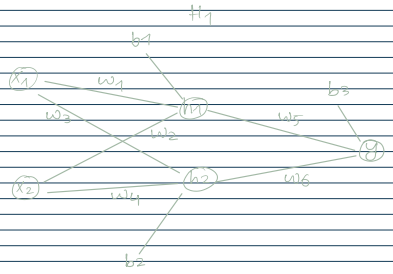
Input (x) = $[x_1, x_2]$

target (Y) = [1]

$h1 = \text{sigmoid}(w_1 * x_1 + w_3 * x_2 + b_1)$

$h2 = \text{sigmoid}(w_2 * x_1 + w_4 * x_2 + b_2)$

$\text{output} = \text{sigmoid}(w_5 * h1 + w_6 * h_2 + b_3)$



MSE loss:

$\text{loss} = 0.5 * (\text{output} - \text{target})^2$

Backward pass——

Computing Gradients:

— $\frac{d(\text{loss})}{d(\text{output})} = \text{output} - \text{target}$ → constant

— $\frac{d(\text{output})}{d(w_5)} = h1 * \text{output}(1 - \text{output})$

— $\frac{d(\text{output})}{d(w_6)} = h2 * \text{output}(1 - \text{output})$

— $\frac{d(\text{output})}{d(h1)} = w_5 * \text{output}(1 - \text{output})$

— $\frac{d(\text{output})}{d(h2)} = w_6 * \text{output}(1 - \text{output})$

— $\frac{d(h1)}{d(w1)} = x1 * h1(1 - h1)$

— $\frac{d(h1)}{d(w2)} = x2 * h1(1 - h1)$

— $\frac{d(h2)}{d(w3)} = x1 * h2(1 - h2)$

— $\frac{d(h2)}{d(w4)} = x2 * h2(1 - h2)$

Back-propagating Gradients:

— updating using learning rate:

$w1 = w1 - \text{learning rate} * \frac{d(\text{loss})}{d(\text{output})} * \frac{d(\text{output})}{d(h1)} * \frac{d(h1)}{d(w1)}$

$w2 = w2 - \text{learning rate} * \frac{d(\text{loss})}{d(\text{output})} * \frac{d(\text{output})}{d(h1)} * \frac{d(h1)}{d(w2)}$

$\text{loss} = \frac{1}{2}(\hat{y} - y)^2$

$\frac{d(\text{loss})}{d(\hat{y})} = \frac{d}{d\hat{y}}\left(\frac{1}{2}(\hat{y} - y)^2\right)$

$= \frac{1}{2} \times 2(\hat{y} - y)$

$= (\hat{y} - y)$

Sigmoid func^n:

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

$\text{output} = \frac{1}{1 + e^{-(w_5 * h1 + w_6 * h_2 + b_3)}}$

$\frac{d(\text{output})}{d(w_5)} = -(1 + e^{-(w_5 * h1 + w_6 * h_2 + b_3)})^{-2} * (e^{-(w_5 * h1 + w_6 * h_2 + b_3)}) * h1$

$= S(x) \cdot (1 - S(x))$; simplified