Aprendizagem 2024/25

Homework III – Group 036

(ist1106468, ist1106324)

I. Pen-and-paper

1)

HomeWork3

Using $\Phi(y_1, y_2) = y_1 \times y_2$:

D	$\Phi(y_1, y_2)$
u_1	1
u_2	3
u_3	6
u_4	9
u_5	8

[1] $\hat{z} = Xw$ (pesos; observações; valor previsto)

Using OLS closed form:

$w = (X^T X)^{-1} X^T z$ (valores reais do output)

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

$$X^T X = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 3 & 6 & 9 & 8 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 6 \\ 1 & 9 \\ 1 & 8 \end{bmatrix} = \begin{bmatrix} 5 & 27 \\ 27 & 191 \end{bmatrix}$$

$$(X^T X)^{-1} = \begin{bmatrix} 0,84513 & -0,11947 \\ -0,11947 & 0,02211 \end{bmatrix}$$

$$(X^T X)^{-1} X^T z = \begin{bmatrix} 3,31593 \\ 0,11372 \end{bmatrix}$$

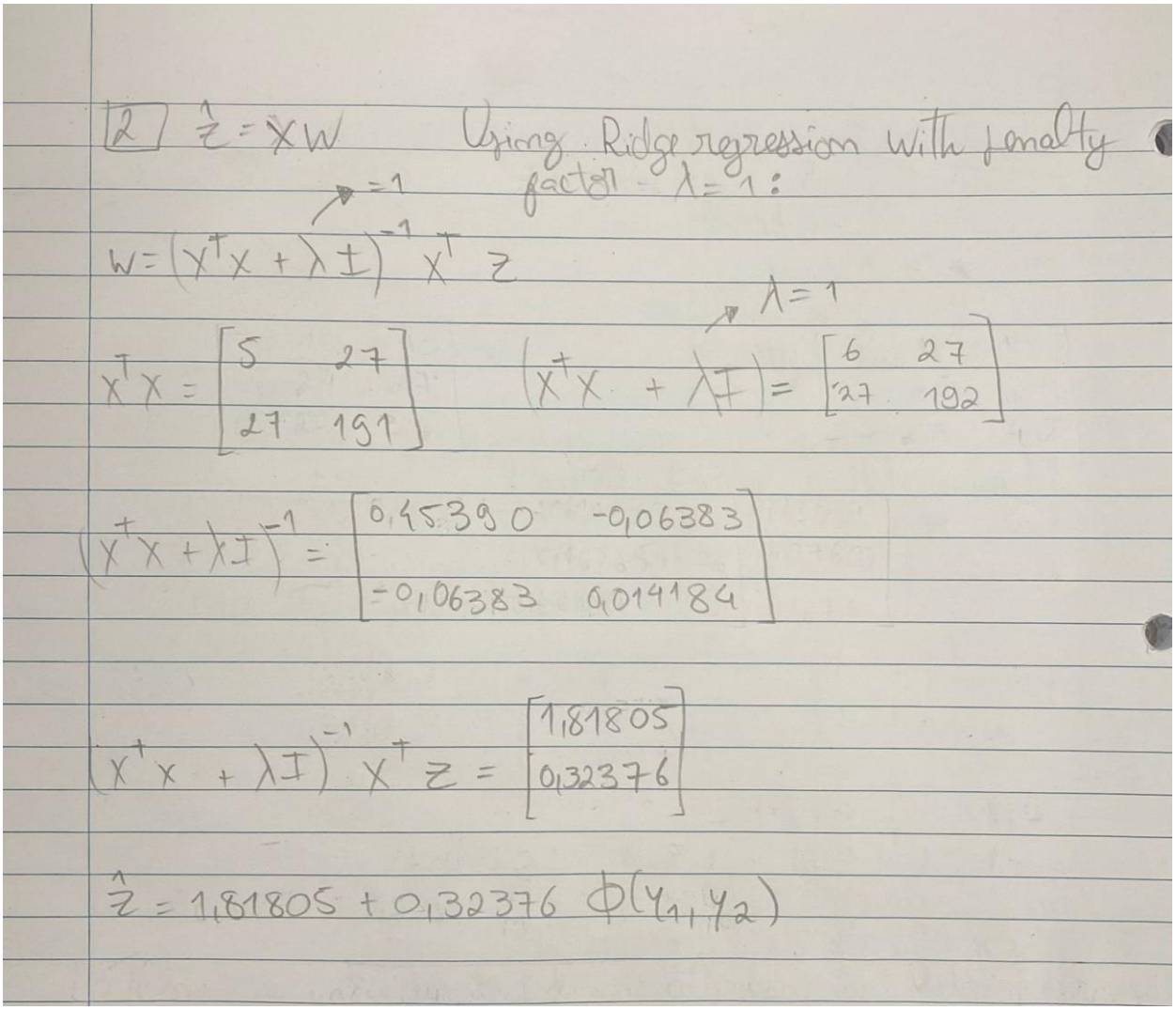
$$\hat{z} = 3,31593 + 0,11372\, \Phi(y_1, y_2)$$

