

## Part I: Pen and paper

For questions in this group, show your numerical results with 5 decimals or scientific notation. Hint: we highly recommend the use of numpy (e.g., `linalg.pinv` for inverse) or other programmatic facilities to support the calculus involved in both questions (1) and (2).

Below is a training dataset  $D$  composed by two input variables and two output variables, one of which is numerical ( $y_{\text{num}}$ ) and the other categorical ( $y_{\text{class}}$ ). Consider a polynomial basis function  $\phi(y_1, y_2) = y_1 \times y_2$  that transforms the original space into a new one-dimensional space.

D	$y_1$	$y_2$	$y_{\text{num}}$	$y_{\text{class}}$
$x_1$	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

1. Learn a regression model on the transformed feature space using the OLS closed form solution to predict the continuous output variable  $y_{\text{num}}$ .

The first step will be to calculate the transformed feature space for each observation using the polynomial basis function:

$$\phi(y_1, y_2) = y_1 \times y_2 \tag{1}$$

By applying (1) to each observation, we get:

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

We end up with the following transformed feature space:  $\phi(y_1, y_2) = [1, 3, 6, 9, 8]$

The regression model in the transformed feature space is given by:

$$z = w_0 + w_1 \cdot \phi(y_1, y_2)$$

The next step is to learn a regression model using the OLS closed form solution which, for this exercise, is given by:

$$W = (\Phi^T \Phi)^{-1} \Phi^T y_{\text{num}}$$

where  $\Phi$  is the design matrix and  $y_{\text{num}}$  is the output variable.

$$\Phi = \begin{bmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 6 \\ 1 & 9 \\ 1 & 8 \end{bmatrix} \quad y_{\text{num}} = \begin{bmatrix} 1.25 \\ 7.0 \\ 2.7 \\ 3.2 \\ 5.5 \end{bmatrix}$$

As suggested in the hint, we will use `numpy` to calculate the weights  $W$ :

```
1 import numpy as np
2
3 M = np.array([[1, 1], [1, 3], [1, 6], [1, 9], [1, 8]])
4 y_num = np.array([1.25, 7.0, 2.7, 3.2, 5.5])
5
6 W = np.linalg.inv(M.T @ M) @ M.T @ y_num
7
8 W = np.round(W, 5) # round to 5 decimal digits
9 print(W)
```

The output of the code above is `[3.31593 0.11372]`.

Therefore, the weights are  $w_0 = 3.31593$  and  $w_1 = 0.11372$ .

The regression model is then given by:

$$y_{\text{num}} = 3.31593 + 0.11372 \cdot \phi(y_1, y_2)$$

2. Repeat the previous exercise, but this time learn a Ridge regression with penalty factor  $\lambda = 1$ . Compare the learnt coefficients with the ones from the previous exercise and discuss how regularization affects them.

The Ridge regression model is given by:

$$W = (\Phi^T \Phi + \lambda I)^{-1} \Phi^T y_{\text{num}}$$

where  $\Phi$  is the design matrix,  $y_{\text{num}}$  is the output variable (which remain the same as in the previous exercise), and  $\lambda$  is the penalty factor.

```
1 import numpy as np
2
3 M = np.array([[1, 1], [1, 3], [1, 6], [1, 9], [1, 8]])
4 lambda_value = 1.0
5 I = np.array([[1, 0], [0, 1]]) # Identity matrix
6 y_num = np.array([1.25, 7.0, 2.7, 3.2, 5.5])
7
8 W = np.linalg.inv(M.T @ M + lambda_value * I) @ M.T @ y_num
9
10 W = np.round(W, 5) # round to 5 decimal digits
11 print(W)
```

The output of the code above is [1.81809 0.32376].

Therefore, the weights are  $w_0 = 1.81809$  and  $w_1 = 0.32376$ .

The Ridge regression model is then given by:

$$y_{\text{num}} = 1.81809 + 0.32376 \cdot \phi(y_1, y_2)$$

Comparing the obtained weights in the OLS and Ridge regression models, we can see that in the Ridge regression model  $w_0$  is smaller and  $w_1$  is larger than the corresponding weights in the OLS model.

