Aprendizagem 2024/25

Homework III – Group 017

(ist1106992, ist1106154)

I. Pen-and-paper

1.

$$w = (x^T x)^{-1} x^T z$$

$$x = \begin{bmatrix} 1 & \phi(1,1) \\ 1 & \phi(1,3) \\ 1 & \phi(3,2) \\ 1 & \phi(3,3) \\ 1 & \phi(2,4) \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 6 \\ 1 & 9 \\ 1 & 8 \end{bmatrix} \qquad z = \begin{bmatrix} 1,25 \\ 3,0 \\ 2,7 \\ 3,2 \\ 5,5 \end{bmatrix}$$

$$x^T = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 3 & 6 & 9 & 8 \end{bmatrix}$$

$$w = \left(\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 3 & 6 & 9 & 8 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 6 \\ 1 & 9 \\ 1 & 8 \end{bmatrix} \right)^{-1} \times \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 3 & 6 & 9 & 8 \end{bmatrix} \times \begin{bmatrix} 1,25 \\ 3,0 \\ 2,7 \\ 3,2 \\ 5,5 \end{bmatrix} =$$

$$= \left(\begin{bmatrix} 5 & 27 \\ 27 & 191 \end{bmatrix} \right)^{-1} \times \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 3 & 6 & 9 & 8 \end{bmatrix} \times \begin{bmatrix} 1,25 \\ 3,0 \\ 2,7 \\ 3,2 \\ 5,5 \end{bmatrix} =$$

$$= \frac{1}{226} \begin{bmatrix} 191 & -27 \\ -27 & 5 \end{bmatrix} \times \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 3 & 6 & 9 & 8 \end{bmatrix} \times \begin{bmatrix} 1,25 \\ 3,0 \\ 2,7 \\ 3,2 \\ 5,5 \end{bmatrix} =$$

$$= \begin{bmatrix} \frac{82}{113} & \frac{55}{113} & \frac{29}{113} & \frac{-26}{113} & \frac{-25}{226} \\ \frac{-11}{113} & \frac{-6}{113} & \frac{3}{226} & \frac{9}{113} & \frac{13}{226} \end{bmatrix} \times \begin{bmatrix} 1,25 \\ 3,0 \\ 2,7 \\ 3,2 \\ 5,5 \end{bmatrix} =$$

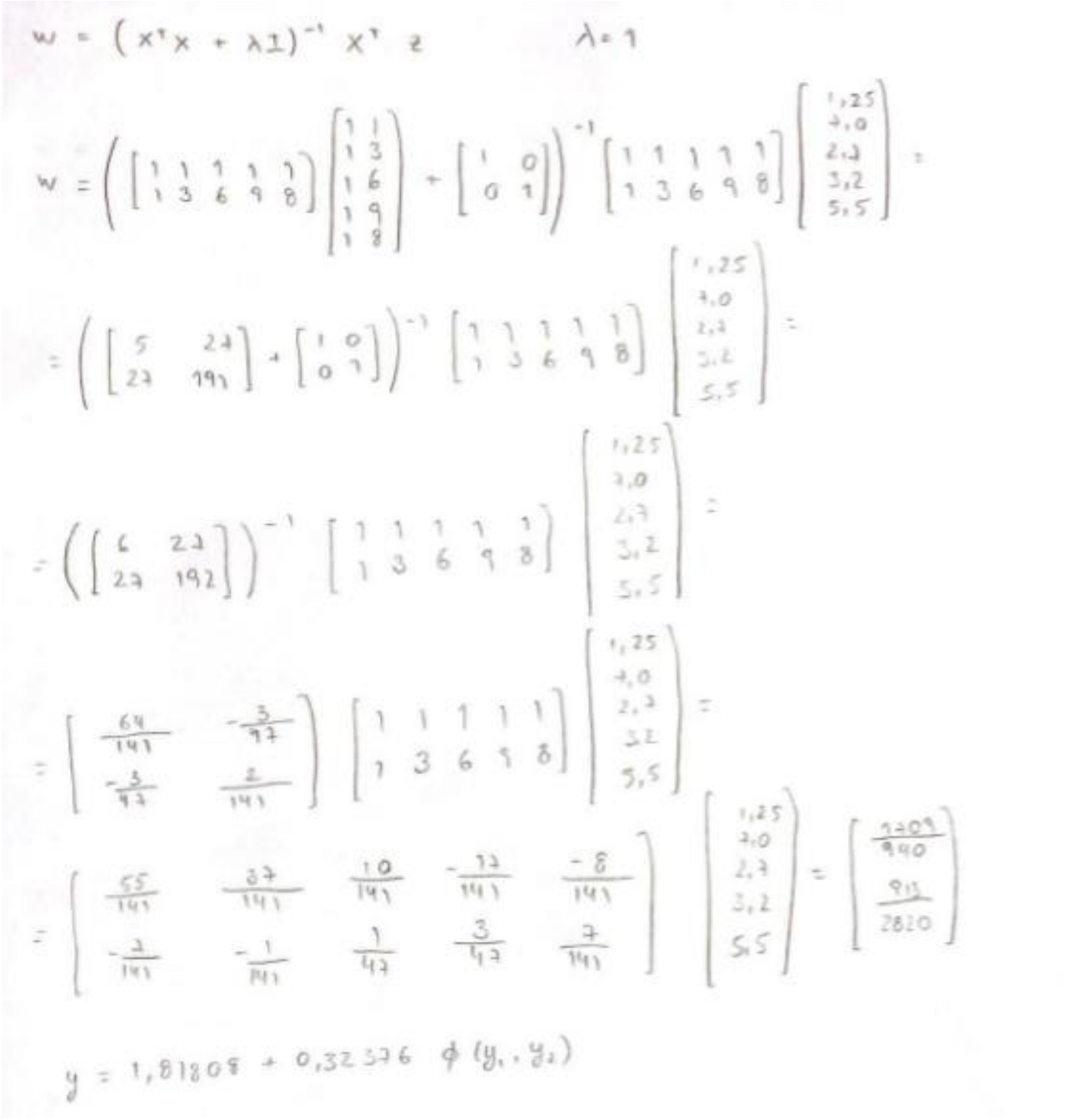
$$= \begin{bmatrix} \frac{3747}{1130} \\ \frac{257}{2260} \end{bmatrix}$$

