Aprendizagem 2023
Homework III – Group 28

Gonçalo Bárias (ist1103124) & Raquel Braunschweig (ist1102624)

Part I: Pen and Paper

1. **Consider the problem of learning a regression model from 4 bivariate observations**

   $\left\{\begin{pmatrix}0.7\\-0.3\end{pmatrix}, \begin{pmatrix}0.4\\0.5\end{pmatrix}, \begin{pmatrix}-0.2\\0.8\end{pmatrix}, \begin{pmatrix}-0.4\\0.3\end{pmatrix}\right\}$ **with targets (0.8, 0.6, 0.3, 0.3)**.

   (a) **Given the radial basis function,** $\varphi_j(x) = \exp(\frac{-\|\mathbf{x}-\mathbf{c}_j\|^2}{2})$ **that transforms the original space onto a new space characterized by the similarity of the original observastions to the following data points** $\left\{c_1 = \begin{pmatrix}0\\0\end{pmatrix}, c_2 = \begin{pmatrix}1\\-1\end{pmatrix}, c_3 = \begin{pmatrix}-1\\1\end{pmatrix}\right\}$**.**
   Learn the Ridge regression (l_2 regularization) using the closed solution with λ = 0.1.

   Gonçalo

   (b) **Compute the training RMSE for the learnt regression.**

   Gonçalo

2. **Consider a MLP classifier of three outcomes - A, B and C - characterized by the weights,**

   $$W^{[1]} = \begin{pmatrix}1111\\1121\\1111\end{pmatrix}, b^{[1]} = \begin{pmatrix}1\\1\\1\end{pmatrix}, W^{[2]} = \begin{pmatrix}141\\111\end{pmatrix}, b^{[2]} = \begin{pmatrix}1\\1\end{pmatrix}, W^{[3]} = \begin{pmatrix}11\\31\\11\end{pmatrix}, b^{[3]} = \begin{pmatrix}1\\1\\1\end{pmatrix}$$

   the activation

   $$f(x) = \frac{e^{0.5x-2} - e^{-0.5x+2}}{e^{0.5x-2} + e^{-0.5x+2}} = \tanh(0.5x - 2)$$

   for every unit, and squared error loss $\frac{1}{2}\|\mathbf{z} - \hat{\mathbf{z}}\|_2^2$**. Perform one back gradient descent update (with learning rate** $\eta = 0.1$**) for training obvervations** $x_1 = \begin{pmatrix}1\\1\\1\\1\end{pmatrix}$ **and** $x_2 = \begin{pmatrix}1\\0\\0\\-1\end{pmatrix}$ **with targets B and A, respectively.**

   To help with the calculations and as suggested in the prompt, we used NumPy to facilitate the calculus of this question. You can find the code in the Jupyter notebook `hw3.ipynb`.

   To begin, we initiate the process with the **forward pass**. Below are the key equations for this step:

   $$\mathbf{z}^{[n]} = \mathbf{W}^{[n]} \cdot \mathbf{X}^{n-1} + \mathbf{b}^{[n]} \tag{1}$$

   $$\mathbf{x}^{[n]} = f(\mathbf{z}^{[n]}) \tag{2}$$

We will commence with x_1:

$$z_1^{[1]} = W^{[1]} \cdot X_1^0 + b^{[1]} = \begin{pmatrix} 1111 \\ 1121 \\ 1111 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 5 \\ 6 \\ 5 \end{pmatrix}$$

$$x_1^{[1]} = f(z_1^{[1]}) \approx \begin{pmatrix} 0.46212 \\ 0.76159 \\ 0.46212 \end{pmatrix}$$

$$z_1^{[2]} = W^{[2]} \cdot X_1^1 + b^{[2]} = \begin{pmatrix} 141 \\ 111 \end{pmatrix} \cdot \begin{pmatrix} 0.46212 \\ 0.76159 \\ 0.46212 \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \end{pmatrix} \approx \begin{pmatrix} 4.97061 \\ 2.68583 \end{pmatrix}$$

$$x_1^{[2]} = f(z_1^{[2]}) \approx \begin{pmatrix} 0.45048 \\ -0.57642 \end{pmatrix}$$

$$z_1^{[3]} = W^{[3]} \cdot X_1^2 + b^{[3]} = \begin{pmatrix} 11 \\ 31 \\ 11 \end{pmatrix} \cdot \begin{pmatrix} 0.45048 \\ -0.57642 \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \approx \begin{pmatrix} 0.87406 \\ 1.77503 \\ 0.87406 \end{pmatrix}$$

$$x_1^{[3]} = f(z_1^{[3]}) \approx \begin{pmatrix} -0.9159 \\ -0.80494 \\ -0.9159 \end{pmatrix}$$

