Aprendizagem 2021/22
Homework III – Group 020

I. Pen-and-paper

1) a.

$f = \tanh(x) = \frac{2}{1+\exp(-2x)} - 1$ $\frac{\partial \tanh(x)}{\partial x} = \frac{\partial}{\partial x}\left(\frac{2}{1+\exp(-2x)} - 1\right) = 1 - \tanh(x)^2$

Connection weights and biases:

$$W^{[1]} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \quad W^{[2]} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \quad W^{[3]} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

$$b^{[1]} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \quad b^{[2]} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad b^{[3]} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Forward propagation:

$$x^{[0]} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \quad z^{[1]} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 6 \\ 1 \\ 6 \end{bmatrix} \quad x^{[1]} = \begin{bmatrix} f(6) \\ f(1) \\ f(6) \end{bmatrix}$$

$$z^{[2]} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}\begin{bmatrix} f(6) \\ f(1) \\ f(6) \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 2f(6) + f(1) + 1 \\ 2f(6) + f(1) + 1 \end{bmatrix} \quad x^{[2]} = \begin{bmatrix} f(2f(6) + f(1) + 1) \\ f(2f(6) + f(1) + 1) \end{bmatrix}$$

$$z^{[3]} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} f(2f(6)+f(1)+1) \\ f(2f(6)+f(1)+1) \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad x^{[3]} = \begin{bmatrix} f(0) \\ f(0) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad t = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

Backward propagation:

$$E(t, x^{[3]}) = \frac{1}{2}\sum_{t=1}^{1} (x^{[3]} - t)^2 = \frac{1}{2}(x^{[3]} - t)^2$$

$$\frac{\partial E}{\partial x^{[3]}}(t, x^{[3]}) = \frac{\partial E}{\partial (x^{[3]} - t)^2} \frac{\partial (x^{[3]} - t)^2}{\partial (x^{[3]} - t)} \frac{\partial (x^{[3]} - t)}{\partial x^{[3]}} = \frac{1}{2}\left[2(x^{[3]} - t)\right] = x^{[3]} - t$$

$$\frac{\partial x^{[l]}}{\partial z^{[l]}}(z^{[l]}) = 1 - f(z^{[l]})^2$$

$$\frac{\partial z^{[l]}}{\partial W^{[l]}}(W^{[l]}, b^{[l]}, x^{[l-1]}) = x^{[l-1]}$$

$$\frac{\partial z^{[l]}}{\partial b^{[l]}}(W^{[l]}, b^{[l]}, x^{[l-1]}) = 1$$

$$\frac{\partial z^{[l]}}{\partial x^{[l-1]}}(W^{[l]}, b^{[l]}, x^{[l-1]}) = W^{[l]}$$

$$\delta^{[3]} = \frac{\partial E}{\partial x^{[3]}} \circ \frac{\partial x^{[3]}}{\partial z^{[3]}} = (x^{[3]} - z) \circ (1 - f(z^{[3]})^2) = \begin{bmatrix} 1 - f(z_1^{[3]})^2 & 0 \\ 0 & 1 - f(z_2^{[3]})^2 \end{bmatrix}\begin{bmatrix} x_1^{[3]} - z_1 \\ x_2^{[3]} - z_2 \end{bmatrix}$$

$$= \begin{bmatrix} 1 - 0^2 & 0 \\ 0 & 1 - 0^2 \end{bmatrix}\begin{bmatrix} 0 - 1 \\ 0 - (-1) \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

$$\delta^{[2]} = \left(\frac{\partial z^{[3]}}{\partial x^{[2]}}\right)^T \cdot \delta^{[3]} \circ \frac{\partial x^{[2]}}{\partial z^{[2]}} = (W^{[3]})^T \cdot \delta^{[3]} \circ (1 - f(z^{[2]})^2) =$$

$$= \begin{bmatrix} 1 - f(z_1^{[2]})^2 & 0 \\ 0 & 1 - f(z_2^{[2]})^2 \end{bmatrix}\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} -1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\delta^{[1]} = \left(\frac{\partial z^{[2]}}{\partial x^{[1]}}\right)^T \cdot \delta^{[2]} \circ \frac{\partial x^{[1]}}{\partial z^{[1]}} = (W^{[2]})^T \cdot \delta^{[2]} \circ (1 - f(z^{[1]})^2) =$$

