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

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

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:

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

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

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.