1.

D	y_1	y_2	y_{num}	y_{class}
x_1	1	1	1.25	B
x_2	1	3	7.0	A
x_3	3	2	2.7	C
x_4	3	3	3.2	A
x_5	2	4	5.5	B

$$\underbrace{w = (X^T \cdot X)^{-1} \cdot X^T \cdot z}_{\text{Moore-Penrose solution}}$$

$$X = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 3 \\ 1 & 3 & 2 \\ 1 & 3 & 3 \\ 1 & 2 & 4 \end{bmatrix} \qquad z = \begin{bmatrix} 1.25 \\ 7.0 \\ 2.7 \\ 3.2 \\ 55 \end{bmatrix}$$

$$w = \left(\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 3 & 3 & 2 \\ 1 & 3 & 2 & 3 & 4 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 3 \\ 1 & 3 & 2 \\ 1 & 3 & 3 \\ 1 & 2 & 4 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 3 & 3 & 2 \\ 1 & 3 & 2 & 3 & 4 \end{bmatrix} \begin{bmatrix} 1.25 \\ 7.0 \\ 2.7 \\ 3.2 \\ 55 \end{bmatrix} =$$

$$= \left(\begin{bmatrix} 5 & 10 & 13 \\ 10 & 24 & 27 \\ 13 & 27 & 39 \end{bmatrix} \right)^{-1} \bullet \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 3 & 3 & 2 \\ 1 & 3 & 2 & 3 & 4 \end{bmatrix} \begin{bmatrix} 1.25 \\ 7.0 \\ 2.7 \\ 3.2 \\ 55 \end{bmatrix} =$$

$$\begin{bmatrix} \frac{23}{11} & \frac{-13}{33} & \frac{-14}{33} \\ \frac{-13}{33} & \frac{26}{99} & \frac{-5}{99} \\ \frac{-14}{33} & \frac{-5}{99} & \frac{20}{99} \end{bmatrix} \bullet \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 3 & 3 & 2 \\ 1 & 3 & 2 & 3 & 4 \end{bmatrix} \begin{bmatrix} 1.25 \\ 7.0 \\ 2.7 \\ 3.2 \\ 55 \end{bmatrix} =$$

$$= \begin{bmatrix} \frac{14}{11} & \frac{14}{33} & \frac{2}{33} & \frac{-4}{11} & \frac{-13}{33} \\ \frac{-2}{11} & \frac{-28}{99} & \frac{29}{99} & \frac{8}{33} & \frac{-7}{99} \\ \frac{-3}{11} & \frac{13}{99} & \frac{-17}{99} & \frac{1}{33} & \frac{28}{29} \end{bmatrix} \begin{bmatrix} 1.25 \\ 7.0 \\ 2.7 \\ 3.2 \\ 55 \end{bmatrix} = \begin{bmatrix} \frac{46}{33} \\ \frac{-1019}{990} \\ \frac{3499}{1980} \end{bmatrix}$$

$$\text{output}(y_1, y_2) = \frac{46}{33} - \frac{1019}{990} \cdot y_1 + \frac{3499}{1980} \cdot y_2$$

2. Ridge regression key difference is in the loss function

$$E(w) = \frac{1}{2}\sum_{i=1}^{N} (z_i - \hat{z}_i)^2 + \frac{\lambda}{2} \|w\|^2$$

$$\Bigg\downarrow \text{Moore Penrose solution}$$

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

$$X = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 3 \\ 1 & 3 & 2 \\ 1 & 3 & 3 \\ 1 & 2 & 4 \end{bmatrix} \quad z = \begin{bmatrix} 1.25 \\ 7.0 \\ 2.7 \\ 3.2 \\ 55 \end{bmatrix} \quad \lambda = 1$$

$$w = \left(\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 3 & 3 & 2 \\ 1 & 3 & 2 & 3 & 4 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 3 \\ 1 & 3 & 2 \\ 1 & 3 & 3 \\ 1 & 2 & 4 \end{bmatrix} + \lambda \cdot I \right)^{-1} \cdot X^T \cdot z$$

$$= \left(\begin{bmatrix} 5 & 10 & 13 \\ 10 & 24 & 27 \\ 13 & 27 & 39 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \right)^{-1} \cdot X^T \cdot z$$

$$= \left(\begin{bmatrix} 6 & 10 & 13 \\ 10 & 25 & 27 \\ 13 & 27 & 40 \end{bmatrix} \right)^{-1} \cdot X^T \cdot z =$$

