List 03 - Result

LIST III - MLP

Construct an MLP (2, 2, 2) according to the architecture shown below, to classify the fruits **Apple** and **Orange**. The fruits will be identified by two features:

- **Size** (0.5 = Apple, 0.8 = Orange)
- **Texture** (smooth = 0.2 = Apple, rough = 0.6 = Orange).

Thus, as initial samples, consider:

- Apple = $[0.5; 0.2]$
- Orange = $[0.8; 0.6]$.

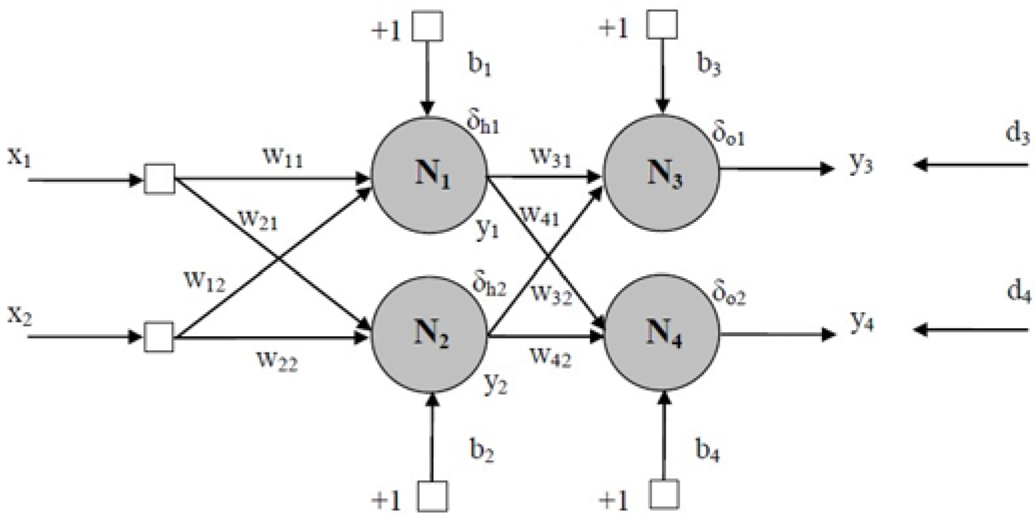


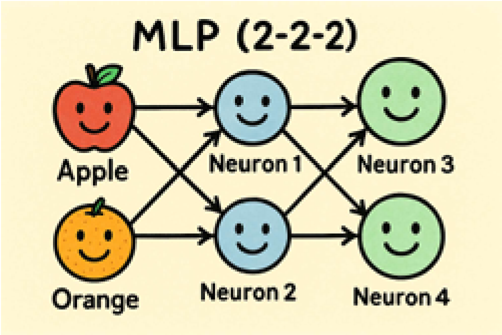
Figura Rede neural MLP(2, 2, 2).

First Part

1.1 Use the backpropagation algorithm to perform the **first training epoch** of the Neural Network.
1.2 Use a **learning rate (eta)** equal to 0.5.
1.3 Use a **momentum rate (alpha)** equal to 0.1.
1.4 Use the **sigmoid activation function** in all layers.
1.5 Initialize all **weights and biases** with 0.1.
1.6 Show all calculations for the first training epoch.
1.7 After completing all the calculations, implement a **Python routine** to replicate the computations and plot an **RMSE curve** (you define the number of epochs and error size).

Second Part

1.8 Redo the First Part using the **ReLU activation function** in the hidden layer.



Result

Use the backpropagation algorithm to perform the **first training epoch** of the Neural Network.

Architecture and data:

- MLP net: $[2, 2, 2]$
- Samples:

$$x^{(a)} = [0.5, 0.2], \quad t^{(a)} = [1, 0] \quad \textbf{(Apple)} \\ x^{(o)} = [0.8, 0.6], \quad t^{(0)} = [0, 1] \quad \textbf{(Orange)} \tag{01}$$

- Hyperparameters:

$$\eta = 0,5 \textbf{(Learning rate)}, \alpha = 0.1 \textbf{(Momentum)} \tag{02}$$

- Activation function: Sigmoidal

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{03}$$

- bias: $b = 0.1$ (explicited in iten 1.4, different to the figure)

Apple

Forward for Apple sample

Hidden State:

$$v_j^{(a_1)} = w_{j1}^{(a_1)} x_1 + w_{j2}^{(a_1)} x_2 + b_j^{(a_1)} \\ v_j^{(a_1)} = 0.1 \cdot 0.5 + 0.1 \cdot 0.2 + 0.1 = 0.17 \\ h_j^{(a_1)} = \sigma(v_j^{(a_1)}) = \sigma(0.17) = 0.542398 \tag{04}$$

Is important to notice that $h_1 = h_2 = h_j$ due to the wieghts which are all equal to 0.1. Thus, the the hidden state output are all equal.