Aprendizagem 2024/25

Homework III – Group 036

(ist1106468, ist1106324)

2)



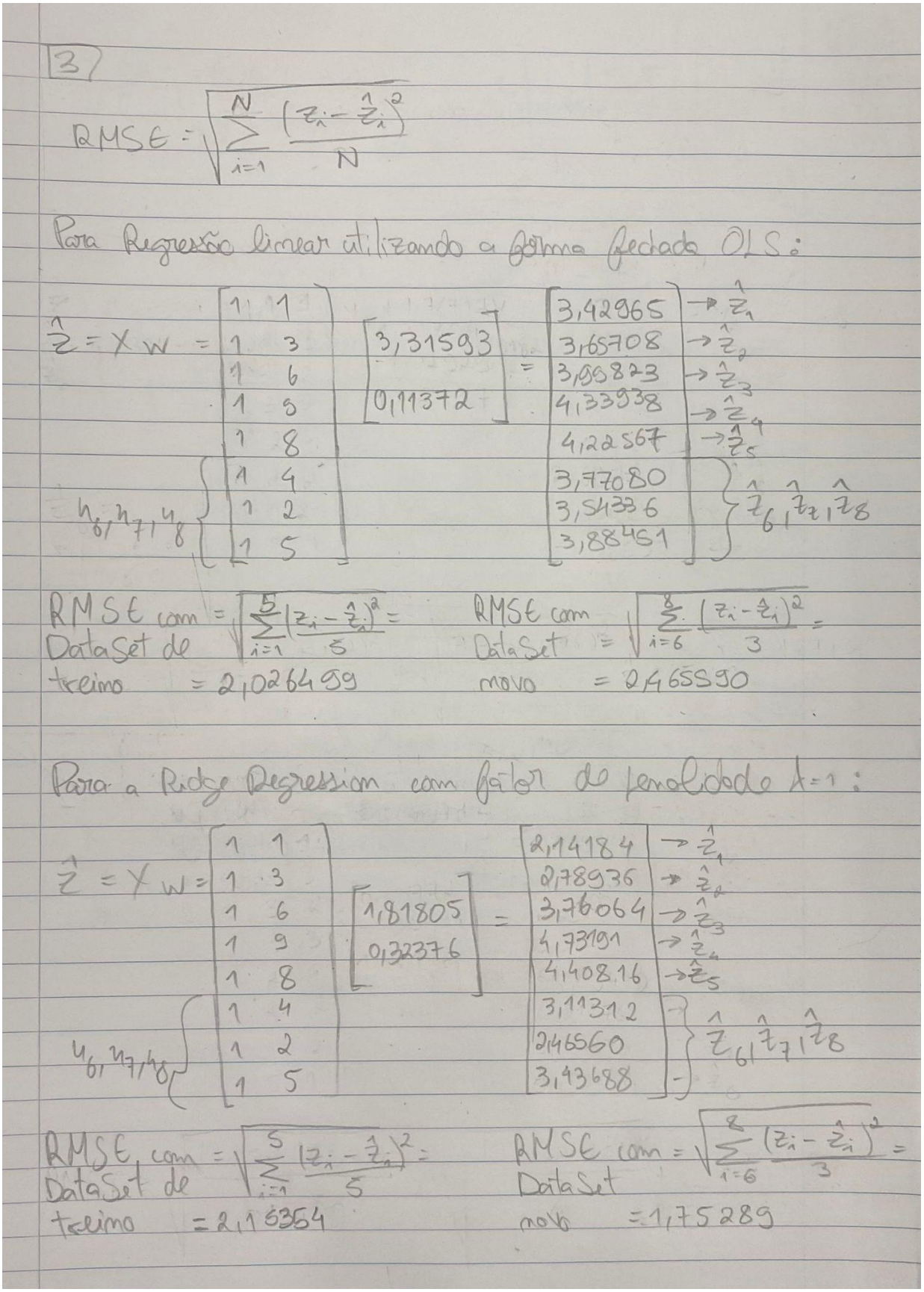
A análise das matrizes de pesos mostra que a regularização da Ridge Regression teve um impacto significativo na redistribuição dos valores dos pesos. No modelo de Regressão Linear, o primeiro peso era bastante elevado (3.3159), enquanto o segundo era muito pequeno (0.1137). Com a aplicação da Ridge Regression, o primeiro peso diminuiu substancialmente (1.8181), e o segundo aumentou ligeiramente (0.3238). Isto acontece porque a regularização de Ridge penaliza pesos grandes para evitar o overfitting, reduzindo a dependência excessiva de variáveis específicas.

