

Figure 5.6 An example of a loss matrix with elements L_{kj} for the cancer treatment problem. The rows correspond to the true class, whereas the columns correspond to the assignment of class made by our decision criterion.

	normal	cancer
normal	0	1
cancer	100	0

they aim to maximize. These are equivalent concepts if we take the utility to be simply the negative of the loss. Throughout this text we will use the loss function convention. Suppose that, for a new value of \mathbf{x} , the true class is \mathcal{C}_k and that we assign \mathbf{x} to class \mathcal{C}_j (where j may or may not be equal to k). In so doing, we incur some level of loss that we denote by L_{kj} , which we can view as the k, j element of a *loss matrix*. For instance, in our cancer example, we might have a loss matrix of the form shown in Figure 5.6. This particular loss matrix says that there is no loss incurred if the correct decision is made, there is a loss of 1 if a healthy patient is diagnosed as having cancer, whereas there is a loss of 100 if a patient having cancer is diagnosed as healthy.

The optimal solution is the one that minimizes the loss function. However, the loss function depends on the true class, which is unknown. For a given input vector \mathbf{x} , our uncertainty in the true class is expressed through the joint probability distribution $p(\mathbf{x}, \mathcal{C}_k)$, and so we seek instead to minimize the average loss, where the average is computed with respect to this distribution and is given by

$$\mathbb{E}[L] = \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} p(\mathbf{x}, \mathcal{C}_k) d\mathbf{x}. \quad (5.22)$$

Each \mathbf{x} can be assigned independently to one of the decision regions \mathcal{R}_j . Our goal is to choose the regions \mathcal{R}_j to minimize the expected loss (5.22), which implies that for each \mathbf{x} , we should minimize $\sum_k L_{kj} p(\mathbf{x}, \mathcal{C}_k)$. As before, we can use the product rule $p(\mathbf{x}, \mathcal{C}_k) = p(\mathcal{C}_k|\mathbf{x})p(\mathbf{x})$ to eliminate the common factor of $p(\mathbf{x})$. Thus, the decision rule that minimizes the expected loss assigns each new \mathbf{x} to the class j for which the quantity

$$\sum_k L_{kj} p(\mathcal{C}_k|\mathbf{x}) \quad (5.23)$$

is a minimum. Once we have chosen values for the loss matrix elements L_{kj} , this is clearly trivial to do.

5.2.3 The reject option

We have seen that classification errors arise from the regions of input space where the largest of the posterior probabilities $p(\mathcal{C}_k|\mathbf{x})$ is significantly less than unity or equivalently where the joint distributions $p(\mathbf{x}, \mathcal{C}_k)$ have comparable values. These are the regions where we are relatively uncertain about class membership. In some applications, it will be appropriate to avoid making decisions on the difficult cases in anticipation of obtaining a lower error rate on those examples for which a classification decision is made. This is known as the *reject option*. For example, in our hypothetical cancer screening example, it may be appropriate to use an automatic