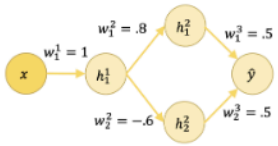
Backpropagation Exercise

September 25, 2024

Problem

On paper, forward and backpropagate to calculate the derivatives of the weights of this network. Calculate new weights after a single step down the gradient with learning rate 0.1



Details:

- Use input $x = -1$ and target $y = 0.5$
- Neurons in the second hidden layer have a ReLU nonlinearity; others do not have a nonlinearity.
- The biases of the neurons in the first and third layer are 0; the biases of the hidden neurons in the second layer are are -0.5
- Leave the biases alone (just find the gradient of the loss with respect to the weights)
- Use loss $L = (\hat{y} - y)^2/2$

Forward propagation

The first step is to forward propagate the input through the network:

$$
\begin{aligned}
x &= -1 \\
h_1^1 &= (-1)1 = -1 \\
h_1^2 &= [h_1^1 w_1^2 - 0.5]^+ = 0 \\
h_2^2 &= [h_1^1 w_2^2 - 0.5]^+ = 0.1 \\
\hat{y} &= 0.5h_1^2 + 0.5h_2^2 = 0.05
\end{aligned}
$$

The loss is $(.05 - .5)^2/2 = 0.10125$.

Backward propagation

Let δ_y be the output neuron's delta. Backpropagating to find the other deltas,

$$\begin{aligned}
\delta_y &= \hat{y} - y = -0.45 \\
\delta_{h_1^2} &= g' w_1^3 \delta_y = 0 \\
\delta_{h_2^2} &= g' w_2^3 \delta_y = 1(0.5)(-0.45) = -0.225 \\
\delta_{h_1^1} &= w_1^2 \delta_{h_1^2} + w_2^2 \delta_{h_2^2} = 0 + -0.6(-0.225) = 0.135
\end{aligned}$$

Then calculate the partial derivatives, each of which is a δ times an input:

$$\begin{aligned}
\frac{\partial L}{\partial w_1^3} &= \delta_y h_1^2 = -0.45(0) = 0 \\
\frac{\partial L}{\partial w_2^3} &= \delta_y h_2^2 = -0.45(0.1) = -0.045 \\
\frac{\partial L}{\partial w_1^2} &= \delta_{h_1^2} h_1^1 = 0(-1) = 0 \\
\frac{\partial L}{\partial w_2^2} &= \delta_{h_2^2} h_1^1 = -0.225(-1) = 0.225 \\
\frac{\partial L}{\partial w_1^1} &= \delta_{h_1^1} x = 0.135(-1) = -0.135
\end{aligned}$$

Weight update

With learning rate 0.1, we then update the the weights by adding $-0.1 \nabla L$. The new values are,

$$\begin{aligned}
w_1^3 &= 0.5 \\
w_2^3 &= 0.5 - 0.1(-0.045) = .5045 \\
w_1^2 &= 0.8 \\
w_2^2 &= -0.6 - 0.1(.225) = -0.6225 \\
w_1^1 &= 1 - 0.1(-0.135) = 1.0135
\end{aligned}$$

These new weights should be a little better at approximating the target output of 0.5 given the input -1. Let's see:

$$
\begin{aligned}
x &= -1 \\
h_1^1 &= (-1)1.0135 = -1.0135 \\
h_1^2 &= [h_1^1 w_1^2 - 0.5]^+ = 0 \\
h_2^2 &= [h_1^1 w_2^2 - 0.5]^+ = 0.1309 \\
\hat{y} &= w_1^3 h_1^2 + w_2^3 h_2^2 = 0.0660
\end{aligned}
$$

So this is indeed an improvement in that it's a bit closer to the target of 0.5.