1 Neural Network

Assume we have the following data point $\boldsymbol{x}$:

$$\boldsymbol{x} = \begin{bmatrix} 2 \\ 5 \end{bmatrix}$$

with corresponding binary label:

$$y = \begin{bmatrix} 1 \end{bmatrix}$$

Which is part of a larger dataset $\boldsymbol{X}$ with binary labels $\boldsymbol{y}$.
We want to create a neural network to solve this classification problem. We want to use squared error as the loss function: $\ell = \frac{1}{2}(y - \hat{y})^2$. (Please take a moment and think about why do we use this loss function.) We will use stochastic gradient descent to train our network by minimizing the loss. Suppose we want only one hidden layer with two neurons, sigmoid activation functions and we want to include all bias terms.

1. Draw what our neural network will look like.

2. What will be the shape of our weight matrices (Hint: we view $\mathbf{x}$ as a column vector and compute the product as $\boldsymbol{W}^T\mathbf{x}$.)

   (a) $\boldsymbol{W}^1 =$

   (b) $\boldsymbol{W}^2 =$

   $\boldsymbol{W}^1 = 3 \times 2$
   $\boldsymbol{W}^2 = 3 \times 1$

Assume our weights are as follows:

$$\boldsymbol{W^1} = \begin{bmatrix} -1 & 2 \\ -2 & 1 \\ 1 & 4 \end{bmatrix} \boldsymbol{W^2} = \begin{bmatrix} 1 \\ 2 \\ -4 \end{bmatrix}$$

3. What are the values being passed into our hidden neurons? (Hint: the bias term is considered the first feature.)

Recall that we need to add a one to our example to represent the bias term.

$$\boldsymbol{\alpha^1} = \boldsymbol{W}^{1T}\boldsymbol{x} = \begin{bmatrix} -1 & -2 & 1 \\ 2 & 1 & 4 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 5 \end{bmatrix} = \begin{bmatrix} -1-4+5 \\ 2+2+20 \end{bmatrix} = \begin{bmatrix} 0 \\ 24 \end{bmatrix}$$

4. What are the outputs of our hidden neurons?

We need to apply the Sigmoid function to this vector to get the output, recall the Sigmoid function is:

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

So our output will be $\boldsymbol{\beta^1} = \begin{bmatrix} 0.5 \\ 0.999 \end{bmatrix}$

Recall we need to add a +1 to the start of this vector to represent the Bias term in this layer!

5. What is the value being passed into our output layer?

$$\alpha^2 = \boldsymbol{W^{2T}}\beta^1 = \begin{bmatrix} 1 & 2 & -4 \end{bmatrix} \begin{bmatrix} 1 \\ 0.5 \\ 0.999 \end{bmatrix} = 1 + 1 - 3.996 = -1.996$$

6. What is our final output $\hat{y}$?

Recall we need to apply the Sigmoid function to this value α^2.
$\hat{y} = \frac{1}{1+e^{1.996}} = 0.119623$

7. What does our model classify $\boldsymbol{x_1}$ as?

0

8. What is the loss ℓ on this example?

$\ell = \frac{1}{2}(y - \hat{y})^2 = \frac{1}{2}(1 - 0.119623)^2 = 0.38753$

Now we do backpropagation to update the weights. I hope you stored your values because you will be reusing them!

First you need to calculate the derivative of ℓ with respect to all of the weights in $\boldsymbol{W^1}$ and $\boldsymbol{W^2}$, then you use gradient descent to update them. Be sure to do it in this order because if you update the weights in $\boldsymbol{W^2}$ before calculating the derivative with respect to $\boldsymbol{W^1}$ then your calculations will be incorrect.

In the interest of time lets just calculate the derivatives: $\frac{d\ell}{d\boldsymbol{W^2_{11}}}$ and $\frac{d\ell}{d\boldsymbol{W^1_{21}}}$.

Let's start with $\frac{d\ell}{d\boldsymbol{W^2_{11}}}$. It helps to examine the path that this weight effects $\boldsymbol{W^2_{11}}$:

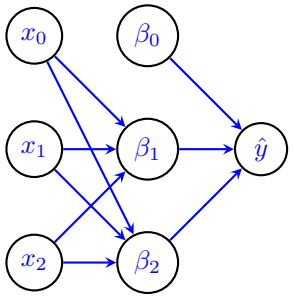

Figure 1: Neural Network

$\boldsymbol{W^2_{11}}$ follows a simple propagation path from $\boldsymbol{\beta^1_1}$ to $\hat{y}$.