$$= \begin{bmatrix} 1-f(z_1^{[1]})^2 & 0 & 0 \\ 0 & 1-f(z_2^{[1]})^2 & \\ 0 & 0 & 1-f(z_3^{[1]})^2 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$\boxed{\text{Updates}}$

$$\frac{\partial E}{\partial W^{[1]}} = \delta^{[1]} \cdot \frac{\partial z^{[1]}}{\partial W^{[1]}} = \delta^{[1]} (x^{[0]})^T = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$W^{[1]} = W^{[1]} - \eta \frac{\partial E}{\partial W^{[1]}} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} - 0.1 \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

$$\frac{\partial E}{\partial b^{[1]}} = \delta^{[1]} \frac{\partial z^{[1]T}}{\partial b^{[1]}} = \delta^{[1]} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$$b^{[1]} = b^{[1]} - \eta \frac{\partial E}{\partial b^{[1]}} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} - 0.1 \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

$$\frac{\partial E}{\partial W^{[2]}} = \delta^{[2]} \frac{\partial z^{[2]}}{\partial W^{[2]}} = \delta^{[2]} (x^{[1]})^T = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \begin{bmatrix} f(6) & f(1) & f(6) \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$W^{[2]} = W^{[2]} - \eta \frac{\partial E}{\partial W^{[2]}} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} - 0.1 \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

$$\frac{\partial E}{\partial b^{[2]}} = \delta^{[2]} \frac{\partial z^{[2]T}}{\partial b^{[2]}} = \delta^{[2]} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$b^{[2]} = b^{[2]} - \eta \frac{\partial E}{\partial b^{[2]}} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} - 0.1 \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$\frac{\partial E}{\partial W^{[3]}} = \delta^{[3]} \cdot \frac{\partial z^{[3]T}}{\partial W^{[3]}} = \delta^{[3]} \cdot (x^{[2]})^T = \begin{bmatrix} -1 \\ 1 \end{bmatrix} \begin{bmatrix} f(2f(6)+f(1)+1) & f(2f(6)+f(1)+1) \end{bmatrix}$$

$$= \begin{bmatrix} -f(2f(6)+f(1)+1) & -f(2f(6)+f(1)+1) \\ f(2f(6)+f(1)+1) & f(2f(6)+f(1)+1) \end{bmatrix} = \begin{bmatrix} -0.99892 & -0.99892 \\ 0.99892 & 0.99892 \end{bmatrix}$$

$$W^{[3]} = W^{[3]} - \eta \frac{\partial E}{\partial W^{[3]}} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} - 0.1 \begin{bmatrix} -0.99892 & -0.99892 \\ 0.99892 & 0.99892 \end{bmatrix} =$$

$$= \begin{bmatrix} 0.099892 & 0.099892 \\ -0.099892 & -0.099892 \end{bmatrix}$$

$$\frac{\partial E}{\partial b^{[3]}} = \delta^{[3]} \frac{\partial z^{[3]T}}{\partial b^{[3]}} = \delta^{[3]} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

$$b^{[3]} = b^{[3]} - \eta \frac{\partial E}{\partial b^{[3]}} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} - 0.1 \begin{bmatrix} -1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0.1 \\ -0.1 \end{bmatrix}$$

Aprendizagem 2021/22

Homework III – Group 020

1) b.

$$g = \text{softmax}([z_1\ z_2 \dots z_d]^T) = [x_1\ x_2 \dots x_d]^T \quad , \quad x_i = \frac{\exp(z_i)}{\sum_{k=1}^{d} \exp(z_k)}$$

$$\frac{\partial x_i}{\partial z_j} = \frac{\partial}{\partial z_j} \frac{\exp(z_i)}{\sum_{k=1}^{d} \exp(z_k)} = \begin{cases} x_i(1-x_i) & \text{se } i=j \\ -x_i x_j & \text{se } i \neq j \end{cases}$$