$$= \begin{bmatrix} \frac{271}{421} & \frac{-49}{421} & \frac{-55}{421} \\ \frac{-49}{421} & \frac{71}{421} & \frac{-32}{421} \\ \frac{-55}{421} & \frac{-32}{421} & \frac{50}{421} \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 3 & 3 & 2 \\ 1 & 3 & 2 & 3 & 4 \end{bmatrix} \begin{bmatrix} 1.25 \\ 7.0 \\ 2.7 \\ 3.2 \\ 55 \end{bmatrix} =$$

$$= \begin{bmatrix} \frac{167}{421} & \frac{57}{421} & \frac{14}{421} & \frac{-41}{421} & \frac{-47}{421} \\ \frac{-10}{421} & \frac{-74}{421} & \frac{100}{421} & \frac{68}{421} & \frac{-35}{421} \\ \frac{-34}{421} & \frac{63}{421} & \frac{-51}{421} & \frac{-1}{421} & \frac{81}{421} \end{bmatrix} \begin{bmatrix} 1.25 \\ 7.0 \\ 2.7 \\ 3.2 \\ 55 \end{bmatrix} =$$

$$= \begin{bmatrix} \frac{5117}{8420} \\ \frac{-1177}{2105} \\ \frac{13987}{8420} \end{bmatrix} \qquad w_1 = \begin{bmatrix} \frac{46}{33} \\ \frac{-1019}{990} \\ \frac{3499}{1980} \end{bmatrix} \quad w_2 = \begin{bmatrix} \frac{5117}{8420} \\ \frac{-1177}{2105} \\ \frac{13987}{8420} \end{bmatrix}$$

Now we compare the norm of weight vectors. Ridge has a lower norm.

$$\|w_1\| = \frac{4136619}{71280} = 6.1254$$

$$\|w_2\| = \frac{121992561}{35418200} = 3.441$$

Even if we remove the bias Ridge also has higher weight values.

$$\|w_1\| = \frac{3279284}{784080} = 4.18$$

$$\|w_2\| = \frac{217801433}{7086400} = 3.07$$

$b = 1.3939$
simple linear regression

$b = 0.6077$
Ridge regression

Both weights and bias are significantly penalized in Ridge regression.

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

$$x_6 = \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix} \quad x_7 = \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix} \quad x_8 = \begin{bmatrix} 1 \\ 5 \\ 1 \end{bmatrix}$$

Simple linear regression:

$$w = \begin{bmatrix} \frac{46}{33} \\ \frac{-1019}{990} \\ \frac{3499}{1980} \end{bmatrix} \quad \hat{z}_6 = \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix} \cdot \begin{bmatrix} \frac{46}{33} \\ \frac{-1019}{990} \\ \frac{3499}{1980} \end{bmatrix} = \frac{947}{330} = 2.8697$$

$$\hat{z}_7 = \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix} \cdot \begin{bmatrix} \frac{46}{33} \\ \frac{-1019}{990} \\ \frac{3499}{1980} \end{bmatrix} = \frac{386}{99} = 3.8990$$

$$\hat{z}_8 = \begin{bmatrix} 1 \\ 5 \\ 1 \end{bmatrix} \cdot \begin{bmatrix} \frac{46}{33} \\ \frac{-1019}{990} \\ \frac{3499}{1980} \end{bmatrix} = \frac{-3931}{1980} = -1.9854$$

$$RMSE = \sqrt{\frac{(0.7-2.8697)^2 + (1.1-3.8990)^2 + (2.2-(-1.9854))^2}{3}}$$

$$= \sqrt{\frac{4.7076 + 7.8344 + 4.1854}{3}} = 2.361$$

Ridge Regression

$$w = \begin{bmatrix} \frac{5117}{8420} \\ \frac{-1177}{2105} \\ \frac{13987}{8420} \end{bmatrix} \quad \hat{z}_6 = \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix} \cdot w = \frac{4735}{1684} = 2.8118$$

$$\hat{z}_7 = \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix} \cdot w = \frac{28383}{8420} = 3.3709$$

$$\hat{z}_8 = \begin{bmatrix} 1 \\ 5 \\ 1 \end{bmatrix} \cdot w = \frac{-1109}{2105} = -0.5268$$

$$\sqrt{\frac{(0.7-2.8118)^2 + (1.1-3.3709)^2 + (2.2+0.5268)^2}{3}} =$$

$$= \sqrt{\frac{4.4597}{3} + \frac{5.1570}{3} + \frac{7.43543824}{3}} =$$

$$= \sqrt{5.6840} = 2.384$$

Simple linear regression ⟶ 2.361

Ridge linear regression ⟶ 2.384

Ridge regression usually prevents overfitting. In this case it did not improve the model's accuracy so it is likely that linear regression was underfitting, that is why Ridge worsened the results.