Next in line is x_2:

$$z_2^{[1]} = W^{[1]} \cdot X_2^0 + b^{[1]} = \begin{pmatrix} 1111 \\ 1121 \\ 1111 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 0 \\ 0 \\ -1 \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

$$x_2^{[1]} = f(z_2^{[1]}) \approx \begin{pmatrix} -0.90515 \\ -0.90515 \\ -0.90515 \end{pmatrix}$$

$$z_2^{[2]} = W^{[2]} \cdot X_2^1 + b^{[2]} = \begin{pmatrix} 141 \\ 111 \end{pmatrix} \cdot \begin{pmatrix} -0.90515 \\ -0.90515 \\ -0.90515 \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \end{pmatrix} \approx \begin{pmatrix} -4.43089 \\ -1.71544 \end{pmatrix}$$

$$x_2^{[2]} = f(z_2^{[2]}) \approx \begin{pmatrix} -0.99956 \\ -0.99343 \end{pmatrix}$$

$$z_2^{[3]} = W^{[3]} \cdot X_2^2 + b^{[3]} = \begin{pmatrix} 11 \\ 31 \\ 11 \end{pmatrix} \cdot \begin{pmatrix} -0.99956 \\ -0.99343 \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \approx \begin{pmatrix} -0.993 \\ -2.99212 \\ -0.993 \end{pmatrix}$$

$$x_2^{[3]} = f(z_2^{[3]}) \approx \begin{pmatrix} -0.98652 \\ -0.99816 \\ -0.98652 \end{pmatrix}$$

Let's initiate the **backpropagation** process. To do this, we first need to take into account the loss function, which, as per the prompt, is defined as follows:

$$\mathcal{L} = \frac{1}{2}\|\mathbf{z} - \hat{\mathbf{z}}\|_2^2 = \frac{1}{2}(\mathbf{t} - \mathbf{o})^2 \tag{3}$$

The next step involves computing its derivative:

$$\frac{d\mathcal{L}}{dW^{[p]}} = \sum_{i=1}^{2} \left(\frac{d\mathcal{L}}{dx_i^{[p]}} \circ \frac{dx_i^{[p]}}{dz_i^{[p]}} \cdot \frac{dz_i^{[p]}}{dW^{[p]}}\right) \tag{4}$$

This can be broken down into:

$$\frac{d\mathcal{L}}{dx_i^{[p]}} \circ \frac{dx_i^{[p]}}{dz_i^{[p]}} = \delta_i^{[p]}$$

$$\frac{dz_i^{[p]}}{dW^{[p]}} = (x_i^{[p-1]})^T$$

After computing δ, we can split it into two separate cases: if it is in the last layer (5), or in non-external layers (6):

$$\delta_i^{[p]} = (x_i^{[p]} - z_i) \circ sech^2(0.5 * z_i^{[p]} - 2) * 0.5 \tag{5}$$

$$\delta_i^{[p]} = (W^{[p+1]^T} \cdot \delta_i^{[p+1]}) \circ sech^2(0.5 * z_i^{[p]} - 2) * 0.5 \tag{6}$$

Therefore, the derivitave of the loss fuction is given by:

$$\frac{d\mathcal{L}}{dW^{[p]}} = \sum_{i=1}^{2} (\delta_i^{[p]} \cdot (x_i^{[p-1]})^T) \tag{7}$$

Since, the $tanh$ function variates from -1 to 1 and since the targets for x_1 and x_2 are B and A, respectively, we can conclude that:

$$z_1 = \begin{pmatrix} -1 \\ 1 \\ -1 \end{pmatrix}, z_2 = \begin{pmatrix} 1 \\ -1 \\ -1 \end{pmatrix}$$

Now, we have all we need to compute the values for the deltas, δ, by replacing the equation on 5 for $p = 3$, and 6 for any $p \neq 3$:

$$\begin{aligned}
\delta_1^{[3]} &= (x_1^{[3]} - z_1) \circ sech^2(0.5 * z_1^{[3]} - 2) * 0.5 = \\
&= (\begin{pmatrix} -0.9159 \\ -0.80494 \\ -0.9159 \end{pmatrix} - \begin{pmatrix} -1 \\ 1 \\ -1 \end{pmatrix}) \circ sech^2(0.5 * \begin{pmatrix} 0.87406 \\ 1.77503 \\ 0.87406 \end{pmatrix} * 0.5) = \\
&= \begin{pmatrix} 0.00678 \\ -0.31773 \\ 0.00678 \end{pmatrix}
\end{aligned}$$