Connection weights and biases:

$$W^{[1]} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \qquad W^{[2]} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \qquad W^{[3]} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

$$b^{[1]} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \qquad b^{[2]} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \qquad b^{[3]} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$\boxed{\text{Forward propagation:}}$

$$x^{[0]} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \qquad z^{[1]} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 6 \\ 1 \\ 6 \end{bmatrix} \qquad x^{[1]} = \begin{bmatrix} f(6) \\ f(1) \\ f(6) \end{bmatrix}$$

$$z^{[2]} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} f(6) \\ f(1) \\ f(6) \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 2f(6) + f(1) + 1 \\ 2f(6) + f(1) + 1 \end{bmatrix} \qquad x^{[2]} = \begin{bmatrix} f(2f(6) + f(1) + 1) \\ f(2f(6) + f(1) + 1) \end{bmatrix}$$

$$z^{[3]} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} f(2f(6)+f(1)+1) \\ f(2f(6)+f(1)+1) \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad x^{[3]} = \begin{bmatrix} g(0) \\ g(0) \end{bmatrix} = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix} \qquad t = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$\boxed{\text{Backward propagation:}}$

$$E(t, x^{[3]}) = -\sum_{i=1}^{d} t_i \log(x_i^{[3]})$$

$$\delta_i^{[3]} = \frac{\partial E(t, x^{[3]})}{\partial z_i} = \frac{\partial}{\partial z_i} \left(-\sum_{k=1}^{d} t_k \log x_k^{[3]} \right) = -\sum_{k=1}^{d} t_k \frac{\partial}{\partial z_i} \log x_k^{[3]} = -\sum_{k=1}^{d} t_k \frac{1}{x_k^{[3]}} \frac{\partial x_k^{[3]}}{\partial z_i}$$

$$= -\sum_{k=i} t_k \frac{1}{x_k^{[3]}} \frac{\partial x_k^{[3]}}{\partial z_i} - \sum_{k \neq i} t_k \frac{1}{x_k^{[3]}} \frac{\partial x_k^{[3]}}{\partial z_i} = -\sum_{k=i} t_k \frac{1}{x_k^{[3]}} \left(x_i^{[3]} (1 - x_i^{[3]}) \right) - \sum_{k \neq i} t_k \frac{1}{x_k^{[3]}} \left(-x_k^{[3]} x_i^{[3]} \right)$$

$$= -t_i \frac{1}{x_i^{[3]}} \left(x_i^{[3]} (1 - x_i^{[3]}) \right) - \sum_{k \neq i} t_k \frac{1}{x_k^{[3]}} \left(-x_k^{[3]} x_i^{[3]} \right) = -t_i (1 - x_i^{[3]}) + \sum_{k \neq i} t_k x_i^{[3]}$$

$$= -t_i + t_i x_i^{[3]} + \sum_{k \neq i} t_k x_i^{[3]} = -t_i + x_i^{[3]} \left(t_i + \sum_{k \neq i} t_k \right) = -t_i + x_i^{[3]} \left(\sum_{k=1}^{d} t_k \right) = -t_i + x_i^{[3]}$$

$$= x_i^{[3]} - t_i$$

$$\frac{\partial x_i^{[l]}}{\partial z_j^{[l]}}(z^{[l]}) = \begin{cases} x_i(1-x_i) & \text{se } i=j \\ -x_i x_j & \text{se } i \neq j \end{cases}$$

$$\frac{\partial z^{[l]}}{\partial W^{[l]}} (W^{[l]}, b^{[l]}, x^{[l-1]}) = x^{[l-1]}$$

$$\frac{\partial z^{[l]}}{\partial b^{[l]}} (W^{[l]}, b^{[l]}, x^{[l-1]}) = 1$$

$$\frac{\partial z^{[l]}}{\partial x^{[l-1]}} (W^{[l]}, b^{[l]}, x^{[l-1]}) = W^{[l]}$$

