HW-3

I

1. OLS : $w_{ols} = (X^T X)^{-1} X^T y$

$$y = \begin{bmatrix} 3.5 \\ 1.0 \\ 3.8 \\ 10.1 \\ 8.5 \end{bmatrix} \qquad X = \begin{bmatrix} 1 & \phi_1 \\ 1 & \phi_2 \\ 1 & \phi_3 \\ 1 & \phi_4 \\ 1 & \phi_5 \end{bmatrix} = \begin{bmatrix} 1 & 4 \\ 1 & 1 \\ 1 & 6 \\ 1 & 18 \\ 1 & 8 \end{bmatrix}$$

$$X^T X = \begin{bmatrix} 5 & 37 \\ 37 & 441 \end{bmatrix} \qquad X^T y = \begin{bmatrix} 26,9 \\ 287,6 \end{bmatrix}$$

$$\det(X^T X) = 836 \qquad (X^T X)^{-1} = \frac{1}{836} \begin{bmatrix} 441 & -37 \\ -37 & 5 \end{bmatrix}$$

$$w_{OLS} = \frac{1}{836} \begin{bmatrix} 441 & -37 \\ -37 & 5 \end{bmatrix} \begin{bmatrix} 26.9 \\ 287.6 \end{bmatrix}$$

$$= \begin{bmatrix} 1,46136 \\ 0,52954 \end{bmatrix}$$

$$\hat{y}_{OLS} = 1,46136 + 0,52954(y_1, y_2)$$

2. $\Lambda = \begin{bmatrix} 0 & 0 \\ 0 & \lambda \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$ $\det[(X^TX + \Lambda)^{-1}] = 841$

Ridge: $w_{ridge} = (X^TX + \Lambda)^{-1} X^T y$

$w_{ridge} = (X^TX + \Lambda)^{-1} X^T y = \frac{1}{841} \begin{bmatrix} 441 & -37 \\ -37 & 5 \end{bmatrix} \begin{bmatrix} 26.7 \\ 287.6 \end{bmatrix}$

$= \begin{bmatrix} 1,48466 \\ 0,526397 \end{bmatrix}$

$\hat{y}_{ridge} = 1,48466 + 0,526397(y_1 y_2)$

Comparison: With ridge the coefficients shrink due to regularization (large coefficients get pulled towards zero).

3.

Train

- OLS: $\hat{y}_{ols} = 1,46136 + 0,52954(y_1 y_2)$
- Ridge: $\hat{y}_{ridge} = 1,48466 + 0,526397(y_1, y_2)$

$Error = y_{mean} - \hat{y}_{predict}$

x	$y_1 y_2$	$\hat{y}_{OLS}$	$\hat{y}_{ridge}$	$\lvert Error \rvert_{ols}$	$\lvert Error \rvert_{ridge}$
x_1	4	3,37952	3,598148	0,07952	0,09025
x_2	7	1,99090	2,011057	0,99090	1,011057
x_3	6	4,63860	4,643042	0,83860	0,84304
x_4	18	10,99308	10,95981	0,89308	0,85981
x_5	8	5,69768	5,69584	2,80232	2,80416

$$MAE_{train,OLS} = \frac{0,07952 + 0,99090 + 0,83860 + 0,89308 + 2,80232}{5} \simeq 1,12088$$

$$MAE_{train,ridge} = \frac{0,09025 + 1,011057 + 0,84304 + 0,85981 + 2,80416}{5} \simeq 1,12066$$

Test

x	$y_1 y_2$	$\hat{y}_{OLS}$	$\hat{y}_{ridge}$	$\lvert Error \rvert_{ols}$	$\lvert Error \rvert_{ridge}$
x_6	0	1,46136	1,48466	0,46136	0,48466
x_7	12	7,81584	7,80142	1,61584	1,601424
x_8	5	4,10906	4,17665	0,90906	0,91665

