Aprendizagem 2022

Homework I – Group 22

Part I: Pen and paper

1. We can find the Ridge regression form solution with the following formula:

$$W = (\phi^T \phi + \lambda I)^{-1} \phi^T z$$

Being λI, z and ϕ the following:

$$\phi = \phi_j((0.8),(1),(1.2),(1.4),(1.6)) = \begin{bmatrix} 1 & 0.8 & 0.64 & 0.512 \\ 1 & 1 & 1 & 1 \\ 1 & 1.2 & 1.44 & 1.728 \\ 1 & 1.4 & 1.96 & 2.744 \\ 1 & 1.6 & 2.56 & 4.096 \end{bmatrix}$$

$$z = \begin{bmatrix} 24 \\ 20 \\ 10 \\ 13 \\ 12 \end{bmatrix}$$

$$\lambda I = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix}$$

Computing now W we get:

$$W = (\phi^T \phi + \lambda I)^{-1} \phi^T z =$$

$$= (\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 0.8 & 1 & 1.2 & 1.4 & 1.6 \\ 0.64 & 1 & 1.44 & 1.96 & 2.56 \\ 0.512 & 1 & 1.728 & 2.744 & 4.096 \end{bmatrix} \begin{bmatrix} 1 & 0.8 & 0.64 & 0.512 \\ 1 & 1 & 1 & 1 \\ 1 & 1.2 & 1.44 & 1.728 \\ 1 & 1.4 & 1.96 & 2.744 \\ 1 & 1.6 & 2.56 & 4.096 \end{bmatrix} + \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix})^{-1} \phi^T z =$$

$$= \begin{bmatrix} 0.342 & -0.121 & -0.075 & -0.009 \\ -0.121 & 0.389 & -0.097 & -0.074 \\ -0.075 & -0.097 & 0.373 & -0.171 \\ -0.009 & -0.074 & -0.171 & 0.18 \end{bmatrix} \phi^T z =$$

$$= \begin{bmatrix} 7.045 \\ 4.641 \\ 1.967 \\ -1.301 \end{bmatrix}$$

We can now write $\hat{z}(x) = 7.045 + 4.641x + 1.967x^2 - 1.301x^3$.

2. The formula to compute the RMSE is the following:

$$RMSE = \sqrt{\frac{\sum i(\hat{z}_i - z_i)^2}{N}}$$

Computing $\hat{z}$ for each observation:

$$\hat{z}(0.8) = 7.045 + 4.641 \times 0.8 + 1.967 \times 1.967 - 1.301 \times 0.512 = 11.351$$

$$\hat{z}(1) = 7.045 + 4.641 \times 1 + 1.967 \times 1 - 1.301 \times 1 = 12.352$$

$$\hat{z}(1.2) = 7.045 + 4.641 \times 1.2 + 1.967 \times 1.44 - 1.301 \times 1.728 = 13.199$$

$$\hat{z}(1.4) = 7.045 + 4.641 \times 1.4 + 1.967 \times 1.96 - 1.301 \times 2.744 = 13.829$$

$$\hat{z}(1.6) = 7.045 + 4.641 \times 1.6 + 1.967 \times 2.56 - 1.301 \times 4.096 = 14.179$$

$$(\hat{z}(0.8) - z(0.8))^2 = (11.351 - 24)^2 = 160.0$$

$$(\hat{z}(1) - z(1))^2 = (12.352 - 20)^2 = 58.492$$

$$(\hat{z}(1.2) - z(1.2))^2 = (13.199 - 10)^2 = 10.234$$

$$(\hat{z}(1.4) - z(1.4))^2 = (13.829 - 13)^2 = 0.687$$

$$(\hat{z}(1.6) - z(1.6))^2 = (14.179 - 12)^2 = 4.748$$

$$RMSE = \sqrt{\frac{160.0 + 58.492 + 10.234 + 0.687 + 4.748}{5}} = 6.84$$

3. Knowing that all weights and biases are initiliazed at 1 we can start by writing:

$$w^{[1]} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \qquad b^{[1]} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

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

We can now do the forward propagation on the 3 observations.