$$y = 3,31593 + 0,11372\ \phi(y_1, y_2)$$

Homework III – Group 017

(ist1106992, ist1106154)

2.



To compare the learnt coefficients with the ones from the previous exercise, we calculated and the vector norms from both solutions.

$||w_{ex1}|| = \sqrt{3.31593^2 + 0.11372^2} = 3.31788$ $||w_{ex2}|| = \sqrt{1.81808^2 + 0.32376^2} = 1.84668$

Given this, as shown in the results, the regularized model has a smaller vector norm which means that its regularization minimizes the error because the model avoids fitting extreme values in the data.

Homework III – Group 017

(ist1106992, ist1106154)

3.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(z_i - \hat{z}_i)^2}$$

test u_6, u_7, u_8

- without Ridge:

$u_{num\,6} = [2 \quad 2] \rightarrow \phi(2,2) = 4$

$y_{num\,6} = w \cdot u = \begin{bmatrix} 3,31543 \\ 0,11372 \end{bmatrix} \begin{bmatrix} 1 \\ 4 \end{bmatrix} = 3,77081$

$u_{num\,7} = [1 \quad 2] \rightarrow \phi(1,2) = 2$

$y_{num\,7} = w \cdot u = \begin{bmatrix} 3,31543 \\ 0,11372 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} = 3,54287$

$u_{num\,8} = [5 \quad 1] \rightarrow \phi(5,1) = 5$

$y_{num\,8} = w \cdot u = \begin{bmatrix} 3,31543 \\ 0,11372 \end{bmatrix} \begin{bmatrix} 1 \\ 5 \end{bmatrix} = 3,88453$

- with Ridge:

$u_{num\,6} = [2 \quad 2] \rightarrow \phi(2,2) = 4$

$y_{num\,6} = w \cdot u = \begin{bmatrix} 1,81808 \\ 0,32376 \end{bmatrix} \begin{bmatrix} 1 \\ 4 \end{bmatrix} = 3,11312$

$u_{num\,7} = [1 \quad 2] = \phi(1,2) = 2$

$y_{num\,7} = w \cdot u = \begin{bmatrix} 1,81808 \\ 0,32376 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} = 2,4656$

$u_{num\,8} = [5 \quad 1] = \phi(5,1) = 5$

$y_{num\,8} = w \cdot u = \begin{bmatrix} 1,81808 \\ 0,32376 \end{bmatrix} \begin{bmatrix} 1 \\ 5 \end{bmatrix} = 3,43688$

