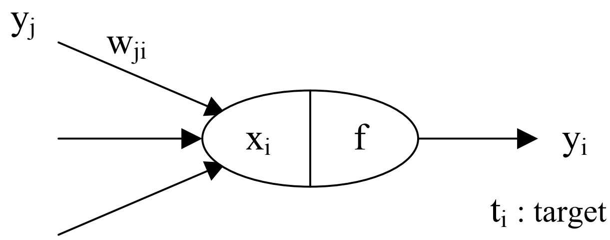
Back Propagation

Basic Notations:

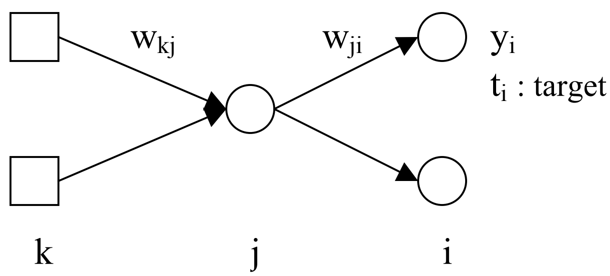


$$x_i = \sum_j w_{ji} y_j$$

$$y_i = f(x_i)$$

If we use the sigmoid as the activation function, then $y_i = f(x_i) = \dfrac{1}{1+e^{-x_i}}$

Network with 1 hidden node:



$$E = \text{Error} = \tfrac{1}{2} \sum_i (t_i - y_i)^2$$

For the output layer, we want to change the weights so that: $W_{ji} \leftarrow W_{ji} - \alpha \times \dfrac{\partial E}{\partial W_{ji}}$

Thus the amount that we want to update is given by $\Delta W_{ji} = -\alpha \times \dfrac{\partial E}{\partial W_{ji}}$

we calculated using the chain rule that $\dfrac{\partial E}{\partial W_{ji}} = \dfrac{\partial E}{\partial y_i} \cdot \dfrac{\partial y_i}{\partial x_i} \cdot \dfrac{\partial x_i}{\partial W_{ji}} = -(t_i - y_i) \cdot f'(x_i) \cdot y_j$

The derivative of the sigmoid is just $y_i(1-y_i)$, so $\Delta W_{ji} = -\alpha \times -(t_i - y_i) \cdot y_i(1-y_i) \cdot y_j$

which we write as $\boxed{\Delta W_{ji} = -\alpha \times -y_j \times \delta_i}$,

where $\boxed{\delta_i = (t_i - y_i) \cdot y_i(1-y_i)}$ (you can think of it as the target amount of adjustment).

For the hidden layer, we want to do something similar: $\Delta W_{kj} = -\alpha \times \frac{\partial E}{\partial W_{ki}}$

So we use the chain rule again: $\frac{\partial E}{\partial W_{kj}} = \frac{\partial E}{\partial y_j} \cdot \frac{\partial y_j}{\partial x_j} \cdot \frac{\partial x_j}{\partial W_{kj}}$

The second and third terms are just like before, just different indices.

The first term is tricky – you want to sum up the errors that this y_j has caused down the line.

So we apply the chain rule again: $\frac{\partial E}{\partial y_j} = \sum_i \frac{\partial E}{\partial y_i} \cdot \frac{\partial y_i}{\partial x_i} \cdot \frac{\partial x_i}{\partial y_j} = \sum_i -(t_i - y_i) \cdot f'(x_i) \cdot W_{ji}$

Plugging this all in, we get $\frac{\partial E}{\partial W_{kj}} = \left(-\sum_i (t_i - y_i) \cdot f'(x_i) \cdot W_{ji} \right) \cdot f'(x_j) \cdot y_k$

Plugging in the sigmoid, we get $\Delta W_{kj} = -\alpha \times \left(-\sum_i (t_i - y_i) \cdot y_i (1 - y_i) \cdot W_{ji} \right) \cdot y_j (1 - y_j) \cdot y_k$

which we again write as $\boxed{\Delta W_{kj} = -\alpha \times -y_k \times \delta_j}$,

where $\delta_j = \left(\sum_i (t_i - y_i) \cdot y_i (1 - y_i) \cdot W_{ji} \right) \cdot y_j (1 - y_j)$

and in fact, you'll notice that this is just $\boxed{\delta_j = \left(\sum_i W_{ji} \cdot \delta_i \right) \cdot y_j (1 - y_j)}$

and the first time is just like a weighted sum of the target adjustment at the output level.