To check the impact of the regularization on the weights, we can calculate the corresponding normalized values:

$$\begin{aligned} w_{0\text{normalized}} &= \sqrt{3.31593^2 + 1.81809^2} = 3.78165 \\ w_{1\text{normalized}} &= \sqrt{0.11372^2 + 0.32376^2} = 0.34315 \end{aligned}$$

The normalized weights show that the Ridge regression model has a smaller overall weight magnitude compared to the OLS model. This is due to the regularization term in the Ridge regression model, which penalizes large weights and helps to prevent overfitting by reducing the complexity of the model. The regularization term shrinks the weights towards zero, which can help to improve the generalization of the model to new data.

### 3. Given three new test observations and their corresponding output

$\mathbf{x}_6 = (2, 2, 0.7)$ ,  $\mathbf{x}_7 = (1, 2, 1.1)$ , and  $\mathbf{x}_8 = (5, 1, 2.2)$ , compare the train and test RMSE of the two models obtained in (1) and (2). Explain if the results go according to what is expected.

The Root Mean Squared Error (RMSE) is given by:

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (z_i - \tilde{z}_i)^2}$$

where  $\tilde{z}_i$  is the predicted value,  $z_i$  is the real value and  $N$  is the number of observations.

OLS model:

Using the values of the weights obtained in the first exercise, we can calculate the predicted values for all observations:

$$\underbrace{\Phi = \begin{bmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 6 \\ 1 & 9 \\ 1 & 8 \end{bmatrix} \wedge w = \begin{bmatrix} 3.31593 \\ 0.11372 \end{bmatrix} \Rightarrow \tilde{z} = \begin{bmatrix} 3.42965 \\ 3.65709 \\ 3.99825 \\ 4.33941 \\ 4.22569 \end{bmatrix}}_{\text{train observations}} \quad \underbrace{\Phi = \begin{bmatrix} 1 & 4 \\ 1 & 2 \\ 1 & 5 \end{bmatrix} \wedge w = \begin{bmatrix} 3.31593 \\ 0.11372 \end{bmatrix} \Rightarrow \tilde{z} = \begin{bmatrix} 3.77081 \\ 3.54337 \\ 3.88453 \end{bmatrix}}_{\text{test observations}}$$

Finally, we take  $z = [1.25, 7.0, 2.7, 3.2, 5.5, 0.7, 1.1, 2.2]$  along with the predicted values to calculate the two RMSE for the OLS model.

$$\text{RMSE}_{\text{train}(x_1-x_5)} = \sqrt{\frac{1}{5} \sum_{i=1}^5 (z_i - \tilde{z}_i)^2} = \sqrt{\frac{1}{5} \times 25.0637837} = 2.2389$$

$$\text{RMSE}_{\text{test}(x_6-x_8)} = \sqrt{\frac{1}{3} \sum_{i=1}^3 (z_i - \tilde{z}_i)^2} = \sqrt{\frac{1}{3} \times 18.23757233} = 2.4656$$

Ridge Regression model:

Repeating the exact same process for the Ridge model, we get the following predicted values:

$$\underbrace{\Phi = \begin{bmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 6 \\ 1 & 9 \\ 1 & 8 \end{bmatrix} \wedge w = \begin{bmatrix} 1.81809 \\ 0.32376 \end{bmatrix} \Rightarrow \tilde{z} = \begin{bmatrix} 2.14185 \\ 2.78937 \\ 3.76065 \\ 4.73193 \\ 4.40817 \end{bmatrix}}_{\text{train observations}} \quad \underbrace{\Phi = \begin{bmatrix} 1 & 4 \\ 1 & 2 \\ 1 & 5 \end{bmatrix} \wedge w = \begin{bmatrix} 1.81809 \\ 0.32376 \end{bmatrix} \Rightarrow \tilde{z} = \begin{bmatrix} 3.11313 \\ 2.46561 \\ 3.43689 \end{bmatrix}}_{\text{test observations}}$$

We take  $z = [1.25, 7.0, 2.7, 3.2, 5.5, 0.7, 1.1, 2.2]$  along with the predicted values to calculate the RMSE for the Ridge model.

