Training a Neural Network with Backpropagation—Mathematics

The backpropagation algorithm has two main phases- forward and backward phase.

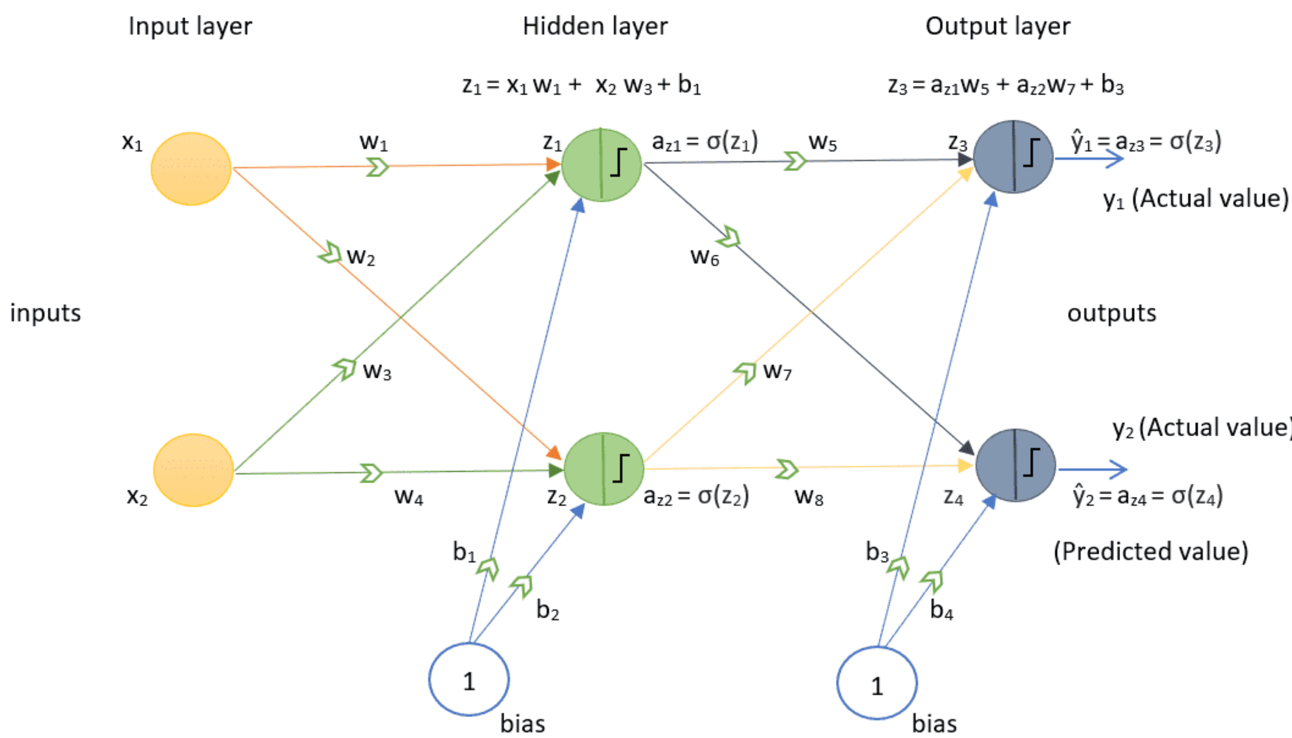

Figure 1 - Artificial Neural Network.

The structure of a simple three-layer neural network is shown in Figure 1. Here, every neuron of one layer is connected to all neurons of the next layer but neurons of the same layer are not interconnected. The information flows from the first layer neurons (input layer), via the second layer neurons (hidden layer) to the third layer neurons (output layer).