$$\begin{aligned}
\delta_2^{[3]} &= (x_2^{[3]} - z_2) \circ sech^2(0.5 * z_2^{[3]} - 2) * 0.5 = \\
&= (\begin{pmatrix} -0.98652 \\ -0.99816 \\ -0.98652 \end{pmatrix} - \begin{pmatrix} 1 \\ -1 \\ -1 \end{pmatrix}) \circ sech^2(0.5 * \begin{pmatrix} -0.993 \\ -2.99212 \\ -0.993 \end{pmatrix} * 0.5) = \\
&= \begin{pmatrix} -0.0266 \\ 0 \\ 0.00018 \end{pmatrix}
\end{aligned}$$

$$\begin{aligned}
\delta_1^{[2]} &= (W^{[3]^T} \cdot \delta_1^{[3]}) \circ sech^2(0.5 * z_1^{[2]} - 2) * 0.5 = \\
&= (\begin{pmatrix} 11 \\ 31 \\ 11 \end{pmatrix}^T \cdot \begin{pmatrix} 0.00678 \\ -0.31773 \\ 0.00678 \end{pmatrix}) \circ sech^2(0.5 * \begin{pmatrix} 4.97061 \\ 2.68583 \end{pmatrix} - 2) * 0.5 = \\
&= \begin{pmatrix} -0.37448 \\ -0.10156 \end{pmatrix}
\end{aligned}$$

$$\delta_2^{[2]} = (W^{[3]^T} \cdot \delta_2^{[3]}) \circ sech^2(0.5 * z_2^{[2]} - 2) * 0.5 =$$
$$= (\begin{pmatrix} 11 \\ 31 \\ 11 \end{pmatrix}^T \cdot \begin{pmatrix} -0.0266 \\ 0 \\ -0.00018 \end{pmatrix}) \circ sech^2(0.5 * \begin{pmatrix} 4.97061 \\ 2.68583 \end{pmatrix} - 2) * 0.5 =$$
$$= \begin{pmatrix} -1 * 10^{-5} \\ -1.7 * 10^{-4} \end{pmatrix}$$

$$\delta_1^{[1]} = (W^{[2]^T} \cdot \delta_1^{[2]}) \circ sech^2(0.5 * z_1^{[1]} - 2) * 0.5 =$$
$$= (\begin{pmatrix} 141 \\ 111 \end{pmatrix}^T \cdot \begin{pmatrix} -0.37448 \\ -0.10156 \end{pmatrix}) \circ sech^2(0.5 * \begin{pmatrix} 5 \\ 6 \\ 5 \end{pmatrix} - 2) * 0.5 =$$
$$= \begin{pmatrix} -0.18719 \\ -0.33587 \\ -0.18719 \end{pmatrix}$$

$$\delta_2^{[1]} = (W^{[2]^T} \cdot \delta_2^{[2]}) \circ sech^2(0.5 * z_2^{[1]} - 2) * 0.5 =$$
$$= (\begin{pmatrix} 141 \\ 111 \end{pmatrix}^T \cdot \begin{pmatrix} -1 * 10^{-5} \\ -1.7 * 10^{-4} \end{pmatrix}) \circ sech^2(0.5 * \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} - 2) * 0.5 =$$
$$= \begin{pmatrix} -2 * 10^{-5} \\ -2 * 10^{-5} \\ -2 * 10^{-5} \end{pmatrix}$$

Next, we will compute the derivatives of the loss function, employing the equation in 7:

$$\frac{d\mathcal{L}}{dW^{[3]}} = \sum_{i=1}^{2} (\delta_i^{[3]} \cdot (x_i^{[2]})^T) =$$
$$= (\delta_1^{[3]} \cdot (x_1^{[2]})^T) + (\delta_2^{[3]} \cdot (x_2^{[2]})^T) =$$
$$= (\begin{pmatrix} 0.00678 \\ -0.31773 \\ 0.00678 \end{pmatrix} \cdot \begin{pmatrix} 0.45048 \\ -0.57642 \end{pmatrix}^T) + (\begin{pmatrix} -0.0266 \\ 0 \\ 0.00018 \end{pmatrix} \cdot \begin{pmatrix} -0.99956 \\ -0.99343 \end{pmatrix}^T) =$$
$$= \begin{pmatrix} 0.02964 & 0.02252 \\ -0.14314 & 0.18315 \\ 0.00287 & -0.00408 \end{pmatrix}$$