$$\text{RMSE}_{\text{train}(x_1-x_5)} = \sqrt{\frac{1}{5} \sum_{i=1}^5 (z_i - \tilde{z}_i)^2} = \sqrt{\frac{1}{5} \times 23.18868212} = 2.1535$$

$$\text{RMSE}_{\text{test}(x_6-x_8)} = \sqrt{\frac{1}{3} \sum_{i=1}^3 (z_i - \tilde{z}_i)^2} = \sqrt{\frac{1}{3} \times 9.217983941} = 1.7529$$

Comparing the RMSE values for the train and test observations, we can see that the Ridge regression model has a lower RMSE for both the train and test data compared to the OLS model. This suggests that the Ridge regression model makes more accurate predictions and fits the data better than the OLS model. The regularization term in the Ridge regression model helps to prevent overfitting and improve the generalization of the model to new data, resulting in lower prediction errors.

#### 4. Consider an MLP to predict the output $y_{\text{class}}$ characterized by the weights

$$W^{[1]} = \begin{bmatrix} 0.1 & 0.1 \\ 0.1 & 0.2 \\ 0.2 & 0.1 \end{bmatrix}, \quad b^{[1]} = \begin{bmatrix} 0.1 \\ 0 \\ 0.1 \end{bmatrix}, \quad W^{[2]} = \begin{bmatrix} 1 & 2 & 2 \\ 1 & 2 & 1 \\ 1 & 1 & 1 \end{bmatrix}, \quad b^{[2]} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

the output activation function

$$\text{softmax}(z_c^{[\text{out}]}) = \frac{e^{z_c^{[\text{out}]}}}{\sum_{l=1}^{|C|} e^{z_c^{[\text{out}]}}}$$

, no activations on the hidden layer(s) and the cross-entropy loss:

$$\text{CE} = - \sum_{i=1}^N \sum_{l=1}^{|C|} t_l^{(i)} \log(X_l^{[\text{out}](i)}) \quad (2)$$

Consider also that the output layer of the MLP gives the predictions for the classes A, B and C in this order. Perform one stochastic gradient descent update to all the weights and biases with learning rate  $\eta = 0.1$  using the training observation  $x_1$ .

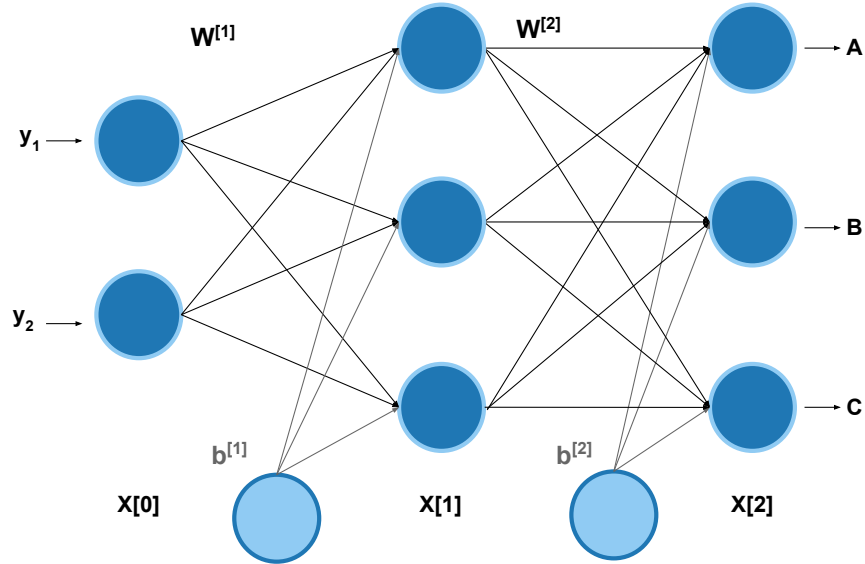


Figure 1: MLP architecture

We may also draw the corresponding scheme:

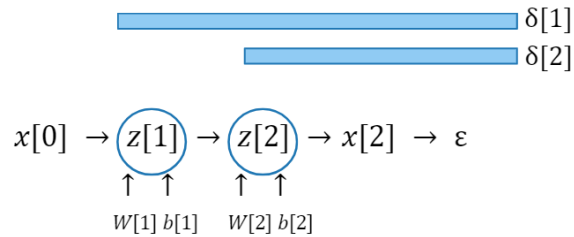


Figure 2: MLP Scheme

To begin, we initiate the process with the **forward propagation**. Below are the key equations to calculate the values of each layer,  $x_i^{[p]}$ :

$$z_i^{[p]} = W^{[p]} x_i^{[p-1]} + b^{[p]} \quad x_i^{[p]} = f(z_i^{[p]})$$

$f$  is the activation function for the output layer, provided in the statement.

$$f = \text{softmax}(z_c^{[\text{out}]}) = \frac{e^{z_c^{[\text{out}]}}}{\sum_{l=1}^{|C|} e^{z_l^{[\text{out}]}}}$$

In this notation, the value  $p$  is the index of the MLP layer.

We may now calculate the values of each node in the network for the training observation:

$$\boxed{x_1}$$

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

$$x^{[1]} = z^{[1]} = \begin{bmatrix} 0.3 \\ 0.3 \\ 0.4 \end{bmatrix}$$

$$z^{[2]} = W^{[2]} z^{[1]} + b^{[2]} = \begin{bmatrix} 1 & 2 & 2 \\ 1 & 2 & 1 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 0.3 \\ 0.3 \\ 0.4 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 2.7 \\ 2.3 \\ 2.0 \end{bmatrix}$$

$$x^{[2]} = f(z^{[2]}) = \begin{bmatrix} 0.461 \\ 0.309 \\ 0.229 \end{bmatrix}$$

Now, we can calculate the **backward propagation** to update the weights and biases.

Let's start by computing the derivatives for the  $z[i]$ :

$$\frac{\partial z^{[i]}(W^{[i]}, b^{[i]}, x^{[i-1]})}{\partial W^{[i]}} = x^{[i-1]} \quad \frac{\partial z^{[i]}(W^{[i]}, b^{[i]}, x^{[i-1]})}{\partial b^{[i]}} = 1 \quad \frac{\partial z^{[i]}(W^{[i]}, b^{[i]}, x^{[i-1]})}{\partial x^{[i-1]}} = W^{[i]}$$

Now we calculate the deltas (for the output layer and the hidden layer):

$$\begin{aligned}
\delta^{[2]} &= \mathbf{x}^{[2]} - t \\
&= \begin{pmatrix} 0.461 \\ 0.309 \\ 0.229 \end{pmatrix} - \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 0.461 \\ -0.691 \\ 0.229 \end{pmatrix} \\
\delta^{[1]} &= \left( \frac{\partial \mathbf{z}^{[2]}}{\partial \mathbf{x}^{[1]}} \right)^T \cdot \delta^{[2]} = (\mathbf{W}^{[2]})^T \cdot \delta^{[2]} \\
&= \begin{pmatrix} 1 & 1 & 1 \\ 2 & 2 & 1 \\ 2 & 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} 0.461 \\ -0.691 \\ 0.229 \end{pmatrix} \circ \begin{pmatrix} 0.3 \\ 0.3 \\ 0.4 \end{pmatrix} = \begin{pmatrix} -0.001 \\ -0.231 \\ 0.460 \end{pmatrix}
\end{aligned}$$

Finally, we can **update the weights and biases** with the gradient descent:

$$\begin{aligned}
\frac{\partial E}{\partial \mathbf{W}^{[1]}} &= \delta^{[1]} \cdot \frac{\partial \mathbf{z}^{[1]}}{\partial \mathbf{W}^{[1]}} = \delta^{[1]} \cdot (\mathbf{x}^{[0]})^T \\
&= \begin{pmatrix} -0.001 \\ -0.231 \\ 0.460 \end{pmatrix} \cdot (1 \quad 1) = \begin{pmatrix} -0.001 & -0.001 \\ -0.231 & -0.231 \\ 0.460 & 0.460 \end{pmatrix} \\
\mathbf{W}^{[1]} &= \mathbf{W}^{[1]} - \eta \cdot \frac{\partial E}{\partial \mathbf{W}^{[1]}} = \begin{pmatrix} 0.1 & 0.1 \\ 0.1 & 0.2 \\ 0.2 & 0.1 \end{pmatrix} - 0.1 \cdot \begin{pmatrix} -0.001 & -0.001 \\ -0.231 & -0.231 \\ 0.460 & 0.460 \end{pmatrix} = \begin{pmatrix} 0.1001 & 0.1001 \\ 0.1231 & 0.2231 \\ 0.1540 & 0.0540 \end{pmatrix} \\
\frac{\partial E}{\partial \mathbf{b}^{[1]}} &= \delta^{[1]} \cdot \frac{\partial \mathbf{z}^{[1]T}}{\partial \mathbf{b}^{[1]}} = \delta^{[1]} \\
\mathbf{b}^{[1]} &= \mathbf{b}^{[1]} - \eta \cdot \frac{\partial E}{\partial \mathbf{b}^{[1]}} = \begin{pmatrix} 0.1 \\ 0 \\ 0.1 \end{pmatrix} - 0.1 \cdot \begin{pmatrix} -0.001 \\ -0.231 \\ 0.460 \end{pmatrix} = \begin{pmatrix} 0.0999 \\ -0.0231 \\ 0.0540 \end{pmatrix}
\end{aligned}$$

$$\begin{aligned}\frac{\partial E}{\partial \mathbf{W}^{[2]}} &= \delta^{[2]} \cdot \frac{\partial \mathbf{z}^{[2]}}{\partial \mathbf{W}^{[2]}} = \delta^{[2]} \cdot (\mathbf{z}^{[1]})^T \\ &= \begin{pmatrix} 0.461 \\ -0.691 \\ 0.229 \end{pmatrix} \cdot (0.3 \quad 0.3 \quad 0.4) = \begin{pmatrix} 0.1383 & 0.1383 & 0.1844 \\ -0.2073 & -0.2073 & -0.2764 \\ 0.0687 & 0.0687 & 0.0916 \end{pmatrix}\end{aligned}$$

$$\begin{aligned}\mathbf{W}^{[2]} &= \mathbf{W}^{[2]} - \eta \cdot \frac{\partial E}{\partial \mathbf{W}^{[2]}} = \begin{pmatrix} 1 & 2 & 2 \\ 1 & 2 & 1 \\ 1 & 1 & 1 \end{pmatrix} - 0.1 \cdot \begin{pmatrix} 0.1383 & 0.1383 & 0.1844 \\ -0.2073 & -0.2073 & -0.2764 \\ 0.0687 & 0.0687 & 0.0916 \end{pmatrix} \\ &= \begin{pmatrix} 0.9862 & 1.9862 & 1.9816 \\ 1.0207 & 2.0207 & 1.0276 \\ 0.9931 & 0.9931 & 0.9908 \end{pmatrix}\end{aligned}$$

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

$$\mathbf{b}^{[2]} = \mathbf{b}^{[2]} - \eta \cdot \frac{\partial E}{\partial \mathbf{b}^{[2]}} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} - 0.1 \cdot \begin{pmatrix} 0.461 \\ -0.691 \\ 0.229 \end{pmatrix} = \begin{pmatrix} 0.9539 \\ 1.0691 \\ 0.9771 \end{pmatrix}$$

## Part II: Programming

Consider the parkinsons.csv dataset (available at the course's webpage), where the goal is to predict a patient's score on the Unified Parkinson's Disease Rating Scale based on various biomedical measurements. To answer question (5), average the performance of the models over 10 separate runs. In each run, use a different 80 – 20 train test split by setting a `random_state = i`, with `i = 1...10`.