4. $w^1 = \begin{bmatrix} 0.1 & 0.1 \\ 0.1 & 0.2 \\ 0.2 & 0.1 \end{bmatrix}$ $b^1 = \begin{bmatrix} 0.1 \\ 0 \\ 0.1 \end{bmatrix}$

$$W_2 = \begin{bmatrix} 1 & 2 & 2 \\ 1 & 2 & 1 \\ 1 & 1 & 1 \end{bmatrix} \quad b^2 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

$$x_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad W^1 x_1 + b^1 = \begin{bmatrix} 0.1 & 0.1 \\ 0.1 & 0.2 \\ 0.2 & 0.1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \begin{bmatrix} 0.1 \\ 0 \\ 0.1 \end{bmatrix} =$$

$$= \begin{bmatrix} 0.2 \\ 0.3 \\ 0.3 \end{bmatrix} + \begin{bmatrix} 0.1 \\ 0 \\ 0.1 \end{bmatrix} = \begin{bmatrix} 0.3 \\ 0.3 \\ 0.4 \end{bmatrix} \to z^1$$

$$W^2 z^1 + b^2 = \begin{bmatrix} 1 & 2 & 2 \\ 1 & 2 & 1 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 0.3 \\ 0.3 \\ 0.4 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} =$$

$$= \begin{bmatrix} \frac{17}{3} \\ \frac{13}{10} \\ 1 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{20}{3} \\ \frac{23}{10} \\ 2 \end{bmatrix} \to z^2$$

$$\sum_{i=1}^{N} e^{z^2_i} = e^{\frac{20}{3}} + e^{\frac{23}{10}} + e^2 = 803.1352$$

$$\frac{e^{\frac{20}{3}}}{803.1352} = 0.97838 \qquad \hat{z}_1 = A$$

$$\frac{e^{\frac{23}{10}}}{803.1352} = 0.01241 \qquad \begin{bmatrix} 0.97838 \\ 0.01241 \\ 0.00920 \end{bmatrix}$$

$$\frac{e^2}{803.1352} = 0.00920$$

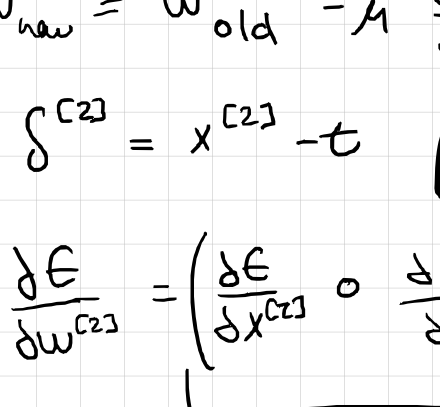
Our targets for x_1 are: $z_1 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$

To perform a stochastic gradient descent, we have to optimize the following function:

$$-\sum_{i=1}^{N} \sum_{c=1}^{|c|} t_c^i \log\left(z_c^{out\, i}\right) =$$

$$= -\left(t_2 \log z_2^{out}\right) = -\log z_2^{out}$$

$t_2 = 1$ $\qquad -\log\left(\text{softmax}\left(x_2^{out}\right)\right)$



Stochastic Gradient descent for x_1

$$W^{[L]}_{new} = W^{[L]}_{old} - \mu \frac{\partial E}{\partial W^{[L]}}$$

$$\delta^{[2]} = x^{[2]} - t \quad \boxed{\text{last layer}}$$

$$\frac{\partial E}{\partial W^{[2]}} = \underbrace{\left(\frac{\partial E}{\partial x^{[2]}} \circ \frac{\partial x^{[2]}}{\partial z^{[2]}}\right)}_{\delta^{[2]}} \cdot \left(\frac{\partial z^{[2]}}{\partial w^{[2]}}\right)^T$$

Last layer:

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

$$= \delta^{[2]} \left(x^{[1]}\right)^T$$

$$\delta^{[2]} = x^{[2]} - t = \begin{bmatrix} 0.97838 \\ 0.01241 \\ 0.00920 \end{bmatrix} - \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0.97838 \\ -0.98759 \\ 0.00920 \end{bmatrix}$$

$$\frac{\partial E}{\partial W^{[2]}} = \begin{bmatrix} 0.97838 \\ -0.98759 \\ 0.00920 \end{bmatrix} \begin{bmatrix} 0.3 & 0.3 & 0.4 \end{bmatrix} =$$

$$= \begin{bmatrix} 0.2935 & 0.2935 & 0.3914 \\ -0.2963 & -0.2963 & -0.3950 \\ 0.0028 & 0.0028 & 0.0037 \end{bmatrix}$$

$$W^{[2]}_{new} = W^{[2]}_{old} - \mu \frac{\partial E}{\partial W^{[2]}} =$$

$$= \begin{bmatrix} 1 & 2 & 2 \\ 1 & 2 & 1 \\ 1 & 1 & 1 \end{bmatrix} - 0.1 \begin{bmatrix} 0.2935 & 0.2935 & 0.3914 \\ -0.2963 & -0.2963 & -0.3950 \\ 0.0028 & 0.0028 & 0.0037 \end{bmatrix}$$

$$= \begin{bmatrix} 0.9706 & 1.9706 & 1.9609 \\ 1.0296 & 2.0296 & 1.0395 \\ 0.9997 & 0.9997 & 0.9996 \end{bmatrix}$$

