function in the form

$$\mathbf{y}(\mathbf{x}) = \widetilde{\mathbf{W}}^{\mathrm{T}}\widetilde{\mathbf{x}} = \mathbf{T}^{\mathrm{T}}\left(\widetilde{\mathbf{X}}^{\dagger}\right)^{\mathrm{T}}\widetilde{\mathbf{x}}. \tag{5.16}$$

An interesting property of least-squares solutions with multiple target variables is that if every target vector in the training set satisfies some linear constraint

$$\mathbf{a}^{\mathrm{T}}\mathbf{t}_n + b = 0 \tag{5.17}$$

for some constants $\mathbf{a}$ and b, then the model prediction for any value of $\mathbf{x}$ will satisfy the same constraint so that

Exercise 5.3

$$\mathbf{a}^{\mathrm{T}}\mathbf{y}(\mathbf{x}) + b = 0. \tag{5.18}$$

Thus, if we use a 1-of-K coding scheme for K classes, then the predictions made by the model will have the property that the elements of $\mathbf{y}(\mathbf{x})$ will sum to 1 for any value of $\mathbf{x}$. However, this summation constraint alone is not sufficient to allow the model outputs to be interpreted as probabilities because they are not constrained to lie within the interval $(0, 1)$.

The least-squares approach gives an exact closed-form solution for the discriminant function parameters. However, even as a discriminant function (where we use it to make decisions directly and dispense with any probabilistic interpretation), it suffers from some severe problems. We have seen that the sum-of-squares error function can be viewed as the negative log likelihood under the assumption of a Gaussian noise distribution. If the true distribution of the data is markedly different from being Gaussian, then least squares can give poor results. In particular, least squares is very sensitive to the presence of *outliers*, which are data points located a long way from the bulk of the data. This is illustrated in Figure 5.4. Here we see that the additional data points in the right-hand figure produce a significant change in the location of the decision boundary, even though these points would be correctly classified by the original decision boundary in the left-hand figure. The sum-of-squares error function gives too much weight to data points that are a long way from the decision boundary, even though they are correctly classified. Outliers can arise due to rare events or may simply be due to mistakes in the data set. Techniques that are sensitive to a very few data points are said to lack *robustness*. For comparison, Figure 5.4 also shows results from a technique called *logistic regression*, which is more robust to outliers.

Section 2.3.4

Section 5.4.3

The failure of least squares should not surprise us when we recall that it corresponds to maximum likelihood under the assumption of a Gaussian conditional distribution, whereas binary target vectors clearly have a distribution that is far from Gaussian. By adopting more appropriate probabilistic models, we can obtain classification techniques with much better properties than least squares, and which can also be generalized to give flexible nonlinear neural network models, as we will see in later chapters.