5. Train a Linear Regression model, an MLP Regressor with 2 hidden layers of 10 neurons each and no activation functions, and another MLP Regressor with 2 hidden layers of 10 neurons each using ReLU activation functions. (Use `random_state=0` on the MLPs, regardless of the run). Plot a boxplot of the test MAE of each model.

```
1 import matplotlib.pyplot as plt, pandas as pd
2 from sklearn.model_selection import train_test_split
3 from sklearn.linear_model import LinearRegression
4 from sklearn.metrics import mean_absolute_error
5 from sklearn.neural_network import MLPRegressor
6
```



```

7 # Read the dataset
8 df = pd.read_csv("./parkinsons.csv")
9 X, y = df.drop("target", axis=1), df["target"]
10
11 linear_mae = []
12
13 # Linear Regression
14 for i in range(1, 11):
15     X_train, X_test, y_train, y_test = train_test_split(X, y, test_size
16                                                         =0.2, random_state=i)
17
18     lr_model = LinearRegression()
19     lr_model.fit(X_train, y_train)
20
21     y_pred = lr_model.predict(X_test)
22     mae = mean_absolute_error(y_test, y_pred)
23     linear_mae.append(mae)
24
25 mlp_no_activation_mae = []
26
27 # mlps with no activation function
28 for i in range(1, 11):
29     X_train, X_test, y_train, y_test = train_test_split(X, y, test_size
30                                                         =0.2, random_state=i)
31
32     mlp_no_activation = MLPRegressor(hidden_layer_sizes=(10, 10),
33                                     activation='identity',
34                                     random_state=0, validation_fraction
35                                     =0.2)
36     mlp_no_activation.fit(X_train, y_train)
37
38     y_pred = mlp_no_activation.predict(X_test)
39     mae = mean_absolute_error(y_test, y_pred)
40     mlp_no_activation_mae.append(mae)
41
42 mlp_relu_mae = []
43
44 # mlps with ReLU activation function
45 for i in range(1, 11):
46     X_train, X_test, y_train, y_test = train_test_split(X, y, test_size
47                                                         =0.2, random_state=i)
48
49     mlp_relu = MLPRegressor(hidden_layer_sizes=(10, 10), activation='relu',
50                             random_state=0, validation_fraction=0.2)
51     mlp_relu.fit(X_train, y_train)
52
53     y_pred = mlp_relu.predict(X_test)
54     mae = mean_absolute_error(y_test, y_pred)
55     mlp_relu_mae.append(mae)
56
57 # Data for plotting
58 b_plot = plt.boxplot(
59     [linear_mae, mlp_no_activation_mae, mlp_relu_mae], patch_artist=True,
60     labels=['Linear Regression', 'MLP No Activation', 'MLP ReLU']
61 )

```

```

57
58 colors = ["#1f77b4", "#E40071"]
59 for patch, color in zip(b_plot["boxes"], colors):
60     patch.set_facecolor(color)
61 for median in b_plot["medians"]:
62     median.set_color("black")
63 # Create boxplot
64 plt.title('Test MAE of Models')
65 plt.ylabel('Mean Absolute Error')
66 plt.grid(axis="y")
67 plt.show()

```

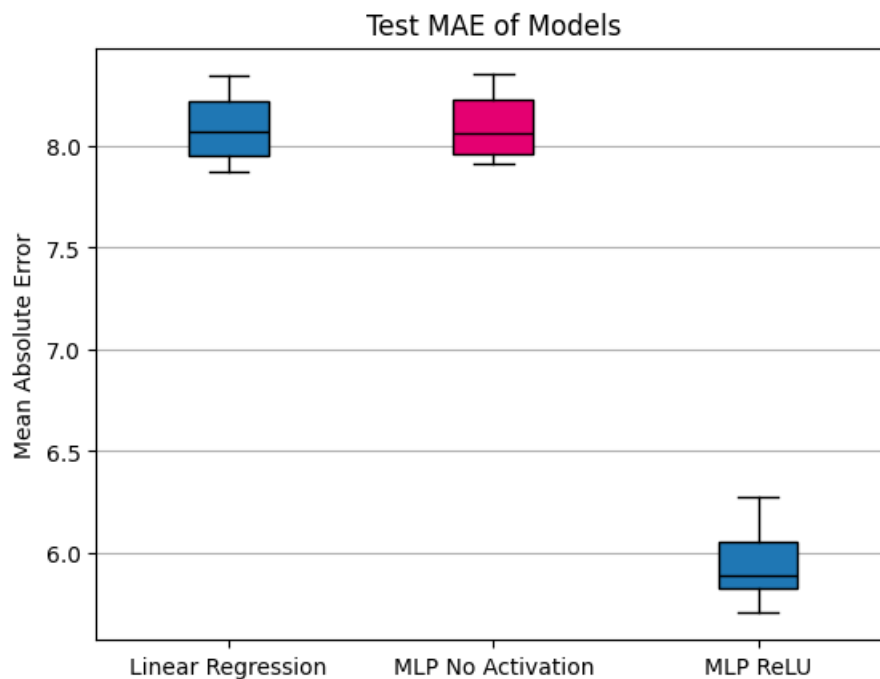


Figure 3: Boxplot of the test MAE for each model

6. Compare a Linear Regression with a MLP with no activations, and explain the impact and the importance of using activation functions in a MLP. Support your reasoning with the results from the boxplots.