Ao mesmo tempo, a Ridge Regression redistribui a importância entre as variáveis, permitindo que o segundo peso, que anteriormente era subestimado, se torne mais relevante. Desta forma, a esta equilibra melhor a influência das variáveis no modelo, melhorando a sua capacidade de generalização.

Homework III – Group 036

(ist1106468, ist1106324)

3)



Aprendizagem 2024/25

Homework III – Group 036

(ist1106468, ist1106324)

Como o RMSE nos dados de teste para o modelo regularizado (Ridge) é menor do que o RMSE do modelo não regularizado, podemos concluir que a regularização melhorou a capacidade de generalização do modelo, reduzindo o erro e tornando as previsões mais precisas, o que era esperado devido à penalização de coeficientes excessivamente altos.

Por outro lado, o RMSE nos dados de treino é maior no modelo regularizado do que no modelo não regularizado. Isto ocorre porque a regularização impõe restrições nos coeficientes, evitando que o modelo se ajuste demasiado aos dados de treino (overfitting), o que leva a um maior erro nesse conjunto.

No entanto, esta ligeira perda de desempenho para os dados de treino é compensada por uma melhor performance nos dados de teste, indicando que o modelo regularizado está menos suscetível a overfitting e, portanto, tem maior capacidade de generalização.

Aprendizagem 2024/25

Homework III – Group 036

(ist1106468, ist1106324)

4)

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

Output activation function:

$$\text{softmax}\left(z_c^{[out]}\right) = \frac{e^{z_c^{[out]}}}{\sum_{l=1}^{|C|} e^{z_l^{[out]}}}$$

Hidden layers activation function: None $\rightarrow z^{[i]} = x^{[i]}$

Error function → Cross-Entropy Loss:

$$\text{Erro} = -\sum_{i=1}^{N} \sum_{l=1}^{|C|} t_l^{(i)} \log\left(z_l^{[out](i)}\right)$$

Back propagation:

$$W^{[i]} = W^{[i]} - \eta \frac{\partial \epsilon}{\partial W^{[i]}} \quad (=0,1)$$