Remaining layer

$$\delta^{[L]} = \left(\frac{\partial z^{[L+1]}}{\partial x^{[L]}}\right)^T \cdot \delta^{[L+1]} \circ \frac{\partial x^{[L]}}{\partial z^{[L]}}$$

$$= W^{[L+1]^T} \cdot \delta^{[L+1]} \circ \frac{\partial x^{[L]}}{\partial z^{[L]}} = W^{[L+1]^T} \cdot \delta^{[L+1]}$$

(no activation function)

$$\delta^{[1]} = W^{[2]T} \cdot \delta^2 =$$

$$= \begin{bmatrix} 0.9706 & 1.0296 & 0.9997 \\ 1.9706 & 2.0296 & 0.9997 \\ 1.9609 & 1.0395 & 0.9996 \end{bmatrix} \begin{bmatrix} 0.97838 \\ -0.98759 \\ 0.00920 \end{bmatrix}$$

$$= \begin{bmatrix} -0.0580 \\ -0.0672 \\ 0.9011 \end{bmatrix}$$

$$\frac{\partial E}{\partial W^{[L]}} = \delta^{[L]} \cdot \left(x^{[L-1]}\right)^T$$

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

$$= \begin{bmatrix} -0.0580 \\ -0.0672 \\ 0.9011 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1.25 \end{bmatrix} = \begin{bmatrix} -0.0580 & -0.0580 \\ -0.0672 & -0.0672 \\ 0.9011 & 0.9011 \end{bmatrix}$$

$$W^{[1]}_{new} = W^{[1]}_{old} - \mu \frac{\partial E}{\partial W^{[1]}} =$$

$$= \begin{bmatrix} 0.1 & 0.1 \\ 0.1 & 0.2 \\ 0.2 & 0.1 \end{bmatrix} - \begin{bmatrix} -0.0580 & -0.0580 \\ -0.0672 & -0.0672 \\ 0.9011 & 0.9011 \end{bmatrix} =$$

$$= \begin{bmatrix} 0.1580 & 0.1580 \\ 0.1672 & 0.2672 \\ -0.7011 & -0.8011 \end{bmatrix}$$

Bias update:

$$\frac{\partial E}{\partial b^{[L]}} = \delta^{[L]} \cdot \frac{\partial z^{[L]}}{\partial b^{[L]}} = \delta^{[L]}$$

$$b^{[2]}_{new} = b^{[2]}_{old} - \mu \frac{\partial E}{\partial b^{[2]}} = b^{[2]}_{old} - \mu \delta^{[2]}$$

$$= \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} - 0.1 \begin{bmatrix} 0.97838 \\ -0.98759 \\ 0.00920 \end{bmatrix} = \begin{bmatrix} 0.9022 \\ 1.9088 \\ 0.9991 \end{bmatrix}$$

$$b^{[1]}_{new} = b^{[1]}_{old} - \mu \frac{\partial E}{\partial b^{[1]}} = b^{[1]}_{old} - \mu \delta^{[1]}$$

$$= \begin{bmatrix} 0.1 \\ 0 \\ 0.1 \end{bmatrix} - 0.1 \begin{bmatrix} -0.0580 \\ -0.0672 \\ 0.9011 \end{bmatrix} = \begin{bmatrix} 0.1058 \\ 0.0067 \\ 0.0099 \end{bmatrix}$$

Summary of updates in ex. 4

$$w^{[1]} = \begin{bmatrix} 0.1580 & 0.1580 \\ 0.1672 & 0.2672 \\ -0.7011 & -0.8011 \end{bmatrix}$$

$$b^{[1]} = \begin{bmatrix} 0.1058 \\ 0.0067 \\ 0.0099 \end{bmatrix} \qquad b^{[2]} = \begin{bmatrix} 0.9022 \\ 1.9088 \\ 0.9991 \end{bmatrix}$$

$$w^{[2]} = \begin{bmatrix} 0.9706 & 1.9706 & 1.9609 \\ 1.0296 & 2.0296 & 1.0395 \\ 0.9997 & 0.9997 & 0.9996 \end{bmatrix}$$