As we can see in the previous plot, the MLP without activation functions behaves similarly to Linear Regression, while the MLP with ReLU activation function performs better than the other two models. Linear Regression and MLP without activation functions have a high MAE, centered around 8.0. On the other hand, the MLP with ReLU activation function has a lower MAE, centered around 6.0 and this shows the power of activation functions. Activation functions enable the MLP to learn from non-linear relationships in the data. Without them, each layer is just performing a linear operation, which severely limits the network's ability to generalize and fit real-world data, especially when the underlying relationships are complex or non-linear.

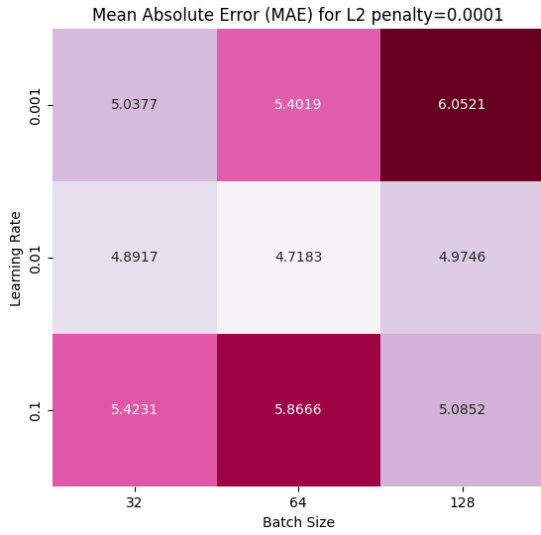
7. Using a 80 – 20 train-test split with `random_state=0`, use a Grid Search to tune the hyperparameters of an MLP regressor with two hidden layers (size 10 each). The parameters to search over are: (i) L2 penalty, with the values  $\{0.0001, 0.001, 0.01\}$ ; (ii) learning rate, with the values  $\{0.001, 0.01, 0.1\}$ ; and (iii) batch size, with the values  $\{32, 64, 128\}$ . Plot the test MAE for each combination of hyperparameters, report the best combination, and discuss the trade-offs between the combinations.

```
1 import pandas as pd
2 from sklearn.model_selection import train_test_split, GridSearchCV
3 from sklearn.neural_network import MLPRegressor
4 from sklearn.metrics import mean_absolute_error
5 import matplotlib.pyplot as plt
6 import seaborn as sns
7
8 data = pd.read_csv('parkinsons.csv')
9
10 df = pd.read_csv("./parkinsons.csv")
11 X, y = df.drop("target", axis=1), df["target"]
12
13 X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
14 random_state=0)
15
16 mlp = MLPRegressor(hidden_layer_sizes=(10, 10), random_state=0)
17
18 # Define the hyperparameters to search
19 param_grid = {
20     'alpha': [0.0001, 0.001, 0.01], # L2 penalty
21     'learning_rate_init': [0.001, 0.01, 0.1], # Learning rate
22     'batch_size': [32, 64, 128], # Batch size
23 }
24
25 # Perform grid search
26 grid_search = GridSearchCV(mlp, param_grid, cv=3, scoring='
27     neg_mean_absolute_error')
28 grid_search.fit(X_train, y_train)
29
30 best_params = grid_search.best_params_
31 best_model = grid_search.best_estimator_
32
33 y_pred = best_model.predict(X_test)
34
35 # Evaluate the model using Mean Absolute Error (MAE)
36 test_mae = mean_absolute_error(y_test, y_pred)
37 print(f"Test MAE: {test_mae}")
38 print(f"Best Hyperparameters: {best_params}")
39
40 results = pd.DataFrame(grid_search.cv_results_)
41 results['mean_test_mae'] = -results['mean_test_score'] # Convert to
42 positive MAE
43
44 # Plot a heatmap for each value of alpha
45 alphas = results['param_alpha'].unique()
46
```