TÉCNICO LISBOA

Aprendizagem 2024/25

Homework III – Group 017

(ist1106992, ist1106154)

$RMSE_{\text{without ridge}} = \sqrt{\frac{1}{3}\left((0,2 - 3,47087)^2 + (3,7 - 3,54332)^2 + (2,2 - 2,88453)^2\right)}$

$= 2,96560$

$RMSE_{\text{with ridge}} = \sqrt{\frac{1}{3}\left((0,2 - 3,11312)^2 + (3,7 - 2,4656)^2 + (2,2 - 2,46888)^2\right)}$

$= 1,75289$

Train: u_1, u_2, u_3, u_4, u_5

$\phi_{u_i}(y_1, y_2) = (1, 3, 6, 9, 8)$

without ridge:

$y_{\text{train } u_i} = \begin{bmatrix} 3,31543 \\ 0,11372 \end{bmatrix} \begin{bmatrix} 1 \\ \phi_{u_i} \end{bmatrix} = (3,42465;\ 3,65709;\ 3,99825;\ 4,33941;\ 4,22569)$

with ridge:

$y_{\text{train } u_i} = \begin{bmatrix} 1,81808 \\ 0,32376 \end{bmatrix} \begin{bmatrix} 1 \\ \phi_{u_i} \end{bmatrix} = (2,14184;\ 2,78936;\ 3,76064;\ 4,73192;\ 4,40818)$

$RMSE_{\text{without ridge}}$:

$\sqrt{\frac{1}{5}\left((1,25 - 3,42465)^2 + (7 - 3,65709)^2 + (2,7 - 3,99825)^2 + (3,2 - 4,33941)^2 + (5,5 - 4,22569)^2\right)}$

$= 2,0265$

$RMSE_{\text{with ridge}}$:

$\sqrt{\frac{1}{5}\left((1,25 - 2,14184)^2 + (7 - 2,78936)^2 + (2,7 - 3,76064)^2 + (3,2 - 4,73192)^2 + (5,5 - 4,40818)^2\right)}$

