Faculty of Engineering and Technology

Electrical and Computer Engineering Department

ENCS5341

Machine Learning and Data Science

Assignment No.2

Model Selection and Hyper-parameters Tunning & Logistic Regression

---

Student's Name: Lojain Abdalrazaq. **ID:** 1190707.

Instructor's Name: Dr. Yazan Abu Farha.

Section: 2.

December 18, 2023

Summery:

Firstly, the data_reg.csv which contains the 200 examples. Each row has the two attributes x1,x2, and continuous target value y.

```
x1,x2,y
0.5488135039273248,0.3117958819941026,0.5478181081735846
0.7151893663724195,0.6963434888154595,0.5760317429256328
0.6027633760716439,0.3777518392924809,0.1134753987640586
0.5448831829968969,0.1796036775596348,1.072285921314013
0.4236547993389047,0.02467872839133123,0.6245245484798667
0.6458941130666561,0.06724963146324858,0.7838764531861209
```

Question 1: Model Selection and Hyper-parameters Tunning

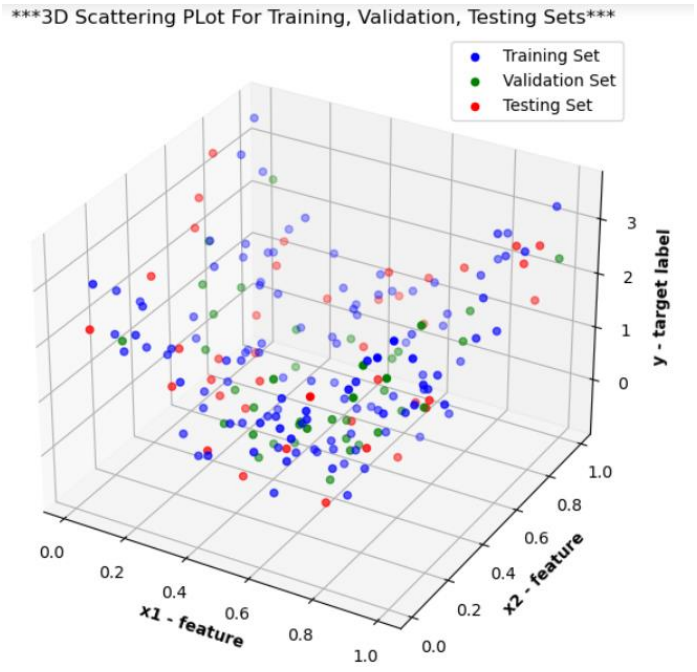
Reading the data from the csv file and split it into:

300 Total Examples
120: Training Set
40: Validating Set
40: Testing Set

Plot the examples form the three sets in a scatter plot (each set with a different color), and the plot will be 3D (x1 = x, x2 = y, z = y label).

Solution:

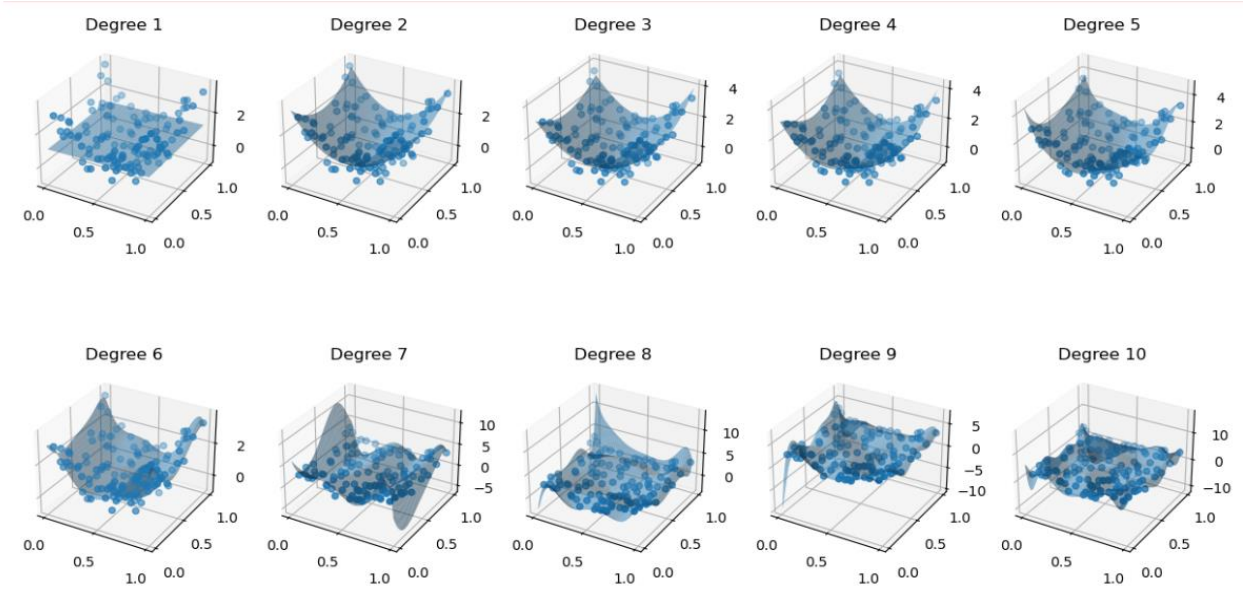
Firstly, the data from the data_reg.csv file was loaded, and divided into 3 sets; Training set, Validation Set, and Testing set. Then, the 3D scatter plot was plotted such that each set is marked in a different color.