For that we need:

$$z^{[1]} = w^{[1]}x^{[0]} + b^{[1]}$$

$$x^{[1]} = f(z^{[1]}) = e^{z^{[1]}}$$
$$z^{[2]} = w^{[2]}x^{[1]} + b^{[2]}$$
$$x^{[2]} = f(z^{[2]}) = e^{z^{[2]}}$$

For the pair (x, z) = (0.8, 24):

$$x^{[0]} = 0.8$$
$$z^{[1]} = \begin{bmatrix}1\\1\end{bmatrix} 0.8 + \begin{bmatrix}1\\1\end{bmatrix} = \begin{bmatrix}1.8\\1.8\end{bmatrix}$$
$$x^{[1]} = f(\begin{bmatrix}1.8\\1.8\end{bmatrix}) = \begin{bmatrix}e^{0.18}\\e^{0.18}\end{bmatrix}$$
$$z^{[2]} = \begin{bmatrix}1 & 1\end{bmatrix} \begin{bmatrix}e^{0.18}\\e^{0.18}\end{bmatrix} + \begin{bmatrix}1\end{bmatrix} = 1 + 2e^{0.18}$$
$$x^{[2]} = f(1 + 2e^{0.18}) = e^{1+2e^{0.18}} = 1.404$$

For the pair (x, z) = (1, 20):

$$x^{[0]} = 1$$
$$z^{[1]} = \begin{bmatrix}1\\1\end{bmatrix} 1 + \begin{bmatrix}1\\1\end{bmatrix} = \begin{bmatrix}2\\2\end{bmatrix}$$
$$x^{[1]} = f(\begin{bmatrix}2\\2\end{bmatrix}) = \begin{bmatrix}e^{0.2}\\e^{0.2}\end{bmatrix}$$
$$z^{[2]} = \begin{bmatrix}1 & 1\end{bmatrix} \begin{bmatrix}e^{0.2}\\e^{0.2}\end{bmatrix} + \begin{bmatrix}1\end{bmatrix} = 1 + 2e^{0.2}$$
$$x^{[2]} = f(1 + 2e^{0.2}) = e^{1+2e^{0.2}} = 1.411$$

For the pair (x, z) = (1.2, 10):

$$x^{[0]} = 1.2$$
$$z^{[1]} = \begin{bmatrix}1\\1\end{bmatrix} 1.2 + \begin{bmatrix}1\\1\end{bmatrix} = \begin{bmatrix}2.2\\2.2\end{bmatrix}$$
$$x^{[1]} = f(\begin{bmatrix}2.2\\2.2\end{bmatrix}) = \begin{bmatrix}e^{0.22}\\e^{0.22}\end{bmatrix}$$
$$z^{[2]} = \begin{bmatrix}1 & 1\end{bmatrix} \begin{bmatrix}e^{0.22}\\e^{0.22}\end{bmatrix} + \begin{bmatrix}1\end{bmatrix} = 1 + 2e^{0.22}$$

$$x^{[2]} = f(1 + 2e^{0.22}) = e^{1+2e^{0.22}} = 1.418$$

Now, for the backward propagation, we are going to compute $\delta^{[1]}$ and $\delta^{[2]}$. As such, we are going to need the formula of the half squared error loss.

$$E(t, x^{[2]}) = \frac{1}{2}(x^{[2]} - t)^2$$

We are also going to need all the derivatives of the functions in the network.