$$\delta^{[3]} = \begin{bmatrix} \delta_1^{[3]} \\ \delta_2^{[3]} \end{bmatrix} = \begin{bmatrix} \frac{\partial E(t, x^{[3]})}{\partial z_1} \\ \frac{\partial E(t, x^{[3]})}{\partial z_2} \end{bmatrix} = \begin{bmatrix} x_1^{[3]} - t_1 \\ x_2^{[3]} - t_2 \end{bmatrix} = \begin{bmatrix} g(0) - 1 \\ g(0) \end{bmatrix}$$

$$\delta^{[2]} = \frac{\partial z^{[3]T}}{\partial x^{[2]}} \cdot \delta^{[3]} \circ \frac{\partial x^{[2]}}{\partial z^{[2]}} = \begin{bmatrix} x_1^{[2]}(1 - x_1^{[2]}) & -x_1^{[2]} x_2^{[2]} \\ -x_2^{[2]} x_1^{[2]} & x_2^{[2]}(1 - x_2^{[2]}) \end{bmatrix} (W^{[3]})^T \delta^{[3]}$$

$$= \begin{bmatrix} x_1^{[2]}(1 - x_1^{[2]}) & -x_1^{[2]} x_2^{[2]} \\ -x_2^{[2]} x_1^{[2]} & x_2^{[2]}(1 - x_2^{[2]}) \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} g(0) - 1 \\ g(0) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\delta^{[1]} = \frac{\partial z^{[2]T}}{\partial x^{[1]}} \cdot \delta^{[2]} \circ \frac{\partial x^{[1]}}{\partial z^{[1]}} = \begin{bmatrix} x_1^{[1]}(1 - x_1^{[1]}) & -x_1^{[1]} x_2^{[1]} & -x_1^{[1]} x_3^{[1]} \\ -x_2^{[1]} x_1^{[1]} & x_2^{[1]}(1 - x_2^{[1]}) & -x_2^{[1]} x_3^{[1]} \\ -x_3^{[1]} x_1^{[1]} & -x_3^{[1]} x_2^{[1]} & x_3^{[1]}(1 - x_3^{[1]}) \end{bmatrix} (W^{[2]})^T \delta^{[2]}$$

$$= \begin{bmatrix} x_1^{[1]}(1 - x_1^{[1]}) & -x_1^{[1]} x_2^{[1]} & -x_1^{[1]} x_3^{[1]} \\ -x_2^{[1]} x_1^{[1]} & x_2^{[1]}(1 - x_2^{[1]}) & -x_2^{[1]} x_3^{[1]} \\ -x_3^{[1]} x_1^{[1]} & -x_3^{[1]} x_2^{[1]} & x_3^{[1]}(1 - x_3^{[1]}) \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$\boxed{\text{Updates}}$

$$\frac{\partial E}{\partial W^{[1]}} = \delta^{[1]} \cdot \frac{\partial z^{[1]}}{\partial W^{[1]}} = \delta^{[1]} (x^{[0]})^T = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$W^{[1]} = W^{[1]} - \eta \frac{\partial E}{\partial W^{[1]}} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} - 0.1 \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

$$\frac{\partial E}{\partial b^{[1]}} = \delta^{[1]} \frac{\partial z^{[1]T}}{\partial b^{[1]}} = \delta^{[1]} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$$b^{[1]} = b^{[1]} - \eta \frac{\partial E}{\partial b^{[1]}} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} - 0.1 \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

$$\frac{\partial E}{\partial W^{[2]}} = \delta^{[2]} \frac{\partial z^{[2]}}{\partial W^{[2]}} = \delta^{[2]} (x^{[1]})^T = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \begin{bmatrix} f(6) & f(1) & f(6) \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$W^{[2]} = W^{[2]} - \eta \frac{\partial E}{\partial W^{[2]}} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} - 0.1 \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

$$\frac{\partial E}{\partial b^{[2]}} = \delta^{[2]} \frac{\partial z^{[2]T}}{\partial b^{[2]}} = \delta^{[2]} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$b^{[2]} = b^{[2]} - \eta \frac{\partial E}{\partial b^{[2]}} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} - 0.1 \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$\frac{\partial E}{\partial W^{[3]}} = \delta^{[3]} \cdot \frac{\partial z^{[3]T}}{\partial W^{[3]}} = \delta^{[3]} \cdot (x^{[2]})^T = \begin{bmatrix} g(0) - 1 \\ g(0) \end{bmatrix} \begin{bmatrix} f(2f(6)+f(1)+1) & f(2f(6)+f(1)+1) \end{bmatrix}$$