$$\frac{d\mathcal{L}}{dW^{[2]}} = \sum_{i=1}^{2}(\delta_i^{[2]} \cdot (x_i^{[1]})^T) =$$
$$= (\delta_1^{[2]} \cdot (x_1^{[1]})^T) + (\delta_2^{[2]} \cdot (x_2^{[1]})^T) =$$
$$= (\begin{pmatrix} -0.37448 \\ 0.10156 \end{pmatrix} \cdot \begin{pmatrix} 0.46212 \\ 0.76159 \\ 0.46212 \end{pmatrix}^T) + (\begin{pmatrix} -1 * 10^{-5} \\ -1.7 * 10^{-4} \end{pmatrix} \cdot \begin{pmatrix} -0.90515 \\ -0.90515 \\ -0.90515 \end{pmatrix}^T) =$$
$$= \begin{pmatrix} -0.17304 & -0.28519 & -0.17304 \\ -0.04678 & -0.07719 & -0.04678 \end{pmatrix}$$

$$\frac{d\mathcal{L}}{dW^{[1]}} = \sum_{i=1}^{2}(\delta_i^{[1]} \cdot (x_i^{[0]})^T) =$$
$$= (\delta_1^{[1]} \cdot (x_1^{[0]})^T) + (\delta_2^{[1]} \cdot (x_2^{[0]})^T) =$$
$$= (\begin{pmatrix} -0.18719 \\ -0.33587 \\ -0.18719 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}^T) + (\begin{pmatrix} -2 * 10^{-5} \\ -2 * 10^{-5} \\ -2 * 10^{-5} \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 0 \\ 0 \\ -1 \end{pmatrix}^T) =$$
$$= \begin{pmatrix} -0.18721 & -0.18719 & -0.18719 & -0.18717 \\ -0.33589 & -0.33587 & -0.33587 & -0.33585 \\ -0.18721 & -0.18719 & -0.18719 & -0.18717 \end{pmatrix}$$

The final step is to calculate the final weights and bias, which shall be given by the following equations:

$$W_{new}^{[i]} = W_{old}^{[i]} - \eta \nabla \mathcal{L} \tag{8}$$

$$b_{new}^{[i]} = b_{old}^{[i]} - \eta \sum_{k=1}^{2}(\delta_k^i) \tag{9}$$

Considering $\eta = 0.1$ we have all the values we need to start computing the results.

By substituting the values into the formula 8, we obtain the following weights:

$$W_{new}^{[1]} = \begin{pmatrix} 1.01872 & 1.01872 & 1.01872 & 1.01872 \\ 1.03359 & 1.03359 & 2.03359 & 1.03359 \\ 1.01872 & 1.01872 & 1.01872 & 1.01872 \end{pmatrix}$$
$$W_{new}^{[2]} = \begin{pmatrix} 1.0173 & 4.02852 & 1.0173 \\ 1.00468 & 1.00772 & 1.00468 \end{pmatrix}$$
$$W_{new}^{[3]} = \begin{pmatrix} 0.99704 & 0.99775 \\ 3.01431 & 0.98169 \\ 0.99971 & 1.00041 \end{pmatrix}$$

Finally, by replacing the values on the equation 9, we get the following bias:

$$b_{new}^{[1]} = \begin{pmatrix} 1.01872 \\ 1.03359 \\ 1.01872 \end{pmatrix}$$
$$b_{new}^{[2]} = \begin{pmatrix} 1.03745 \\ 1.01017 \end{pmatrix}$$
$$b_{new}^{[3]} = \begin{pmatrix} 1.00198 \\ 1.03177 \\ 0.9993 \end{pmatrix}$$

Part II: Programming and critical analysis

Considering the `winequality-red.csv` dataset (available at the webpage) where the goal is to estimate the quality (sensory appreciation) of a wine based on physicochemical inputs.

Using a 80-20 training-test split with a fixed seed (`random_state = 0`), you are asked to learn MLP regressors to answer the following questions.
Given their stochastic behavior, average the performance of each MLP from 10 runs (for reproducibility consider seeding the MLPs with `random_state` $\in$ `{1..10}`).

1. **Learn a MLP regressor with 2 hidden layers of size 10, rectifier linear unit activation on all nodes, and early stopping with 20% of training data set aside for validation. All remaining parameters (e.g., loss, batch size, regularization term, solver) should be set as default. Plot the distribution of the residues (in absolute value) using a histogram.**

   Blah

2. **Since we are in the presence of a integer regression task, a recommended trick is to round and bound estimates. Assess the impact of these operations on the MAE of the MLP learnt in previous question.**

   Blah

3. **Similarly assess the impact on RMSE from replacing early stopping by a well-defined number of iterations in {20,50,100,200} (where one iteration corresponds to a batch).**

   Blah

4. **Critically comment the results obtained in previous question, hypothesizing at least one reason why early stopping favors and/or worsens performance**

   Blah