Output:

$$\begin{aligned} v_k^{(a_2)} &= w_{k1}^{(a_2)} h_1 + w_{k2}^{(a_2)} h_2 + b_k^{(a_2)} \\ v_k^{(a_2)} &= 0.1 \cdot 0.542398 + 0.1 \cdot 0.542398 + 0.1 = 0.208480 \\ \hat{y_k}^{(a_2)} &= \sigma(v_k^{(a_2)}) = \sigma(0.208480) = 0.551932 \end{aligned} \tag{05}$$

Local Gradient

Mean square error (MSE):

$$E = \frac{1}{2} \sum_k (t_k - \hat{y}_k)^2 \tag{6}$$

Output layer

Local gradient: δ_k

$$\begin{aligned} \delta_k &= \frac{\partial E}{\partial v_k^{(a_2)}}, \\ \delta_k &= \frac{\partial E}{\partial \hat{y}_k} \cdot \frac{\partial \hat{y}_k}{\partial v_k^{(a_2)}} \end{aligned} \tag{07}$$

Therefore, the local gradient of error in relation to neuron k output.

Thus, $\hat{y}_k = \sigma(v_k^2)$ we have:

$$\begin{aligned} \hat{y}_k = \sigma(v_k^2) &= \frac{1}{1 + e^{-(v_k^{(2)})}}, \\ \frac{\partial \hat{y}_k}{\partial v_k^{(2)}} = \frac{0 - 1(-1(e^{-(v_k^{(2)})})}{(1 + e^{-(v_k^{(2)})})^2} = \frac{e^{-(v_k^{(2)})}}{(1 + e^{-(v_k^{(2)})})^2} = \frac{1}{1 + e^{-(v_k^{(2)})}} &\cdot \frac{e^{-(v_k^{(2)})}}{1 + e^{-(v_k^{(2)})}}, \\ \frac{\partial \hat{y}_k}{\partial v_k^{(2)}} &= \hat{y}_k(1 - \hat{y}_k). \end{aligned} \tag{07}$$

$$\begin{aligned} \frac{\partial E}{\partial \hat{y}_k} = \frac{\partial}{\partial \hat{y}_k}[\frac{1}{2} \sum_k (t_k - \hat{y}_k)^2] &= -1(t_k - \hat{y}_k), \\ \frac{\partial E}{\partial \hat{y}_k} &= (\hat{y}_k - t_k) \end{aligned} \tag{08}$$

So:

$$\delta_k = \frac{\partial E}{\partial \hat{y}_k} \cdot \frac{\partial \hat{y}_k}{\partial v_k^{(a_2)}} = (\hat{y}_k - t_k)\hat{y}_k(1 - \hat{y}_k) \tag{09}$$

Hidden layer

In the Hidden layer, the neurons does not compare directly with the target t which influence the error indirectly feeding the other neurons in the output layer. Thus, in the backpropagation sum up all the contributions h_j.

Local gradient: δ_j

$$\begin{gathered}\delta_j = \frac{\partial E}{\partial v_j^{(a_1)}}, \\ \frac{\partial E}{\partial v_j^{(a_1)}} = \frac{\partial E}{\partial \hat{y}_k} \cdot \frac{\partial \hat{y}_k}{\partial v_k^{(a_2)}} \cdot \frac{\partial v_k^{(a_2)}}{\partial h_j} \cdot \frac{\partial h_j}{\partial v_j^{(a_1)}}\end{gathered} \tag{10}$$

where:

$$\delta_k = \frac{\partial E}{\partial \hat{y}_k} \cdot \frac{\partial \hat{y}_k}{\partial v_k^{(a_2)}} = (\hat{y}_k - t_k)\hat{y}_k(1 - \hat{y}_k) \tag{09}$$