Let's consider, the inputs, outputs, the initial weights and biases as:

Input values: $x_1 = 0.05$, $x_2 = 0.10$

Output values: $y_1 = 0.01$, $y_2 = 0.99$

Initial weights: $w_1=0.15$, $w_2=0.20$, $w_3=0.25$, $w_4=0.30$, $w_5=0.40$, $w_6=0.45$, $w_7=0.50$, $w_8=0.55$

Initial bias: $b_1=0.40$, $b_2=0.35$, $b_3=0.25$, $b_4=0.60$

Forward Pass

The input layer receives signals and without performing any computation simply transmits the information to the hidden layer. The net input to a neuron of the hidden layer is calculated as the summation of each output of the input layer multiplied by weights (weights are initialized as small random numbers) and an additional bias is incorporated. Then sigmoid activation function is applied to learn complex patterns in the data and to normalize the output of each neuron to a range between 1 and 0. In each successive layer, every neuron sums its inputs and then applies an activation function to compute its output. The output layer of the network then produces the final response, i.e., the predicted value.

$z_1 = x_1 w_1 + x_2 w_3 + 1*b_1 = (0.05 *0.15) + (0.10 * 0.25) + (1 * 0.40) = 0.4325$

Let's consider, Activation function = "sigmoid",
$a_{z1} = \sigma(z_1) = 1/(1+\exp(-z_1)) = 1/(1+\exp(- 0.4325)) = 0.606470487$

$z_2 = x_1 w_2 + x_2 w_4 + 1*b_2 = (0.05 *0.20) + (0.10 * 0.30) + (1 * 0.35) = 0.39$

$a_{z2} = \sigma(z_2) = 1/(1+\exp(-z_2)) = 1/(1+\exp(- 0.39)) = 0.596282699$

$z_3 = a_{z1}w_5 + a_{z2} w_7 + 1*b_3 = 0.790729544$

$\hat{y}_1 = a_{z3} = \sigma(z_3) = 1/(1+\exp(-z_3)) = 0.687987956$

$z_4 = a_{z1}w_6 + a_{z2} w_8 + 1*b_4 = 1.200867204$

$\hat{y}_2 = a_{z4} = \sigma(z_4) = 1/(1+\exp(-z_4)) = 0.768679018$

Total Error Calculation

Now, we need to calculate the total error using the mean squared error loss function. Loss function describes how efficient the model performs with respect to the expected outcome.

Consider, loss function= "mean squared error"

$E_{total} = ½ *[(y_1 - \hat{y}_1)^2 + (y_2 – \hat{y}_2)^2] = ½ * [(0.01 - 0.687987956)^2 + (0.99 - 0.768679018)^2]$

$= 0.254325322 \approx 0.2543$

The derivative of this loss now needs to be computed with respect to the weights and bias in all layers in the backward phase.

First Backward Pass

The main goal of the backward phase is to learn the gradient of the loss function with respect to the different weights and bias by using the chain rule of differential calculus. These gradients are used to update the weights and bias. Since these gradients are learned in the backward direction, starting from the output node, this learning process is referred to as the backward propagation.

Weight updation – w_5

$$\frac{\partial Etotal}{\partial w5} = \frac{\partial Etotal}{\partial \hat{y}1}\,\frac{\partial \hat{y}1}{\partial z3}\,\frac{\partial z3}{\partial w5}$$

$$= \frac{\partial}{\partial \hat{y}1}[½ *(y_1 - \hat{y}_1)^2] \;\; \frac{\partial}{\partial z3}[1/(1+\exp(-z_3)] \;\; \frac{\partial}{\partial w5}[a_{z1}w_5 + a_{z2} w_7 + 1*b_3]$$

(Since, $\hat{y}_1 = a_{z3}$)

$= (y_1 - \hat{y}_1)(-1) \,.\, \hat{y}_1 (1 - \hat{y}_1) \,.\, a_{z1}$

(∵ if the sigmoid function is defined as, $\sigma(z) = 1/(1+\exp(-z)$
it's differentiation is, $d\sigma(z)/d(z) = \sigma(z)\cdot(1-\sigma(z))$)

$= (\hat{y}_1 - y_1) \cdot \hat{y}_1 (1 - \hat{y}_1) \cdot a_{z1}$

$= (0.687987956 – 0.01)(0.687987956)(1- 0.687987956)(0.606470487)$

$= 0.088264048$

$w_{5(new)} = w_5 – \eta \frac{\partial Etotal}{\partial w5}$

$= 0.40 – (0.01 * 0.088264048)$

(Consider, $\eta = 0.01$)

