1 ANN Back Propagation

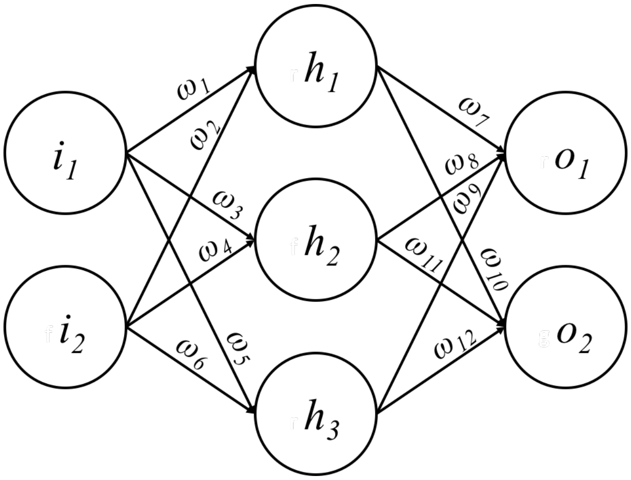

Figure 1: Example of a trivial artificial neural network with two input nodes, one hidden layer with three nodes, and two output nodes.

The net input to a particular node k in the hidden layer(s) or output layer with N different input streams, each with a value x that could be either the raw input or the output of previous nodes is:

$$net_k = \sum_{i=1}^{N} \omega_i x_i \tag{1}$$

The output for a node k in the hidden layer(s) or output layer is a function of the net input to that node. There are many possibilities for this *activation function*, but the logistic function is a common choice:

$$out_k = f(net_k) = \frac{1}{1+e^{-net_k}} \tag{2}$$

The total output error, E_T, for an artificial neural network is the summation of the M individual output node errors. As with the activation function, there are different possibilities for calculating the error on a particular node. The squared deviation of the node's output from the target for that node is a typical choice:

$$E_T = \sum_{i=1}^{M} (target_{oi} - out_{oi})^2 \tag{3}$$

The method of back-propagation gradient descent has the goal of adjusting the network's weights such that the total output error is minimized. This is accomplished by adjusting individual weights based on the gradient of the total error with respect to an individual weight. In practice, this is done by adding the value of the partial derivative of the total error function with respect to an individual weight, multiplied by some learning rate η, to the current weight value:

$$\omega_i^+ = \omega_i + \eta \frac{\partial E_T}{\partial \omega_i} \tag{4}$$

The total error is composed of three separate functions: the total error function itself, the net input function, and the activation function. Therefore, taking the partial derivative of the total error with respect to a particular weight must make use of the chain rule. For example, to update ω_7 in the network depicted in Fig. 1, we'd use the following formulation:

$$\frac{\partial E_T}{\partial \omega_7} = \frac{\partial E_T}{\partial out_{o1}} \frac{\partial out_{o1}}{\partial net_{o1}} \frac{\partial net_{o1}}{\partial w_7} \tag{5}$$

The partial with respect to output node 1 would start with the total error function. Note that we modify the total error function slightly by multiplying by 1/2 just to make the math a bit cleaner. Since this linear transformation preserves the error proportionality, and because we end up multiplying by η anyway, this has no effect:

$$E_T = E_{o1} + E_{o2} = \frac{1}{2}(target_{o1} - out_{o1})^2 + \frac{1}{2}(target_{o2} - out_{o2})^2 \tag{6}$$

$$\frac{\partial E_T}{\partial out_{o1}} = -(target_{o1} - out_{o1}) \tag{7}$$

Next we consider how out_{o1} changes as a function of its net input, net_{o1}. Taking the activation function to be the logistic function and remembering that the derivative of the logistic function is: $f(x)(1 - f(x))$:

$$out_{o1} = \frac{1}{1 + e^{-net_{o1}}} \tag{8}$$

$$\frac{\partial out_{o1}}{\partial net_{o1}} = out_{o1}(1 - out_{o1}) \tag{9}$$

Finally, we consider how the net_{o1} changes as ω_7 changes:

$$net_{o1} = \omega_7 out_{h1} + \omega_8 out_{h2} + \omega_9 out_{h3} \tag{10}$$

$$\frac{\partial net_{o1}}{\partial w_7} = out_{h1} \tag{11}$$

Finally, expanding Eq. 5 with results from Eqs. 7, 9, and 11:

$$\frac{\partial E_T}{\partial \omega_7} = \frac{\partial E_T}{\partial out_{o1}} \frac{\partial out_{o1}}{\partial net_{o1}} \frac{\partial net_{o1}}{\partial w_7} = [-(target_{o1} - out_{o1})][out_{o1}(1 - out_{o1})][out_{h1}] \tag{12}$$

