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:

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