1 The feedforward algorithm

The activation of a neural network is iteratively defined by

$$\begin{aligned} y_n &= f(x_n) \\ x_{n+1} &= w_n y_n \end{aligned}$$

where y_n is the output vector at layer n, f is the activation function, x_n is the input vector at layer n, and w_n is the weight matrix between layers n and $n+1$. The first layer is $n = 1$ and the last layer is $n = N$.

2 The backpropagation algorithm

The error of the network is defined by

$$c = \frac{1}{2}(y_N - t)^2$$

The error gradient of the input vector at a layer n is defined as

$$\delta_n = \frac{\partial c}{\partial x_n}$$

The error gradient of the input vector at the last layer N is

$$\begin{aligned} \delta_N &= \frac{\partial c}{\partial x_N} \\ &= \frac{\partial}{\partial x_N}\frac{1}{2}(y_N - t)^2 \\ &= \left(\frac{\partial}{\partial y_N}\frac{1}{2}(y_N - t)^2\right)\frac{\partial y_N}{\partial x_N} \\ &= (y_N - t)\frac{\partial f(x_N)}{\partial x_N} \\ &= (y_N - t)f'(x_N) \end{aligned}$$

The error gradient of the input vector at an inner layer n is

$$
\begin{aligned}
\delta_n &= \frac{\partial c}{\partial x_n} \\
&= \frac{\partial c}{\partial x_{n+1}} \frac{\partial x_{n+1}}{\partial x_n} \\
&= \delta_{n+1} \frac{\partial x_{n+1}}{\partial x_n} \\
&= \delta_{n+1} \frac{\partial w_n y_n}{\partial x_n} \\
&= \delta_{n+1} \frac{\partial w_n y_n}{\partial y_n} \frac{\partial y_n}{\partial x_n} \\
&= \delta_{n+1} \frac{\partial w_n y_n}{\partial y_n} \frac{\partial f(x_n)}{\partial x_n} \\
&= \delta_{n+1} w_n f'(x_n)
\end{aligned}
$$

Therefore, the error gradient of the input vector at a layer n is

$$
\delta_n = f'(x_n) \begin{cases} (y_N - t) & \text{if } n = N \\ \delta_{n+1} w_n & \text{if } n < N \end{cases}
$$

Hence, the error gradient of the weight matrix w_n is

$$
\begin{aligned}
\frac{\partial c}{\partial w_n} &= \frac{\partial c}{\partial x_{n+1}} \frac{\partial x_{n+1}}{\partial w_n} \\
&= \delta_{n+1} \frac{\partial w_n y_n}{w_n} \\
&= \delta_{n+1} y_n
\end{aligned}
$$

Therefore, the change in weight should be

$$
\begin{aligned}
\Delta w_n &= -\alpha \frac{\partial c}{\partial w_n} \\
&= -\alpha \delta_{n+1} y_n
\end{aligned}
$$

where α is the learning rate (or rate of gradient descent).