$= 0.399117359$

Weight updation – w_6

$\frac{\partial Etotal}{\partial w6} = \frac{\partial Etotal}{\partial \hat{y}2} \frac{\partial \hat{y}2}{\partial z4} \frac{\partial z4}{\partial w6}$

$= (\hat{y}_2 – y_2) \cdot \hat{y}_2 (1 – \hat{y}_2) \cdot a_{z1} = - 0.023866696$

$w_{6(new)} = w_6 – \eta \frac{\partial Etotal}{\partial w6}$

$= 0.45 – (0.01 * - 0.023866696) = 0.450238667$

Weight updation – w_7

$\frac{\partial Etotal}{\partial w7} = \frac{\partial Etotal}{\partial \hat{y}1} \frac{\partial \hat{y}1}{\partial z3} \frac{\partial z3}{\partial w7}$

$= (\hat{y}_1 - y_1) \cdot \hat{y}_1 (1 - \hat{y}_1) \cdot a_{z2} = 0.186047383$

$w_{7(new)} = w_7 – \eta \frac{\partial Etotal}{\partial w7}$

$= 0.50 – (0.01 * 0.186047383) = 0.498139526$

Weight updation – w_8

$\frac{\partial Etotal}{\partial w8} = = \frac{\partial Etotal}{\partial \hat{y}2} \frac{\partial \hat{y}2}{\partial z4} \frac{\partial z4}{\partial w8}$

$= (\hat{y}_2 – y_2) \cdot \hat{y}_2 (1 – \hat{y}_2) \cdot a_{z2} = - 0.023465772$

$w_{8(new)} = w_8 – \eta \frac{\partial Etotal}{\partial w8}$

$= 0.55 – (0.01 * - 0.023465772) = 0.550234657$

Bias constant (usually 1) has its own weight for different nodes. The weight of the bias in a layer is updated in the same fashion as all the other weights are updated.

Bias updation – b_3

$$\frac{\partial Etotal}{\partial b3} = \frac{\partial Etotal}{\partial \hat{y}1}\ \frac{\partial \hat{y}1}{\partial z3}\ \frac{\partial z3}{\partial b3}$$

$$= (\hat{y}_1 - y_1)\ .\ \hat{y}_1 (1 - \hat{y}_1)\ .\ 1 = 0.145537252$$

$$(\because \frac{\partial}{\partial b3}(a_{z1}w_5 + a_{z2}\ w_7 + b_3) = 1)$$

$$b_{3(new)} = b_3 - \eta \frac{\partial Etotal}{\partial b3}$$

$$= 0.25 - (0.01 * 0.145537252) = 0.248544627$$

Bias updation – b_4

$$\frac{\partial Etotal}{\partial b4} = \frac{\partial Etotal}{\partial \hat{y}2}\ \frac{\partial \hat{y}2}{\partial z4}\ \frac{\partial z4}{\partial b4}$$

$$= (\hat{y}_2 - y_2)\ .\ \hat{y}_2 (1 - \hat{y}_2)\ .\ 1 = -\ 0.039353434$$

$$b_{4(new)} = b_4 - \eta \frac{\partial Etotal}{\partial b4}$$

$$= 0.60 - (0.01 * -\ 0.039353434) = 0.60039353434$$

Next, we will continue the backwards pass to update the values of w1, w2, w3, w4 and b1, b2. The gradient with respect to these weights and bias depends on w5 and w8, and we will be using the old values, not the updated ones.

Weight updation – w_1

$$\frac{\partial Etotal}{\partial w1} = \left(\frac{\partial Etotal}{\partial az1}\right)\ \frac{\partial az1}{\partial z1}\ \frac{\partial z1}{\partial w1}$$

$$= \left(\frac{\partial Etotal}{\partial \hat{y}1}\ \frac{\partial \hat{y}1}{\partial z3}\ \frac{\partial z3}{\partial az1}\right)\frac{\partial az1}{\partial z1}\ \frac{\partial z1}{\partial w1}$$

$$(\because, \frac{\partial Etotal}{\partial az1} = \frac{\partial Etotal}{\partial \hat{y}1}\ \frac{\partial \hat{y}1}{\partial z3}\ \frac{\partial z3}{\partial az1})$$