$$b^{[i]} = b^{[i]} - \eta \frac{\partial \epsilon}{\partial b^{[i]}}$$

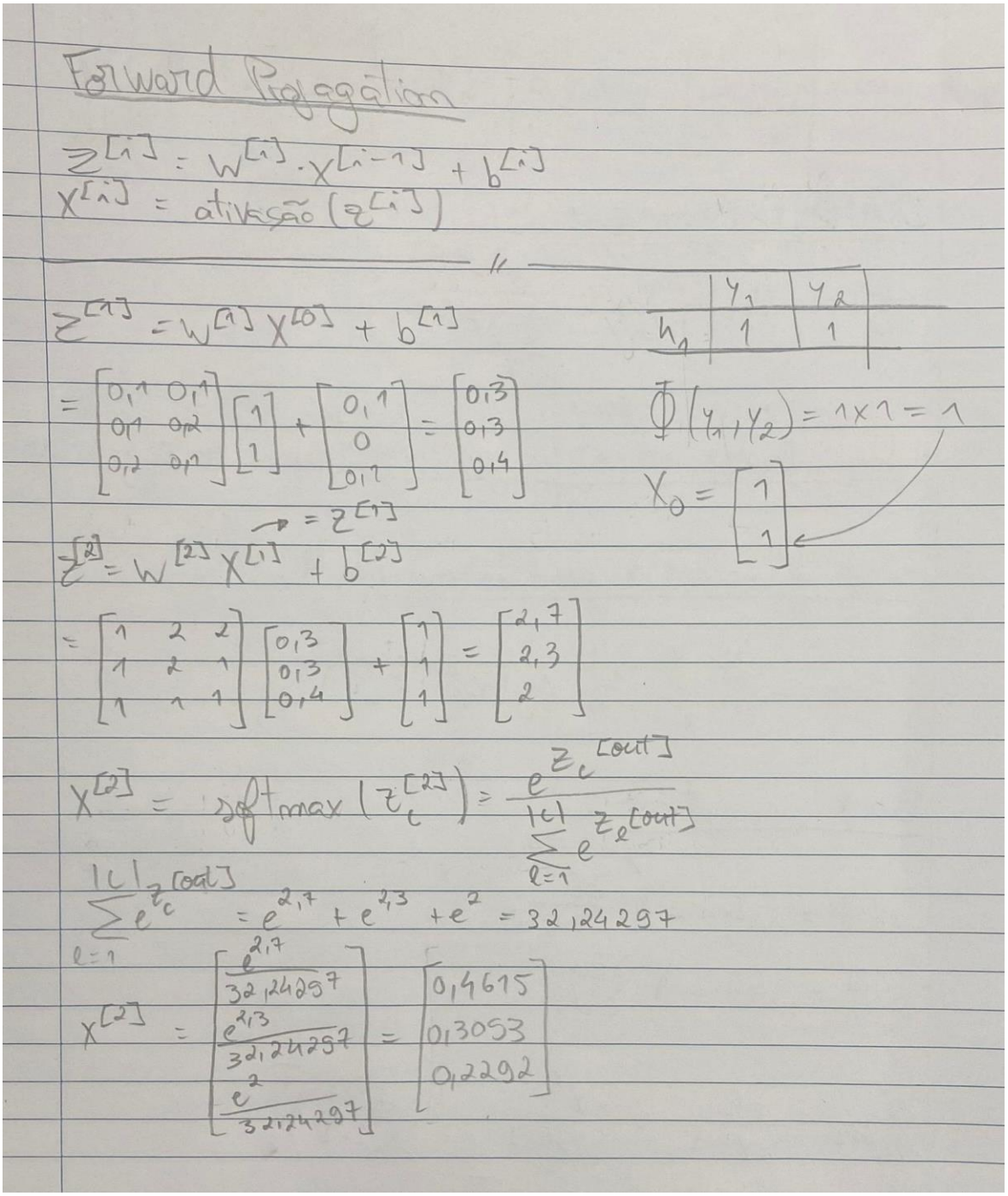
$$\frac{\partial \epsilon}{\partial W^{[i]}} = \frac{\partial \epsilon}{\partial x^{[i]}} \circ \frac{\partial x^{[i]}}{\partial z^{[i]}} \cdot \left(\frac{\partial z^{[i]}}{\partial W^{[i]}}\right)^T = \delta^{[i]} \cdot \left(x^{[i-1]}\right)^T$$

$$\frac{\partial \epsilon}{\partial b^{[i]}} = \frac{\partial \epsilon}{\partial x^{[i]}} \circ \frac{\partial x^{[i]}}{\partial z^{[i]}} \cdot \frac{\partial z^{[i]}}{\partial b^{[i]}} = \delta^{[i]}$$

