Aprendizagem

LEIC IST-UL

RELATÓRIO - HOMEWORK 3

Grupo 10:

Gabriel Ferreira	107030
Irell Zane	107161

2024/2025 – 1st Semester, P1

Part I: Pen and paper

1. Given the polynomial basis function:

$$\phi(y_1, y_2) = y_1 \times y_2$$

We apply this to our input data:

$$\begin{aligned} x_1 &: \phi(1,1) = 1 \times 1 = 1 \\ x_2 &: \phi(1,3) = 1 \times 3 = 3 \\ x_3 &: \phi(3,2) = 3 \times 2 = 6 \\ x_4 &: \phi(3,3) = 3 \times 3 = 9 \\ x_5 &: \phi(2,4) = 2 \times 4 = 8 \end{aligned}$$

For OLS, we need **X** (input) and **y** (output) matrices:

$$\mathbf{X} = \begin{bmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 6 \\ 1 & 9 \\ 1 & 8 \end{bmatrix}$$

$$\mathbf{z} = \begin{bmatrix} 1.25 \\ 7.0 \\ 2.7 \\ 3.2 \\ 5.5 \end{bmatrix}$$

OLS closed form solution calculation

The OLS closed form solution is given by:

$$\boldsymbol{w} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{z}$$

Calculation of the separate components of formula:

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

$$(\mathbf{X}^T\mathbf{X})^{-1} \approx \begin{bmatrix} 0.845 & -0.119 \\ -0.119 & 0.022 \end{bmatrix}$$

$$\mathbf{X}^T\mathbf{z} = \begin{bmatrix} 19.65 \\ 111.25 \end{bmatrix}$$

Finally,

$$\begin{aligned} \boldsymbol{w} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{z} \\ &= \begin{bmatrix} 0.845 & -0.119 \\ -0.119 & 0.022 \end{bmatrix} \times \begin{bmatrix} 19.65 \\ 111.25 \end{bmatrix} \\ &\approx \begin{bmatrix} 3.316 \\ 0.114 \end{bmatrix} \end{aligned}$$

Therefore, the regression model in the transformed space is:

$$y_{num} = 3.316 + 0.114 \times \phi(y_1, y_2)$$

2. **Ridge regression closed form solution calculation**
The ridge regression closed form solution with penalty factor $\lambda = 1$ is given by:

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

Calculation of the separate components of formula:

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

$$\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I} = \begin{bmatrix} 5 & 27 \\ 27 & 191 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 6 & 27 \\ 27 & 192 \end{bmatrix}$$

$$(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1} \approx \begin{bmatrix} 0.454 & -0.064 \\ -0.064 & 0.014 \end{bmatrix}$$

$$\mathbf{X}^T\mathbf{z} = \begin{bmatrix} 19.65 \\ 111.25 \end{bmatrix}$$

Finally,

$$\begin{aligned} \boldsymbol{w} &= (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{z} \\ &= \begin{bmatrix} 0.454 & -0.064 \\ -0.064 & 0.014 \end{bmatrix} \times \begin{bmatrix} 19.65 \\ 111.25 \end{bmatrix} \\ &\approx \begin{bmatrix} 1.818 \\ 0.324 \end{bmatrix} \end{aligned}$$

Therefore, the ridge regression model in the transformed space with $\sigma = 1$ is:

$$y_{num} = 1.818 + 0.324 \times \phi(y_1, y_2)$$

The impact of ridge regression on the coefficients can be observed by comparing the Ordinary Least Squares (OLS) solution with the ridge regression solution (using $\lambda = 1$):

$$\text{OLS coefficients:} \quad \boldsymbol{w}_{\text{OLS}} \approx \begin{bmatrix} 3.316 \\ 0.114 \end{bmatrix}$$

$$\text{Ridge coefficients:} \quad \boldsymbol{w}_{\text{Ridge}} \approx \begin{bmatrix} 1.818 \\ 0.324 \end{bmatrix}$$

The effects of ridge regression are as follows: The larger coefficient (intercept) has been substantially reduced from 3.316 to 1.818. The smaller coefficient increased from. This aligns with the primary goal of ridge regression to shrink large coefficients, which are more heavily penalized due to the quadratic regularization term.

