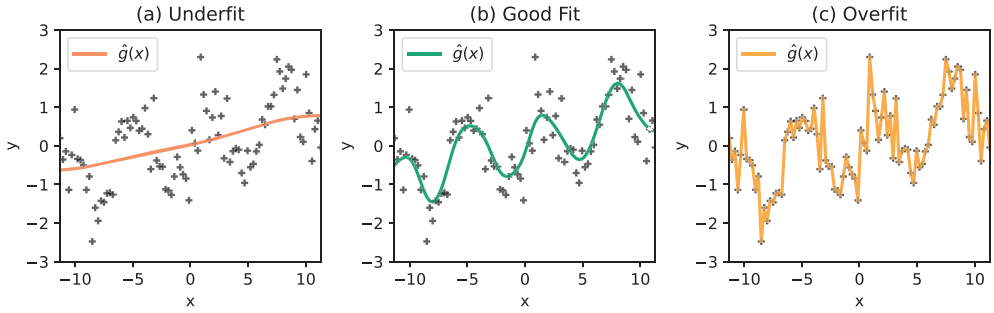


Figure 1.1 Three different candidates for prediction function $\hat{g}$, fitted to the same set of predictor–response (x, y) pairs. Intuitively, it seems clear that in graph (a), $\hat{g}$ is too simple to capture the systematic relationship between x and y. The prediction function in graph (b) seems to do a good job, while in graph (c), the prediction function appears to have fit to noise in the data.

avoid it. Broadly speaking, the trick will be to constrain – or at least to encourage – $\hat{g}$ to take a reasonably simple form. (Of course, what counts as 'reasonably simple' will often depend on the context.)

The process of searching for an optimal $\hat{g}$ is known as *supervised learning*, because for each predictor value we have a response y available to help us decide how well we have done. Supervised learning tasks are called *regression* tasks when the response is a continuously varying quantity, and *classification* tasks when the response is a categorical label.

■ ***Exercise*** 1.3 For each of the risky inference tasks listed at the start of this section, state whether it is best thought of as a regression or a classification task.

■ ***Exercise*** 1.4 You have reason to believe that the distance travelled by a cannonball, y, is directly proportional to the volume of gunpowder used to fire it, x. You perform three experiments, obtaining the (x, y) pairs $(1, 5)$, $(2, 11)$ and $(3, 14)$. How much gunpowder would you use if you wanted the cannonball to travel a distance $y = 7$? This is an example of risky inference: there isn't a provably correct answer. How would you justify yours?

We will study a regression task not too different from that of Exercise 1.4 in Chapter 2, and we will see many further examples of regression in Chapters 5, 7, 9 and 13. In Chapters 11 and 13, we will also study classification.

1.3 Unsupervised Learning

Supervised learning is not the only kind of inference we will be interested in. Consider the following example. As a bank customer, you make purchases of all sorts of sizes in different geographical locations or online stores. It is in your bank's interest to detect if someone has

https://doi.org/10.1017/9781009630696.002 Published online by Cambridge University Press

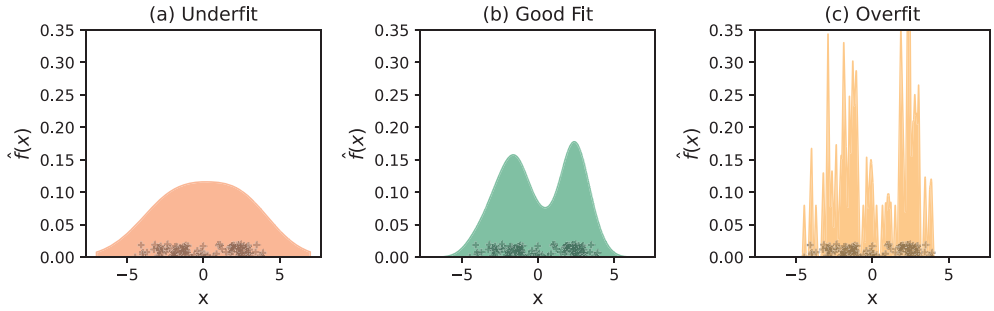


Figure 1.2 Three different candidates for a density function $\hat{f}$, fitted to the same set of data $x_1, x_2, \ldots, x_{50}$ (plotted as crosses at the base of each plot with some vertical jitter to aid visualisation). The data forms two visible clusters on either side of the origin. In graph (a), $\hat{f}$ is too simple to capture the two clusters. The density function in graph (b) seems to do a good job at capturing the clusters, while in graph (c), the density function has fit to noise in the data.