Aprendizagem 2024/25

Homework III – Group 036

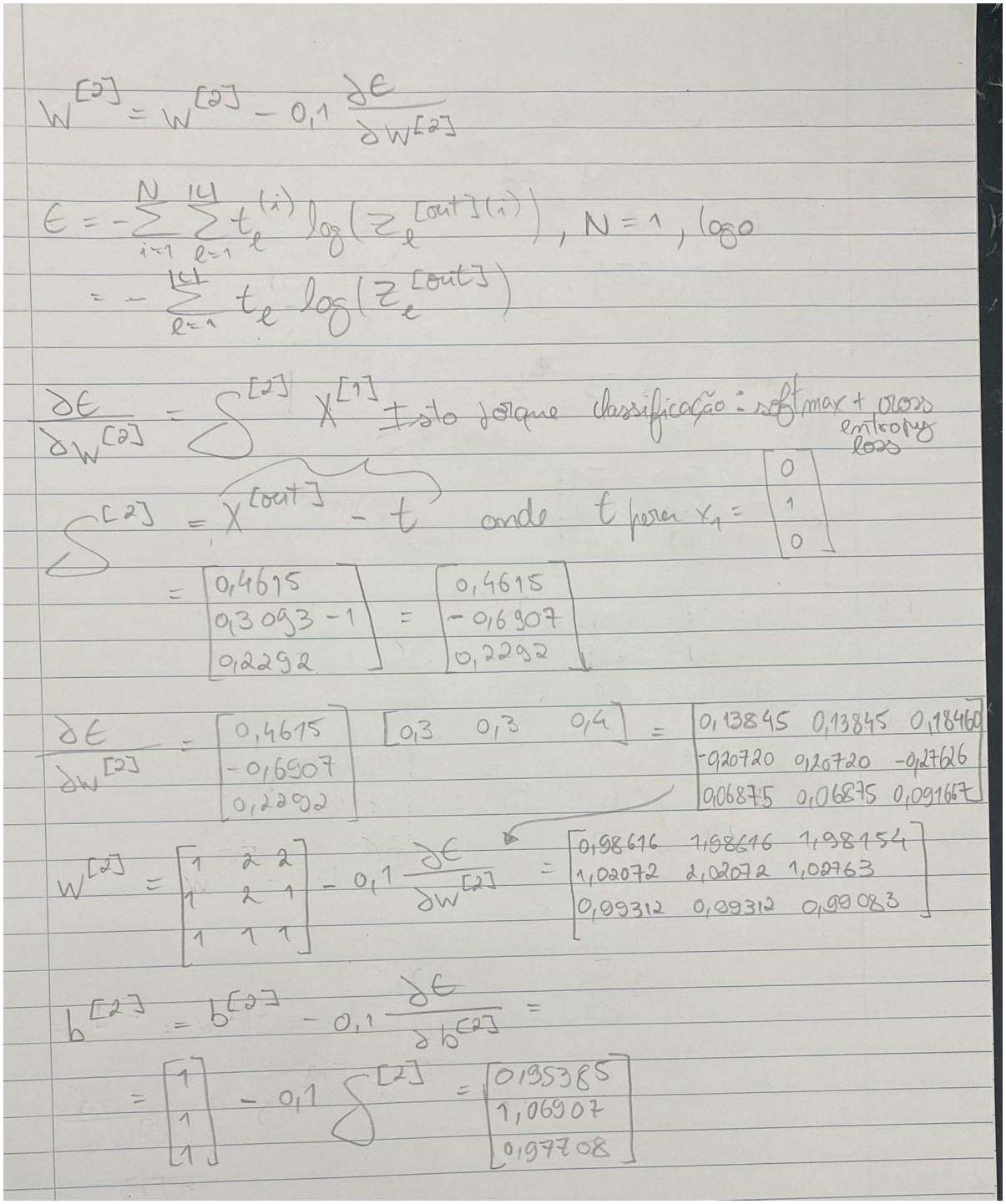
(ist1106468, ist1106324)