Ultimately the Ridge regression has reduced the sum of squared coefficients:

$$\begin{aligned} \text{OLS:} \quad & 3.316^2 + 0.114^2 \approx 11.00 \\ \text{Ridge:} \quad & 1.818^2 + 0.324^2 \approx 3.41 \end{aligned}$$

This significant reduction in the sum of squared coefficients indicates a less complex model, which is likely to generalize better to unseen data.

3. **OLS and Ridge prediction calculations**

For $x_6 = (2, 2, 0.7)$:

$$\begin{aligned} \phi(2, 2) &= 2 \times 2 = 4 \\ y_{OLS} &= 3.316 + 0.114 \times 4 = 3.772 \\ y_{Ridge} &= 1.818 + 0.324 \times 4 = 3.114 \end{aligned}$$

For $x_7 = (1, 2, 1.1)$:

$$\begin{aligned} \phi(1, 2) &= 1 \times 2 = 2 \\ y_{OLS} &= 3.316 + 0.114 \times 2 = 3.544 \\ y_{Ridge} &= 1.818 + 0.324 \times 2 = 2.466 \end{aligned}$$

For $x_8 = (5, 1, 2.2)$:

$$\begin{aligned}\phi(5,1) &= 5 \times 1 = 5\\ y_{OLS} &= 3.316 + 0.114 \times 5 = 3.886\\ y_{Ridge} &= 1.818 + 0.324 \times 5 = 3.438\end{aligned}$$

Observation	y_1	y_2	$\phi(y_1, y_2)$	y_{OLS}	y_{Ridge}
x_6	2	2	4	3.772	3.114
x_7	1	2	2	3.544	2.466
x_8	5	1	5	3.886	3.438

Table 1: Predicted values for test observations

RMSE calculation of both models

$$\begin{aligned}RMSE_{OLS} &= \sqrt{\frac{1}{3}\sum_{i=6}^{8}(y_i - \hat{y}_i)^2}\\ &= \sqrt{\frac{1}{3}[(0.7 - 3.772)^2 + (1.1 - 3.544)^2 + (2.2 - 3.886)^2]}\\ &\approx 2.467\end{aligned}$$

$$\begin{aligned}RMSE_{Ridge} &= \sqrt{\frac{1}{3}\sum_{i=6}^{8}(y_i - \hat{y}_i)^2}\\ &= \sqrt{\frac{1}{3}[(0.7 - 3.114)^2 + (1.1 - 2.466)^2 + (2.2 - 3.438)^2]}\\ &\approx 1.748\end{aligned}$$

The results show that the Ridge regression model has a lower RMSE (1.748) compared to the OLS model (2.467) for the given test data. This aligns with our expectations as the objective of the regularization in ridge regression was so that the model overfits the training data less and generalize better for unseen data.

4. MLP Propagation:

$$x^{[1]} = z^{[1]} = W[1]x^{[0]} + b^{[1]} = \begin{pmatrix} 0.1 & 0.1 \\ 0.1 & 0.2 \\ 0.2 & 0.1 \end{pmatrix} \begin{pmatrix} 1 & 1 \end{pmatrix} + \begin{pmatrix} 0.1 \\ 0 \\ 0.1 \end{pmatrix} = \begin{pmatrix} 0.3 \\ 0.3 \\ 0.4 \end{pmatrix}$$

$$z^{[2]} = W[2]x^{[1]} + b^{[2]} = \begin{pmatrix} 1 & 2 & 2 \\ 1 & 2 & 1 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 0.3 \\ 0.3 \\ 0.4 \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 2.7 \\ 2.3 \\ 2 \end{pmatrix}$$

$$x^{[2]} = softmax(z^{[2]}) = \begin{pmatrix} 0.46 \\ 0.31 \\ 0.23 \end{pmatrix}$$

Backpropagation:

$$\delta^{[2]} = \frac{\partial E}{\partial x^{[2]}} \circ \frac{\partial x^{[2]}}{\partial z^{[2]}} = x^{[2]} - t = \begin{pmatrix} 0.46 - 0 \\ 0.31 - 1 \\ 0.23 - 0 \end{pmatrix} = \begin{pmatrix} 0.46 \\ -0.69 \\ 0.23 \end{pmatrix}$$