stolen your card or otherwise obtained access to your account. At first sight, this kind of fraud detection might look like a supervised problem: the predictors are the transaction sizes and locations, and the response is either 'fraudulent' or 'genuine'. In reality, fraud is quite rare, and there are lots of different ways in which it is committed, so it would be difficult to collect enough examples to learn an effective prediction function. Let's think about this another way. We do have plenty of examples of legitimate transactions for every cardholder. To keep things simple, we'll imagine that we know only the sizes of these transactions, which we could write as a long list $x_1, x_2, x_3, \ldots$. Suppose we could find a function $\hat{f}$ which measured how *typical* a transaction size was, with $\hat{f}(x)$ taking large values for typical values of x and small values for unusual ones. We could use this to detect *anomalies*: transaction sizes which stand out from the others. If x can be regarded as a continuous variable, it is natural to constrain $\hat{f}$ to be a probability density. The task of learning $\hat{f}$ from the data will then be an example of *density estimation*. This is an *unsupervised* task because we aren't given 'true' values of the density to compare our estimates to. Despite this, it is still possible to devise methods for learning good densities. Our density should describe the existing observations well, but not too well. For example, it would clearly be foolish to assume that in order to be typical, a new transaction would need to be exactly the same size as one of the previous transactions. That assumption would amount to *overfitting* – that is, matching previous observations too closely, and not allowing for the possibility of future deviations. (For stylised examples, see Figure 1.2.)

More generally, unsupervised learning involves the search for structure or patterns within a set of observations. Density estimation is one way to discover structure. However, there are other approaches which can often yield more practically useful and interpretable results. As an example, consider the problem of learning the sets of distinct sounds, or *phones*, that humans use to communicate. Each human language uses a subset of the possible sounds that the vocal tract is physically capable of making, and as a child learns to speak and understand, they must learn what their language's set is. How might we, as scientists, discover how

https://doi.org/10.1017/9781009630696.002 Published online by Cambridge University Press

many phones a given language uses? An experienced speaker of standard English knows instinctively that the vowel sounds in 'had', 'head', 'hit', 'heed', 'hot', 'hood', 'hud' and 'heard' are all different, but are these *all* the possible vowels? How would we go about counting how many there are in the language as a whole? One approach would be to measure frequency spectra, represented as high-dimensional vectors, emitted from the vocal tract when uttering a wide range of vowels. We could then attempt to identify groups of similar spectra. Identifying groups in this way is an inference problem known as *clustering*.

A typical clustering method is to assign each observation a group label and then iteratively relabel so that observations with the same label are close to each other. Unfortunately, if we tried to cluster vowel spectra using the above method, we would soon run into 'the curse of dimensionality'. In low dimensions, simple measures of distance between observations are usually meaningful. For example, we might be interested in identifying clusters of disease outbreaks based on a set of geographical coordinates of individual cases. In this case, the Euclidean distance between cases makes sense as a measure. With high-dimensional observations such as vowel spectra, standard distance measures may cease to be useful, as Exercise 1.5 illustrates.

■ ***Exercise*** 1.5 Consider a three-dimensional vector $\boldsymbol{U} = (U_1, U_2, U_3)$, where each coordinate U_i is selected uniformly and independently at random from $[0, 1]$. The vector $\boldsymbol{U}$ represents the location of a random point inside a unit cube. By adding extra coordinates, so that we have d dimensions, we can represent a random point inside a d-dimensional hypercube,

$$\boldsymbol{U} = (U_1, U_2, \ldots, U_d).$$

For a given value of d, imagine that you have randomly selected many such points. Consider any pair of them. The square of the difference in their ith coordinates is a random variable with some mean μ and standard deviation σ.[a] (Of course, neither μ nor σ depends on d.)

(a) Explain why the square of the distance between two such points has mean μd and standard deviation $\sigma\sqrt{d}$. (Hint: recall that the variance of the sum of independent random variables is the sum of their variances.)
(b) Explain why the ratio of the largest interpoint distance to the smallest interpoint distance tends to one as the dimension d increases (but the number of points is held fixed).

[a] The interested reader may wish to show that $\mu = 1/6$ and $\sigma^2 = 7/180$.

The above exercise reveals that in high dimensions, the contrast between narrowly and distantly separated observations begins to disappear, making it difficult to find groups. One way to tackle this problem is to look for ways of describing our observations using a smaller number of 'factors of variation'. Vowel sounds, for example, can often be described in terms of a small number of resonant frequencies. In general, however, we want automatic methods for boiling high-dimensional data down to a small number of features which describe its essence as efficiently as possible. Such *dimensionality reduction* techniques, in common with other unsupervised methods, are about finding patterns and structure in data. They have

https://doi.org/10.1017/9781009630696.002 Published online by Cambridge University Press

broad applicability, from visualising and explaining observations to improving predictions made using them.

We will play with a simple density estimation task in Chapter 3 and consider density estimation problems using frequentist and Bayesian methods in Chapters 6 and 10. We take a closer look at unsupervised learning in general in Chapter 12.

