Notes on neural networks

Leonardo Mascelli

May 9, 2025

Contents

1 Network equation

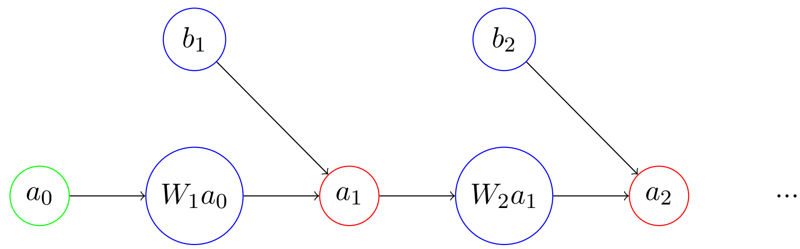
1.1 Simple single neuron network

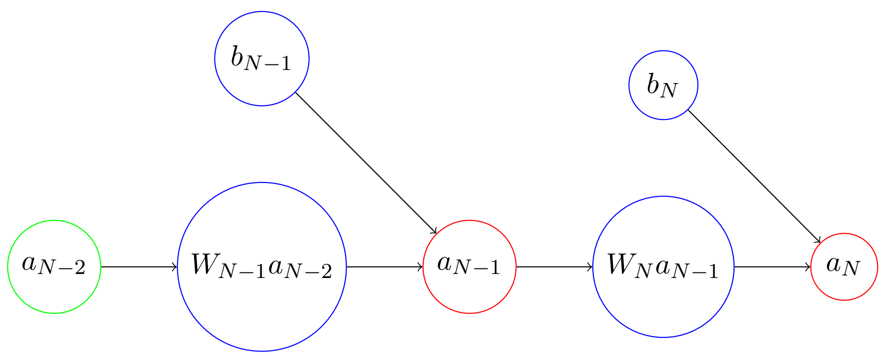
I'll start with a simple network where each layer has only one neuron, one input and one output.

Denote:

- a_0 is the input of the network,
- i is the i_{th} layer of the network, varing from 1 to N,
- W_i with the i_{th} matrix (in this case scalar) of weights of the i_{th} layer,
- b_i with the i_{th} vector (in this case scalar) of bias of the i_{th} layer,
- a_i with the i_{th} vector (in this case scalar) of outputs of the i_{th} layer,
- $\sigma(x) = \frac{1}{1+e^{-x}}$ so that

$$a_i = \sigma(W_i a_{i-1} + b_i) \tag{1}$$





and denote the cost function

$$C(W) = \sum_{c=1}^{C} (a_{N,c} - y_c)^2 \tag{2}$$

The goal is to find the parameters of the network $W_1, W_2, ..W_n$ that minimize the cost function.

$$\frac{\partial C}{\partial W_1} = \sum_{c=1}^{N} 2 \tag{3}$$

2 Complete network

2.1 Activation function of one layer

$$a_n^k = \sigma(\sum_{l=1}^{N(k-1)} w_{n,l}^k a_l^{k-1} + b_n) \tag{4}$$

2.2 Cost function

Let:

- T: the number of trials,
- K: the number of layers of the network,
- $N(k)$: the number of nodes in the k layer of the network,

$$C = \sum_{i=1}^{T} \sum_{n=1}^{N(K)} (a_n^{N(K)} - y_n)^2 \tag{5}$$

2.2.1 Layer K: Last layer

let's try finding the derivatives with the weights of the last layer, K and lets denote the error of the n_{th} output with:

$\epsilon_n = (a_n^K - y_n)$

$$\frac{\partial C}{\partial w_{c,p}^K} = \sum_{i=1}^{T} \sum_{n=1}^{N(K)} 2\epsilon_n \frac{\partial a_n^K}{\partial w_{c,p}^K} \tag{6}$$

$$\frac{\partial C}{\partial w_{c,p}^K} = \sum_{i=1}^{T} \sum_{n=1}^{N(K)} 2\epsilon_n \frac{\partial \sigma(\sum_{l=1}^{N(K-1)} w_{n,l}^K a_l^{K-1} + b_n)}{\partial w_{c,p}^K} \tag{7}$$

$$\frac{d\sigma(x)}{dx} = \sigma(x)(1 - \sigma(x)) \tag{8}$$

$$z_n^K = \sum_{l=1}^{N(K-1)} w_{n,l}^K a_l^{K-1} + b_n \tag{9}$$

$$\frac{\partial C}{\partial w_{c,p}^K} = \sum_{i=1}^{T} \sum_{n=1}^{N(K)} 2\epsilon_n a_n^K (1 - a_n^K) \frac{\partial z_n^K}{\partial w_{c,p}^K} \tag{10}$$

$$\frac{\partial C}{\partial w_{c,p}^K} = \sum_{i=1}^{T} 2\epsilon_c a_c^K (1 - a_c^K) a_p^{K-1} \tag{11}$$

We've found the the derivatives for the weights of the last layer are:

$$\frac{\partial C}{\partial w_{c,p}^K} = \sum_{i=1}^{T} 2\epsilon_c a_c^K (1 - a_c^K) a_p^{K-1} \tag{12}$$