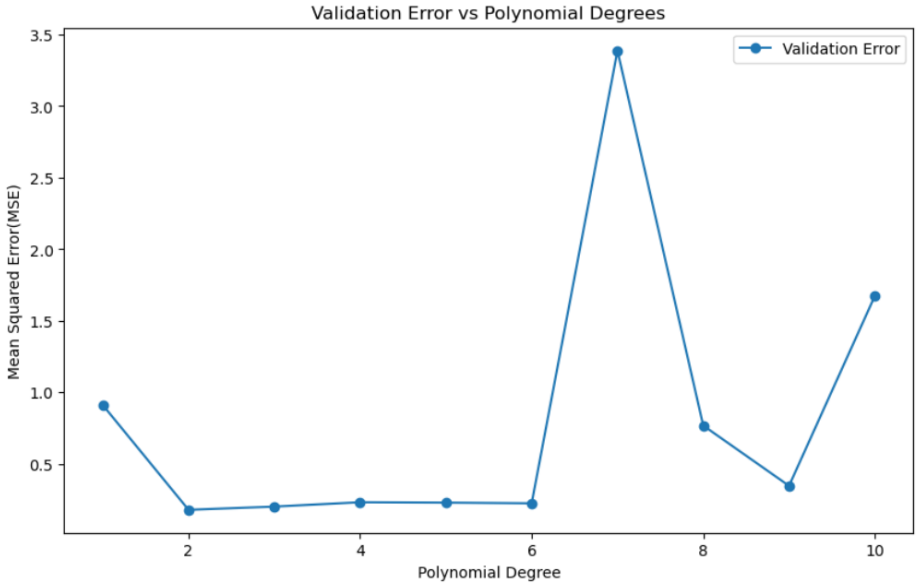
Applying the polynomial regression on the Training Set with degrees in the range of 1 to 10.

The objective of this part applying the polynomial regression on the training set with the degree range (1 – 10) degrees in order to find the best polynomial degree. In addition, the surface of each model was plotted. The following figures show the results:

- The surfaces for each polynomial degree are as the following:



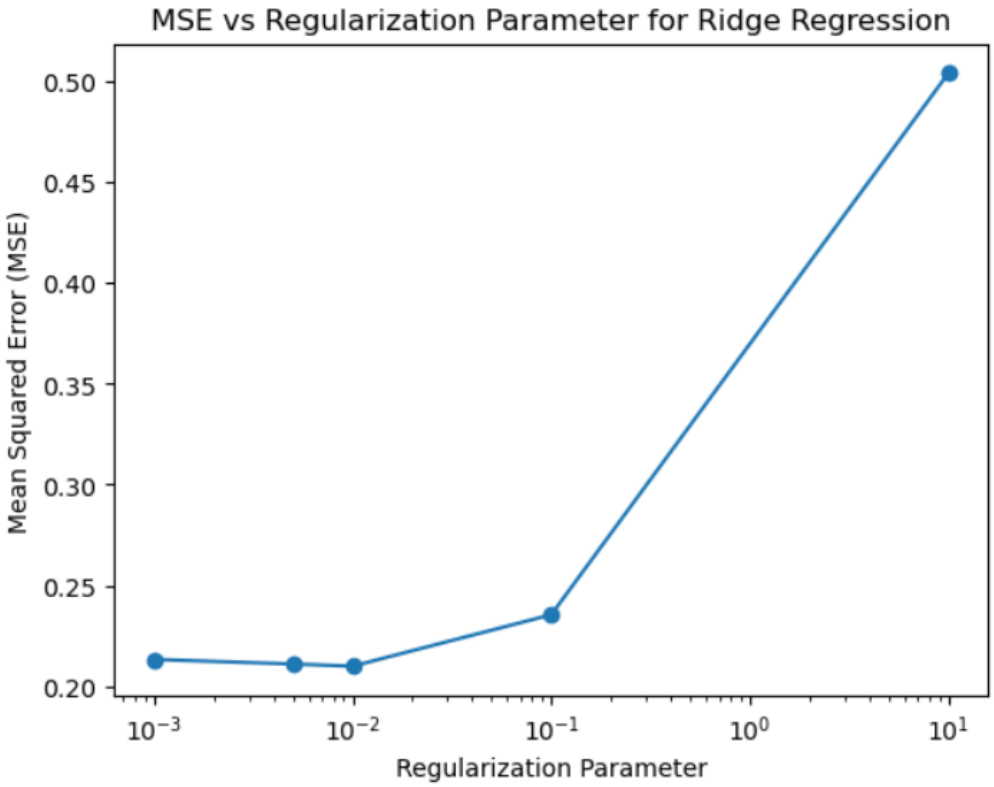
- The plot of Mean Square Error versus different polynomial degrees:



From the previous results, it is noticed that the best **degree is 2,** since it is the degree that minimize the Mean Square Error Value more than other degree values.

Applying the ridge regression on the training set to fit a polynomial of degree 8 given that regularization parameters {0.001, 0.005, 0.01, 0.1, and 10}. In addition, plotting the Mean Square Error on the validation versus regularization parameters to find the best value.

After applying the ridge regression, the following figure represents the plot of the Mean Square Value and the Regularization parameters. It is is noticed that the value 10^-2 is the best value among other values since it achieves the lowest mean square error value compared to other parameters.

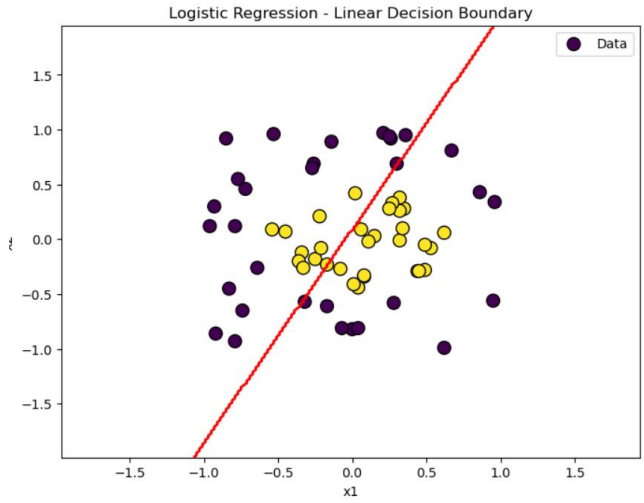


Question 2: Logistic Regression

Firstly, we have two files the training **train_cls.csv** for a binary classification problem, and testing **test_cls.csv**. Using the logistic regression implementation, it is required to learn a logistic regression model with a linear and quadratic decision boundary with training and testing accuracy calculations.

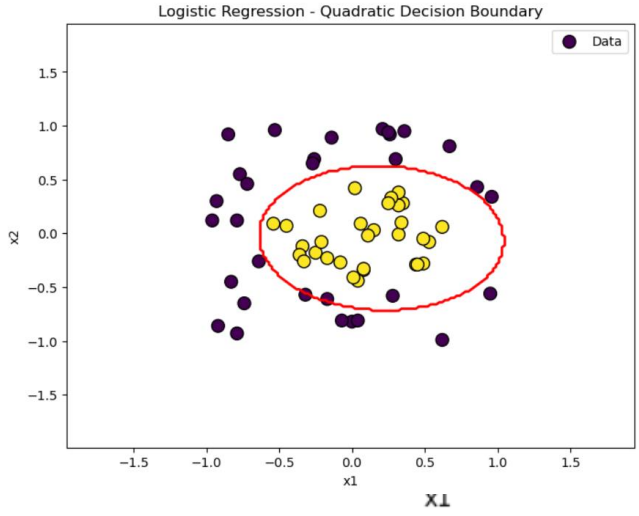
Solution:

- **Linear decision boundary:**




```
Training Accuracy for Linear Decision Boundary : 0.66
Testing Accuracy for Linear Decision Boundary: 0.68
```

- **Quadratic decision boundary:**




```
Training Accuracy for Quadratic Decision Boundary: 0.97
Testing Accuracy for Quadratic Decision Boundary: 0.95
```

From the previous results, it can be noticed that:

1. The Linear Decision Boundary model represents the under fitting since it did not capture the whole data well, while the Quadratic Decision Boundary model represents the over fitting since it learned the training data too well but included the noise from the Class 1 (C1). As a result the Linear is considered as underfitting while the Quadratic is overfitting.

2. After computing the accuracy of the training and testing sets for both models (linear and quadratic), it is noticed that the accuracy of the quadratic boundary model is higher than the linear one for both training and testing cases.