Aprendizagem 2024/25

Homework III – Group 036

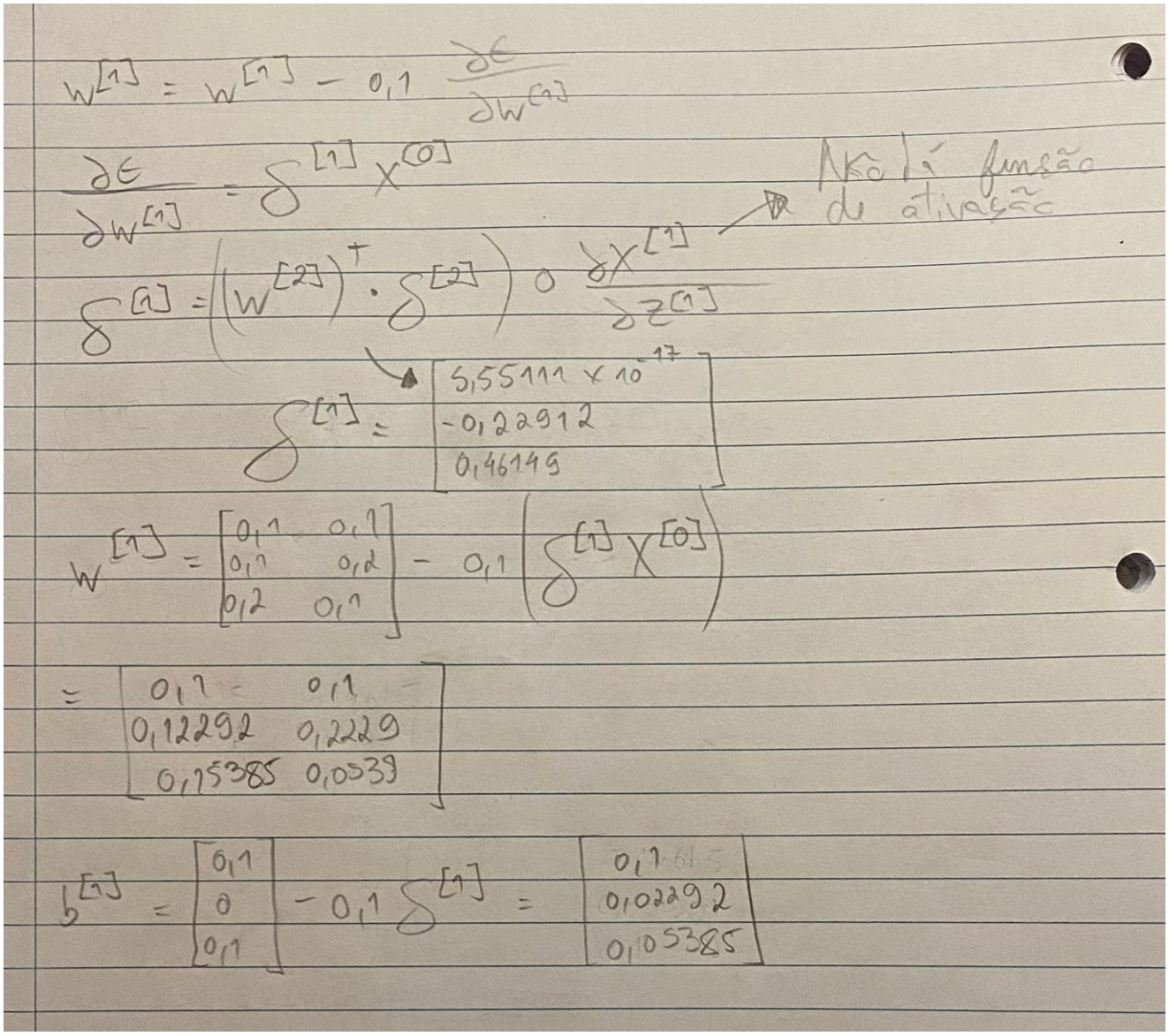
(ist1106468, ist1106324)



