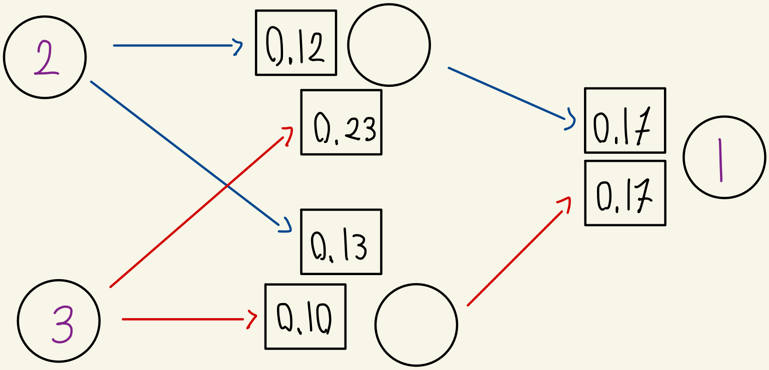
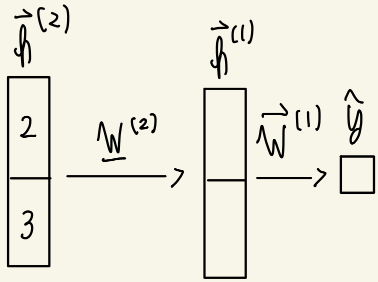



Which can be rewritten into Martha's version, which I'll be using here:




Our network is initialized with:

| | | |
|---|---|---|
| $y$ | True value of the labeled input | $y = 1$ |
| $\hat{y}$ | Model output | $\hat{y} = 0.2533$ |
| $\vec{h}^{(1)} = [h_1^{(1)} \quad h_2^{(1)}]^T$ | Input values for our labeled input | $\vec{h}^{(1)} = [0.93 \quad 0.56]^T$ |
| $\vec{h}^{(2)} = [h_1^{(2)} \quad h_2^{(2)}]^T$ | Input values for our labeled input | $\vec{h}^{(2)} = [2 \quad 3]^T$ |
| $\vec{W}^{(1)} = [w_1^{(1)} \quad w_2^{(1)}]$ | Weights of the linear model at the end | $\vec{W}^{(1)} = [0.17 \quad 0.17]$ |
| $\underline{W}^{(2)} = \begin{bmatrix} w_1^{(2)} & w_3^{(2)} \\ w_2^{(2)} & w_4^{(2)} \end{bmatrix}$ | Weights in the input layer | $\underline{W}^{(2)} = \begin{bmatrix} 0.12 & 0.13 \\ 0.23 & 0.10 \end{bmatrix}$ |
| $a$ | The learning rate we use for backprop | $a = 0.05$ |
| $l : \mathbb{R}^2 \to \mathbb{R}$ | Loss function for training | $l(\hat{y}, y) = \frac{1}{2}(\hat{y} - y)^2$ |

Since we'll need it later, we should write out our prediction y_hat in terms of the inputs

$$\hat{y} = \vec{w}^{(1)} \vec{h}^{(1)}$$

$$= \vec{w}^{(1)} \underline{W}^{(2)} \vec{h}^{(2)}$$

$$= \begin{bmatrix} w_1^{(1)} & w_2^{(1)} \end{bmatrix} \begin{bmatrix} w_1^{(2)} & w_3^{(2)} \\ w_2^{(2)} & w_4^{(2)} \end{bmatrix} \begin{bmatrix} h_1^{(2)} \\ h_2^{(2)} \end{bmatrix}$$

$$= \begin{bmatrix} w_1^{(1)} & w_2^{(1)} \end{bmatrix} \begin{bmatrix} w_1^{(2)} h_1^{(2)} + w_3^{(2)} h_2^{(2)} \\ w_2^{(2)} h_1^{(2)} + w_4^{(2)} h_2^{(2)} \end{bmatrix}$$

$$= w_1^{(1)} \left( w_1^{(2)} h_1^{(2)} + w_3^{(2)} h_2^{(2)} \right) + w_2^{(1)} \left( w_2^{(2)} h_1^{(2)} + w_4^{(2)} h_2^{(2)} \right)$$