Let's define $2\epsilon_c a_c^K(1 - a_c^K)$, the back propagation of the error in c node of the last layer as e_c^K. Then:

$$\frac{\partial C}{\partial w_{c,p}^K} = \sum_{i=1}^{T} e_c a_p \tag{13}$$

2.2.2 Layer K - 1

let's try now with the derivatives of the layer before the last, $K - 1$:

$$\frac{\partial C}{\partial w_{c,p}^{K-1}} = \sum_{i=1}^{T} \sum_{n=1}^{N(K)} 2\epsilon_n \frac{\partial a_n^K}{\partial w_{c,p}^{K-1}} \tag{14}$$

$$\frac{\partial C}{\partial w_{c,p}^{K-1}} = \sum_{i=1}^{T} \sum_{n=1}^{N(K)} 2\epsilon_n \frac{\partial \sigma(\sum_{l=1}^{N(K-1)} w_{n,l}^K a_l^{K-1} + b_n)}{\partial w_{c,p}^{K-1}} \tag{15}$$

$$z_n^K = \sum_{l=1}^{N(K-1)} w_{n,l}^K a_l^{K-1} + b_n \tag{16}$$

$$\frac{\partial C}{\partial w_{c,p}^{K-1}} = \sum_{i=1}^{T} \sum_{n=1}^{N(K)} 2\epsilon_n a_n^K(1 - a_n^K) \frac{\partial z_n^K}{\partial w_{c,p}^{K-1}} \tag{17}$$

$$\frac{\partial z_n^K}{\partial w_{c,p}^{K-1}} = \frac{\partial \sum_{l=1}^{N(K-1)} w_{n,l}^K a_l^{K-1} + b_n}{\partial w_{c,p}^{K-1}} = \sum_{l=1}^{N(K-1)} w_{n,l}^K \frac{\partial a_l^{K-1}}{\partial w_{c,p}^{K-1}} \tag{18}$$

$$z_l^{K-1} = \sum_{m=1}^{N(K-2)} w_{l,m}^{K-1} a_m^{K-2} + b_l \tag{19}$$

$$\frac{\partial z_n^K}{\partial w_{c,p}^{K-1}} = \sum_{l=1}^{N(K-1)} w_{n,l}^K a_l^{K-1}(1 - a_l^{K-1}) \frac{\partial z_l^{K-1}}{\partial w_{c,p}^{K-1}} \tag{20}$$

$$\frac{\partial z_n^{K-1}}{\partial w_{c,p}^{K-1}} = a_p^{K-2} \tag{21}$$

$$\frac{\partial z_n^K}{\partial w_{c,p}^{K-1}} = w_{n,c}^K a_c^{K-1}(1 - a_c^{K-1}) a_p^{K-2} \tag{22}$$

$$\frac{\partial C}{\partial w_{c,p}^{K-1}} = \sum_{i=1}^{T} \sum_{n=1}^{N(K)} 2\epsilon_n a_n^K(1 - a_n^K) w_{n,c}^K a_c^{K-1}(1 - a_c^{K-1}) a_p^{K-2} \tag{23}$$

We've found the the derivatives for the weights of layer before the last layer are:

$$\frac{\partial C}{\partial w_{c,p}^{K-1}} = \sum_{i=1}^{T} \sum_{n=1}^{N(K)} 2\epsilon_n a_n^K (1 - a_n^K) w_{n,c}^K a_c^{K-1} (1 - a_c^{K-1}) a_p^{K-2} \tag{24}$$

You can recognize in the equation the term $2e_n a_n^K(1 - a_n^K)$ to be what we first had defined as the propagation of the error in the K layer, e_n^K.

$$\frac{\partial C}{\partial w_{c,p}^{K-1}} = \sum_{i=1}^{T} \sum_{n=1}^{N(K)} e_n^K w_{n,c} a_c^{K-1} (1 - a_c^{K-1}) a_p^{K-2} \tag{25}$$

and define

$$e_c^{K-1} = \sum_{i=1}^{T} \sum_{n=1}^{N(K)} e_n^K w_{n,c} a_c^{K-1} (1 - a_c^{K-1}) \tag{26}$$

so that:

$$\frac{\partial C}{\partial w_{c,p}^{K-1}} = \sum_{i=1}^{T} e_c^{K-1} a_p^{K-2} \tag{27}$$

