Assignment 1 - COL 865

Due on Oct 31, 2022 11:59

Marking scheme:

- Marks are equally distributed among the three datasets.
- Within each dataset, Oblique Trees carry 40%, other two carry 30% marks.
- Marks will be awarded based on the quality and correctness of the implementation, efforts made to improve the performance of models, submitted report and the performance during viva.

1 Description

The goal of this assignment is as follows: you are given three different tabular datasets. You should evaluate and compare the performance of

1. Random Forests
2. Oblique Decision Trees and
3. TabNet

on these datasets.

You are not expected to implement these models from scratch – for eg., you can use their scikit or TensorFlow implementations (or any other).

2 Datasets

Two of the datasets can be downloaded from `https://openml.org`. Please use the dataset ID to locate the right version to download. The third one can be downloaded from the given URL.

The dataset descriptions at the given locations also contain the task description. Please follow the task description as given.

2.1 Dataset1: BitcoinHeist Ransomware Dataset

Openml ID: 42553
Ref: Akcora, C.G., Li, Y., Gel, Y.R. and Kantarcioglu, M., 2019. BitcoinHeist. Topological Data Analysis for Ransomware Detection on the Bitcoin Blockchain. IJCAI-PRICAI 2020.

2.2 Dataset2: Traffic Violations

Openml ID: 42132
Note that this dataset has missing values in certain columns.

2.3 Dataset3: German Credit data

Available from `https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)` Note that this is a complex dataset – and some values may be in German.

3 Metrics

- Accuracy. If there are multiple labels to predict, then report accuracy for each target separately.
- Computational cost of training (how long it took on a specific hardware).
- Computational cost of testing.

Based on the approach you take, you are encouraged to provide ablation studies, but this is not mandatory.

4 Submission Method

- Every submission must have an accompanying Github repository containing all the code, scripts, and a report (LaTeXor a Markdown document) that presents all the results.
- Data cleaning, normalization, and any other pre-processing has to be included as part of the code release.
- Each method implementation should be in separate subdirectory.
- All dependencies have to be listed precisely. It is mandatory to create the environment.yml so that it can be set up in a separate conda environment without any hassle.
- Only the commits done **before** the deadline will be used in evaluation. **It will be strictly followed.**
- It is mandatory to provide the following in addition:
  - For deep learning models, it is necessary to provide logs for loss tracking, validation performance at each epoch.
  - Document the random seed values that were used so that results can be reproduced.
  - For decision trees, the model visualization is necessary.

All submissions have to be made by sending an email to the instructor. The email should simply contain the link to the GitHub repository. Note that the repository contains the LaTeXor markdown report.
You can write in the email any additional details you may want to share with the instructors.

4.1 Demo

Each submission will also get a demo slot where, apart from showing model execution (inference only), you will have a viva on the results you have obtained.