Part 3 Backpropagation Example 1 (Part 2)

Input: $x = \begin{pmatrix} 0 & 1 \\ 1 & 1 \\ 2 & 1 \end{pmatrix}$, Output: $y^* = \begin{pmatrix} \frac{9}{4} \\ \frac{3}{2} \\ \frac{3}{4} \end{pmatrix}$

We consider to build a 2 Layer Neural Network, where each layer consists of a single Neuron. For the first layer, we choose the sigmoid activation function which yields the output

$$f_{1,1} = \phi_{1,1}(x'_{[d]} \cdot w_{1,1}) = \frac{1}{1 + \mathrm{e}^{-(x^{(d1)} w_{1,1}^{(1)} + x^{(d2)} w_{1,1}^{(2)})}} \ ,$$

where $w_{1,1} = \left(w_{1,1}^{(1)}, w_{1,1}^{(2)}\right)'$ is the weight (vector) for the first layer.
For the second layer, i.e. the output layer, we set the identity as the activation function, which gives us

$$f_{2,1} = \phi_{2,1}(f_{1,1} \cdot w_{2,1}) = f_{1,1} \cdot w_{2,1} \ .$$

Since $f_{2,1}$ yields the output of our Neural Network, it holds that

$$f_{2,1} = f = y \text{ (just notation)}$$

Choose $w_{1,1} = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$, $w_{2,1} = 2$:

FEED FORWARD

$$f_{1,1}\left(\begin{pmatrix} 0 \\ 1 \end{pmatrix}, w_{1,1}\right) = \frac{1}{1+\mathrm{e}^{-(0\cdot(-1)+1\cdot 1)}} = \frac{1}{1+\mathrm{e}^{-1}} \quad ; \quad f\left(\frac{1}{1+\mathrm{e}^{-1}}, w_{2,1}\right) = \frac{2}{1+\mathrm{e}^{-1}}$$

$$f_{1,1}\left(\begin{pmatrix} 1 \\ 1 \end{pmatrix}, w_{1,1}\right) = \frac{1}{1+\mathrm{e}^{-(1\cdot(-1)+1\cdot 1)}} = \frac{1}{2} \quad ; \quad f\left(\frac{1}{2}, w_{2,1}\right) = 1$$

$$f_{1,1}\left(\begin{pmatrix} 2 \\ 1 \end{pmatrix}, w_{1,1}\right) = \frac{1}{1+\mathrm{e}^{-(2\cdot(-1)+1\cdot 1)}} = \frac{1}{1+\mathrm{e}} \quad ; \quad f\left(\frac{1}{1+\mathrm{e}}, w_{2,1}\right) = \frac{2}{1+\mathrm{e}}$$

Loss (squared $\mathcal{L}_2$):

$$L = \frac{1}{3}\left[\left(\frac{9}{4} - \frac{2}{1+\mathrm{e}^{-1}}\right)^2 + \left(\frac{3}{2} - 1\right)^2 + \left(\frac{3}{4} + \frac{2}{1+\mathrm{e}}\right)^2\right] \approx 0.31 \ .$$

Gradient:

$$\nabla_w L(w) = \frac{1}{3}\sum_{d=1}^{3} \nabla_w (f_{[d]} - f^*_{[d]})^2 = \frac{1}{3}\sum_{d=1}^{3} 2 \cdot (f_{[d]} - f^*_{[d]}) \cdot (\nabla_w f_{[d]} - f^*_{[d]}) \ .$$

Since all the terms, except $\nabla_w f_{[d]}$, were calculated previously in the FEED FORWARD-Part, only $\nabla_w f_{[d]}$ remains. Applying the nice property for the derivative of the sigmoid function, $\sigma(x)$,

$$\frac{\partial}{\partial x}\sigma(x) = \sigma(x)(1 - \sigma(x)) \ ,$$

we get the Gradient using the chain rule:

$$\nabla_w f_{[d]} = \begin{pmatrix} \dfrac{\partial f_{[d]}}{\partial w_{1,1}^{(1)}} \\ \dfrac{\partial f_{[d]}}{\partial w_{1,1}^{(2)}} \\ \dfrac{\partial f_{[d]}}{\partial w_{2,1}} \end{pmatrix} = \begin{pmatrix} \dfrac{\partial f_{[d]}}{\partial f_{1,1}} \cdot \dfrac{\partial f_{1,1}}{\partial w_{1,1}^{(1)}} \\ \dfrac{\partial f_{[d]}}{\partial f_{1,1}} \cdot \dfrac{\partial f_{1,1}}{\partial w_{1,1}^{(2)}} \\ \dfrac{\partial f_{[d]}}{\partial w_{2,1}} \end{pmatrix} = \begin{pmatrix} -w_{2,1} x^{(d1)} (1 - f_{1,1_{[d]}}) f_{1,1_{[d]}} \\ -w_{2,1} x^{(d2)} (1 - f_{1,1_{[d]}}) f_{1,1_{[d]}} \\ f_{1,1_{[d]}} \end{pmatrix} .$$

