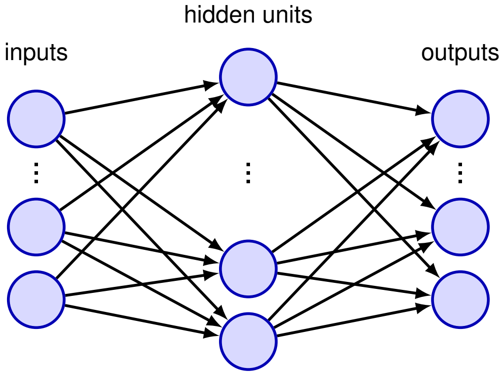

Figure 1.15 A neural network having two layers of parameters in which arrows denote the direction of information flow through the network. Each of the hidden units and each of the output units computes a function of the form given by (1.5) and (1.6) in which the activation function $f(\cdot)$ is differentiable.

Section 1.2.3

used the sum-of-squares error function (1.2) to fit polynomials.

With these changes, we now have an error function whose derivatives with respect to each of the parameters in the network can be evaluated. We can now consider networks having more than one layer of parameters. Figure 1.15 shows a simple network with two processing layers. Nodes in the middle layer called *hidden units* because their values do not appear in the training set, which only provides values for inputs and outputs. Each of the hidden units and each of the output units in Figure 1.15 computes a function of the form given by (1.5) and (1.6). For a given set of input values, the states of all of the hidden and output units can be evaluated by repeated application of (1.5) and (1.6) in which information is flowing forward through the network in the direction of the arrows. For this reason, such models are sometimes also called *feed-forward neural networks*.

Chapter 8

Chapter 7

To train such a network the parameters are first initialized using a random number generator and are then iteratively updated using gradient-based optimization techniques. This involves evaluating the derivatives of the error function, which can be done efficiently in a process known as *error backpropagation*. In backpropagation, information flows backwards through the network from the outputs towards the inputs (Rumelhart, Hinton, and Williams, 1986). There exist many different optimization algorithms that make use of gradients of the function to be optimized, but the one that is most prevalent in machine learning is also the simplest and is known as *stochastic gradient descent*.

The ability to train neural networks having multiple layers of weights was a breakthrough that led to a resurgence of interest in the field starting around the mid-1980s. This was also a period in which the field moved beyond a focus on neurobiological inspiration and developed a more rigorous and principled foundation (Bishop, 1995b). In particular, it was recognized that probability theory, and ideas from the field of statistics, play a central role in neural networks and machine learning. One key insight is that learning from data involves background assumptions, sometimes called *prior knowledge* or *inductive biases*. These might be incorporated explicitly, for example by designing the structure of a neural network such that the classification of a skin lesion does not depend on the location of the lesion within the image, or they might take the form of implicit assumptions that arise from the mathematical