BIRZEIT UNIVERSITY

Electrical and Computer Engineering Department

Machine Learning and Data Science - ENCS5341

Assignment #3

Submission deadline: 22.01.2024

This assignment is intended for teams of two students.

In this assignment, you will test different machine learning models to perform a predictive task on a real-world problem. Furthermore, you need to communicate the results of your project by writing a comprehensive report.

You need to complete the following tasks:

1. Pick a dataset

The first task is to pick a topic for your project and find a dataset to apply machine learning models on. You are free to choose either a regression task or a classification task.

There are many websites that contain datasets. Here are some popular options:

- UCI Machine Learning Repository: https://archive.ics.uci.edu/
- Kaggle: https://www.kaggle.com/datasets
- Google Dataset Search: https://datasetsearch.research.google.com/
- Microsoft Research Open Data: https://www.microsoft.com/en-us/research/project/microsoft-research-open-data/

2. Baseline Model

As a baseline model, evaluate a nearest neighbor baseline using a distance of your choice. Report the performance of this baseline using both k=1 and k=3.

3. The proposed ML models

Try to achieve better performance by evaluating two additional models on the task. Discuss and motivate your model selection, and comment on why the performance has improved (or potentially did not improve). If the chosen models have hyper-parameters, make sure to tune at least one hyper-parameter for each model. To tune a parameter, you need to test at least 4 different values.

4. Performance analysis

In this task, you are required to analyze the performance of your best model from the previous part. To this end, examine instances in the test set where your model exhibits errors (classification errors in case of classification tasks, or high MSE error in case of regression). Try to find any interesting pattern in these examples.

5. Report

Besides your code, you need to submit a report (max. 8 pages) to analyze and discuss your results. Your project will be mainly evaluated based on the quality of the report. The report should have the following sections:

- Introduction: describe the task you are addressing in the project and the different models you have tried. Also, define the evaluation metrics that you have used.

- Dataset: describe the dataset that you used for this project, provide some statistics, and apply *exploratory data analysis*. Make sure to include both quantitative measures (descriptive statistics) and visualizations as well.

- Experiments and Results: discuss all the experiments you have done (baselines, model evaluation, hyper-parameters selection, …. etc) and the reported results in each case. Comment on all the results.

- Analysis: In this section, you need to discuss the process you used to analyze the performance of your best model. Highlight any interesting findings.

- Conclusions and Discussion: discuss your conclusions and also comment on the limitations of the used models and the evaluation metrics.