$$\frac{\partial E}{\partial x^{[2]}}(t, x^{[2]}) = x^{[2]} - t$$

$$\frac{\partial x^{[l]}}{\partial z^{[l]}} = \frac{\partial e^{0.1z^{[l]}}}{\partial z^{[l]}} = 0.1e^{0.1z^{[l]}}$$

$$z^{[l]} = w^{[l]}x^{[l-1]} + b^{[l]}$$

$$\frac{\partial z^{[l]}}{\partial w^{[l]}} = x^{[l-1]}$$

$$\frac{\partial z^{[l]}}{\partial x^{[l-1]}} = w^{[l]}$$

$$\frac{\partial z^{[l]}}{\partial b^{[l]}} = 1$$

We can now define the expressions for $\delta^{[1]}$ and $\delta^{[2]}$

$$\delta^{[2]} = \frac{\partial E}{\partial x^{[2]}} \cdot \frac{\partial x^{[2]}}{\partial z^{[2]}} = (x^{[2]} - t) \cdot 0.1e^{0.1z^{[2]}}$$

$$\delta^{[1]} = \frac{\partial z^{[2]}}{\partial x^{[1]}}^T \cdot \delta^{[2]} \cdot \frac{\partial x^{[1]}}{\partial z^{[1]}} = (w^{[2]})^T \cdot \delta^{[2]} \cdot 0.1e^{0.1z^{[1]}}$$

Computing the value for the observations:

For the pair (x, z) = (0.8, 24):

$$\delta^{[2]} = (1.404 - 24) \cdot 0.1e^{0.1(1+2e^{0.18})} = -3.173$$

$$\delta^{[1]} = \begin{bmatrix}1\\1\end{bmatrix} \cdot (-3.173) \cdot \begin{bmatrix}0.1e^{0.1\times1.8}\\0.1e^{0.1\times1.8}\end{bmatrix} = \begin{bmatrix}-0.3799\\-0.3799\end{bmatrix}$$

For the pair (x, z) = (1, 20):

$$\delta^{[2]} = (1.411 - 20) \cdot 0.1e^{0.1(1+2e^{0.2})} = -2.623$$

$$\delta^{[1]} = \begin{bmatrix}1\\1\end{bmatrix} \cdot (-2.623) \cdot \begin{bmatrix}0.1e^{0.1\times2}\\0.1e^{0.1\times2}\end{bmatrix} = \begin{bmatrix}-0.3204\\-0.3204\end{bmatrix}$$

For the pair (x, z) = (1.2, 10):

$$\delta^{[2]} = (1.418 - 10) \cdot 0.1e^{0.1(1+2e^{0.22})} = -1.217$$

$$\delta^{[1]} = \begin{bmatrix}1\\1\end{bmatrix} \cdot (-1.217) \cdot \begin{bmatrix}0.1e^{0.1\times2.2}\\0.1e^{0.1\times2.2}\end{bmatrix} = \begin{bmatrix}-0.1516\\-0.1516\end{bmatrix}$$

Now we can update the weights values. For that we will need to compute the sum of $\frac{\partial E}{\partial w^{[1]}}$, $\frac{\partial E}{\partial w^{[2]}}$, $\frac{\partial E}{\partial b^{[1]}}$ and $\frac{\partial E}{\partial b^{[2]}}$ of all the observations.

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

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

$$\frac{\partial E}{\partial b^{[1]}} = \delta^{[1]} \cdot \frac{\partial z^{[1]}}{\partial b^{[1]}}^T = \delta^{[1]}$$

$$\frac{\partial E}{\partial b^{[2]}} = \delta^{[2]} \cdot \frac{\partial z^{[2]}}{\partial b^{[2]}}^T = \delta^{[2]}$$

For the pair (x, z) = (0.8, 24):

$$\frac{\partial E}{\partial w^{[1]}} = \begin{bmatrix}-0.3799\\-0.3799\end{bmatrix} \cdot 0.8 = \begin{bmatrix}-0.3039\\-0.3039\end{bmatrix}$$

$$\frac{\partial E}{\partial w^{[2]}} = (-3, 172) \cdot \begin{bmatrix}e^{0.18}\\e^{0.18}\end{bmatrix} = \begin{bmatrix}-3.799 & -3.799\end{bmatrix}$$

$$\frac{\partial E}{\partial b^{[1]}} = \begin{bmatrix}-0.3799\\-0.3799\end{bmatrix}$$