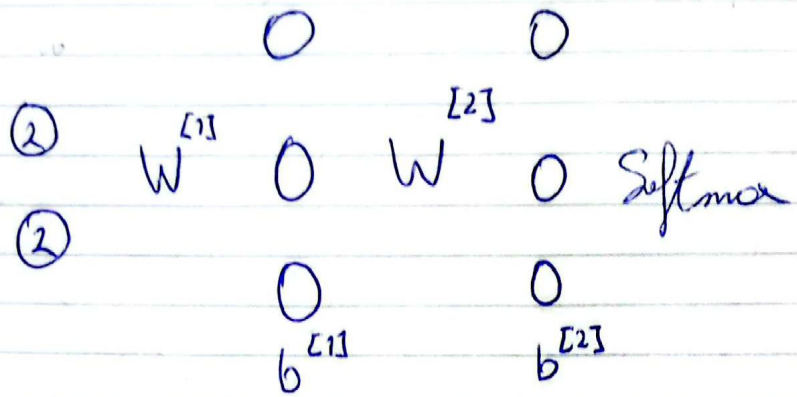
$$MAE_{test,OLS} = \frac{0,46136 + 1,61584 + 0,90906}{3} \simeq 0,86207$$

$$MAE_{test,ridge} = \frac{0,48466 + 1,601424 + 0,91665}{3} \simeq 0,86758$$

Explanation: The ridge model shows a slightly higher training MAE and a similar test MAE compared to OLS. This behavior is what we expect from regularization (Ridge adds a small bias aiming to reduce overfitting and improve stability on new data [although, that is not prominent with these results]).

HW-III

4.



Forward propagation:

$$Z^{[i]} = w^{[i]} \times X^{[i-1]} + b^{[i]}$$

1st label: $Z^{[1]} = \begin{bmatrix} 0,1 & 0.1 \\ 0.1 & 0,2 \\ 0.2 & 0.1 \end{bmatrix} \cdot \begin{bmatrix} 2 \\ 2 \end{bmatrix} + \begin{bmatrix} 0.1 \\ 0 \\ 0.1 \end{bmatrix} = \begin{bmatrix} 0,4 \\ 0,6 \\ 0,6 \end{bmatrix} + \begin{bmatrix} 0.1 \\ 0 \\ 0.1 \end{bmatrix}$

$$= \begin{bmatrix} 0.5 \\ 0,6 \\ 0.7 \end{bmatrix}$$

Sigmoid: $\sigma(x) = \frac{e^x}{1+e^x}$

Output: $\sigma\left(\begin{bmatrix} 0.5 \\ 0.6 \\ 0.7 \end{bmatrix}\right) = \begin{bmatrix} 0,62246 \\ 0,64566 \\ 0,66819 \end{bmatrix}$

$$Z^{[2]} = \begin{bmatrix} 1 & 2 & 2 \\ 1 & 2 & 1 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 0,62246 \\ 0,64566 \\ 0,66819 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 3.2502 \\ 2.58197 \\ 1.93631 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

$$= \begin{bmatrix} 4.2502 \\ 358197 \\ 2.93631 \end{bmatrix} \rightarrow \text{Softmax}$$

Softmax: $\frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$

$e^{4.2502} + e^{3.58197} + e^{2.93631} \approx 124.9099$

$e^{z_{[2]}}$

Find $\rightarrow X$: $\begin{bmatrix} 0.56136 \\ 0.28776 \\ 0.150878 \end{bmatrix}$

Back propagation:

Usando $x_1 \rightarrow t = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \begin{matrix} A \\ B \\ C \end{matrix}$

Atualização: $W^{[i]}_{novo} = W^{[i]} - \eta \frac{\partial \mathcal{E}}{\partial W^{[i]}}$

$\frac{\partial \mathcal{E}}{\partial W^{[i]}} = \delta^{[i]} (X^{[i-1]})^T$ $\qquad \frac{\partial \mathcal{E}}{\partial b^{[i]}} = \delta^{[i]}$

Output layer:

$\delta^{[2]} = \hat{y} - t = Soft\left(\begin{bmatrix} 4.2502 \\ 3.58197 \\ 2.93631 \end{bmatrix}\right) - \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0.56136 \\ 0.28776 \\ 0.150878 \end{bmatrix} - \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$

$= \begin{bmatrix} -0.43864 \\ 0.28776 \\ 0.150878 \end{bmatrix} = \frac{\partial \mathcal{E}}{\partial b^{[2]}}$

$\frac{\partial \mathcal{E}}{\partial W^{[2]}} = \begin{bmatrix} -0.43864 \\ 0.28776 \\ 0.150878 \end{bmatrix} \cdot$ [0.62246 0.64566 0.66819]

