ECE4179 Assignment

Lucas Hou-wen Liu

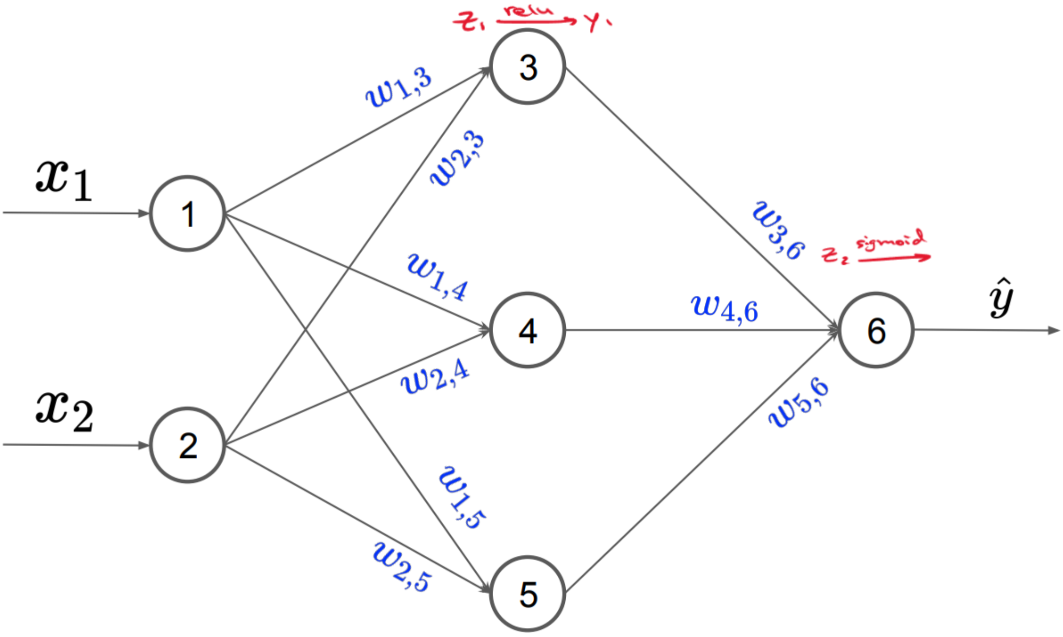
31445179

lliu0060@student.monash.edu

Calculation Exercise 1: MLP

1.1 Compute the output of the network for $x = (1, 2)^T$

$$w_1 = \begin{bmatrix} 0.4 & 0.0 \\ -0.2 & 0.7 \\ -0.3 & 0.1 \end{bmatrix}$$
$$w_2 = \begin{bmatrix} -0.2 & 0.5 & -0.6 \end{bmatrix}$$



$$z_1 = w_1 x = (0.4, 1.2, -0.1)^T$$
$$y_1 = \text{relu}(z_1) = (0.4, 1.2, 0.0)^T$$
$$z_2 = w_2 y_1 = 0.52$$
$$\hat{y} = \text{sigmoid}(z_2) = 0.627$$

1.2 Assume the label of $x = (1,2)^T$ is $y = 0$. If we use the BCE loss to train our MLP ,what will be the value of the loss?

BCE is defined as

$$\text{BCE}(y, \hat{y}) = -\big(y \ln(\hat{y}) + (1 - y) \ln(1 - \hat{y})\big)$$

For $\hat{y} = 0.627$ and $y = 0$, the loss is then

$$\begin{aligned} \text{BCE}(0, 0.627) &= -\ln(1 - 0.627) \\ &= 0.987 \end{aligned}$$

1.3 Now assume the label of $x = (1,2)^T$ is $y = 1$. Do you expect the loss to be bigger or smaller compared to the previous part?

We would expect the loss to be *smaller* than in the previous case. This is because the output of 0.627 is closer to 1 than it is to 0, which means the loss should be smaller.

For $\hat{y} = 0.627$ and $y = 1$, the loss is then

$$\begin{aligned} \text{BCE}(0, 0.627) &= -\ln(0.627) \\ &= 0.467 \end{aligned}$$

This confirms what was expected.

1.4 Assume the learning rate of the SGD is $lr = 0.1$. For a training sample $x = (1,2)^T$ and $y = 0$, obtain the updated value of $w_{3,6}$

For this, we need to find $\frac{\partial \mathcal{L}}{\partial w_{3,6}}$ using the chain rule

$$\frac{\partial \mathcal{L}}{\partial w_{3,6}} = \frac{\partial \mathcal{L}}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z_2} \frac{\partial z_2}{\partial w_{3,6}}$$

From the definition of BCE loss, we have

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \hat{y}} &= -\frac{y}{\hat{y}} + \frac{1 - y}{1 - \hat{y}} \\ &= \frac{\hat{y} - y}{\hat{y}(1 - \hat{y})} \end{aligned}$$

