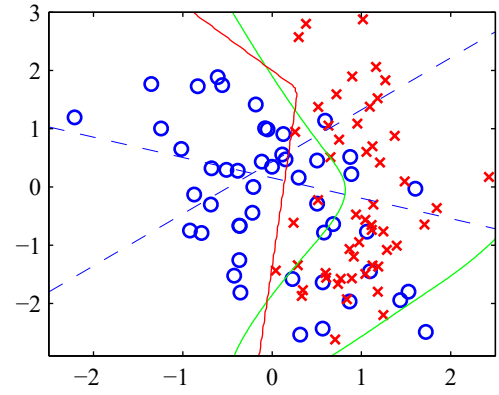

Figure 6.11 Example of the solution of a simple two-class classification problem involving synthetic data using a neural network having two inputs, two hidden units with tanh activation functions, and a single output having a logistic-sigmoid activation function. The dashed blue lines show the $z = 0.5$ contours for each of the hidden units, and the red line shows the $y = 0.5$ decision surface for the network. For comparison, the green lines denote the optimal decision boundary computed from the distributions used to generate the data.

sibilities. In most cases, all the hidden units in a network will be given the same activation function, although in principle there is no reason why different choices could not be applied in different parts of the network.

The simplest option for a hidden unit activation function is the identity function, which means that all the hidden units become linear. However, for any such network, we can always find an equivalent network without hidden units. This follows from the fact that the composition of successive linear transformations is itself a linear transformation, and so its representational capability is no greater than that of a single linear layer. However, if the number of hidden units is smaller than either the number of input or output units, then the transformations that such a network can generate are not the most general possible linear transformation from inputs to outputs because information is lost in the dimensionality reduction at the hidden units. Consider a network with N inputs, M hidden units, and K outputs, and where all activation functions are linear. Such a network has $M(N + K)$ parameters, whereas a linear transformation of inputs directly to outputs would have NK parameters. If M is small relative to N or K, or both, this leads to a two-layer linear network having fewer parameters than the direct linear mapping, corresponding to a rank-deficient transformation. Such 'bottleneck' networks of linear units corresponds to a standard data analysis technique called *principal component analysis*. In general, however, there is limited interest in using multilayer networks of linear units since the overall function computed by such a network is still linear.

Chapter 16

A simple, nonlinear differentiable function is the logistic sigmoid given by

$$\sigma(a) = \frac{1}{1 + \exp(-a)}, \tag{6.13}$$

which is plotted in Figure 5.12. This was widely used in the early years of work on multilayer neural networks and was partly inspired by studies of the properties of