Pedro Nunes 109368
Andrei Barb 109 762

1. $w = (X^T \cdot X)^{-1} \cdot X^T \cdot z$

$$X = \begin{bmatrix} 1 & \phi(2,2) \\ 1 & \phi(1,1) \\ 1 & \phi(3,2) \\ 1 & \phi(6,3) \\ 1 & \phi(8,1) \end{bmatrix} = \begin{bmatrix} 1 & 4 \\ 1 & 1 \\ 1 & 6 \\ 1 & 18 \\ 1 & 8 \end{bmatrix} \qquad z = y_{num} = \begin{bmatrix} 3,5 \\ 1,0 \\ 3,8 \\ 10,1 \\ 8,5 \end{bmatrix}$$

$$X^T = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 4 & 1 & 6 & 18 & 8 \end{bmatrix}$$

$$X^T \cdot X = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 4 & 1 & 6 & 18 & 8 \end{bmatrix} \begin{bmatrix} 1 & 4 \\ 1 & 1 \\ 1 & 6 \\ 1 & 18 \\ 1 & 8 \end{bmatrix} = \begin{bmatrix} 5 & 37 \\ 37 & 441 \end{bmatrix}$$

$$(X^T \cdot X)^{-1} = \frac{1}{836} \begin{bmatrix} 441 & -37 \\ -37 & 5 \end{bmatrix} = \begin{bmatrix} 0,52751 & -0,04426 \\ -0,04426 & 0,00598 \end{bmatrix}$$

$$(X^T \cdot X)^{-1} \cdot X^T = \begin{bmatrix} 0,52751 & -0,04426 \\ -0,04426 & 0,00598 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 4 & 1 & 6 & 18 & 8 \end{bmatrix}$$

$$= \begin{bmatrix} 0,35047 & 0,48325 & 0,26195 & -0,26917 & 0,17343 \\ -0,02034 & -0,03828 & -0,00838 & 0,06338 & 0,00358 \end{bmatrix}$$

$$w = \begin{bmatrix} 0,35047 & 0,48325 & 0,26195 & -0,26917 & 0,17343 \\ -0,02034 & -0,03828 & -0,00838 & 0,06338 & 0,00358 \end{bmatrix} \begin{bmatrix} 3,5 \\ 1,0 \\ 3,8 \\ 10,1 \\ 8,5 \end{bmatrix}$$

$$= \begin{bmatrix} 1,46136 \\ 0,52955 \end{bmatrix}$$

$$\hat{y}_{mum} = 1,46136 + 0,52955 \cdot \phi(y_1, y_2)$$

2. $w = (X^T \cdot X + \lambda \cdot I)^{-1} \cdot X^T \cdot z \qquad \lambda = 1$

$$X = \begin{bmatrix} 1 & \phi(2,2) \\ 1 & \phi(1,1) \\ 1 & \phi(3,2) \\ 1 & \phi(6,3) \\ 1 & \phi(8,1) \end{bmatrix} = \begin{bmatrix} 1 & 4 \\ 1 & 1 \\ 1 & 6 \\ 1 & 18 \\ 1 & 8 \end{bmatrix} \qquad z = y_{muma} = \begin{bmatrix} 3{,}5 \\ 1{,}0 \\ 3{,}8 \\ 10{,}1 \\ 8{,}5 \end{bmatrix}$$

$$X^T = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 4 & 1 & 6 & 18 & 8 \end{bmatrix}$$

$$X^T \cdot X = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 4 & 1 & 6 & 18 & 8 \end{bmatrix} \begin{bmatrix} 1 & 4 \\ 1 & 1 \\ 1 & 6 \\ 1 & 18 \\ 1 & 8 \end{bmatrix} = \begin{bmatrix} 5 & 37 \\ 37 & 441 \end{bmatrix}$$

$$X^T \cdot X + \lambda I = \begin{bmatrix} 5 & 37 \\ 37 & 441 \end{bmatrix} + 1 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 6 & 37 \\ 37 & 442 \end{bmatrix}$$