$$= \begin{bmatrix} (g(0)-1) f(2f(6)+f(1)+1) & (g(0)-1) f(2f(6)+f(1)+1) \\ g(0) f(2f(6)+f(1)+1) & g(0) f(2f(6)+f(1)+1) \end{bmatrix} = \begin{bmatrix} -3.19981 & -3.19981 \\ 3.19981 & 3.1998 \end{bmatrix}$$

$$W^{[3]} = W^{[3]} - \eta \frac{\partial E}{\partial W^{[3]}} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} - 0.1 \begin{bmatrix} -3.19981 & -3.19981 \\ 3.19981 & 3.19981 \end{bmatrix} = \begin{bmatrix} 0.0319981 & 0.0319981 \\ -0.0319981 & -0.0319981 \end{bmatrix}$$

$$\frac{\partial E}{\partial b^{[3]}} = \delta^{[3]} \frac{\partial z^{[3]T}}{\partial b^{[3]}} = \delta^{[3]} = \begin{bmatrix} g(0) - 1 \\ g(0) \end{bmatrix} = \begin{bmatrix} -0.5 \\ 0.5 \end{bmatrix}$$

$$b^{[3]} = b^{[3]} - \eta \frac{\partial E}{\partial b^{[3]}} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} - 0.1 \begin{bmatrix} -0.5 \\ 0.5 \end{bmatrix} = \begin{bmatrix} 0.05 \\ -0.05 \end{bmatrix}$$

Aprendizagem 2021/22

Homework III – Group 020

II. Programming and critical analysis

2) Seguem abaixo as matrizes de confusão do MLP na ausência (esquerda) e presença (direita) de *early stopping*:

Predicted \ True	N	P
N	85	2
P	3	46

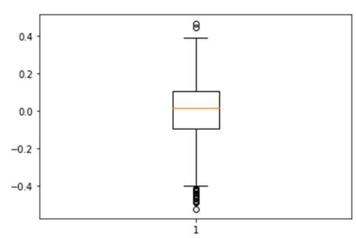
Sem *early stopping*

Predicted \ True	N	P
N	66	21
P	1	48

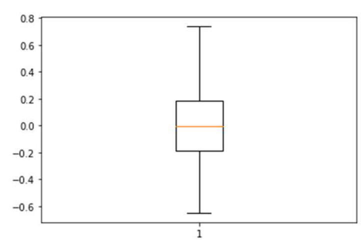
Com *early stopping*

Verifica-se que com *early stopping* a *accuracy* é inferior para a classe N e superior para a classe P, apesar de muito semelhante. No caso da classe N, esta diferença de performances pode ter acontecido devido ao facto de que com *early stopping* o modelo ainda não ter treinado o suficiente quando a paragem acontece e como não faz uso de toda a *training data* disponível pode dar origem ao fenómeno de *underfitting*, ou seja, o modelo não consegue capturar bem a relação entre as variáveis de *input* e a de *output* para o *training set* e por isso não consegue generalizar para o *testing set*. No caso da classe S pelo contrário, sem *early stopping* o modelo faz uso de todo o *training set* disponível, o que pode levar a *overfitting*, ou seja, bom treino para o *training set* mas má generalização para o *testing set*. Outro fator que pode ter dado origem à grande diferença de *accuracies* é se ter usado *early stopping* com *cross-validation*. Isto porque o *early stopping* está designado a monitorizar a generalização do erro de um modelo e parar de treinar quando esta começa a piorar, enquanto que o *cross-validation* assume que não sabemos esta generalização do erro.

3) Seguem abaixo as distribuições de resíduos usando *boxplots* na ausência (esquerda) e presença (direita) de regularização, bem como os erros médios absolutos em percentagem para o *training* e *testing set*:



Sem regularização
Erro *training set* = 22%
Erro *testing set* = 23%



Com regularização
Erro *training set* = 42%
Erro *testing set* = 43%

Verifica-se que com regularização o erro tanto no *training* como no *testing set* é maior. Isto pode acontecer devido ao facto de a regularização servir para combater a complexidade do modelo penalizando grandes coeficientes de peso e eliminando peculiaridades no *data set* que neste caso podiam não existir.

