BIRZEIT UNIVERSITY

Electrical and Computer Engineering Department

Machine Learning and Data Science - ENCS5341

Assignment #2

Submission deadline: 22.12.2023

Model Selection and Hyper-parameters Tunning

The `data_reg.csv` file contains a set of 200 examples. Each row represents one example which has two attributes `x1` and `x2`, and a continuous target label y.
Using python, implement the solution of the following tasks:

1- Read the data from the csv file and split it into training set (the first 120 examples), validation set (the next 40 examples), and testing set (the last 40 examples). Plot the examples from the three sets in a scatter plot (each set encoded with a different color). Note that the plot here will be 3D plot where the x and y axes represent the x1 and x2 features, whereas the z-axis is the target label y.

2- Apply polynomial regression on the training set with degrees in the range of 1 to 10. Which polynomial degree is the best? Justify your answer by plotting the validation error vs polynomial degree curve. For each model plot the surface of the learned function alongside with the training examples on the same plot.
(hint: you can use `PolynomialFeatures` and `LinearRegression` from `scikit-learn` library)

3- Apply ridge regression on the training set to fit a polynomial of degree 8. For the regularization parameter, choose the best value among the following options: {0.001, 0.005, 0.01, 0.1, 10}. Plot the MSE on the validation vs the regularization parameter.
(hint: you can use `Ridge` regression implementation from `scikit-learn`)

Logistic Regression

The `train_cls.csv` file contains a set of training examples for a binary classification problem, and the testing examples are provided in the `test_cls.csv` file. The following figures show these examples.

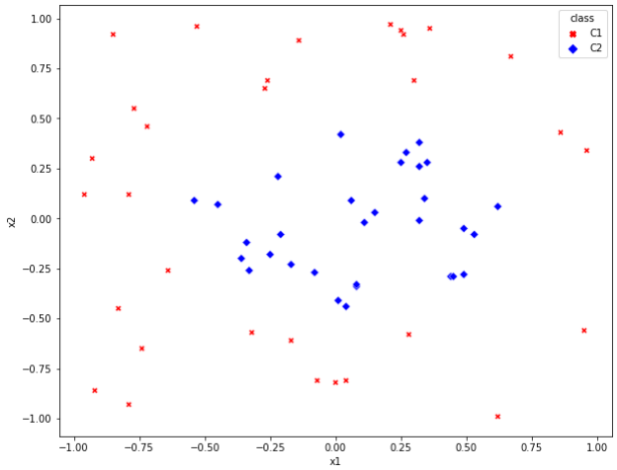

Figure 1 Training Set

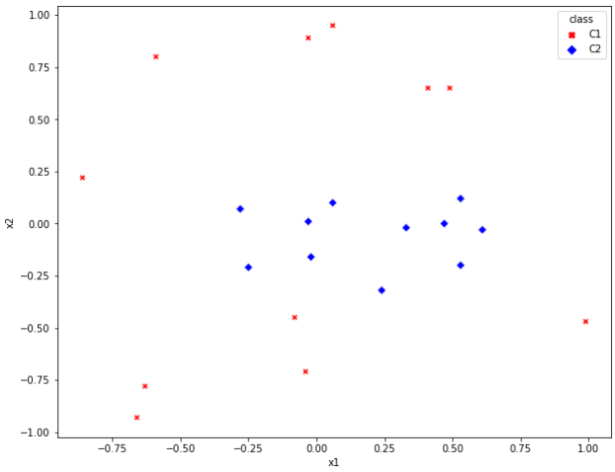

Figure 2 Testing Set

1. using the logistic regression implementation of `scikit-learn` library, Learn a logistic regression model with a linear decision boundary. Draw the decision boundary of the learned model on a scatterplot of the training set (similar to Figure 1). Compute the training and testing accuracy of the learned model.

2. Repeat part 1 but now to learn a logistic regression model with quadratic decision boundary.

3. Comment on the learned models in 1 and 2 in terms of overfitting/underfitting.

Submission instructions

Create a folder with the name [StudentID_FirstName] and put all the files for your solution in this folder, then compress it and submit only the **.zip** file.

For the solution files, you can either submit both a code (**.py**) and a short report (**.pdf**) that summarizes the results. Or alternatively, you can use `Jupyter Notebook` to prepare your solution in one (**.ipynb**) file that contains both *code cells* and *text cells* that discusses the results. (If you use colab, do not submit links to your notebook. only the .ipynb file is required)

Any violations to the submission instructions will be penalized by 10% of the assignment grade.

Good Luck!