$$\frac{\partial v_k^{(a_2)}}{\partial h_j} = \frac{\partial}{\partial h_j}[\sum_k w_k h_j + b_k] = w_k \tag{11}$$

$$\begin{gathered}\frac{\partial h_j}{\partial v_j^{(a_1)}} = \frac{\partial}{\partial v_j^{(a_1)}}[\sigma(v_j)] = \frac{0 - 1(-1(e^{-(v_j^{(2)})}))}{(1 + e^{-(v_j^{(2)})})^2} = \frac{e^{-(v_j^{(2)})}}{(1 + e^{-(v_j^{(2)})})^2} = \frac{1}{1 + e^{-(v_j^{(2)})}} \cdot \frac{e^{-(v_j^{(2)})}}{1 + e^{-(v_j^{(2)})}}, \\ \frac{\partial h_j}{\partial v_j^{(a_1)}} = h_j(1 - h_j)\end{gathered} \tag{12}$$

So:

$$\begin{gathered}\delta_j = \frac{\partial E}{\partial v_j^{(a_1)}}, \\ \frac{\partial E}{\partial v_j^{(a_1)}} = \frac{\partial E}{\partial \hat{y}_k} \cdot \frac{\partial \hat{y}_k}{\partial v_k^{(a_2)}} \cdot \frac{\partial v_k^{(a_2)}}{\partial h_j} \cdot \frac{\partial h_j}{\partial v_j^{(a_1)}} \\ \frac{\partial E}{\partial v_j^{(a_1)}} = (\hat{y}_k - t_k)\hat{y}_k(1 - \hat{y}_k) w_k h_j(1 - h_j)\end{gathered} \tag{13}$$

Each hidden neuron j affects multiple neurons in the next layer - delivering the hidden state h_j to all those neurons in the next layer.

Thus:

$$\boxed{\delta_j = h_j(1 - h_j) \sum_k (y_k - t_k)\, y_k(1 - y_k)\, w_{kj}^{(2)}} \tag{14}$$

Apple - Delta Calculation

Output layer:

$$\begin{aligned}\delta_1^{(2)} &= (\hat{y}_k - t_k)\hat{y}_k(1 - \hat{y}_k) = (0.551932 - 1)0.551932(1 - 0.551932) = -0.110809 \\ \delta_2^{(2)} &= (\hat{y}_k - t_k)\hat{y}_k(1 - \hat{y}_k) = (0.551932 - 0)0.551932(1 - 0.551932) = 0.136494\end{aligned} \tag{10}$$

Hidden layer:

Local gradient: δ_j

$$\delta_k = \frac{\partial E}{\partial v_k^{(a_2)}},$$
$$\delta_j = h_j(1-h_j)\sum_k (y_k - t_k) y_k (1-y_k) w_{kj}^{(2)} \tag{15}$$

$$\delta_1^{(1)} = \delta_2^{(1)} = h_j(1-h_j)\sum_k (y_k - t_k) y_k (1-y_k) w_{kj}^{(2)} = (0.542398) \cdot (1 - 0.542398) \cdot (0.1 \cdot -0.110$$

$h_1 = h_2 = h_j$

Updating weights

given a generic weight w with gradient $\partial E / \partial w$, the updating tax Δw_t will be:

$$\boxed{\Delta w_t \;=\; -\eta \frac{\partial E}{\partial w} \;+\; \alpha\, \Delta w_{t-1}}, \qquad \boxed{w \;\leftarrow\; w + \Delta w_t}. \tag{17}$$

Where z_i is the input provided by the previous neuron.

Thus:

$$\frac{\partial E}{\partial w_{ji}} = \frac{\partial E}{\partial v_j} \cdot \frac{\partial v_j}{\partial w_{ji}} = \delta_j \cdot z_i. \tag{18}$$

Bias b, will be the same way.

