Data Protection – Project

Year : 2024-2025
Lecturer : Côme Frappé - - Vialatoux

Organisation

The planning is as follows:

- Quickoff: 23/10/2024
- Defence: 22/11/2024

Organize your work with best practices:

- Work in groups of 4 (total of 7-8 groups for the whole promo)
- Share the work (and don't forget to mention task allocation in your report)

Project objectives

The objective of this project is to apply the data analysis chain on a cyber-physical dataset:

1) Using only network data
2) Using only physical data

Your code must run on an average laptop (16Go ram – no gpu) in a reasonable time, optimisation (Ex: GPU support, memory optimisations, etc.) are welcome and useful.

Evaluation required:

- Compare following algorithms: KNN, CART, Random Forest, XGBoost, MLP. You are free to replace on of the models by another one of your choice. If so, explain your choice in your report.

- Compare the metrics for balanced data (*precision, recall, TPR, TNR, accuracy*), metrics for unbalanced data (*F1-score, balanced accuracy, Matthews Correlation Coefficient*) and confusion matrices for each algorithm, and each attack class.

- Evaluate resources consumption for learning and for detection. (Fit time, prediction time, RAM)

- Compare the performance of your models to the ones published in the paper associated.

- You are free to use oversampling and/or undersampling in order to increase your models performances

- ***BONUS :***
  - *find a way to combine the information present in both the physical and network dataset*
  - *You are free and encouraged to test any idea you have of novel detection methods. Even if the results are bad, do not hesitate to include them !*

The dataset :

The dataset and associated paper can be accessed there:

- Link

/!\ Network datas are heavy !

Project deliverables

The deliverables of the project are:

- A streamlit webapp providing an interactive interface to explore your results (models performances, data visualisations, exploratory data analysis results worth showing, etc.)

- For treatments outside the webapp, the associated notebook containing said treatments (Ex: model training, data prep etc.)

- Project report (10-20 pages) explaining :
  - The results of your exploratory data analysis and their consequences on how you handled the data
  - for each model
    - Your data preparation steps (can be identical for multiple models)
    - How you trained your models (parameter tuning, improvements made, model architectures, computational resources, etc.)
    - Their performances with their analysis
    - If applicable, your choice of model
  - **BONUS:**
    - *How you combined network and physical information (and the benefits associated, if any)*
    - *Your novel detection methods (if any)*

- Conclusion
- personal sections (1 page / per member)
  - your contribution
  - your takeaways on this project.

Final presentation

For the final presentation:

- 10 minutes + ~3 minutes questions
- For all datasets
  - Data exploration **highlight** (Select your top findings/Key information explaining dataprep choices)
  - Dataprep
  - Compare algorithms performances
  - Evaluate the algorithms resource consumption

And as always, nice and clear visualisations to support what you are saying !

It is recommended to use your streamlit webapp directly as a support for your presentation, but you are not bound to it and can still use slides. If your presentation is on slides, then you will need to reserve part of your presentation time to do a demonstration of the streamlit webapp.

/!\ rehearse your presentation ! Only 1 group respected the 10 minutes last year ! → The report is exhaustive, the presentation can't fit all your results so you must do a selection !

Evaluation criteria

The final grade is evaluated as follow

Presentation	Timing	1
	Fluidity	1
	Clarity	1
Results	EDA of datasets	3
	Algorithms application	3
	Performance Analysis	3
	Handling of datasets	2
Webapp	Completeness	2
	User Experience	2
	Pertinence	2
Bonus		1
Total	Evaluation	20 (+1)

The global mark is conditioned to having sent the deliverables.