Evaluated at $\hat{y} = 0.627$ and $y = 0$ we get $\frac{\partial \mathcal{L}}{\partial \hat{y}} = 2.682$

From the definition of the sigmoid, we have

$$\frac{\partial \hat{y}}{\partial z_2} = \hat{y}(1-\hat{y})$$

Evaluated at $\hat{y} = 0.627$ we get $\frac{\partial \hat{y}}{\partial z_2} = 0.234$

From matrix multiplication, we have

$$\frac{\partial z_2}{\partial w_{3,6}} = y_{1_1}$$

From the calculations in part 1.1, we get $\frac{\partial z_2}{\partial w_{3,6}} = 0.4$

Putting these numbers together

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial w_{3,6}} &= 2.682 \cdot 0.234 \cdot 0.4 \\ &= 0.251\end{aligned}$$

The updated value for $w_{3,6}$ is given by

$$\begin{aligned}w_{3,6} &= w_{3,6} - lr \cdot \frac{\partial \mathcal{L}}{\partial w_{3,6}} \\ &= -0.2 - 0.1 \cdot 0.251 \\ &= -0.225\end{aligned}$$

1.5 Using the assumptions from the previous part, obtain the updated value of $w_{2,5}$

Using the chain rule

$$\frac{\partial \mathcal{L}}{\partial w_{2,5}} = \frac{\partial \mathcal{L}}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z_2} \frac{\partial z_2}{\partial y_{1_3}} \frac{\partial y_{1_3}}{\partial z_{1_3}} \frac{\partial z_{1_3}}{\partial w_{2,5}}$$

The first 2 terms in the product are the same as in the previous part.

$\frac{\partial z_2}{\partial y_{1_3}}$ is given by the weight $w_{5,6} = -0.6$

$\frac{\partial y_{1_3}}{\partial z_{1_3}}$ is given by the derivative of ReLU for z_{1_3} which happens to be 0 (since $z_{1_3} \leq 0$).

Since $\frac{\partial y_{1_3}}{\partial z_{1_3}} = 0$, the entire term becomes zero and the updated value is $w_{2,5} = 0.1$, which is the same as its previous value.

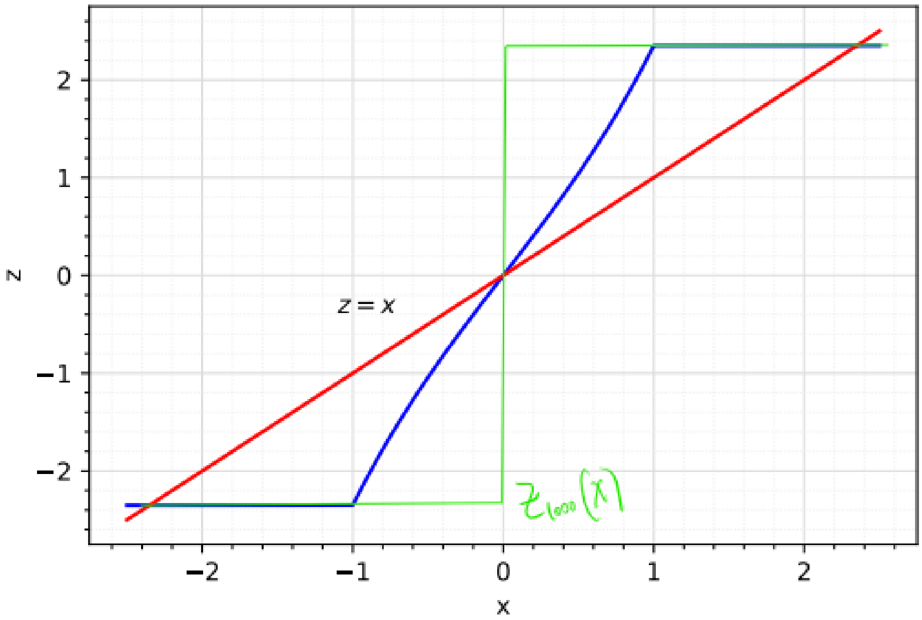
Calculation Exercise 2: Activation Function

For input values of $|x| \approx 0.5$ and greater, after one pass through the the activation function (z), they will be taken to the outer boundaries of the function, after which another pass through z will lock them at the constant outer boundary value. The output of the outer values is greater than the piecewise boundary so those values will always stay there.

As for inputs $|x| < 0.5$, we can see that the gradient of z is greater than 1, which means those values will shift outwards after each pass through z and eventually also get locked at the constant outer values.

For $x = 0$, the output will remain at exactly 0, and for extremely minisule inputs they would lie somewhere between -2.35 and 2.35.

The graph of z_{1000} would essentially be a step function with the low output being -2.35, a high output of 2.35 and a very small continuous section in between those two levels.


z
x
2
1
0
−1
−2
−2
−1
0
1
2
z = x
$z_{1000}(x)$