$$= \frac{\partial}{\partial \hat{y}1}[\tfrac{1}{2} *(y_1 - \hat{y}_1)^2]\ \frac{\partial}{\partial z3}[1/(1+\exp(-z_3))]\ \frac{\partial}{\partial az1}[a_{z1}w_5 + a_{z2}\ w_7 + 1*b_3]$$
$$\frac{\partial}{\partial z1}[1/(1+\exp(-z_1))]\ \frac{\partial}{\partial w1}[x_1\,w_1 + x_2\,w_3 + 1*b_1]$$

$$= (\hat{y}_1 - y_1)\ .\ \hat{y}_1 (1 - \hat{y}_1)\ .\ w_5\ .\ a_{z1} (1 - a_{z1}).\ x_1$$

$$= 0.145537252 * 0.40 * 0.606470487 * (1 - 0.606470487) * 0.05$$

$$= 0.694690157 * 10^{-3}$$

$$w_{1(new)} = w_1 - \eta \frac{\partial Etotal}{\partial w1}$$

$$= 0.15 - (0.01 * 0.694690157 * 10^{-3}) = 0.149993053$$

Weight updation – w_2

$$\frac{\partial Etotal}{\partial w2} = \left(\frac{\partial Etotal}{\partial \hat{y}2} \frac{\partial \hat{y}2}{\partial z4} \frac{\partial z4}{\partial az2} \right) \frac{\partial az2}{\partial z2} \frac{\partial z2}{\partial w2}$$

$= (\hat{y}_2 - y_2) \, . \, \hat{y}_2 \, (1 - \hat{y}_2). \, w_8 \, . \, a_{z2} \, (1 - a_{z2}) \, . \, x_1$

$= - 0.039353434 * 0.55 * 0.596282699 * (1 - 0.596282699) * 0.05$

$= - 2.60522297 * 10^{-4}$

$w_{2(new)} = w_2 - \eta \frac{\partial Etotal}{\partial w2}$

$= 0.20 - (0.01 * - 2.60522297 * 10^{-4}) = 0.200002605$

Weight updation – w_3

$$\frac{\partial Etotal}{\partial w3} = \left(\frac{\partial Etotal}{\partial \hat{y}1} \frac{\partial \hat{y}1}{\partial z3} \frac{\partial z3}{\partial az1}\right) \frac{\partial az1}{\partial z1} \frac{\partial z1}{\partial w3}$$

$= (\hat{y}_1 - y_1) \, . \, \hat{y}_1 \, (1 - \hat{y}_1) \, . \, w_5 \, . \, a_{z1} \, (1- a_{z1}). \, x_2$

$= 0.145537252 * 0.40 * 0.606470487 * (1- 0.606470487) * 0.10$

$= 1.389380314 * 10^{-3}$

$w_{3(new)} = w_3 - \eta \frac{\partial Etotal}{\partial w3}$

$= 0.25 - (0.01 * 1.389380314 * 10^{-3}) = 0.249986106$

Weight updation – w_4

$$\frac{\partial Etotal}{\partial w4} = \left(\frac{\partial Etotal}{\partial \hat{y}2} \frac{\partial \hat{y}2}{\partial z4} \frac{\partial z4}{\partial az2} \right) \frac{\partial az2}{\partial z2} \frac{\partial z2}{\partial w4}$$

$= (\hat{y}_2 - y_2) \, . \, \hat{y}_2 \, (1 - \hat{y}_2). \, w_8 \, . \, a_{z2} \, (1 - a_{z2}) \, . \, x_2$

$= - 0.039353434 * 0.55 * 0.596282699 * (1 - 0.596282699) * 0.10$

$= - 5.21044594 * 10^{-4}$

$w_{4(new)} = w_4 - \eta \frac{\partial Etotal}{\partial w4}$

$= 0.30 - (0.01 * - 5.21044594 * 10^{-4}) = 0.30000521$

Bias updation – b_1

$$\frac{\partial Etotal}{\partial b1} = \left(\frac{\partial Etotal}{\partial \hat{y}1} \frac{\partial \hat{y}1}{\partial z3} \frac{\partial z3}{\partial az1}\right) \frac{\partial az1}{\partial z1} \frac{\partial z1}{\partial b1}$$

