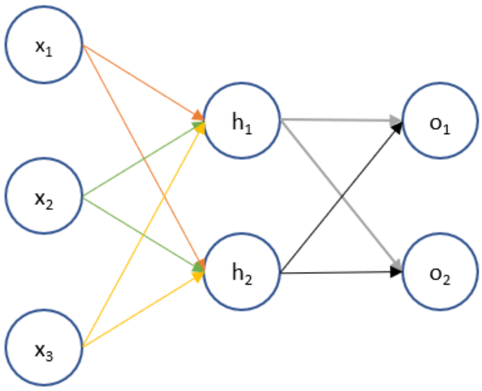
Backpropagation by Hand



Assume we have the above network to be trained. The following are the (very) high level steps to train this network.

(1) Initialize weights for the parameters we want to train

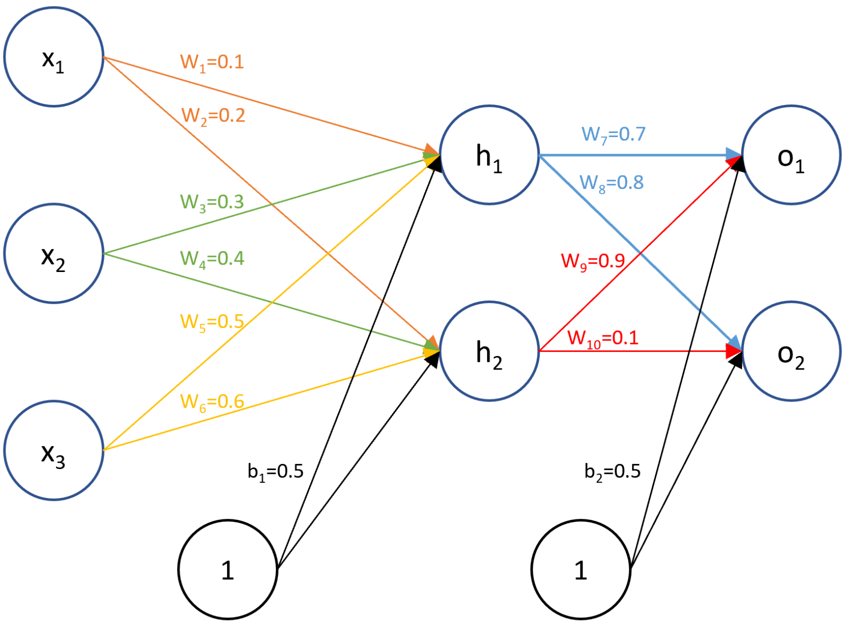
(2) Forward propagate through the network to get the output values

(3) Define the error or cost function and its first derivatives

(4) Backpropagate through the network to determine the error derivatives

(5) Update the parameter estimates using the error derivative and the current value

1. Initialize



The input and target values for this problem are $x_1=1$, $x_2=4$, $x_3=5$ and $t_1=0.1$, $t_2=0.05$. Initialize weights as shown in the diagram. Generally, you will assign them randomly but for calculation purposes, I've chosen these numbers.

2. Forward Propagate

Mathematically, we have the following relationships between nodes in the networks. For the input and output layer, let's use the convention of denoting z_{h1}, z_{h2}, z_{o1}, and z_{o2} to denote the value before the activation function is applied and the notation of h_1, h_2, o_1, and o_2 to denote the values after application of the activation function.

Input to hidden layer

$$w_1x_1 + w_3x_2 + w_5x_3 + b_1 = z_{h_1}$$

$$w_2x_1 + w_4x_2 + w_6x_3 + b_1 = z_{h_2}$$

$$h_1 = \sigma(z_{h_1})$$

$$h_2 = \sigma(z_{h_2})$$

Hidden layer to output layer

$$w_7h_1 + w_9h_2 + b_2 = z_{o_1}$$

$$w_8h_1 + w_{10}h_2 + b_2 = z_{o_2}$$

$$o_1 = \sigma(z_{o_1})$$

$$o_2 = \sigma(z_{o_2})$$

We can use the formulas above to forward propagate through the network.
Calculate the numerical values.