Aprendizagem 2024/25

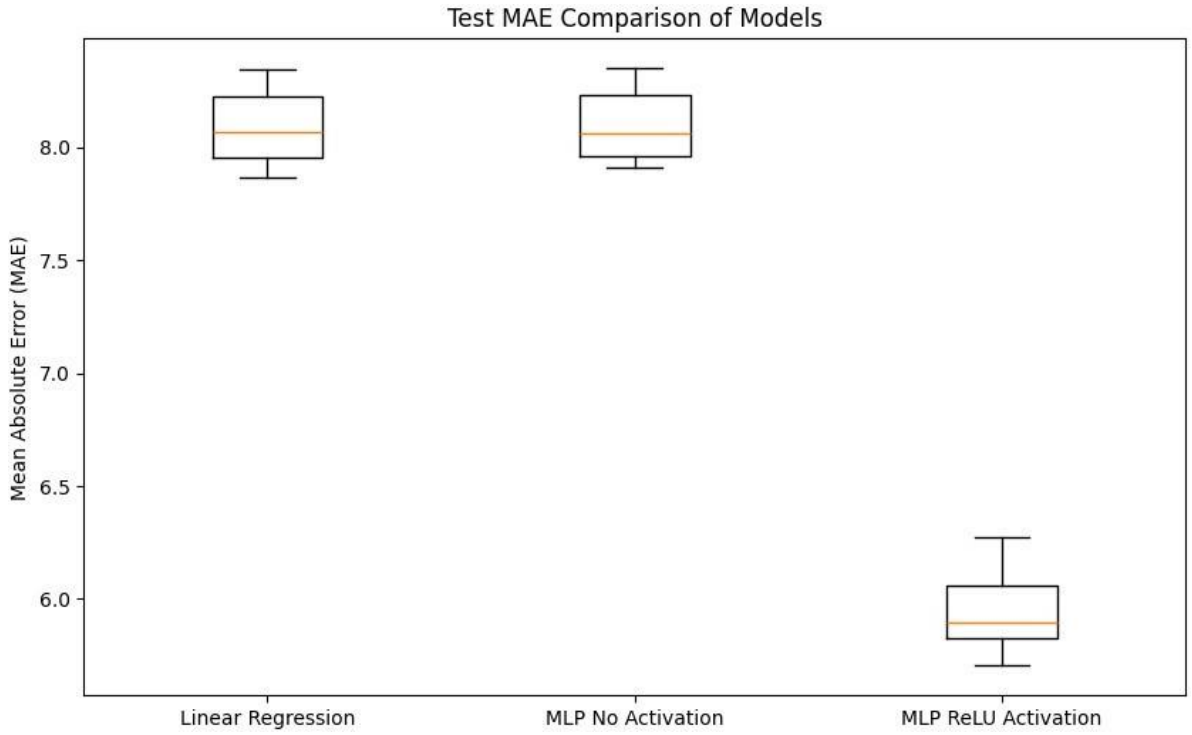
Homework III – Group 036

(ist1106468, ist1106324)



II. Programming and critical analysis

5)



6)

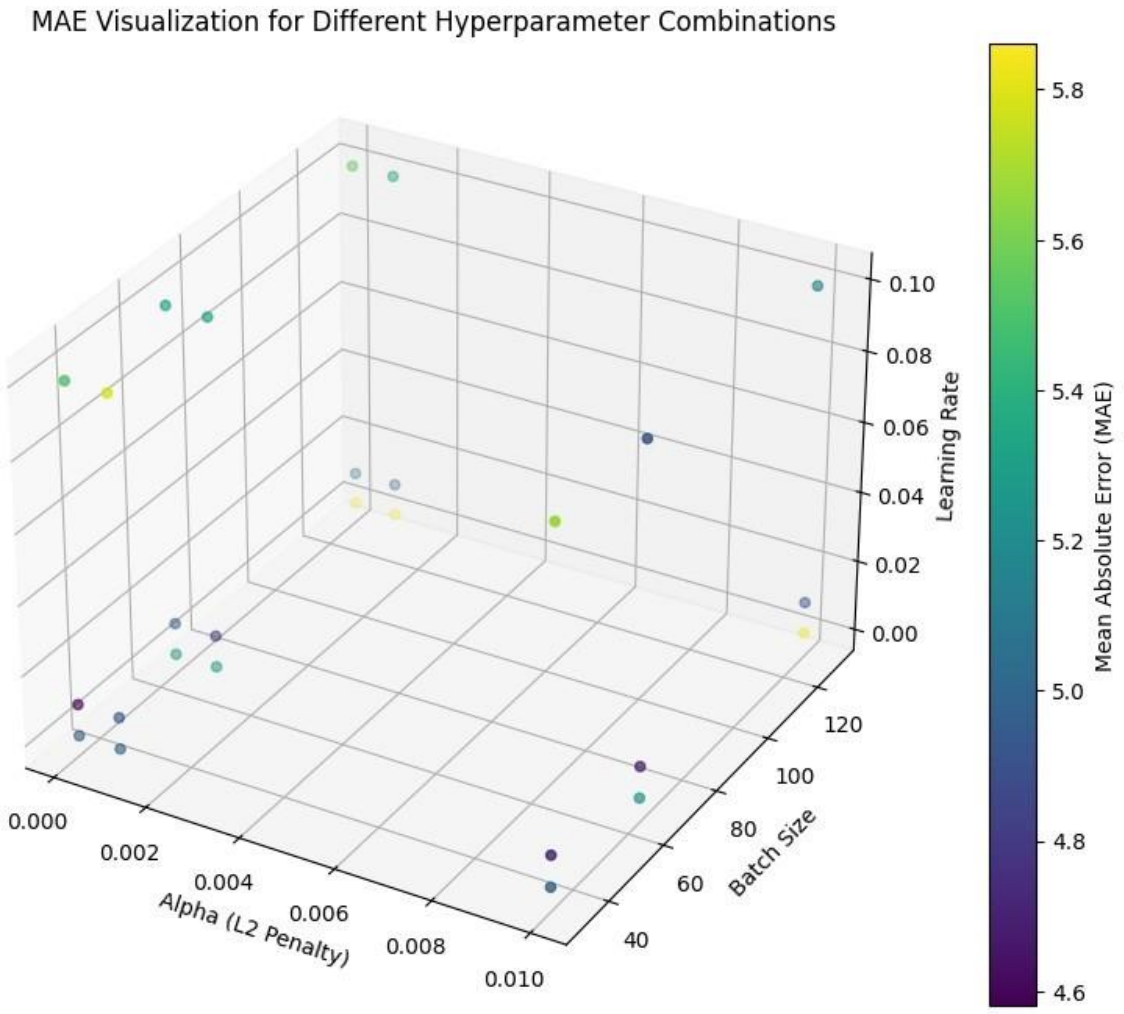
Uma regressão linear é um modelo simples capaz de estabelecer relações entre as features e o target. Este modelo assume que os dados podem ser modelados através de uma reta (ou plano, etc. para dimensões maiores). Está, portanto, limitada para capturar relações não lineares complexas.

