Attention

1. Implement the models in Python. Use the independent libraries matplotlib, NumPy and pandas for implementation and display.

2. It is not allowed to use ready-made libraries to implement models such as scikit-learn.

3. Please, in addition to attaching the code, analyze the results and attach them to your report file.

4. The code can be in .py format or Jupyter, but note that you use relative addressing to access the data (your code should be able to run without modification)

The first question

The dataset folder for the first question and a dataset for the linear regression problem are provided Data into two training sections. The first column is the independent variable and the second column is the dependent variable.

1. Implement a linear regression model based on the normal equation.
2. Perform degree 5 polynomial feature extraction on the data.

3. Train 5 independent polynomial regression models of degree 5 with 10, 25, 50, and 200 train samples (preserving the order of the data). Report MSE error on training and test data respectively.

4. Based on the results, talk about the underfit and overfit of each model, then explain the reason for the superiority of some models or their proximity to each other.

The second question

You are provided with a data set for a linear regression problem. The data is divided into two parts, training and testing, in the second question dataset folder. In this data set, the first 13 columns are the independent variables and the last column is the dependent variable, but the test section is not labeled, and you are only provided to make a prediction on it and send it. Consider MSE as the evaluation criterion.

1. First, standardize the data based on the training dataset.
2. Implement the linear regression model based on the normal equation with regularization capability.
3. evaluate polynomial regressions of order 1, order 3, and order 5 with 3 regularization values 0.0, 1, and 10 using Repeated 5-Fold Cross-Validation with 10 repetitions (9 unique model should be checked and each model is trained 50 times, which produces 50 training and evaluation errors) then display the training and evaluation error using Boxplot.

4. Analyze the results, which of the models are overfit or underfit. Also, according to the results, can better settings be provided? Explain.

5. Determine the optimal setting and predict the test data through the model with the optimal setting. Save and attach the results in a file called prediction.csv. Note that each row must be a prediction of the row corresponding to the test data.