3. Sum of Squares Error

Define the sum of squares error using the target values and the results from the last layer from forward propagation.

$$E = \frac{1}{2}[(o_1 - t_1)^2 + (o_2 - t_2)^2$$

$$\frac{dE}{d_{o_1}} = o_1 - t_1$$

$$\frac{dE}{d_{o_2}} = o_2 - t_2$$

BackPropagate

We are now ready to backpropagate through the network to compute all the error derivatives with respect to the parameters. Note that although there will be many long formulas, we are not doing anything fancy here. We are just using the basic principles of calculus such as the chain rule.

First, derivative of a Sigmoid function:

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

$$\frac{d\sigma}{dx} = \frac{e^{-x}}{(1+e^{-x})^2}$$

We can write $\frac{e^{-x}}{1+e^{-x}}$ as $1 - \frac{1}{1+e^{-x}}$. The derivative can be written as

$$\frac{d\sigma}{dx} = \frac{1}{1+e^{-x}}(1 - \sigma(x))$$

$$\frac{d\sigma}{dx} = \sigma(x)(1 - \sigma(x))$$

Also, given that $w_7h_1 + w_9h_2 + b_2 = z_{o_1}$ and $w_8h_1 + w_{10}h_2 + b_2 = z_{o_2}$, we have $\frac{dz_{o_1}}{dw_7} = h_1$, $\frac{dz_{o_2}}{dw_8} = h_1$, $\frac{dz_{o_1}}{dw_9} = h_2$, $\frac{dz_{o_2}}{dw_{10}} = h_2$, $\frac{dz_{o_1}}{db_2} = 1$, and $\frac{dz_{o_2}}{db_2} = 1$.

We are now ready to calculate $\frac{dE}{dw_7}$, $\frac{dE}{dw_8}$, $\frac{dE}{dw_9}$, and $\frac{dE}{dw_{10}}$ using the derivatives we have already discussed.

$$\frac{dE}{dw_7} = \frac{dE}{do_1}\frac{do_1}{dz_{o_1}}\frac{dz_{o_1}}{dw_7}$$

$$\frac{dE}{dw_7} = (o_1 - t_1)(o_1(1 - o_1))h_1$$

Similarly find the other three and find the numerical values w.r.t w_8, w_9, w_{10}:

The error derivative of b_2 is a little bit more involved since changes to b_2 affect the error through both o_1 and o_2.

$$\frac{dE}{db_2} = \frac{dE}{do_1}\frac{do_1}{dz_{o_1}}\frac{dz_{o_1}}{db_2} + \frac{dE}{do_2}\frac{do_2}{dz_{o_2}}\frac{dz_{o_2}}{db_2}$$

Propagate to the previous layer first w.r.t to h_1 and then h_2

To summarize, we have computed numerical values for the error derivatives with respect to w_7, w_8, w_9, w_{10}, and b_2. We will now backpropagate one layer to compute the error derivatives of the parameters connecting the input layer to the hidden layer. These error derivatives are $\frac{dE}{dw_1}, \frac{dE}{dw_2}, \frac{dE}{dw_3}, \frac{dE}{dw_4}, \frac{dE}{dw_5}, \frac{dE}{dw_6}$, and $\frac{dE}{db_1}$.

I will calculate $\frac{dE}{dw_1}, \frac{dE}{dw_3}$, and $\frac{dE}{dw_5}$ first since they all flow through the h_1 node.

$$\frac{dE}{dw_1} = \frac{dE}{dh_1}\frac{dh_1}{dz_{h_1}}\frac{dz_{h_1}}{dw_1}$$

The calculation of the first term on the right hand side of the equation above is a bit more involved than previous calculations since h_1 affects the error through both o_1 and o_2.

$$\frac{dE}{dh_1} = \frac{dE}{do_1}\frac{do_1}{dz_{o_1}}\frac{dz_{o_1}}{dh_1} + \frac{dE}{do_2}\frac{do_2}{dz_{o_2}}\frac{dz_{o_2}}{dh_1}$$

Proceed with the numerical values for the error derivatives above. These derivatives have already been calculated above or are similar in style to those calculated above.