Um modelo MLP sem funções de ativação é equivalente a uma regressão linear. Se não tiver funções de ativação não lineares, o output da rede neuronal será sempre uma transformação linear do input (independentemente do número de camadas).

Assim, e tal como podemos observar no gráfico do exercício anterior, uma rede neuronal e um modelo MLP sem funções de ativação vão ter resultados idênticos.

Entende-se, portanto, a importância de funções de ativação não lineares nos modelos MLP. Ao incluir funções de ativação, o modelo torna-se capaz de capturar padrões mais complexos, incluindo relações não lineares (interações entre as variáveis que seriam impossíveis de identificar com uma simples transformação linear). Isto permite que o modelo MLP aprenda representações mais completas e flexíveis dos dados, melhorando a sua capacidade de generalizar para novos inputs. Embora esta flexibilidade possa aumentar o risco de overfitting, também possibilita ao modelo extrair padrões complexos que a regressão linear não consegue capturar.

7)



Nesta representação em 3D, é possível observar como o desempenho do modelo, medido pelo Erro Absoluto Médio (MAE), varia em função de diferentes combinações de hiperparâmetros: alpha, learning rate e batch size. A barra de cores representa os valores de MAE, com tons mais escuros indicando erros menores - melhor desempenho do modelo.

Uma learning rate mais alta permite que o modelo convirja rapidamente, mas pode resultar em "saltos" sobre a solução ótima, prejudicando a generalização. Em contrapartida, uma learning rate mais baixa torna o treino mais demorado, mas frequentemente resulta num desempenho superior ao explorar soluções de forma mais detalhada.

Valores elevados de alpha proporcionam uma forte regularização, ajudando a reduzir o overfitting ao penalizar pesos excessivos. No entanto, se o valor de alpha for demasiado alto, o modelo pode deixar de identificar padrões relevantes nos dados. Por outro lado, valores baixos de alpha diminuem a regularização, permitindo um melhor ajuste aos dados de treinamento, mas aumentando o risco de overfitting.

Quanto ao batch size, um valor reduzido atualiza os pesos com mais frequência, gerando uma aprendizagem mais "ruidosa" que pode ajudar o modelo a evitar mínimos locais, embora demore mais para convergir. Em

contraste, valores grandes de batch size processam mais dados antes de realizar atualizações, resultando em mudanças mais precisas e estáveis, mas que podem tornar a aprendizagem menos variável, possivelmente fazendo com que o modelo fique preso em mínimos locais.

No gráfico, observamos claramente a vantagem de utilizar valores mais baixos de batch size e learning rate, como esperado. No caso de alpha, no entanto, abordagens mais extremas parecem funcionar melhor. Tanto um alpha baixo (que permite ao modelo aprender relações mais complexas entre os dados) como um alpha alto (que simplifica o modelo) produzem melhores resultados do que valores intermédios.

END