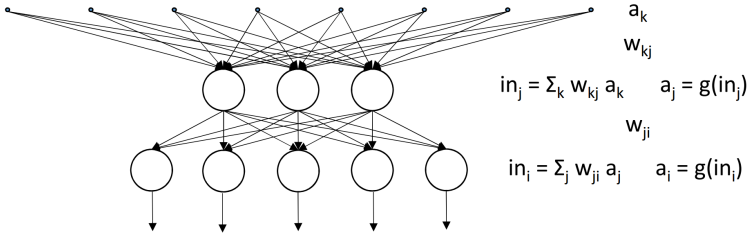
Russel and Norvig, Pages 733ff



k = index of inputs
j = index of hidden units
i = index of outputs

We use $g(x) = \frac{1}{(1+e^{-x})}$. Then $g'(x) = \frac{e^{-x}}{(1+e^{-x})^2} = g(x)(1-g(x))$.
We use error Error=$\frac{1}{2}\sum_i (y_i - a_i)^2$ for a *single* labeled example, where y_i is the desired i^{th} output for this labeled example. Then,

$$\begin{aligned}
\frac{d\text{Error}}{dw_{ji}} &= -(y_i - a_i)\frac{da_i}{dw_{ji}} = -(y_i - a_i)\frac{dg(in_i)}{dw_{ji}} \\
&= -(y_i - a_i)g'(in_i)\frac{din_i}{dw_{ji}} \\
&= -(y_i - a_i)g'(in_i)\frac{d\sum_j w_{ji}a_j}{dw_{ji}} \\
&= -(y_i - a_i)g'(in_i)a_j \\
&= -\Delta[i]a_j, \text{ where } \Delta[i] = (y_i - a_i)g'(in_i)
\end{aligned}$$

Thus, the weights w_{ji} get updated by the approximation of gradient descent as follows, where $\alpha > 0$ is a small learning rate:
$w_{ji} := w_{ji} - \alpha\frac{d\text{Error}}{dw_{ji}} = w_{ji} + \alpha a_j \Delta[i]$
(The derivation up to here is basically the one for a single perceptron with inputs a_j.)

Russel and Norvig, Pages 733ff

Furthermore,

$$\begin{aligned}
\frac{d\text{Error}}{dw_{kj}} &= -\sum_i (y_i - a_i)\frac{da_i}{dw_{kj}} \\
&= -\sum_i (y_i - a_i)\frac{dg(in_i)}{dw_{kj}} \\
&= -\sum_i (y_i - a_i)g'(in_i)\frac{din_i}{dw_{kj}} \\
&= -\sum_i \Delta[i]\frac{d(\sum_j w_{ji}a_j)}{dw_{kj}} \\
&= -\sum_i \Delta[i]w_{ji}\frac{da_j}{dw_{kj}} \\
&= -\sum_i \Delta[i]w_{ji}\frac{dg(in_j)}{dw_{kj}} \\
&= -\sum_i \Delta[i]w_{ji}g'(in_j)\frac{din_j}{dw_{kj}} \\
&= -\sum_i \Delta[i]w_{ji}g'(in_j)\frac{d(\sum_k w_{kj}a_k)}{dw_{kj}} \\
&= -\sum_i \Delta[i]w_{ji}g'(in_j)a_k \\
&= -\Delta[j]a_k, \text{ where } \Delta[j] = \sum_i \Delta[i]w_{ji}g'(in_j)
\end{aligned}$$

Thus, the weights w_{kj} get updated by the approximation of gradient descent as follows:
$w_{kj} := w_{kj} - \alpha\frac{d\text{Error}}{dw_{kj}} = w_{kj} + \alpha a_k\Delta[j]$