So our derivative chain will be:

$$\frac{d\ell}{d\boldsymbol{W^2_{11}}} = \frac{d\ell}{d\hat{y}} \frac{d\hat{y}}{d\alpha^2} \frac{d\alpha^2}{d\boldsymbol{W^2_{11}}}$$

1. What is $\frac{d\ell}{d\hat{\boldsymbol{y}}}$?

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

2. What is $\frac{d\hat{y}}{d\boldsymbol{\alpha^2}}$?

Recall that the derivative of the Sigmoid function has a very simple form! $\sigma(\alpha^2)(1 - \sigma(\alpha^2))$
Notice that we can simplify this further as $\hat{y}(1 - \hat{y})$

3. What is $\frac{d\alpha^2}{d\boldsymbol{W_{11}^2}}$?

Recall: $\alpha^2 = \boldsymbol{W_{01}^2}\beta_0^1 + \boldsymbol{W_{11}^2}\beta_1^1 + \boldsymbol{W_{21}^2}\beta_2^1$
$\beta_1^1 = 0.5$

4. Finally, what is $\frac{d\ell}{d\boldsymbol{W_{11}^2}}$?

$\frac{d\ell}{d\boldsymbol{W_{11}^2}} = -(y - \hat{y})\hat{y}(1 - \hat{y})0.5 = -(1 - 0.119623)0.119623(1 - 0.119623)0.5 = -0.04635772$

Now we will calculate the derivative with respect to $\boldsymbol{W_{21}^1}$: $\frac{d\ell}{d\boldsymbol{W_{21}^1}}$

$$\frac{d\ell}{d\boldsymbol{W_{21}^1}} = \frac{d\ell}{d\hat{\boldsymbol{y}}}\frac{d\hat{y}}{d\boldsymbol{\alpha^2}}\frac{d\alpha^2}{d\boldsymbol{\beta_1^1}}\frac{d\beta_1^1}{d\boldsymbol{\alpha_1^1}}\frac{d\alpha_1^1}{d\boldsymbol{W_{21}^1}}$$

We already had the first two derivatives calculated from the previous question.

5. What is $\frac{d\alpha^2}{d\beta_1^1}$?

Recall: $\alpha^2 = \boldsymbol{W_{01}^2}\beta_0^1 + \boldsymbol{W_{11}^2}\beta_1^1 + \boldsymbol{W_{21}^2}\beta_2^1$
The answer is $\boldsymbol{W_{11}^2} = 2$

6. What is $\frac{d\beta_1^1}{d\alpha_1^1}$?

This is again the derivative of a sigmoid function, so our answer is: $\sigma(\alpha_1^1)(1-\sigma(\alpha_1^1)) = \beta_1^1(1-\beta_1^1)$

7. What is $\frac{d\alpha_1^1}{d\boldsymbol{W_{21}^1}}$?

Recall $\alpha_1^1 = \boldsymbol{W_{01}^1}x_0 + \boldsymbol{W_{11}^1}x_1 + \boldsymbol{W_{21}^1}x_2$
Hence: $x_2 = 5$

8. Finally what is $\frac{d\ell}{d\boldsymbol{W_{21}^1}}$?

$\frac{d\ell}{d\boldsymbol{W_{21}^1}} = -(y-\hat{y})\hat{y}(1-\hat{y})2\beta_1^1(1-\beta_1^1)5 = -(1-0.119623)0.119623(1-0.119623)2\cdot 0.5\cdot 0.5\cdot 5 = -0.23178$

9. What is our updated $\boldsymbol{W_{11}^2}$ and $\boldsymbol{W_{21}^1}$ if we use learning rate $\lambda = 2$?

Recall gradient descent rule: $w_{new} = w_{old} - \lambda\frac{d\ell}{d\boldsymbol{w_{old}}}$
$\boldsymbol{W_{11}^2} = 2 - 2\cdot -0.04635772 = 2.0927$
$\boldsymbol{W_{21}^1} = 1 - 2\cdot -0.23178 = 1.4636$