$$(X^T \cdot X + \lambda I)^{-1} = \frac{1}{1283} \begin{bmatrix} 442 & -37 \\ -37 & 6 \end{bmatrix} = \begin{bmatrix} 0.34451 & -0{,}02884 \\ -0{,}02884 & 0{,}00468 \end{bmatrix}$$

$$(X^T \cdot X + \lambda I)^{-1} \cdot X^T = \begin{bmatrix} 0.34451 & -0{,}02884 \\ -0{,}02884 & 0{,}00468 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 4 & 1 & 6 & 18 & 8 \end{bmatrix}$$

$$= \begin{bmatrix} 0.22915 & 0{,}31567 & 0{,}17147 & -0{,}17461 & 0{,}11379 \\ -0{,}01012 & -0{,}02416 & -0{,}00076 & 0{,}05540 & 0{,}00860 \end{bmatrix}$$

$$w = \begin{bmatrix} 0.22915 & 0,31567 & 0,17147 & -0,17461 & 0,11379 \\ -0,01012 & -0,02416 & -0,00076 & 0,05540 & 0,00860 \end{bmatrix} \begin{bmatrix} 3,5 \\ 1,0 \\ 3,8 \\ 10,1 \\ 8,5 \end{bmatrix}$$

$$= \begin{bmatrix} 0,97319 \\ 0,56921 \end{bmatrix}$$

$$\hat{y}_{num} = 0,97319 + 0,56921\, \phi(y_1, y_2)$$

In the Ridge regression model, the first coefficient decreases significantly, showing that regularization effectively penalizes large weights to prevent overfitting. Meanwhile, the second coefficient increases slightly, indicating that Ridges balances the contribution of both features by redistributing the weights, thus achieving a more stable model.

3. $MAE = \frac{1}{m} \sum |z_i - \hat{z}_i|$

	y_{num}	ϕ	$\hat{y}$ OLS	$\hat{y}$ Ridge
x_1	3,5	4	3,57956	3,25003
x_2	1,0	1	1,99091	1,5424
x_3	3,8	6	4,63866	4,38845
x_4	10,1	18	10,99326	11,21897
x_5	8,5	8	5,69776	5,52687

	y_{num}	ϕ	$\hat{y}$ OLS	$\hat{y}$ Ridge
x_6	1	0	1,46136	0,97319
x_7	6,2	12	7,81596	7,80371
x_8	3,6	5	4,10911	3,81924

$\hat{y}_{OLS}(x_1) = 1,46136 + 0,52955 \times 4 = 3,57956$
$\hat{y}_{OLS}(x_2) = 1,46136 + 0,52955 \times 1 = 1,99091$
...

$\hat{y}_{RIDGE}(x_1) = 0,97319 + 0,56921 \times 4 = 2,2157979 2$
$\hat{y}_{Ridge}(x_2) = 0,97319 + 0,56921 \times 1 = 0,5539494$
...

$$MAE_{TrainOLS} = \frac{1}{5}\left(|3,5 - 3,57956| + |1 - 1,99091| + \dots\right) = 1,120926$$

$$MAE_{TrainRidge} = \frac{1}{5}\left(|3,5 - 3,25003| + |1 - 1,5424| + \dots\right) = 1,094584$$

$$MAE_{Test\,OLS} = \frac{1}{3}\left(|1 - 1{,}46136| + \cdots \right) = 0{,}862143$$

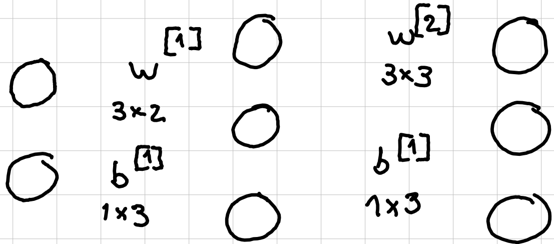
$$MAE_{Test\,Ridge} = \frac{1}{3}\left(|1 - 0{,}97319| + \cdots \right) = 0{,}616587$$

MAE(OLS) > MAE (Ridge), in both training and testing.