$$\frac{\partial E}{\partial b^{[2]}} = (-3, 172)$$

For the pair (x, z) = (1, 20):

$$\frac{\partial E}{\partial w^{[1]}} = \begin{bmatrix}-0.3203\\-0.3203\end{bmatrix} \cdot 1 = \begin{bmatrix}-0.3203\\-0.3203\end{bmatrix}$$

$$\frac{\partial E}{\partial w^{[2]}} = (-2.623) \cdot \begin{bmatrix}e^{0.2}\\e^{0.2}\end{bmatrix} = \begin{bmatrix}-3.204 & -3.204\end{bmatrix}$$

$$\frac{\partial E}{\partial b^{[1]}} = \begin{bmatrix} -0.3203 \\ -0.3203 \end{bmatrix}$$

$$\frac{\partial E}{\partial b^{[2]}} = (-2.623)$$

For the pair (x, z) = (1.2, 10):

$$\frac{\partial E}{\partial w^{[1]}} = \begin{bmatrix} -0.1516 \\ -0.1516 \end{bmatrix} \cdot 1.2 = \begin{bmatrix} -0.1820 \\ -0.1820 \end{bmatrix}$$

$$\frac{\partial E}{\partial w^{[2]}} = (-1.217) \cdot \begin{bmatrix} e^{0.22} \\ e^{0.22} \end{bmatrix} = \begin{bmatrix} -1.516 & -1.516 \end{bmatrix}$$

$$\frac{\partial E}{\partial b^{[1]}} = \begin{bmatrix} -0.1516 \\ -0.1516 \end{bmatrix}$$

$$\frac{\partial E}{\partial b^{[2]}} = (-1.217)$$

Computing the sum of the parameters of all observations:

$$\frac{\partial E}{\partial w^{[1]}{}_{total}} = \begin{bmatrix} -0.3039 \\ -0.3039 \end{bmatrix} + \begin{bmatrix} -0.3203 \\ -0.3203 \end{bmatrix} + \begin{bmatrix} -0.1820 \\ -0.1820 \end{bmatrix} = \begin{bmatrix} -0.8062 \\ -0.8062 \end{bmatrix}$$

$$\frac{\partial E}{\partial w^{[2]}{}_{total}} = \begin{bmatrix} -3.799 & -3.799 \end{bmatrix} + \begin{bmatrix} -3.204 & -3.204 \end{bmatrix} + \begin{bmatrix} -1.516 & -1.516 \end{bmatrix} = \begin{bmatrix} -8.518 & -8.518 \end{bmatrix}$$

$$\frac{\partial E}{\partial b^{[1]}{}_{total}} = \begin{bmatrix} -0.3799 \\ -0.3799 \end{bmatrix} + \begin{bmatrix} -0.3203 \\ -0.3203 \end{bmatrix} + \begin{bmatrix} -0.1516 \\ -0.1516 \end{bmatrix} = \begin{bmatrix} -0.8518 \\ -0.8518 \end{bmatrix}$$

$$\frac{\partial E}{\partial b^{[2]}{}_{total}} = -3,172 - 2.623 - 1.217 = -7.013$$

We can now update the weights. Being $\eta = 0.1$:

$$W^{[1]} = w^{[1]} - \eta \frac{\partial E}{\partial w^{[1]}{}_{total}} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} - 0.1 \begin{bmatrix} -0.8062 \\ -0.8062 \end{bmatrix} = \begin{bmatrix} 1.08 \\ 1.08 \end{bmatrix}$$

$$W^{[2]} = w^{[2]} - \eta \frac{\partial E}{\partial w^{[2]}{}_{total}} = \begin{bmatrix} 1 & 1 \end{bmatrix} - 0.1 \begin{bmatrix} -8.518 & -8.518 \end{bmatrix} = \begin{bmatrix} 1.85 & 1.85 \end{bmatrix}$$

$$B^{[1]} = b^{[1]} - \eta \frac{\partial E}{\partial b^{[1]}{}_{total}} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} - 0.1 \begin{bmatrix} -0.8518 \\ -0.8518 \end{bmatrix} = \begin{bmatrix} 1.09 \\ 1.09 \end{bmatrix}$$

$$B^{[2]} = b^{[2]} - \eta \frac{\partial E}{\partial b^{[2]}{}_{total}} = 1 - 0.1 \times (-7.013) = 1.70$$

Part II: Programming

4. The MAE of the Linear Regression is 0.162830, of the MLP with early stoppinig is 0.068041, and of the MLP without early stopping is 0.097807. Code solution is provided in Appendix(1).