It is left to plug in the input values, in order to receive the actual values for the gradient:

$$x = \begin{pmatrix} 0 \\ 1 \end{pmatrix}: \quad \nabla_w f_{[1]} = \begin{pmatrix} -2 \cdot 0 \cdot (1 - \frac{1}{1+e^{-1}}) \frac{1}{1+e^{-1}} \\ -2 \cdot 1 \cdot (1 - \frac{1}{1+e^{-1}}) \frac{1}{1+e^{-1}} \\ \frac{1}{1+e^{-1}} \end{pmatrix} = \begin{pmatrix} 0 \\ -(1 - \frac{1}{1+e^{-1}}) \frac{2}{1+e^{-1}} \\ \frac{1}{1+e^{-1}} \end{pmatrix}$$

$$x = \begin{pmatrix} 1 \\ 1 \end{pmatrix}: \quad \nabla_w f_{[2]} = \begin{pmatrix} -2 \cdot 1 \cdot (1 - \frac{1}{2}) \frac{1}{2} \\ -2 \cdot 1 \cdot (1 - \frac{1}{2}) \frac{1}{2} \\ \frac{1}{2} \end{pmatrix} = \begin{pmatrix} -\frac{1}{2} \\ -\frac{1}{2} \\ \frac{1}{2} \end{pmatrix}$$

$$x = \begin{pmatrix} 2 \\ 1 \end{pmatrix}: \quad \nabla_w f_{[3]} = \begin{pmatrix} -2 \cdot 2 \cdot (1 - \frac{1}{1+e}) \frac{1}{1+e} \\ -2 \cdot 1 \cdot (1 - \frac{1}{1+e}) \frac{1}{1+e} \\ \frac{1}{1+e} \end{pmatrix} = \begin{pmatrix} -(1 - \frac{1}{1+e}) \frac{4}{1+e} \\ -(1 - \frac{1}{1+e}) \frac{2}{1+e} \\ \frac{1}{1+e} \end{pmatrix}$$

<u>Insert into the Gradient of the Loss</u>:

$$\nabla_w L(w) = \frac{2}{3} \left[\left(f_{[1]} - f_{[1]}^* \right) \cdot \left(\nabla_w f_{[1]} - f_{[1]}^* \right) + \left(f_{[2]} - f_{[2]}^* \right) \cdot \left(\nabla_w f_{[2]} - f_{[2]}^* \right) + \left(f_{[3]} - f_{[3]}^* \right) \cdot \left(\nabla_w f_{[3]} - f_{[3]}^* \right) \right]$$
$$\approx \begin{pmatrix} 4.10 \\ 4.38 \\ 2.35 \end{pmatrix}$$

<u>Gradient Descent Algorithm</u>:

In general: $w_{t+1} = w_t - \eta \cdot \nabla_w L(w_t)$. We will set $\eta = 0.1$ and since we chose the initial weights to be

$$w_0 = \begin{pmatrix} w_{1,1}^{(1)} \\ w_{1,1}^{(2)} \\ w_{2,1} \end{pmatrix} = \begin{pmatrix} -1 \\ 1 \\ 2 \end{pmatrix}$$

we get that

$$w_1 = w_0 + 0.1 \cdot \nabla_w L(w_0) \approx \begin{pmatrix} -1 \\ 1 \\ 2 \end{pmatrix} - 0.1 \cdot \begin{pmatrix} 4.10 \\ 4.38 \\ 2.35 \end{pmatrix} = \begin{pmatrix} -1 - 0.41 \\ 1 - 0.438 \\ 2 - 0.235 \end{pmatrix} = \begin{pmatrix} -1.41 \\ 0.562 \\ 1.765 \end{pmatrix} .$$

Now, we can start again at the Feed Forward-Part using our updated weights $w_{1,1} = \begin{pmatrix} -1.41 \\ 0.562 \end{pmatrix}$ and $w_{2,1} = 1.765$.