7.4.1 Data normalization

Sometimes we encounter data sets in which different input variables span very different ranges. For example, in health data, a patient's height might be measured in meters, such as 1.8m, whereas their blood platelet count might be measured in platelets per microliter, such as 300,000 platelets per μL. Such variations can make gradient descent training much more challenging. Consider a single-layer regression network with two weights in which the two corresponding input variables have very different ranges. Changes in the value of one of the weights produce much larger changes in the output, and hence in the error function, than would similar changes in the other weight. This corresponds to an error surface with very different curvatures along different axes as illustrated in Figure 7.3.

For continuous input variables, it can therefore be very beneficial to re-scale the input values so that they span similar ranges. This is easily done by first evaluating the mean and variance of each input:

$$\mu_i = \frac{1}{N}\sum_{n=1}^{N} x_{ni} \tag{7.48}$$

$$\sigma_i^2 = \frac{1}{N}\sum_{n=1}^{N}(x_{ni}-\mu_i)^2, \tag{7.49}$$

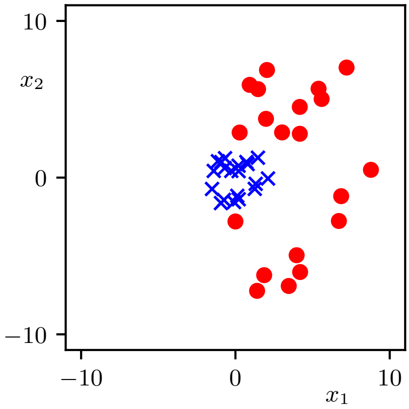
which is a calculation that is performed once, before any training is started. The input values are then re-scaled using

$$\widetilde{x}_{ni} = \frac{x_{ni}-\mu_i}{\sigma_i} \tag{7.50}$$

Exercise 7.14

so that the re-scaled values $\{\widetilde{x}_{ni}\}$ have zero mean and unit variance. Note that the same values of μ_i and σ_i must be used to pre-process any development, validation, or test data to ensure that all inputs are scaled in the same way. Input data normalization is illustrated in Figure 7.7.

Figure 7.7 Illustration of the effect of input data normalization. The red circles show the original data points for a data set with two variables. The blue crosses show the data set after normalization such that each variable now has zero mean and unit variance across the data set.


10
x_2
0
−10
−10
0
10
x_1