1.4 Reasoning with Probability: Frequentist and Bayesian Approaches

Risky inference involves guesswork. Given a number of alternative explanations of our observations, or alternative predictions, we want to find methods to choose between them and to quantify how certain we are about our guesses. The mathematical tool which allows us to quantify uncertainty and chance is probability. It is invaluable for doing inference. However, defining exactly what we mean when we make statements like 'the probability that the next coin toss will yield heads is 50%' or 'the probability that the defendant is innocent is only 1%' is difficult. There are two alternative ways to do it: *frequentist* and *Bayesian*. People sometimes talk about *frequentist inference* and *Bayesian inference* as opposed schools of thought. We prefer to see them as two conceptual toolkits, both of which are useful; but it is true that there are two distinct concepts of probability in play.

Frequentist inference relies on a definition of probability as long-run relative frequency. When construed frequentistically, the statement 'my coin lands heads with probability 0.5 when tossed' *means* that the proportion of heads in a sequence of N tosses of my coin would converge to 0.5 as N tends to infinity. Bayesian inference, by contrast, uses probabilities to represent 'degrees of belief' in *anything* that we are not sure about, be that the value of a constant, a prediction, or a scientific hypothesis. The starting point of Bayesian inference is a precise probabilistic statement of the prior beliefs we are bringing to the inference problem. Inference then consists of updating our beliefs in light of observations. Some people feel that the Bayesian interpretation of probability yields a more unified and satisfying approach to inference. It allows us to address a wider range of questions in a way that is closer to how we naturally think about the process of learning from observations. On the downside, we lose the objectivity of the frequentist definition of probability: degrees of belief are inherently subjective. However, we will see that the conclusions reached using the two approaches are often similar. Box 1.1 summarises the two interpretations of probability.

Chapters 4–6 develop some conceptual and mathematical machinery for thinking about inference probabilistically. Chapter 6 introduces core ideas of frequentist and Bayesian inference, and contrasts them. We will draw on both throughout the remainder of the book. Bayesian statistical modelling is covered in detail in Chapter 10.

★ Box 1.1 Frequentist and Bayesian probability

According to the **frequentist** view, the probability of an event A is the long run relative frequency with which it occurs over a repeated sequence of trials. If $N(A)$ is the number of times that A occurs out of N trials, then $\mathbf{Pr}(A) = \lim_{N\to\infty} N(A)/N$. According to the **Bayesian** view, $\mathbf{Pr}(A)$ is our degree of belief that A will occur or has occurred.

https://doi.org/10.1017/9781009630696.002 Published online by Cambridge University Press

1.5 Deep Learning

Neural networks – architectures for building complex functions loosely inspired by the brain – are currently driving a revolution in artificial intelligence. The received view in the 1990s was that neural networks were difficult to train. Precisely because they were such supremely flexible function approximators, it seemed inevitable that large networks would overfit to their training data. Now, however, we routinely see neural networks with hundreds of billions of tunable parameters trained successfully on just a few tens of gigabytes of data. Exactly why contemporary architectures work so well remains to some extent an open question, but depth appears to be an important factor. Networks built out of many layers of neurons, with the ith layer's output serving as the $(i + 1)$th layer's input, are often easier to train than shallower networks with fewer layers. 'Deep learning' has therefore become the favoured catch-all label for designing, training and applying large neural networks.

This is our topic in Chapter 13. It is, of course, a *big* topic. We aim merely to introduce deep learning and to show readers who have worked through Chapters 1–12 that they are ready to explore it in detail. Much of the terrain will be familiar: training methods are partly inspired by probabilistic inference ideas; constraining network complexity is essential to avoid overfitting; and neural networks often learn most effectively when their architectures reflect prior knowledge about the domains in which they will be applied.

Inference has been the engine of science for generations, but there has never been a more exciting time to study it as a subject in its own right.

1.6 Chapter Summary

The purpose of this chapter has been to give an overview of the kinds of inference problems we are going to solve, and of the different ways that inference is done. Let us summarise the main points, with key words in bold.

- We are interestedin **risky inference**: drawing conclusions, learning, and making predictions when we cannot be certain about the truth.
- Learning how to predict a *response* from one or more *predictors* based on previous predictor-response pairs is called **supervised learning**. Learning to predict a continuous response is called **regression**, whereas learning to predict a categorical label is **classification**.
- Inference problems in which there is no response variable (and we simply seek patterns and structure in observations) are called **unsupervised learning** problems. These include **density estimation**, **clustering** and **dimensionality reduction**.
- **Overfitting** occurs where we model data in excessive detail and fail to distinguish systematic patterns from noise.
- The **Bayesian** and **frequentist** interpretations of probability are associated with two distinct approaches to inference.
- Large **neural networks** can be used for supervised and unsupervised inference tasks. Modern architectures are remarkably effective, confounding the naive expectation that models with billions of parameters will inevitably overfit to their training data.

https://doi.org/10.1017/9781009630696.002 Published online by Cambridge University Press