5. Code solution is provided in Appendix(2). The histogram and the boxplot are provided in Appendix(3).

6. MLP_1 takes 452 iterations to converge. MLP_2 takes 77 iterations to converge. Code solution is provided in Appendix(4).

7. The unexpected differences on the number of iterations can be explained by the nature of the models. While the MLP without early stopping converges when the weights are no longer updated for a full cycle, the early stopping strategy doesn't converge until the validation score did not improve by at least 1×10^{-4} during the last 10 iterations. Moreover, while the first is only fitted to the training data (thus easier to reach convergence), the second model splits the training data into a training and a validation set. The model is then fitted on the training set and the prediction score is computed on the validation set. This means the first model is more likely to over-fit while the second has to adjust to the verification set every iteration, thus taking more to converge.

   Additionally, because the MLP with early stopping monitors convergence on a validation score, based on a validation set which doesn't belong to the training set, the models generalizes better to unseen data and reduces the chance of over-fitting the training data. This might explain the observed performance differences between the MLPs.

Appendix

1.

```python
import pandas as pd
from scipy.io.arff import loadarff

# Reading the ARFF file
data = loadarff('data/kin8nm.arff')
df = pd.DataFrame(data[0])

X = df.drop('y', axis=1)
y = df['y']
# Splitting the data into 70 30 training and test sets
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(
                                    X.values, y.values, test_size=0.3, random_state=0)
from sklearn.linear_model import Ridge

# linear regression with Ridge regularization
ridge = Ridge(alpha=0.1)
ridge.fit(X_train, y_train)
from sklearn.neural_network import MLPRegressor

# MLP1 with early stopping
mlp1 = MLPRegressor(hidden_layer_sizes=(10, 10),
                    activation='tanh',
                    max_iter=500,
                    random_state=0,
                    early_stopping=True)
mlp1.fit(X_train, y_train)

# MLP2 without early stopping
mlp2 = MLPRegressor(hidden_layer_sizes=(10, 10),
                    activation='tanh',
                    max_iter=500,
                    random_state=0,
                    early_stopping=False)
mlp2.fit(X_train, y_train)
from sklearn.metrics import mean_absolute_error

# compute mean absolute error
y_pred = ridge.predict(X_test)
print('Linear Regression MAE: %f' % mean_absolute_error(y_test, y_pred))

y_pred = mlp1.predict(X_test)
print('MLP MAE with early stopping: %f' % mean_absolute_error(y_test, y_pred))
```

```
y_pred = mlp2.predict(X_test)
print('MLP MAE without early stopping: %f' % mean_absolute_error(y_test, y_pred))
```

```
>> Linear Regression MAE: 0.162830
>> MLP MAE with early stopping: 0.068041
>> MLP MAE without early stopping: 0.097807
```