$= \begin{bmatrix} -0,27303 & -0.283212 & -0.293095 \\ 0,17912 & 0.185795 & 0.192278 \\ 0,093915 & 0.097416 & 0.100815 \end{bmatrix}$

$$W^{[2]}_{new} = \begin{bmatrix} 1 & 2 & 2 \\ 1 & 2 & 1 \\ 1 & 1 & 1 \end{bmatrix} - 0.5 \begin{bmatrix} -0,27203 & -0,283212 & -0.293095 \\ 0,17912 & 0.185795 & 0.192278 \\ 0,093915 & 0.097416 & 0.100815 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 2 & 2 \\ 1 & 2 & 1 \\ 1 & 1 & 1 \end{bmatrix} - \begin{bmatrix} -0.136515 & -0.14161 & 0.14655 \\ 0.08956 & 0.092898 & 0.09614 \\ 0.046958 & 0.048708 & 0.05041 \end{bmatrix}$$

$$\approx \begin{bmatrix} 1.136515 & 2.14161 & 2.14655 \\ 0.91044 & 1.907102 & 0.90386 \\ 0.953042 & 0.951292 & 0.94959 \end{bmatrix}$$

$$b^{[2]}_{new} = b^{[2]} - \eta \frac{\partial \mathcal{E}}{\partial b^{[2]}} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} - 0.5 \begin{bmatrix} 0.43864 \\ 0.28776 \\ 0.150878 \end{bmatrix}$$

$$= \begin{bmatrix} 1.21932 \\ 0.85612 \\ 0.924561 \end{bmatrix}$$

$$W^{[1]}_{new} = W^{[1]} - \eta \frac{\partial \mathcal{E}}{\partial W^{[1]}} \quad , \quad \frac{\partial \mathcal{E}}{\partial W^{[1]}} = \delta^{[1]} \left(x^{[0]}\right)^T$$

In this layer:

$$\delta^{[1]} = \left(W^{[2]}_{new}\right)^T \delta^{[2]} \cdot f'\left(z^{[1]}\right) \quad f' \rightarrow \text{derivada do sigmoid}$$

$$f'(z) = \sigma(z)(1 - \sigma(z))$$

$$\delta^{[1]} = \begin{bmatrix} 1.136515 & 2.14161 & 2.14655 \\ 0.91044 & 1.907102 & 0.90386 \\ 0.953042 & 0.951292 & 0.94959 \end{bmatrix}^T \begin{bmatrix} -0.43864 \\ 0.28776 \\ 0.150878 \end{bmatrix} \cdot f'\left(z^{[1]}\right)$$

$$= \begin{bmatrix} -0.09274 \\ -0.24708 \\ -0.53819 \end{bmatrix} \cdot \begin{bmatrix} 0,62246 \\ 0,64566 \\ 0,66819 \end{bmatrix} \cdot \left(\begin{bmatrix} 0.37754 \\ 0.35434 \\ 0.33181 \end{bmatrix} \right) =$$

$$= \begin{bmatrix} -0.021794 \\ -0.05653 \\ -0.119323 \end{bmatrix}$$

$$\frac{\partial E}{\partial W^{[1]}} = \begin{bmatrix} -0.021794 \\ -0.05653 \\ -0.119323 \end{bmatrix} \begin{bmatrix} 2 & 2 \end{bmatrix} = \begin{bmatrix} -0.04359 & -0.04359 \\ -0.11306 & -0.11306 \\ -0.23865 & -0.23865 \end{bmatrix}$$

$$W^{[1]}_{New} = \begin{bmatrix} 0.1 & 0.1 \\ 0.1 & 0.2 \\ 0.2 & 0.1 \end{bmatrix} - 0.5 \begin{bmatrix} -0.04359 & -0.04359 \\ -0.11306 & -0.11306 \\ -0.23865 & -0.23865 \end{bmatrix}$$

$$= \begin{bmatrix} 0.1218 & 0.1218 \\ 0.18653 & 0.25653 \\ 0.31933 & 0.21933 \end{bmatrix} //$$

To use a sigmoid activation, allows the MLP to learn non-linear patterns and relationships, allowing for better prediction and representation capabilities.