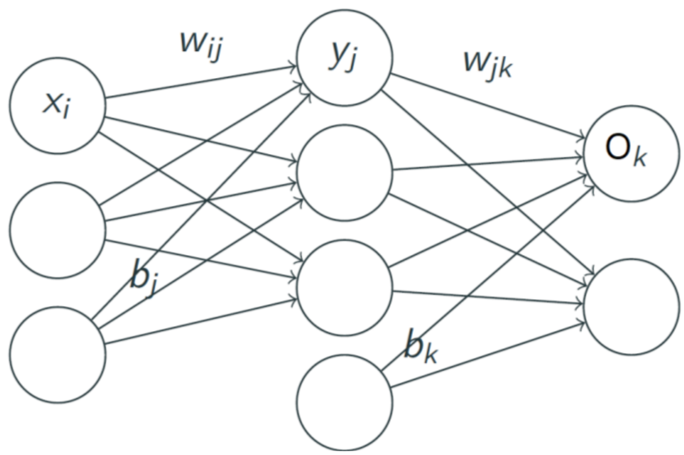
Determine $\frac{\partial E}{\partial w_{ij}^{O}}$ and $\frac{\partial E}{\partial w_{ij}^{H}}$ of loss function

$$E(w, b) = \frac{1}{2} \sum_{k \in N_o} (O_k - t_k)^2$$

for a network with one input layer (with N_I neurons), output layer (with N_O neurons) and hidden layer (with N_H neurons). Note that every neuron is assumed to be connected to every neuron of the next layer, i.e., a Multi-Layer Perceptron is considered. Further the sigmoid function is the considered action function for every neuron in the hidden and output layer.

Solution:



Here,

w_{ij}: weights connecting node i in layer $(l - 1)$ to node j in layer l.

b_j, b_k: bias for nodes j and k.

u_j, u_k: inputs to nodes j and k (where $u_j = b_j + \sum x_i w_{ij}$).

g_j, g_k: activation function for node j (applied to u_j) and node k.

$y_j = g_j(u_j)$, $O_k = g_k(u_k)$: output/activation of nodes j and k.

t_k: target value for node k in the output layer.

Nodes in the output layer:

Forward-propagate for each output O_k

$$O_k = g_k(u_k) = g_k(b_k + \sum y_j w_{jk}) = g_k(b_k + \sum g_j(b_j + \sum x_i w_{ij})\, w_{jk})$$

Error function,

$$E(w, b) = \frac{1}{2} \sum_{k \in N_o} (O_k - t_k)^2$$

Let's start at the output layer with weight W_{jk}, $u_j = b_j + \sum W_{ij} y_i$ and $u_k = b_k + \sum W_{jk} y_j$

Now,

$$\frac{\partial E}{\partial W_{ij}^{O}} = \frac{\partial E}{\partial O_K}\frac{\partial O_K}{\partial u_K}\frac{\partial u_K}{\partial y_j}\frac{\partial y_j}{\partial u_j}\frac{\partial u_j}{\partial W_{ij}^{O}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots. \quad (1)$$

Now,

$$\frac{\partial E}{\partial O_K} = \frac{\partial}{\partial O_K}\left(\frac{1}{2}\sum_{k\,\in\,N_O}(O_k - t_k)^2\right) = (O_k - t_k)\ldots\ldots\ldots\ldots\ldots \quad (2)$$

$$\frac{\partial O_K}{\partial u_K} = g_k'(u_k) \quad \ldots\ldots\ldots\ldots\ldots\ldots. \quad (3)$$

$$\frac{\partial u_K}{\partial y_j} = W_{jk} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (4)$$

$$\frac{\partial y_j}{\partial u_j} = g_j'(u_j) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots. \quad (5)$$

$$\frac{\partial u_j}{\partial W_{ij}^{O}} = \frac{\partial}{\partial W_{ij}^{O}}\left(b_j + \sum_i w_{ij}{}^{O} y_i\right) = y_i \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (6)$$

Using the value of (2), (3), (4), (5), and (6) we can write (1) as follows:

$$\frac{\partial E}{\partial W_{jk}^{O}} = \underbrace{(O_k - t_k)g_k'(u_k)W_{jk}\, g_j'(u_j)}_{\delta_j} y_j = \delta_j\, y_j$$

Here, $\delta_j = g_j'(u_j) \sum_{k\,\in K}(O_k - t_k)\, g_k'(u_k)W_{jk}$, the error in u_j.