Layer 2 weights:

$$\frac{\partial E}{\partial W^{[2]}} = \delta^{[2]}(x^{[1]})^T = \begin{pmatrix} 0.46 \\ -0.69 \\ 0.23 \end{pmatrix} \begin{pmatrix} 0.3 & 0.3 & 0.4 \end{pmatrix} = \begin{pmatrix} 0,14 & 0,14 & 0,18 \\ -0,21 & -0,21 & -0,28 \\ 0,07 & 0,07 & 0,09 \end{pmatrix}$$

$$W^{[2]}_{new} = W^{[2]}_{old} - \eta \frac{\partial E}{\partial W^{[2]}} = \begin{pmatrix} 1 & 2 & 2 \\ 1 & 2 & 1 \\ 1 & 1 & 1 \end{pmatrix} - 0.1 \begin{pmatrix} 0.81 & 0.81 & 1.08 \\ 0.39 & 0.39 & 0.52 \\ 0.6 & 0.6 & 0.8 \end{pmatrix} = \begin{pmatrix} 0.99 & 1.99 & 1.98 \\ 1.02 & 2.02 & 1.03 \\ 0.99 & 0.99 & 0.99 \end{pmatrix}$$

Layer 2 biases:

$$\frac{\partial E}{\partial b^{[2]}} = \delta^{[2]} \frac{\partial z^{[2]^T}}{\partial b^{[2]}} = \delta^{[2]} = \begin{pmatrix} 0.46 \\ -0.69 \\ 0.23 \end{pmatrix}$$

$$b^{[2]}_{new} = b^{[2]}_{old} - \eta \frac{\partial E}{\partial b^{[2]}} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} - 0.1 \begin{pmatrix} 0.46 \\ -0.69 \\ 0.23 \end{pmatrix} = \begin{pmatrix} 0.95 \\ 1.07 \\ 0.98 \end{pmatrix}$$

Layer 1 weights:

$$\delta^{[1]} = \left(\frac{\partial z^{[2]}}{\partial x^{[1]}}\right)^T \cdot \delta^{[2]} \circ \frac{\partial x^{[1]}}{\partial z^{[1]}} = (W^{[2]})^T \delta^{[2]} \times 1 = \begin{pmatrix} 0.00 \\ -0.23 \\ 0.46 \end{pmatrix}$$

$$\frac{\partial E}{\partial W^{[1]}} = \delta^{[1]}(x^{[0]})^T = \begin{pmatrix} 0.00 \\ -0.23 \\ 0.46 \end{pmatrix} \begin{pmatrix} 1 & 1 \end{pmatrix} = \begin{pmatrix} 0.00 & 0.00 \\ -0.23 & -0.23 \\ 0.46 & 0.46 \end{pmatrix}$$

$$W^{[1]}_{new} = W^{[1]}_{old} - \eta \frac{\partial E}{\partial W^{[1]}} = \begin{pmatrix} 0.1 & 0.1 \\ 0.1 & 0.2 \\ 0.2 & 0.1 \end{pmatrix} - 0.1 \begin{pmatrix} 0.00 & 0.00 \\ -0.23 & -0.23 \\ 0.46 & 0.46 \end{pmatrix} = \begin{pmatrix} 0.10 & 0.10 \\ 0.12 & 0.22 \\ 0.15 & 0.05 \end{pmatrix}$$

Layer 1 biases:

$$\frac{\partial E}{\partial b^{[1]}} = \delta^{[1]} \frac{\partial z^{[1]^T}}{\partial b^{[1]}} = \delta^{[1]} = \begin{pmatrix} 0.00 \\ -0.23 \\ 0.46 \end{pmatrix}$$

$$b^{[1]}_{new} = b^{[1]}_{old} - \eta \frac{\partial E}{\partial b^{[1]}} = \begin{pmatrix} 0.1 \\ 0 \\ 0.1 \end{pmatrix} - 0.1 \begin{pmatrix} 0.00 \\ -0.23 \\ 0.46 \end{pmatrix} = \begin{pmatrix} 0.10 \\ 0.02 \\ 0.05 \end{pmatrix}$$

Part II: Programming

5. Solution to the programming questions here.

End note: do not forget to also submit your Jupyter notebook