Tendo em conta que o erro do *training* e do *testing set* são semelhantes entre si em ambos os casos não estamos na presença de overfitting mas como ambos são valores elevados (especialmente com regularização) existe *underfitting*. Para minimizar o erro observado no regressor MLP existem várias estratégias como a utilização de um data set maior, aumento do tamanho e número de parâmetros (mais camadas), aumento da complexidade do modelo, reduzir a regularização ou mudar a função de ativação.

Aprendizagem 2021/22

Homework III – Group 020

III. APPENDIX

```
# -------------------- EXERCISE 2) --------------------
from scipy.io import arff
import numpy as np
import pandas as pd
from sklearn.model_selection import KFold
from sklearn.neural_network import MLPClassifier
from sklearn.metrics import confusion_matrix

data = arff.loadarff(r'/home/sara/apre/tpc-3/breast.w.arff')
df = pd.DataFrame(data[0])

data_array = df.to_numpy()

x = np.empty((0,9))
y = np.empty((0,1))

for row in data_array:
    array_sum = np.sum(row[:-1])
    array_has_nan = np.isnan(array_sum)
    if(not array_has_nan):
        x = np.append(x, [row[:-1]], axis=0)
        y = np.append(y, row[-1])

kf = KFold(n_splits=5, random_state=0, shuffle=True)

clf = MLPClassifier(activation='relu', hidden_layer_sizes=(3,2), max_iter=2000,
early_stopping=False, shuffle=False, random_state=False)

# clf = MLPClassifier(activation='relu', hidden_layer_sizes=(3,2), max_iter=2000,
# early_stopping=True, shuffle=False, random_state=False)

y_pred = np.empty((0,1))

for train_index , test_index in kf.split(x):
    x_train , x_test = x[train_index], x[test_index]
    y_train , y_test = y[train_index], y[test_index]

    clf.fit(x_train,y_train)
    y_pred = clf.predict(x_test)

tn, fp, fn, tp = confusion_matrix(y_test, y_pred).ravel()
print(confusion_matrix(y_test, y_pred))
print(tn, fp, fn, tp)


# -------------------- EXERCISE 3) --------------------
from scipy.io import arff
import numpy as np
import pandas as pd
from sklearn.model_selection import KFold
from sklearn.neural_network import MLPRegressor
from matplotlib import pyplot as plt
from sklearn.metrics import mean_absolute_percentage_error

data = arff.loadarff(r'/home/sara/apre/tpc-3/kin8nm.arff')
df = pd.DataFrame(data[0])

data_array = df.to_numpy()

x = np.empty((0,9))
y = np.empty((0,1))
```

```
for row in data_array:
    array_sum = np.sum(row[:-1])
    array_has_nan = np.isnan(array_sum)
    if(not array_has_nan):
        x = np.append(x, [row[:-1]], axis=0)
        y = np.append(y, row[-1])

kf = KFold(n_splits=5, random_state=0, shuffle=True)

clf = MLPRegressor(activation='relu', hidden_layer_sizes=(3,2), max_iter=2000, shuffle=False,
random_state=False, alpha=0)

# clf = MLPRegressor(activation='relu', hidden_layer_sizes=(3,2), max_iter=2000, shuffle=False,
# random_state=False, alpha=10)

y_pred = np.empty((0,1))
y_pred_train = np.empty((0,1))

for train_index , test_index in kf.split(x):
    x_train , x_test = x[train_index], x[test_index]
    y_train , y_test = y[train_index], y[test_index]

    clf.fit(x_train,y_train)
    y_pred = clf.predict(x_test)
    y_pred_train = clf.predict(x_train)

plt.boxplot(y_test-y_pred)
plt.show()
print(mean_absolute_percentage_error(y_train,y_pred_train))
print(mean_absolute_percentage_error(y_test,y_pred))
```

END