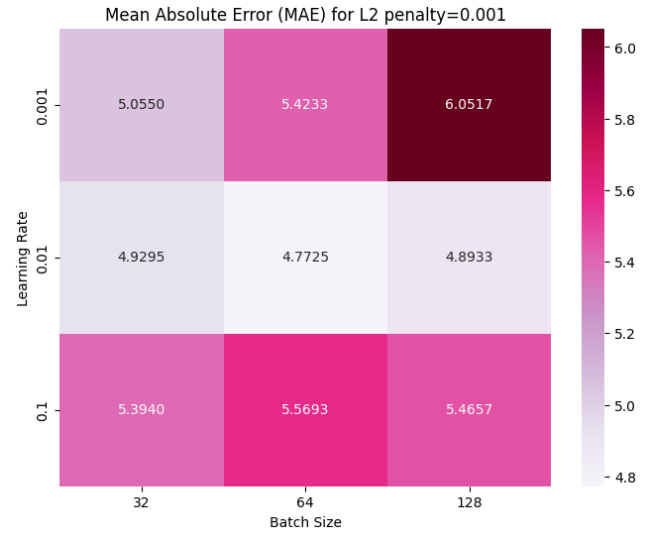
```

44 for alpha in alphas:
45     subset = results[results['param_alpha'] == alpha]
46     pivot_table = subset.pivot(index='param_learning_rate_init', columns='
param_batch_size', values='mean_test_mae')
47
48     # Plot the heatmap
49     fig, ax = plt.subplots(figsize=(8, 6))
50     sns.heatmap(pivot_table, annot=True, fmt=".4f", cmap="PuRd", ax=ax)
51     plt.title(f'Mean Absolute Error (MAE) for L2 penalty={alpha}')
52     plt.xlabel('Batch Size')
53     plt.ylabel('Learning Rate')
54     plt.show()

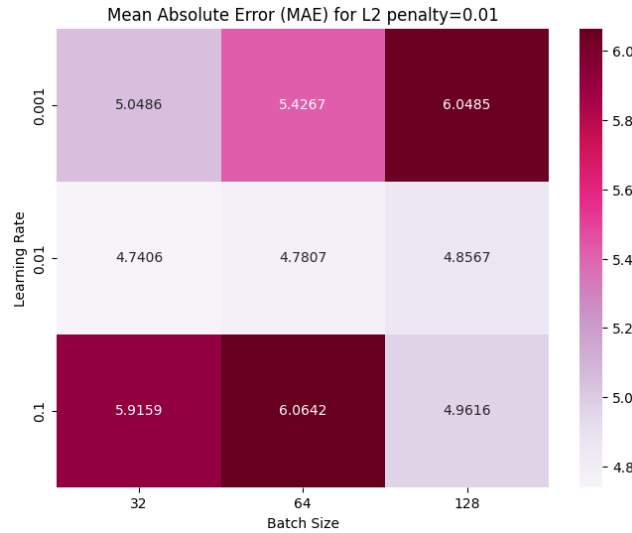
```



(a) MAE plot for L2 penalty = 0.0001



(b) MAE plot for L2 penalty = 0.001



(c) MAE plot for L2 penalty = 0.01

Figure 4: MAE plot for each combination of hyperparameters

When analyzing MAE values, the smaller the MAE the better the models performance. With that being said, the best combination for each heatmap is:

- Learning Rate = 0.01 and Batch Size = 64 for L2 penalty = 0.0001
- Learning Rate = 0.01 and Batch Size = 64 for L2 penalty = 0.001
- Learning Rate = 0.01 and Batch Size = 32 for L2 penalty = 0.01

The best combination of hyperparameters for the MLP regressor, achieving the lowest Mean Absolute Error (MAE) of 4.7183, consists of an L2 penalty (alpha) of 0.0001, a learning rate of 0.01, and a batch size of 64. Lower alphas help avoid over-penalizing the model, while the 0.01 learning rate offers a good balance between convergence speed and avoiding overshooting. Batch size 64 provides the best generalization. Higher alphas and learning rates led to higher errors due to over-regularization and poor convergence.