$= (\hat{y}_1 - y_1) \, . \, \hat{y}_1 \, (1 - \hat{y}_1) \, . \, w_5 \, . \, a_{z1} \, (1- a_{z1}). \, 1$

= 0.145537252 * 0.40 * 0.606470487 * (1- 0.606470487)

= 0.013893803

$b_{1(new)} = b_1 – \eta \frac{\partial Etotal}{\partial b1}$

= 0.40 - (0.01 * 0.013893803) = 0.399861062

Bias updation – b_2

$\frac{\partial Etotal}{\partial b2} = \left(\frac{\partial Etotal}{\partial \hat{y}2} \frac{\partial \hat{y}2}{\partial z4} \frac{\partial z4}{\partial az2} \right) \frac{\partial az2}{\partial z2} \frac{\partial z2}{\partial b2}$

= $(\hat{y}_2 – y_2)$. $\hat{y}_2 (1 – \hat{y}_2)$. w_8 . $a_{z2} (1 – a_{z2})$. 1

= - 0.039353434 * 0.55 * 0.596282699 * (1 - 0.596282699)

= - 5.21044594 * 10^{-3}

$b_{2(new)} = b_2 – \eta \frac{\partial Etotal}{\partial b2}$

= 0.35 - (0.01 * - 5.21044594 * 10^{-3}) = 0.350052104

Updated Weights and Bias,

$w_{1(new)}$ = 0.149993053
$w_{2(new)}$ = 0.200002605
$w_{3(new)}$ = 0.249986106
$w_{4(new)}$ = 0.30000521
$w_{5(new)}$ = 0.399117359
$W_{6(new)}$ = 0.450238667
$W_{7(new)}$ = 0.498139526
$W_{8(new)}$ = 0.550234657

$b_{1(new)}$ = 0.399861062
$b_{2(new)}$ = 0.350052104
$b_{3(new)}$ = 0.248544627
$b_{4(new)}$ = 0.60039353434

Forward Pass with Updated Weights and Bias

$z_1' = x_1\, w_{1(new)} + x_2\, w_{3(new)} + 1*b_{1(new)}$

= (0.05 *0.149993053) + (0.10 * 0.249986106) + (1 * 0.399861062) = 0.432359325

Since, Activation function = “sigmoid”
a_{z1}'= σ(z_1') = 1/(1+exp(-z_1')) = 1/(1+exp(- 0.432359325)) = 0.606436912

$z_2' = x_1\, w_{2(new)} + x_2\, w_{4(new)} + 1*b_{2(new)}$

= (0.05 *0.200002605) + (0.10 * 0.30000521) + (1 * 0.350052104) = 0.390052755

a_{z2}'= σ(z_2') = 1/(1+exp(-z_2')) = 1/(1+exp(- 0.390052755)) = 0.596295398

$z_3' = a_{z1}' \, w_{5(new)} + a_{z2}' \, w_{7(new)} + 1 * b_{3(new)}$

= (0.606436912 * 0.399117359) + (0.596295398 * 0.498139526) + 0.248544627 = 0.787622432

$\hat{y}_1' = a_{z3}' = \sigma(z_3') = 1/(1+\exp(-z_3')) = 1/(1+\exp(-0.787622432)) = 0.687320592$

$z_4' = a_{z1}' \, w_{6(new)} + a_{z2}' \, w_{8(new)} + 1 * b_{4(new)}$

= (0.606436912 * 0.450238667) + (0.596295398 * 0.550234657) + 0.60039353434 = 1.201537275

$\hat{y}_2' = a_{z4}' = \sigma(z_4') = 1/(1+\exp(-z_4')) = 1/(1+\exp(-1.201537275)) = 0.768798143$

Error Calculation

$E_{total}' = ½ * [(y_1 - \hat{y}_1')^2 + (y_2 - \hat{y}_2')^2] = ½ * [(0.01 - 0.687320592)^2 + (0.99 - 0.768798143)^2]$

$= 0.253846722 \approx 0.2539$

After the first round of backpropagation, the total error has decreased to 0.2539 (approximately).