$= 2,1535$

Aprendizagem 2024/25

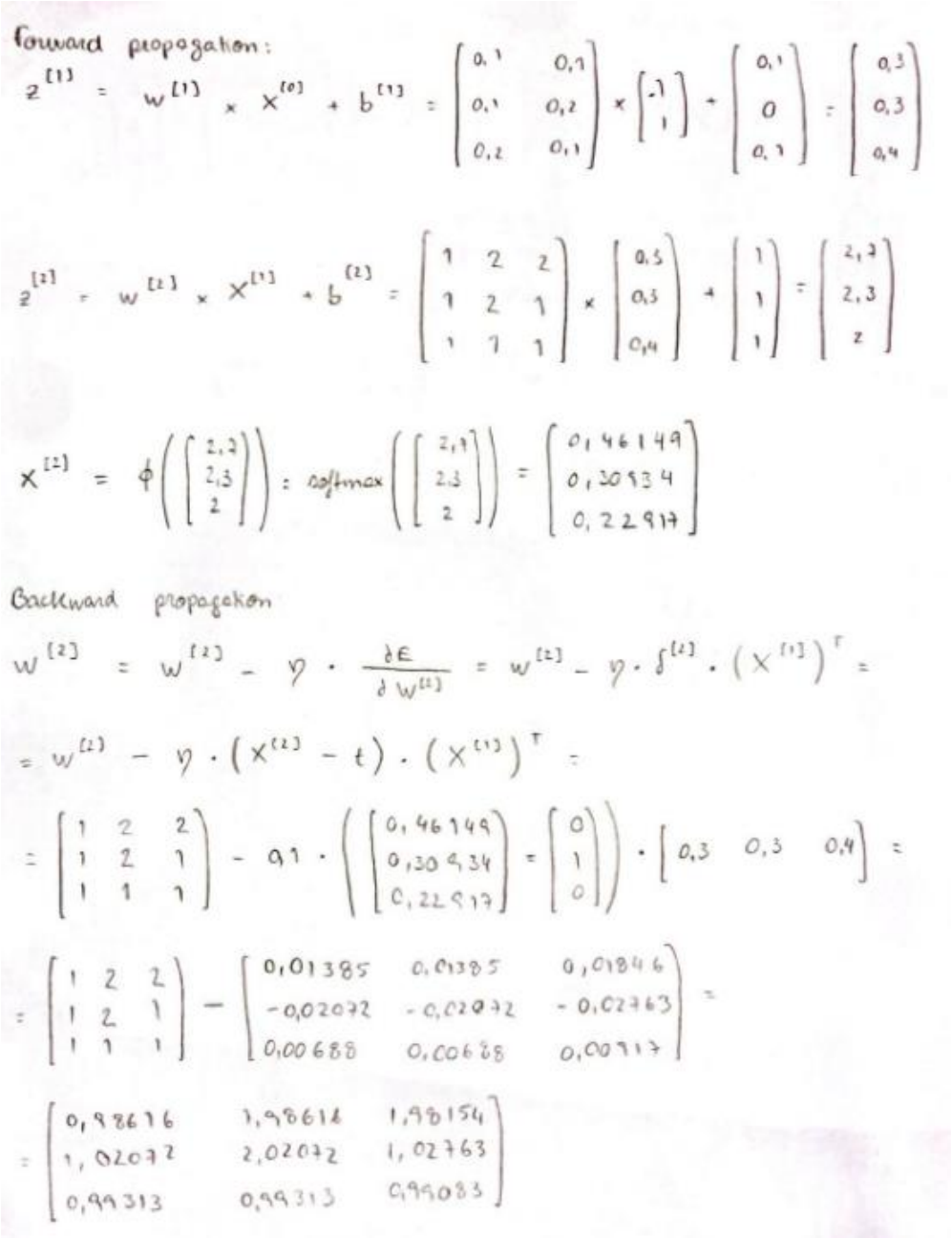
Homework III – Group 017

(ist1106992, ist1106154)

The test RMSE of the model without Ridge is greater than the test RMSE of the model with Ridge. This is expected because Ridge regularization helps prevent overfitting by adding a penalty term that discourages overly complex models, leading to better generalization capacities on the test data (lower RMSE). Without Ridge, the model may fit the training data too closely (overfitting), which results in poorer performance (higher RMSE) on unseen test data.

The training RMSE of the model without Ridge is smaller than the training RMSE of the model with Ridge. This is also expected because without regularization, the model is free to fit the training data more precisely, leading to better performance on the training data (lower RMSE). On the other hand, Ridge regularization increases training RMSE slightly since it imposes constraints on the model, to avoid overfitting, making it to fit the training data less precisely (higher RMSE). In conclusion, the results are according to what is expected.

4.



$$b^{[2]} = b^{[2]} - \eta \cdot \frac{\partial E}{\partial b^{[2]}} = b^{[2]} - \eta\, \delta^{[2]} =$$

$$= b^{[2]} - \eta \cdot (x^{[2]} - t) = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} - 0{,}1 \cdot \left(\begin{bmatrix} 0{,}46149 \\ 0{,}30934 \\ 0{,}22917 \end{bmatrix} - \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \right) =$$

$$= \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} - \begin{bmatrix} 0{,}04615 \\ -0{,}06907 \\ 0{,}02292 \end{bmatrix} = \begin{bmatrix} 0{,}95385 \\ 1{,}06907 \\ 0{,}97708 \end{bmatrix}$$

$$w^{[1]} = w^{[1]} - \eta \cdot \frac{\partial E}{\partial w^{[1]}} = w^{[1]} - \eta \cdot \delta^{[1]} \cdot (x^{[0]})^T =$$

$$= w^{[1]} - \eta \cdot \left((w^{[2]})^T \cdot \delta^{[2]} \right) \cdot (x^{[0]})^T =$$

$$= \begin{bmatrix} 0{,}1 & 0{,}1 \\ 0{,}1 & 0{,}2 \\ 0{,}2 & 0{,}1 \end{bmatrix} - 0{,}1 \cdot \left(\begin{bmatrix} 1 & 1 & 1 \\ 2 & 2 & 1 \\ 2 & 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} 0{,}46149 \\ -0{,}69066 \\ 0{,}22917 \end{bmatrix} \right) \cdot [1 \;\; 1] =$$

$$= \begin{bmatrix} 0{,}1 & 0{,}1 \\ 0{,}1 & 0{,}2 \\ 0{,}2 & 0{,}1 \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ -0{,}02292 & -0{,}02292 \\ 0{,}04615 & 0{,}04615 \end{bmatrix} = \begin{bmatrix} 0{,}1 & 0{,}1 \\ 0{,}12292 & 0{,}22292 \\ 0{,}15385 & 0{,}05385 \end{bmatrix}$$