Output layer:

$k = 1$:

$$\begin{aligned} \Delta w_{k1}^{(2)} &= -\eta \frac{\partial E}{\partial w} + \alpha \Delta w_{k1-1} = -\eta \cdot \delta_1^{(2)} \cdot h_j + \alpha \Delta w_{k1-1}, \\ \Delta w_{k1}^{(2)} &= -0.5 \cdot -0.110809 \cdot 0.542398 + 0.1 \cdot 0 = 0.030051 \\ \Delta b_1^{(2)} &= -0.5 \cdot (-0.110809) = +0.055404. \end{aligned} \tag{19}$$

$k = 2$:

$$\begin{aligned} \Delta w_{k2}^{(2)} &= -\eta \frac{\partial E}{\partial w} + \alpha \Delta w_{k2-1} = -\eta \cdot \delta_2^{(2)} \cdot h_j + \alpha \Delta w_{k2-1} \\ \Delta w_{k2}^{(2)} &= -0.5 \cdot (+0.136494) \cdot 0.542398 = -0.037017 \\ \Delta b_2^{(2)} &= -0.5 \cdot (+0.136494) = -0.068247. \end{aligned} \tag{20}$$

Hidden layer:

$j = 1$:

$$
\begin{aligned}
\Delta w_{j1}^{(1)} &= -\eta \frac{\partial E}{\partial w} + \alpha \Delta w_{j1-1} = -\eta \cdot \delta_j^{(1)} \cdot x_i + \alpha \Delta w_{j1-1}, \\
\Delta w_{j1}^{(1)} &= -0.5 \cdot (0.000638) \cdot 0.5 = -0.000159 \\
\Delta b_j^{(1)} &= -0.5 \cdot (0.000638) = -0.000319.
\end{aligned} \tag{21}
$$

$j = 2$:

$$
\begin{aligned}
\Delta w_{j2}^{(1)} &= -\eta \frac{\partial E}{\partial w} + \alpha \Delta w_{j2-1} = -\eta \cdot \delta_j^{(1)} \cdot x_i + \alpha \Delta w_{j2-1}, \\
\Delta w_{j2}^{(1)} &= -0.5 \cdot (0.000638) \cdot 0.2 = -0.000064
\end{aligned} \tag{22}
$$

Output Layer:

Weights	**Updated Values**
$w_{1j}^{(2)}$	$0.1 + 0.030051 = \mathbf{0.130051} (j = 1, 2)$
$b_1^{(2)}$	$0.1 + 0.055404 = \mathbf{0.155404}$
$w_{2j}^{(2)}$	$0.1 - 0.037017 = \mathbf{0.062983} (j = 1, 2)$
$b_2^{(2)}$	$0.1 - 0.068247 = \mathbf{0.031753}$

Hidden Layer:

Weights	**Updated Values**
$w_{j1}^{(1)}$	$0.1 + (-0.000159) = 0.099841$
$w_{j2}^{(1)}$	$0.1 + (-0.000064) = 0.099936$
$b_j^{(1)}$	$0.1 + (-0.000319) = 0.099681$

Orange

Forward for Orange sample

Hidden state

$$
\begin{aligned}
v_j^{(a_1)} &= w_{j1}^{(a_1)} x_1 + w_{j2}^{(a_1)} x_2 + b_j^{(a_1)} \\
&= 0.099841 \cdot 0.8 + 0.099936 \cdot 0.6 + 0.099681 \\
&= 0.239515 \\
h_j^{(a_1)} &= \sigma(v_j^{(a_1)}) = \sigma(0.239515) = 0.559594 \qquad (j = 1, 2)
\end{aligned} \tag{23}
$$

Output

$$
\begin{aligned}
v_1^{(a_2)} &= w_{11}^{(a_2)} h_1 + w_{12}^{(a_2)} h_2 + b_1^{(a_2)} = 0.130051(0.559594 + 0.559594) + 0.155404 = 0.300956, \\
\hat{y}_1^{(a_2)} &= \sigma(0.300956) = 0.574676, \\
v_2^{(a_2)} &= w_{21}^{(a_2)} h_1 + w_{22}^{(a_2)} h_2 + b_2^{(a_2)} = 0.062983(0.559594 + 0.559594) + 0.031753 = 0.102242, \\
\hat{y}_2^{(a_2)} &= \sigma(0.102242) = 0.525538.
\end{aligned}
\quad (24
$$

Local Gradients (MSE)

Output layer

$$\delta_k^{(2)} = (\hat{y}_k - t_k)\, \hat{y}_k (1 - \hat{y}_k).$$

So, for Orange ($t = [0, 1]$):

$$\delta_1^{(2)} = (0.574676 - 0) \cdot 0.574676 \cdot 0.425324 = \boxed{+0.140464}, \qquad \delta_2^{(2)} = (0.525538 - 1) \cdot 0.525538$$

