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