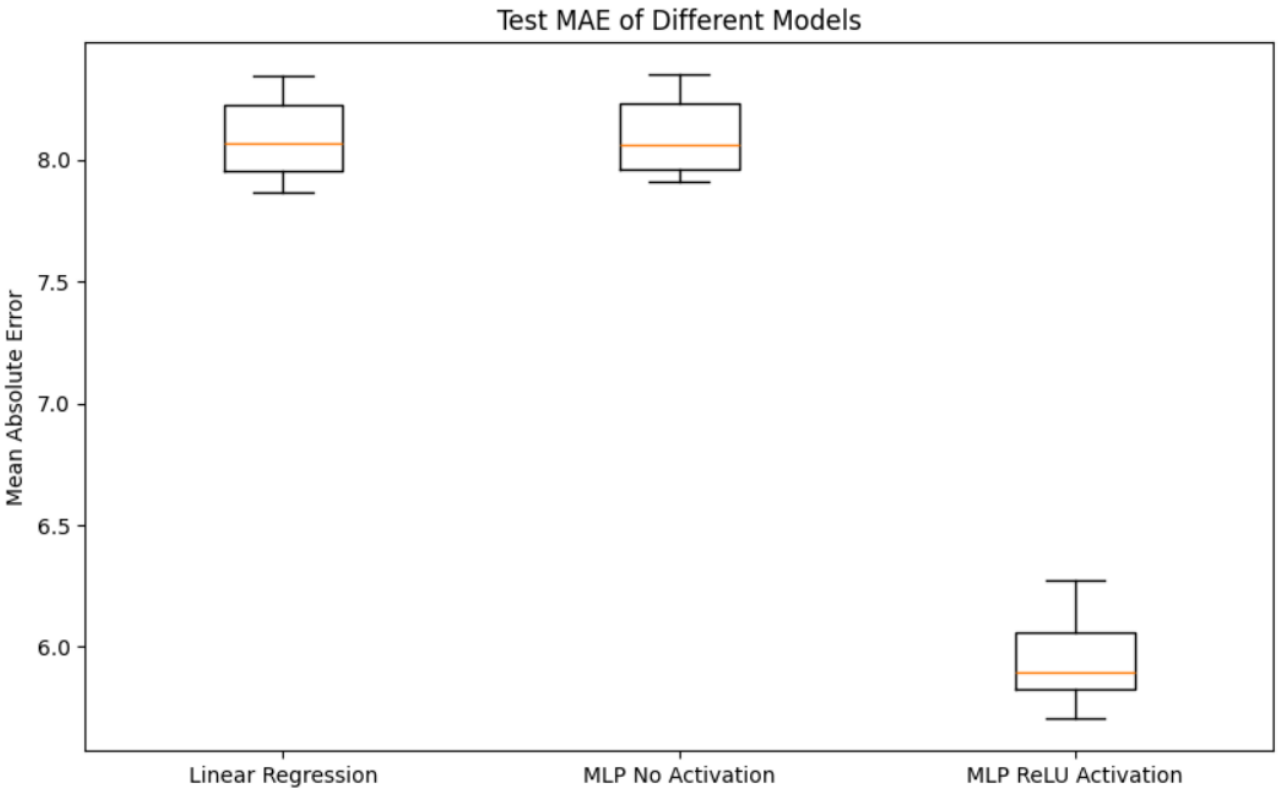
$$b^{[1]} = b^{[1]} - \eta \cdot \frac{\partial E}{\partial b^{[1]}} = b^{[1]} \cdot \eta\, \delta^{[1]} =$$

$$= \begin{bmatrix} 0{,}1 \\ 0 \\ 0{,}1 \end{bmatrix} - 0{,}1 \cdot \begin{bmatrix} 0 \\ -0{,}22917 \\ 0{,}46749 \end{bmatrix} = \begin{bmatrix} 0{,}1 \\ 0 \\ 0{,}1 \end{bmatrix} - \begin{bmatrix} 0 \\ -0{,}02292 \\ 0{,}04615 \end{bmatrix} =$$

$$= \begin{bmatrix} 0{,}1 \\ 0{,}02292 \\ 0{,}05385 \end{bmatrix}$$

II. Programming and critical analysis

5.



6.