The calculations for $\frac{dE}{dw_3}$ and $\frac{dE}{dw_5}$ are below

$$\frac{dE}{dw_3} = \frac{dE}{dh_1}\frac{dh_1}{dz_{h_1}}\frac{dz_{h_1}}{dw_3}$$

$$\frac{dE}{dw_5} = \frac{dE}{dh_1}\frac{dh_1}{dz_{h_1}}\frac{dz_{h_1}}{dw_5}$$

I will now calculate $\frac{dE}{dw_2}$, $\frac{dE}{dw_4}$, and $\frac{dE}{dw_6}$ since they all flow through the h_2 node.

$$\frac{dE}{dw_2} = \frac{dE}{dh_2}\frac{dh_2}{dz_{h_2}}\frac{dz_{h_2}}{dw_2}$$

The calculation of the first term on the right hand side of the equation above is a bit more involved since h_2 affects the error through both o_1 and o_2.

$$\frac{dE}{dh_2} = \frac{dE}{do_1}\frac{do_1}{dz_{o_1}}\frac{dz_{o_1}}{dh_2} + \frac{dE}{do_2}\frac{do_2}{dz_{o_2}}\frac{dz_{o_2}}{dh_2}$$

Find these and other derivates mentioned above w.r.t w_2, w_4, w_6. Some hint below:

The calculations for $\frac{dE}{dw_4}$ and $\frac{dE}{dw_6}$ are below

$$\frac{dE}{dw_4} = \frac{dE}{dh_2}\frac{dh_2}{dz_{h_2}}\frac{dz_{h_2}}{dw_4}$$

$$\frac{dE}{dw_4} = (0.0818)(0.0049)(4) = 0.0016$$

$$\frac{dE}{dw_6} = \frac{dE}{dh_2}\frac{dh_2}{dz_{h_2}}\frac{dz_{h_2}}{dw_6}$$

The final error derivative we have to calculate is $\frac{dE}{db_1}$, which is done next

$$\frac{dE}{db_1} = \frac{dE}{do_1}\frac{do_1}{dz_{o_1}}\frac{dz_{o_1}}{dh_1}\frac{dh_1}{dz_{h_1}}\frac{dz_{h_1}}{db_1} + \frac{dE}{do_2}\frac{do_2}{dz_{o_2}}\frac{dz_{o_2}}{dh_2}\frac{dh_2}{dz_{h_2}}\frac{dz_{h_2}}{db_1}$$

Update the weights and bias

We now have all the error derivatives and we're ready to make the parameter updates after the first iteration of backpropagation. We will use the learning rate of \alpha = 0.01

$$w_1 := w_1 - \alpha\frac{dE}{dw_1} =$$

$$w_2 := w_2 - \alpha\frac{dE}{dw_2} =$$

$$w_3 := w_3 - \alpha\frac{dE}{dw_3} =$$

$$w_4 := w_4 - \alpha\frac{dE}{dw_4} =$$

$$w_5 := w_5 - \alpha\frac{dE}{dw_5} =$$

$$w_6 := w_6 - \alpha\frac{dE}{dw_6} =$$

$$w_7 := w_7 - \alpha\frac{dE}{dw_7} =$$

$$w_8 := w_8 - \alpha\frac{dE}{dw_8} =$$

$$w_9 := w_9 - \alpha\frac{dE}{dw_9} =$$

$$w_{10} := w_{10} - \alpha\frac{dE}{dw_{10}} =$$

$$b_1 := b_1 - \alpha\frac{dE}{db_1} =$$

$$b_2 := b_2 - \alpha\frac{dE}{db_2} =$$

Repeat the iterations starting from the beginning.