2.2.3 Layer K-2

let's try now with the derivatives of two layers before the last, $K-2$:

$$\frac{\partial C}{\partial w_{c,p}^{K-2}} = \sum_{i=1}^{T} \sum_{n=1}^{N(K)} 2\epsilon_n \frac{\partial a_n^K}{\partial w_{c,p}^{K-2}} \tag{28}$$

$$\frac{\partial C}{\partial w_{c,p}^{K-2}} = \sum_{i=1}^{T} \sum_{n=1}^{N(K)} 2\epsilon_n \frac{\partial \sigma(\sum_{l=1}^{N(K-1)} w_{n,l}^K a_l^{K-1} + b_n)}{\partial w_{c,p}^{K-2}} \tag{29}$$

$$\frac{\partial C}{\partial w_{c,p}^{K-2}} = \sum_{i=1}^{T} \sum_{n=1}^{N(K)} 2\epsilon_n a_n^K (1-a_n^K) \frac{\partial (\sum_{l=1}^{N(K-1)} w_{n,l}^K a_l^{K-1} + b_n)}{\partial w_{c,p}^{K-2}} \tag{30}$$

$$\frac{\partial C}{\partial w_{c,p}^{K-2}} = \sum_{i=1}^{T} \sum_{n=1}^{N(K)} 2\epsilon_n a_n^K (1-a_n^K) (\sum_{l=1}^{N(K-1)} w_{n,l}^K \frac{\partial a_l^{K-1}}{\partial w_{c,p}^{K-2}}) \tag{31}$$

$$\frac{\partial C}{\partial w_{c,p}^{K-2}} = \sum_{i=1}^{T} \sum_{n=1}^{N(K)} 2\epsilon_n a_n^K (1-a_n^K) (\sum_{l=1}^{N(K-1)} w_{n,l}^K a_l^{K-1}(1-a_l^{K-1}) \frac{\partial \sum_{m=1}^{N(K-2)} w_{l,m}^{K-1} a_m^{K-2} + b_m^{K-1}}{\partial w_{c,p}^{K-2}}) \tag{32}$$

$$\frac{\partial C}{\partial w_{c,p}^{K-2}} = \sum_{i=1}^{T} \sum_{n=1}^{N(K)} 2\epsilon_n a_n^K (1-a_n^K) (\sum_{l=1}^{N(K-1)} w_{n,l}^K a_l^{K-1}(1-a_l^{K-1}) \sum_{m=1}^{N(K-2)} w_{l,m}^{K-1} \frac{\partial a_m^{K-2}}{\partial w_{c,p}^{K-2}}) \tag{33}$$

$$\frac{\partial C}{\partial w_{c,p}^{K-2}} = \sum_{i=1}^{T} \sum_{n=1}^{N(K)} 2\epsilon_n a_n^K (1-a_n^K) \sum_{l=1}^{N(K-1)} w_{n,l}^K a_l^{K-1}(1-a_l^{K-1}) \sum_{m=1}^{N(K-2)} w_{l,m}^{K-1} a_m^{K-2}(1-a_m^{K-2}) \frac{\partial \sum_{o=1}^{N(K-3)} w_{m,o}^{K-2} a_o^{K-3} + b_m^{K-2}}{\partial w_{c,p}^{K-2}} \tag{34}$$

$$\frac{\partial C}{\partial w_{c,p}^{K-2}} = \sum_{i=1}^{T} \sum_{n=1}^{N(K)} 2\epsilon_n a_n^K (1-a_n^K) \sum_{l=1}^{N(K-1)} w_{n,l}^K a_l^{K-1}(1-a_l^{K-1}) w_{l,c}^{K-1} a_c^{K-2}(1-a_c^{K-2}) a_p^{K-3} \tag{35}$$

So the derivative of the layer $K-2$ is:

$$\frac{\partial C}{\partial w_{c,p}^{K-2}} = \sum_{i=1}^{T} \sum_{n=1}^{N(K)} 2\epsilon_n a_n^K (1-a_n^K) \sum_{l=1}^{N(K-1)} w_{n,l}^K a_l^{K-1}(1-a_l^{K-1}) w_{l,c}^{K-1} a_c^{K-2}(1-a_c^{K-2}) a_p^{K-3} \tag{36}$$

switching the summatories with n and l:

$$\frac{\partial C}{\partial w_{c,p}^{K-2}} = \sum_{i=1}^{T} \sum_{l=1}^{N(K-1)} \sum_{n=1}^{N(K)} 2\epsilon_n a_n^K (1-a_n^K) w_{n,l}^K a_l^{K-1} (1-a_l^{K-1}) w_{l,c}^{K-1} a_c^{K-2} (1-a_c^{K-2}) a_p^{K-3} \tag{37}$$

you can see that the term

$$\sum_{n=1}^{N(K)} 2\epsilon_n a_n^K (1-a_n^K) w_{n,l}^K a_l^{K-1} (1-a_l^{K-1}) \tag{38}$$

is the propagation of the error to the l element of the $K-1$ layer, e_l^{K-1}. Then you can write:

$$\frac{\partial C}{\partial w_{c,p}^{K-2}} = \sum_{i=1}^{T} \sum_{l=1}^{N(K-1)} e_l^{K-1} w_{l,c}^{K-1} a_c^{K-2} (1-a_c^{K-2}) a_p^{K-3} \tag{39}$$

and denote

$$e_c^{K-2} = \sum_{l=1}^{N(K-1)} e_l^{K-1} w_{l,c}^{K-1} a_c^{K-2} (1-a_c^{K-2}) \tag{40}$$

as the propagation of the error to the c node of the $K-2$ layer so that:

$$\frac{\partial C}{\partial w_{c,p}^{K-2}} = e_c^{K-2} a_p^{K-3} \tag{41}$$