New Uzbekistan University
2023/24 Term 1
Assigned Date: 08/11/2023
Due Date: 19/11/2023

Assignment 1
Introduction to Machine Learning

Due Date

Late assignments will not be accepted and will receive ZERO mark.

Objectives

The objectives of this assignment are as follows.

- You will practice Linear regression by applying it to solve the problem of Software Defects Prediction and Logistic regression on the problem of Malware Detection both in Scikit-Learn library in Python.
- You will perform the necessary preprocessing steps such as: imputation, rescaling and encoding of categorical features.
- You will also evaluate the learned models using appropriate metrics from the Scikit-Learn package.
- Finally, you will solve a case that needs some research on your behalf which would help us to assess your understanding of how to handle a learning problem when there are some unwanted, significantly different data points.

1 Linear Regression [3.5 Points]

In this task, you will implement a Linear Regression model to predict the of number of defects in a software. You will work with the dataset of KC1-class-level from the PROMISE repository. You can visit the dataset link to read more about it, such as the description of different features.

More specifically, you are required to perform the following tasks on the “defects.csv" dataset:

1. Read data in Python. Then explore and clean it; display datatypes, and check for missing values.
2. Apply imputations for the missing values.
3. (optional) Some features can be filtered out. Remove them and provide justification(s) for their removal.
4. Rescale the dataset with a scaler of your choice.
5. Split your data into train and test sets (80% and 20% respectively).
6. Train a linear regression model, validate it on the test set, and report its performance.
7. Train a polynomial regression model (try different degrees). Plot the relationship between the degree vs the test loss and the degree vs the training loss on the same plot.
8. Answer these questions: Which degree would you choose for your model? Would it be better to choose a higher degree?

9. Write your own code to split the the dataset into train and test sets, and use it to split the dataset three times, such that each split has different train and test sets. Use these splits to train three Linear Regression models. Did the models turn out to be the same, or were there any differences? Why?
10. List the MSE for the test set in each case and mention the most and the least important feature according to the three built models.

2 Logistic Regression [1.5 Points]

Android is one of the most famous mobile OS worldwide thanks to its open-source code and its technological impact. However, due to the possibility of installing applications from third parties without any intensive central monitoring, Android has recently become a malware target.

To prevent malware attacks, researchers and developers have proposed different security solutions including static analysis, dynamic analysis, and artificial intelligence. Indeed, data science has become a promising area in cybersecurity, since analytical models based on data allow for the discovery of patterns that can help to predict malicious activities.

In this task, you will use some network layer features to build a machine learning classification model that will detect malware applications, using public benchmarks (datasets).

More specifically, you are required to perform the following tasks on the "android_traffic.csv" dataset:

1. Read data in Python. Then explore it; display datatypes, and check for missing values.
2. Rescale the dataset with a scaler of your choice.
3. Split your data into train and test sets (80% and 20% respectively).
4. Train a logistic regression model and evaluate it on test set. Report accuracy, precision, recall, and f1-score.
5. Answer the following question: Which metric from the previous step is more appropriate for this task and why?

3 Outliers Experiment [1 Point]

On the reduced version of the defects dataset "defects_small.csv" that has only one feature, perform the following task:

- Train a Linear regression model on it and report the MSE on the train set.

- Plot the dataset on a 2D graph, the feature vs the # of defects. Do you see any outliers? If you have found any such points, remove them manually and repeat step 1.
- Do some research to find (and then explain) the technique(s) that can help us automatically discover outliers in the dataset, and remove them or mitigate their effect on the learning process. Explain your findings in your own words; do not copy and paste from a source and cite your sources.

Deliverables

You are required to submit your solutions as a single ipynb file. Please, put your name and university email as the first line in the notebook.

Notes

- Cheating is a serious academic offense and will be strictly treated for all parties involved. So delivering nothing is always better than delivering a copy.
- Organize your notebook appropriately. Divide it into sections and cells with clear titles for each task and subtask.
- Make your code clean with the appropriate naming conventions. So maybe you want to take a look at some references: Link 1 and Link 2.
- Make sure your plots are descriptive and self-contained.