Additionally,

$$\frac{\partial E}{\partial W_{jk}^{O}} = \frac{\partial E}{\partial O_K}\frac{\partial O_K}{\partial u_K}\frac{\partial u_K}{\partial W_{jk}^{O}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots. \quad (7)$$

Now,

$$\frac{\partial E}{\partial O_K} = \frac{\partial}{\partial O_K}\left(\frac{1}{2}\sum_{k\,\in\,N_O}(O_k - t_k)^2\right) = (O_k - t_k)\ldots\ldots\ldots\ldots\ldots \quad (8)$$

$$\frac{\partial O_K}{\partial u_K} = g_k'(u_k) \quad \ldots\ldots\ldots\ldots\ldots\ldots. \quad (9)$$

$$\frac{\partial u_K}{\partial W_{jk}^O} = \frac{\partial}{\partial W_{jk}^O}\left(b_k + \sum_j w_{jk}{}^O y_j\right) = y_j \quad \dots\dots\dots\dots\dots\dots\dots\dots \qquad (10)$$

Using the value of (8), (9), and (10) we can write (10) as follows:

$$\frac{\partial E}{\partial W_{jk}^O} = \underbrace{(O_k - t_k)g_k'(u_k)}_{\delta_k} y_j = \delta_k\, y_j$$

Here, $\delta_k = (O_k - t_k)g_k'(u_k)$ is called the error in u_k.

Nodes in the hidden layer:

Now we know,

$$u_j = b_i + \sum w_{jk} x_i$$

$$u_k = b_k + \sum w_{jk} g_j(u_i)$$

$$O_k = g_k(u_k)$$

Now,

$$\frac{\partial E}{\partial W_{ij}^H} = \sum_{k \in K} \frac{\partial E}{\partial u_K}\frac{\partial u_K}{\partial y_j}\frac{\partial y_j}{\partial u_j}\frac{\partial u_j}{\partial W_{ij}^H} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots . (11)$$

Now,

$$\frac{\partial E}{\partial u_K} = \delta_k \qquad \dots\dots\dots\dots \qquad (12)$$

$$\frac{\partial u_K}{\partial y_j} = W_{jk} \qquad \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \qquad (13)$$

$$\frac{\partial y_j}{\partial u_j} = g_j'(u_j) \qquad \dots\dots\dots\dots\dots\dots\dots. \qquad (14)$$

$$\frac{\partial u_j}{\partial W_{ij}^H} = x_i \qquad \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \qquad (15)$$

Using the value of (12), (13), (14), and (15) we can write (11) as follows

$$\frac{\partial E}{\partial W_{ij}^H} = \sum_{k \in K} \underbrace{\delta_k W_{jk} g_j'(u_j)}_{\delta_j} x_i = \delta_j x_i \dots\dots\dots\dots\dots\dots\dots (16)$$

Here, $\delta_j = g_j'(u_j) \sum_{k \in K}(O_k - t_k)\, g_k'(u_k) W_{jk}$, the error in u_j

Now since we know the O_k, y_j, x_i, u_k and u_j for a given set of parameter values w, b, we can use these expressions to calculate the gradients at each iteration and update them.

Update the weights and biases with learning rate η. For example

$$w_{jk} \leftarrow w_{jk} - \eta \frac{\partial E}{\partial W_{jk}^{O}} \; and \; w_{ij} \leftarrow w_{ij} - \eta \frac{\partial E}{\partial W_{ij}^{H}}$$