FACULDADE DE
CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE D
COIMBRA

Advanced Machine Learning

2022/2023

The Net Wars: Attack of the Bots

Nuno Lourenço

Ernesto Costa

João Correia

February 27, 2023

1 Introduction

We present here the first practical projetct, part of the students' evaluation process of the Advanced Machine Learning course of the Master in Engineering and Data Science of the University of Coimbra. This work is to be done autonomously by a group of **two** students. The deadline for delivering the work is **07 of April** via Inforestudante.

The quality of your work will be judged as a function of the value of the technical work, the written description, and the **public defence**. All sources used to perform the work (including the code) must be clearly identified. The document may be written in Portuguese or in English, using a word processor of your choice[1]. The written report is limited to **8** pages long, but in special, justified, cases (e.g., the need of presenting many images and/or tables), that number may be increased accordingly. The document should be well structured, including a general introduction, a description of the problem, the experimental setup, the analysis of the results, and a conclusion. The report should follow the Springer LNCS format. The Latex and Word templates are available in the Support Material of the course. The final mark will be given to each member of the group individually.

To do the work the student may consult any source he/she wants. Nevertheless, plagiarism will not be allowed and, if detected, it will imply failing the course. While doing the work and when submitting it, you should pay particular attention to the following aspects (whose relative importance depends on the type of work done):

- description of the approach to the problem
- description of the general architecture of the methods used;
- description of the experiment, including a table with the parameters used which should allow full replication;
- description of the evaluation metrics used for the validation: quality of the final result, efficacy, efficiency, diversity, or any other most appropriate;

Do not forget, besides what was just said, that it is fundamental: (1) to do a correct experimental analysis; (2) to do an informed discussion about the results obtained; (3) to put in evidence the advantages of the chosen alternative.

[1] We strongly suggest the use of LaTeX.

2 Problem Statement

Nowadays we depend on online services to perform many of our daily activities or relax such as paying bills, playing games, or streaming entertainment content. These services are delivered to us through the Internet, and we expect them to be working most of the time. With such a demand, and given that some of the platforms deal with sensitive data, security is of the utmost importance. All computer systems suffer from some type of security vulnerability that, if left unchecked, can be explored by bad-intentioned parties, disrupting the normal functioning of the services and causing damage to the reputation of companies [1].

To prevent incidents from happening, companies have invested resources in the research and development of Intrusion Detection Systems (IDSs) as tools to detect anomalies and attacks. The development of IDSs has been focused mostly on misused and/or anomaly detection. The former tends to be favored in commercial products due to its predictability, whilst the latter is seen as more powerful due to its theoretical potential for tackling novel disruptions.

In this work, we are going to embrace this theme by creating an IDS system. For this a dataset will be provided for you to train and test your solutions.

3 Objective

In general terms, the main objective is to analyse and explore the dataset provided and create an approach that can detect if a system is being attacked, and if attacked, which type of attack was used. To fulfil this task, you should tend to the following objectives:

- Prepare the pipeline necessary to process the data and create a model to distinguish between different types of attacks. You should compare the performance of different algorithms such as:
  - Regression
  - Ensembles
  - Support Vector Machines
  - Artificial Neural Networks
- Use model selection and parameterisation techniques to improve your models.

Take into account that you should not be limited to these algorithms as they are just examples of approaches that you can use. You should apply what you have learned during the course to solve this particular problem.

Dataset

The dataset is divided into a training dataset and test dataset. The training dataset should be used to prepare your approach and models and then you should evaluate the generalization ability of your system with the test dataset. You will not have access directly to the ground truth (see Competition section). You are in charge of preparing the data, analyse it and pre-process it as you see fit. Based on this step of data preparation and analysis, you should then train your models.

Description The datasets are composed of 41 columns of features that describe the traffic in the network and 1 column that classifies that sample into 4 categories (3 types of attacks or no attack). Table 1 details the names of the features and their possible values.

The first dataset, called dataset_train should be used to train and evaluate the models.

The second dataset, called dataset_test should be used to assess the generalisation ability of the models. You should use it to create a .csv file containing the predicted labels and submit it to the competition. The file will be automatically evaluated using the Accuracy Metric.

Evaluation Metrics

Given the training dataset, you should split it into train, validation, and test to see how fit the models that we are training/creating. Thus, the validation part of this work is crucial and you should select the most appropriate set of metrics (and justify your choice).

4 Competition

To evaluate the generalisation ability we are going to use a Kaggle competition. The competition **will not impact the final mark**, but rather will act as a way for you access the progress you are making and evaluate the generalisation performance of your models. The competition is available at

Table 1: Feature names and types. The feature attack_type corresponds to the labels, i.e., what we want to classify. The value normal in the attack_type means no attack is being performed.

Feature Name	**Feature Possible Value**
duration	numeric
protocol_type	categorical
service	categorical
flag	categorical
src_bytes	numeric
dst_bytes	numeric
land	0, 1
wrong_fragment	numeric
urgent	numeric
hot	numeric
num_failed_logins	numeric
logged_in	0, 1
num_compromised	numeric
root_shell	numeric
su_attempted	numeric
num_root	numeric
num_file_creations	numeric
num_shells	numeric
num_access_files	numeric
num_outbound_cmds	numeric
is_host_login	0, 1
is_guest_login	0, 1
count	real
srv_count	numeric
serror_rate	numeric
srv_serror_rate	numeric
rerror_rate	numeric
srv_rerror_rate	numeric
same_srv_rate	numeric
diff_srv_rate	numeric
srv_diff_host_rate	numeric
dst_host_count	numeric
dst_host_srv_count	numeric
dst_host_same_srv_rate	numeric
dst_host_diff_srv_rate	numeric
dst_host_same_src_port_rate	numeric
dst_host_srv_diff_host_rate	numeric
dst_host_serror_rate	numeric
dst_host_srv_serror_rate	numeric
dst_host_rerror_rate	numeric
dst_host_srv_rerror_rate	numeric
attack_type (label)	normal (0), Dos (1) , R2L (2) , U2R (3) , Probe (4)

the following address:
https://www.kaggle.com/t/e9c3e8bf88f6491db29bd6936be73659

To participate in the competition, you should prepare a csv file with two columns: the first column contains the Id of the sample that you are classifying, and the second column should contain the corresponding classification label. An example of a submission file is provided along with the project statement.

5 Conclusion

A few short comments. First, the control of the progression of your work will be done during the classes (T and PL). Moreover, you can discuss eventual problems by presenting yourself during office hours. Second, the projects reflect for the most part your actual knowledge. The rest will be object of lecturing soon after Easter. Third, we try to balance the difficulty of all the work, but we are aware that this is not an easy task and it is somehow a subjective matter. Fourth, we try to ask a workload compatible with the value of the work for the final mark.

Methodological issues, like the statistical background, were elucidated during the previous lectures. You may use the statistical tool you feel at ease with, including the Python code that was provided. Finally, even if this is a work that asks you to do simulations and analyze the results, i.e., it has a practical flavor, there is however a theory behind the work, and you are advised to consult the necessary literature.

Good luck!

References

[1] Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani. A detailed analysis of the kdd cup 99 data set. In *2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications*, pages 1–6, 2009.