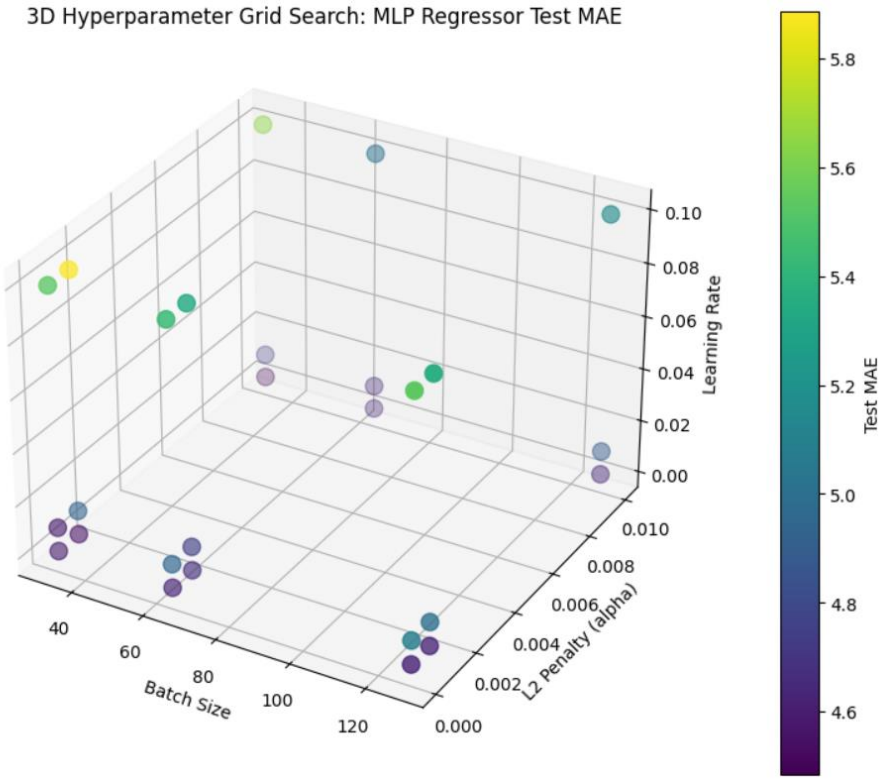
With the analysis of the boxplot above, we discovered that the results for mean absolute error are very similar for the linear regression model and the MLP model without activation.

This happens because the linear regression model assumes a linear relationship between the input features and the target variable. Similarly, the Multilayer Perceptron (MLP) without activation functions consists in compositions of linear transformations of inputs; these compositions still behave as linear functions so, even though the model's neural network has multiple layers, gives similar results to the linear regression model, as seen in the boxplot.

Additionally, when analysing the boxplot, we assessed that the mean absolute error decreases significantly when using a MLP model with activation (such as ReLU, sigmoid, or tanh). This happens because non-linearity is introduced into the network, enabling it to learn complex patterns and relationships and to approximate any continuous function, resulting in a more comprehensive and accurate learning of the data.

In conclusion, the MLP model with activation functions is more efficient than the previous 2, as confirmed by the lower mean absolute error displayed in the boxplot, revealing the importance of using these functions in the MLP model.

7.



When analyzing the plot and its scale, we conclude the relation between the alterations in batch sizes, L2 Penalty and learning rate in the MAE. To get a better overview of the results we first decided to analyze each parameter separately:

When it comes to the L2 Penalty, the values influence the amount of regularization applied to the model. This means that for smaller values, the model fits the training data better when compared to larger values. In more extreme cases the smaller alpha values lead to overfitting while the larger lead to underfitting, both decreasing the model's efficiency.

For the learning rate, we know that higher values lead to faster convergences of the model. This can lead to an overshooting of the optimal solution so, it’s better to have smaller learning rates that allow a more stable training even though they increase the training times (that aren't considered in the plot).

On the other hand, batch sizes control the frequency of updates to the model's weights. For smaller batch sizes, these are more frequent which helps with finer adjustments while larger sizes smooth the training processes but decrease the model's ability to capture detailed patterns.

Given this, the best combination of hyperparameters - when the MAE is as low as possible - is 0.01 for L2 Penalty, 32 for batch size and 0.001 for learning rate. This is due to the batch's size and the learning rate being as small as possible and the L2 Penalty being the largest value tested. With these values, the model doesn't overshoot the optimal solution and is able to capture most of the details in the data but still not get overfitted.

END