This means that Ridge has a better performance, therefore generalizing better.

The expected was that the MAE in Training would be better for OLS, because OLS is more prone to overfitting. We achived the oposite, because regularization smooths the coeficients, reducing the influence of outliers or extreme feature values. With a small dataset, this affect can lead to a improvement even on the training set.

4.



$$\eta = 0{,}5 \qquad x^{[0]} = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$$

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

Forward Propagation:

$$z^{[1]} = w^{[1]} \cdot x^{[0]} + b^{[1]} = \begin{bmatrix} 0{,}1 & 0{,}1 \\ 0{,}1 & 0{,}2 \\ 0{,}2 & 0{,}1 \end{bmatrix} \begin{bmatrix} 2 \\ 2 \end{bmatrix} + \begin{bmatrix} 0{,}1 \\ 0 \\ 0{,}1 \end{bmatrix} = \begin{bmatrix} 0{,}5 \\ 0{,}6 \\ 0{,}7 \end{bmatrix}$$

$$x^{[1]} = \text{Sigmoid}\left(\begin{bmatrix} 0{,}5 \\ 0{,}6 \\ 0{,}7 \end{bmatrix} \right) = \begin{bmatrix} 0{,}62246 \\ 0{,}64566 \\ 0{,}66819 \end{bmatrix}$$

$$z^{[2]} = w^{[2]} \cdot x^{[1]} + b^{[2]} = \begin{bmatrix} 1 & 2 & 2 \\ 1 & 2 & 1 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 0{,}62246 \\ 0{,}64566 \\ 0{,}66819 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 4{,}25016 \\ 3{,}58197 \\ 2{,}93631 \end{bmatrix}$$

$$x^{[2]} = \text{softmax}\left(z^{[2]} \right) = \begin{bmatrix} 0{,}56135 \\ 0{,}28777 \\ 0{,}15088 \end{bmatrix}$$

Back Propagation :

$$t = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \qquad x^{[2]} = \begin{bmatrix} 0{,}56135 \\ 0{,}28777 \\ 0{,}15088 \end{bmatrix}$$

$$w^{[2]} = w^{[2]} - \eta \frac{\partial E}{\partial w^{[2]}}$$

$$\frac{\partial E}{\partial w^{[2]}} = \frac{\partial E}{\partial z^{[2]}} \frac{\partial E}{\partial w^{[2]}} = \delta^{[2]} \cdot (x^{[1]})^T$$

$$\delta^{[2]} = \frac{\partial E}{\partial z^{[2]}} = x^{[2]} - t = \begin{bmatrix} 0{,}56135 \\ 0{,}28777 \\ 0{,}15088 \end{bmatrix} - \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} -0{,}43865 \\ 0{,}28777 \\ 0{,}15088 \end{bmatrix}$$

$$\frac{\partial E}{\partial w^{[2]}} = \begin{bmatrix} -0{,}43865 \\ 0{,}28777 \\ 0{,}15088 \end{bmatrix} \begin{bmatrix} 0{,}62246 & 0{,}64566 & 0{,}66819 \end{bmatrix}$$

$$= \begin{bmatrix} -0{,}27304 & -0{,}28322 & -0{,}29310 \\ 0{,}17913 & 0{,}18580 & 0{,}19229 \\ 0{,}09392 & 0{,}09742 & 0{,}10082 \end{bmatrix}$$

$$w^{[2]} = \begin{bmatrix} 1 & 2 & 2 \\ 1 & 2 & 1 \\ 1 & 1 & 1 \end{bmatrix} - 0{,}5 \begin{bmatrix} -0{,}27304 & -0{,}28322 & -0{,}29310 \\ 0{,}17913 & 0{,}18580 & 0{,}19229 \\ 0{,}09392 & 0{,}09742 & 0{,}10082 \end{bmatrix}$$

$$= \begin{bmatrix} 1{,}13652 & 2{,}14161 & 2{,}14655 \\ 0{,}91044 & 1{,}90710 & 0{,}90386 \\ 0{,}95304 & 0{,}95129 & 0{,}94959 \end{bmatrix}$$