Hidden layer (sigmoid)

$$\delta_j^{(1)} = h_j (1 - h_j) \sum_k \delta_k^{(2)}\, w_{kj}^{(a_2)}.$$

Here $h_j(1 - h_j) = 0.559594 \cdot 0.440406 = \ 0.246711$, and

$$\sum_k \delta_k^{(2)}\, w_{kj}^{(a_2)} = (+0.140464) \cdot 0.130051 + (-0.118306) \cdot 0.062983 = \boxed{+0.010808},$$

Thus:

$$\delta_j^{(1)} = 0.246711 \cdot 0.010808 = \boxed{+0.002666} \qquad (j = 1, 2). \quad (26)$$

Weight Updates with Momentum

$$\boxed{\Delta w_t = -\eta \frac{\partial E}{\partial w} + \alpha\, \Delta w_{t-1}}, \qquad \frac{\partial E}{\partial w} = \delta \cdot (\text{input}).$$

Output layer ($h \to y$)

For each j (same h_j for both $j = 1, 2$):

- For $k = 1$:

$$\Delta w_{1j}^{(a_2)} = -\eta\, \delta_1^{(2)}\, h_j + \alpha\, \Delta w_{1j}^{(a_2)}(\text{prev}) = -0.5 \cdot (0.140464) \cdot 0.559594 + 0.1 \cdot (0.030051) = \boxed{-0.036296}$$

$$\Delta b_1^{(a_2)} = -\eta\, \delta_1^{(2)} + \alpha\, \Delta b_1^{(a_2)}(\text{prev}) = -0.5 \cdot 0.140464 + 0.1 \cdot 0.055404 = \boxed{-0.064692}.$$

- For $k = 2$:

$$\Delta w_{2j}^{(a_2)} = -\eta\, \delta_2^{(2)}\, h_j + \alpha\, \Delta w_{2j}^{(a_2)}(\text{prev}) = -0.5 \cdot (-0.118306) \cdot 0.559594 + 0.1 \cdot (-0.037017) = \boxed{+0.029}$$

$$\Delta b_2^{(a_2)} = -\eta\, \delta_2^{(2)} + \alpha\, \Delta b_2^{(a_2)}(\text{prev}) = -0.5 \cdot (-0.118306) + 0.1 \cdot (-0.068247) = \boxed{+0.052328}.$$

Hidden layer

Inputs are $x_1 = 0.8,\ x_2 = 0.6$.

$$\Delta w_{j1}^{(a_1)} = -\eta\,\delta_j^{(1)}\,x_1 + \alpha\,\Delta w_{j1}^{(a_1)}(\text{prev}) = -0.5 \cdot (0.002666) \cdot 0.8 + 0.1 \cdot (-0.000159) = \boxed{-0.001082}.$$

$$\Delta w_{j2}^{(a_1)} = -\eta\,\delta_j^{(1)}\,x_2 + \alpha\,\Delta w_{j2}^{(a_1)}(\text{prev}) = -0.5 \cdot (0.002666) \cdot 0.6 + 0.1 \cdot (-0.000064) = \boxed{-0.000806}.$$

$$\Delta b_j^{(a_1)} = -\eta\,\delta_j^{(1)} + \alpha\,\Delta b_j^{(a_1)}(\text{prev}) = -0.5 \cdot (0.002666) + 0.1 \cdot (-0.000319) = \boxed{-0.001365}.$$

Updated Weights (after Orange)

Output layer

Weights	New value
$w_{1j}^{(a_2)}$	$0.130051 + (-0.036296) = \mathbf{0.093755}$ (for $j = 1, 2$)
$b_1^{(a_2)}$	$0.155404 + (-0.064692) = \mathbf{0.090713}$
$w_{2j}^{(a_2)}$	$0.062983 + 0.029400 = \mathbf{0.092383}$ (for $j = 1, 2$)
$b_2^{(a_2)}$	$0.031753 + 0.052328 = \mathbf{0.084081}$

Hidden layer

Weights	New value
$w_{j1}^{(a_1)}$	$0.099841 + (-0.001082) = \mathbf{0.098758}$
$w_{j2}^{(a_1)}$	$0.099936 + (-0.000806) = \mathbf{0.099130}$
$b_j^{(a_1)}$	$0.099681 + (-0.001365) = \mathbf{0.098317}$