We have values for each of the variables in the above expression: out_{o1}, out_{h1}, and $target_{o1}$ so updating ω_7 is a simple matter of plugging these values in and applying the update formula: $\omega_7^+ = \omega_7 + \eta \frac{\partial E_T}{\partial \omega_7}$.

Things are only slightly more complicated for adjusting weights between input nodes and the hidden layer nodes (or between multiple hidden layers). While the same principle is employed (i.e. the chain rule), these weights affect the value of multiple output nodes. Consider updating ω_1: we can write the chain rule enabled expansion as above:

$$\frac{\partial E_T}{\partial \omega_1} = \frac{\partial E_T}{\partial out_{h1}} \frac{\partial out_{h1}}{\partial net_{h1}} \frac{\partial net_{h1}}{\partial w_1} \tag{13}$$

For which the latter two partials are straightforward:

$$out_{h1} = \frac{1}{1 + e^{-net_{h1}}} \tag{14}$$

$$\frac{\partial out_{h1}}{\partial net_{h1}} = out_{h1}(1 - out_{h1}) \tag{15}$$

and:

$$net_{h1} = \omega_1 out_{i1} + \omega_2 out_{i2} \tag{16}$$

$$\frac{\partial net_{h1}}{\partial w_1} = out_{i1} \tag{17}$$

Now we turn our attention to the partial derivative of the total error with respect to the output of node $h1$. Node h_1 affects the output, and therefore E_T, of *both* nodes o_1 and o_2. This is different than the previously considered case of ω_7 which influenced only o_1. Therefore, the partial derivative of E_T with respect to out_{h1} is the sum of the partial for E_{o1} and E_{o2}:

$$\frac{\partial E_T}{\partial out_{h1}} = \frac{\partial (E_{o1} + E_{o2})}{\partial out_{h1}} = \frac{\partial E_{o1}}{\partial out_{h1}} + \frac{\partial E_{o2}}{\partial out_{h1}} \tag{18}$$

Applying the chain rule again we can arrive at an expression for the first partial:

$$\frac{\partial E_{o1}}{\partial out_{h1}} = \frac{\partial E_{o1}}{\partial net_{o1}} \frac{\partial net_{o1}}{\partial out_{h1}} \tag{19}$$

using previous results (see Eqs. 6, 7, and 9):

$$\frac{\partial E_{o1}}{\partial net_{o1}} = \frac{\partial E_{o1}}{\partial out_{o1}} \frac{\partial out_{o1}}{\partial net_{o1}} = [-(target_{o1} - out_{o1})][out_{o1}(1 - out_{o1})] \tag{20}$$

and

$$\frac{\partial net_{o1}}{\partial out_{h1}} = \omega_7 \tag{21}$$

so:

$$\frac{\partial E_{o1}}{\partial out_{h1}} = [-(target_{o1} - out_{o1})][out_{o1}(1 - out_{o1})]\omega_7 \tag{22}$$

Follow the same procedure for the partial of E_{o2} to obtain:

$$\frac{\partial E_{o2}}{\partial out_{h1}} = [-(target_{o2} - out_{o2})][out_{o2}(1 - out_{o2})]\omega_8 \tag{23}$$

Thus, we can fully expand Eq. 13 with the results from Eqs. 15, 17, 18, 22, and 23:

$$\frac{\partial E_T}{\partial \omega_1} = [[-(target_{o2} - out_{o2})][out_{o2}(1 - out_{o2})]\omega_8 + [-(target_{o1} - out_{o1})][out_{o1}(1 - out_{o1})]\omega_7][out_{h1}(1 - out_{h1})][out_{i1}] \tag{24}$$