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

$$b^{[2]} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} - 0{,}5 \begin{bmatrix} -0{,}43865 \\ 0{,}28777 \\ 0{,}15088 \end{bmatrix} = \begin{bmatrix} 1{,}21933 \\ 0{,}85612 \\ 0{,}92456 \end{bmatrix}$$

$$w^{[1]} = w^{[1]} - \eta \frac{\partial E}{\partial w^{[1]}}$$

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

derivada da sigmoid

$$\delta^{[1]} = \frac{\partial E}{\partial z^{[1]}} = (w^{[2]})^T \cdot \delta^{[2]} \circ \frac{\partial x^{[1]}}{\partial z^{[1]}}$$

$$= (w^{[2]})^T \cdot \delta^{[2]} \circ \sigma(z^{[1]}) \circ (1 - \sigma(z^{[1]}))$$

$$= (w^{[2]})^T \cdot \delta^{[2]} \circ x^{[1]} \circ (1 - x^{[1]})$$

$$= \begin{bmatrix} 1{,}13652 & 0{,}91044 & 0{,}95304 \\ 2{,}14161 & 1{,}907110 & 0{,}95129 \\ 2{,}14655 & 0{,}90386 & 0{,}94959 \end{bmatrix} \begin{bmatrix} -0{,}43865 \\ 0{,}28777 \\ 0{,}15088 \end{bmatrix} \circ x^{[1]} \circ (1 - x^{[1]})$$

$$= \begin{bmatrix} -0{,}09274 \\ -0{,}24708 \\ -0{,}53821 \end{bmatrix} \circ \begin{bmatrix} 0{,}62246 \\ 0{,}64566 \\ 0{,}66819 \end{bmatrix} \circ \left(\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} - \begin{bmatrix} 0{,}62246 \\ 0{,}64566 \\ 0{,}66819 \end{bmatrix} \right)$$

$$= \begin{bmatrix} -0{,}09274 \\ -0{,}24708 \\ -0{,}53821 \end{bmatrix} \circ \begin{bmatrix} 0{,}23514 \\ 0{,}22878 \\ 0{,}22171 \end{bmatrix} = \begin{bmatrix} -0{,}02181 \\ -0{,}05653 \\ -0{,}11933 \end{bmatrix}$$

$$\frac{\partial E}{\partial w^{[1]}} = \begin{bmatrix} -0,02181 \\ -0,05653 \\ -0,11933 \end{bmatrix} \begin{bmatrix} 2 & 2 \end{bmatrix} = \begin{bmatrix} -0,04368 & -0,04368 \\ -0,11306 & -0,11306 \\ -0,23866 & -0,23866 \end{bmatrix}$$

$$w^{[1]} = \begin{bmatrix} 0,1 & 0,1 \\ 0,1 & 0,2 \\ 0,2 & 0,1 \end{bmatrix} - 0,5 \begin{bmatrix} -0,04368 & -0,04368 \\ -0,11306 & -0,11306 \\ -0,23866 & -0,23866 \end{bmatrix}$$

$$= \begin{bmatrix} 0,12184 & 0,12184 \\ 0,15653 & 0,25653 \\ 0,31933 & 0,21933 \end{bmatrix}$$

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

$$b^{[1]} = \begin{bmatrix} 0,1 \\ 0 \\ 0,1 \end{bmatrix} - 0,5 \begin{bmatrix} -0,02181 \\ -0,05653 \\ -0,11933 \end{bmatrix} = \begin{bmatrix} 0,110905 \\ 0,028265 \\ 0,159665 \end{bmatrix}$$

Using a sigmoid activation function allows the MLP to learn non-linear relationships between the input features and the output classes, thus increasing its representational capacities. In addiction, the sigmoid outputs values in the range (0,1), which can be interpreted as class probabilities, making it suitable for classification tasks.

Without any activation function, the model reduces to a purely linear transformation an can only handle lineary seperable problems.