$$= w_1^{(1)} w_1^{(2)} h_1^{(2)} + w_1^{(1)} w_3^{(2)} h_2^{(2)} + w_2^{(1)} w_2^{(2)} h_1^{(2)} + w_2^{(1)} w_4^{(2)} h_2^{(2)}$$

$$\vec{W}_*^{(1)} = \vec{W}^{(1)} - \alpha \frac{\partial \ell(\hat{y}, y)}{\partial \vec{w}^{(1)}}$$

$$= \vec{W}^{(1)} - \alpha \frac{\partial \ell(\hat{y}, y)}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \vec{w}^{(1)}}$$

$$= \vec{W}^{(1)} - \alpha \frac{\partial \ell(\hat{y}, y)}{\partial \hat{y}} \begin{bmatrix} \frac{\partial \hat{y}}{\partial w_1^{(1)}} & \frac{\partial \hat{y}}{\partial w_2^{(1)}} \end{bmatrix}^T$$

$$= \vec{W}^{(1)} - \alpha (\hat{y} - y) \begin{bmatrix} w_1^{(2)} h_1^{(2)} + w_3^{(2)} h_2^{(2)} & w_2^{(2)} h_1^{(2)} + w_4^{(2)} h_2^{(2)} \end{bmatrix}^T$$

$$= \vec{W}^{(1)} - 0.05(-0.7467) \begin{bmatrix} 0.1581 & 0.0952 \end{bmatrix}^T$$

$$= \begin{bmatrix} 0.17 \\ 0.17 \end{bmatrix} + 0.0373 \begin{bmatrix} 0.1581 \\ 0.0952 \end{bmatrix}$$

$$= \begin{bmatrix} 0.1759 \\ 0.1736 \end{bmatrix}$$

$$\underline{W}_*^{(2)} = \underline{W}^{(2)} - \alpha \frac{\partial \ell(\hat{y}, y)}{\partial \underline{w}^{(2)}}$$

$$= \underline{W}^{(2)} - \alpha \frac{\partial \ell(\hat{y}, y)}{\partial (\vec{w}^{(1)} \vec{h}^{(1)})} \frac{\partial (\vec{w}^{(1)} \vec{h}^{(1)})}{\partial \underline{w}^{(2)}}$$

$$= \underline{W}^{(2)} - \alpha \frac{\partial \ell(\hat{y}, y)}{\partial (\vec{w}^{(1)} \vec{h}^{(1)})} \frac{\partial (\vec{w}^{(1)} \vec{h}^{(1)})}{\partial \underline{w}^{(2)}}$$

$$= \underline{W}^{(2)} - a\left(\hat{y} - y\right) \begin{bmatrix} \partial\hat{y}/\partial w_1^{(2)} & \partial\hat{y}/\partial w_3^{(2)} \\ \partial\hat{y}/\partial w_2^{(2)} & \partial\hat{y}/\partial w_4^{(2)} \end{bmatrix}$$

$$= \underline{W}^{(2)} - a\left(\hat{y} - y\right) \begin{bmatrix} w_1^{(1)} h_1^{(2)} & w_1^{(1)} h_2^{(2)} \\ w_2^{(1)} h_1^{(2)} & w_2^{(1)} h_2^{(2)} \end{bmatrix}$$

$$= \begin{bmatrix} 0.12 & 0.13 \\ 0.23 & 0.10 \end{bmatrix} + 0.0373 \begin{bmatrix} 0.34 & 0.51 \\ 0.34 & 0.51 \end{bmatrix}$$

$$= \begin{bmatrix} 0.1327 & 0.1490 \\ 0.2427 & 0.1190 \end{bmatrix}$$

Now we do a forward pass with these new weights

$$\vec{h}^{(2)} = \begin{bmatrix} 2 \\ 3 \end{bmatrix} \qquad y = 1$$

$$\hat{y} = \vec{w}_*^{(1)} \, \underline{w}_*^{(2)} \, \vec{h}^{(2)}$$

$$= \begin{bmatrix} 0.1759 \\ 0.1736 \end{bmatrix} \begin{bmatrix} 0.1327 & 0.1490 \\ 0.2427 & 0.1190 \end{bmatrix} \begin{bmatrix} 2 \\ 3 \end{bmatrix}$$

$$= 0.2715$$

This is about 7% better than our previous output of 0.2533, so the backpropagation update clearly is working