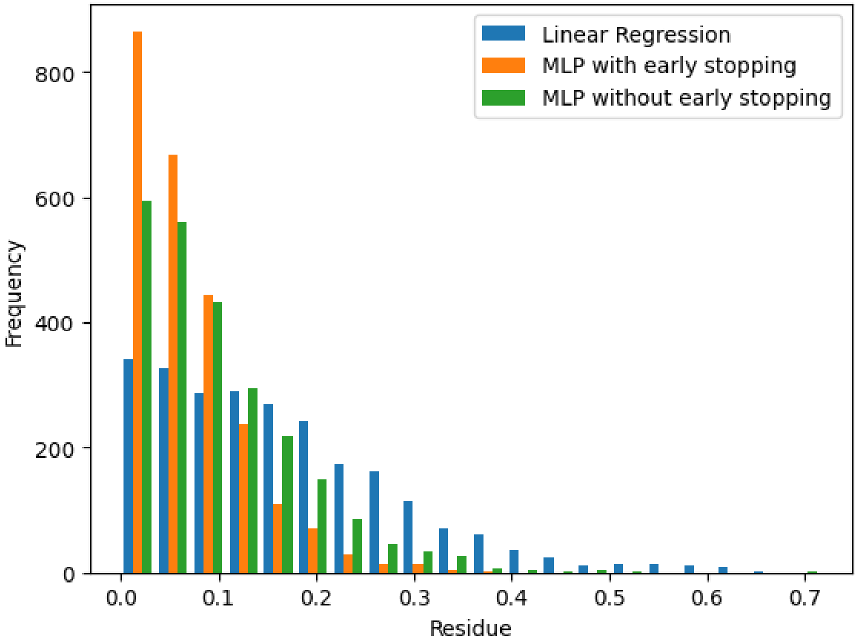
2.

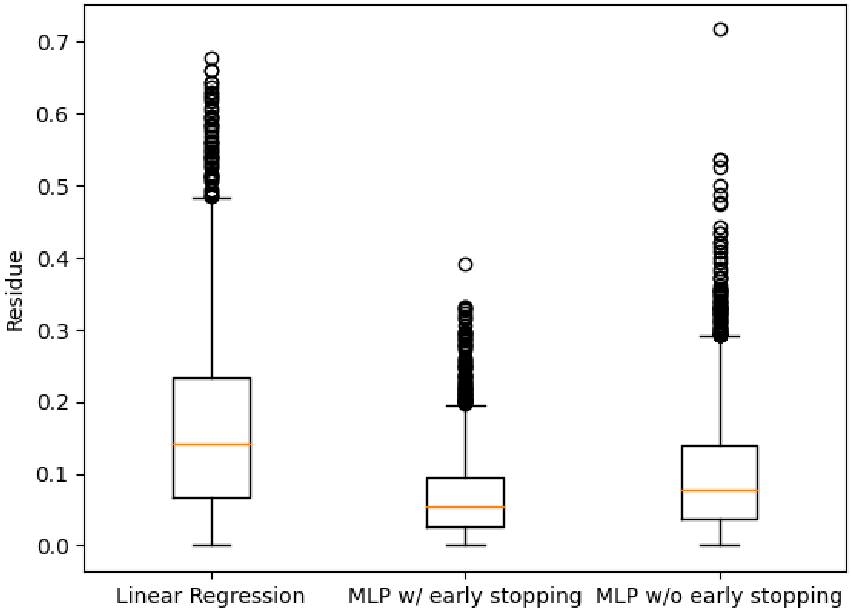
```
# Plot the residues (in absolute value) using two visualizations: boxplots and histograms
import matplotlib.pyplot as plt

# boxplot
plt.boxplot([abs(y_test - ridge.predict(X_test)),
                abs(y_test - mlp1.predict(X_test)),
                abs(y_test - mlp2.predict(X_test))])
plt.xticks([1, 2, 3], ['Linear Regression', 'MLP w/ early stopping', 'MLP w/o early stopping'])
plt.ylabel('Residue')
plt.show()

# histogram
plt.hist([abs(y_test - ridge.predict(X_test)),
            abs(y_test - mlp1.predict(X_test)),
            abs(y_test - mlp2.predict(X_test))],
        bins=20,
        label=['Linear Regression', 'MLP with early stopping', 'MLP without early stopping'])
plt.legend()
plt.xlabel('Residue')
plt.ylabel('Frequency')
plt.show()
```

3.





4.

```
#count number of iterations of mlps
print('MLP1 iterations: %d' % mlp1.n_iter_)
print('MLP2 iterations: %d' % mlp2.n_iter_)
```

```
>> MLP1 iterations: 452
>> MLP2 iterations: 77
```