Interpretable and Differentially Private Machine Learning

U. Aïvodji, J. Ferry, S. Gambs, M.-J. Huguet and M. Siala

November 22, 2022

1 Context

The recent advances in artificial intelligence as well as the growing use of algorithmic systems to assist humans or even make decisions autonomously, have been accompanied by major ethical challenges. For instance, these challenges include the interpretability, fairness and privacy of the machine learning models that form the basis of such systems.

This internship will focus on machine learning models that are *interpretable by design* [Lip18, Rud19] while ensuring strong privacy guarantees, expressed in terms of *differential privacy* [NA21, DR14].

The internship is part of an international scientific collaboration on "Operational Research for Fairness, Privacy and Interpretability in Machine Learning" (co-funded by the LabEx CIMI[1]). The internship will take place at LAAS (ROC team[2]) and will be supervised by:

- J. Ferry, M.-J. Huguet and M. Siala, LAAS-CNRS (Toulouse, France)
- U. Aïvodji, ETS (Montréal, Canada) and S. Gambs, UQAM (Montréal, Canada)

2 Proposed Subject

Background (to be completed during the internship)

Interpretability. Basic notions of interpretability for machine learning and motivations can be found in [Lip18, Rud19]. More precisely, in this internship we will consider *rule list* models [Riv87]. To learn such models, we will leverage on the CORELS[3] algorithm [ALSA+17, ALSA+18], which builds Certifiably Optimal RulE ListS (in terms of accuracy and sparsity) for categorical data: `https://corels.eecs.harvard.edu/corels/`.

Pricvacy risks of machine learning models. Machine learning models have been shown to be vulnerable to privacy and security attacks targeting the confidentiality [RG20] of the model and its environment (*e.g.*, training data or hyperparameters), namely membership inference [SSSS17], property inference [AFM+13], model inversion [FJR15], training-data reconstruction [C+21] as well as model extraction [TZJ+16]. Several frameworks implementing these attacks have been developed in recent years, including ART[4]. This internship will investigate exclusively the design solutions to defend against membership inference attacks [SSSS17], which are considered as one of the fundamental privacy attacks.

Differential privacy. To achieve a privacy-preserving learning, we will consider a standard rigorous notion named *differential privacy* [NA21, DR14]. Various mechanisms (*e.g.,* Laplace, Gaussian or Exponential) were proposed to obtain the *differential privacy* property. Some of them are, for instance, implemented in public libraries such as `https://github.com/google/differential-privacy/` or `https://github.com/IBM/differential-privacy-library`.

[1] `https://cimi.univ-toulouse.fr/`
[2] `https://www.laas.fr/public/fr/roc`
[3] `https://github.com/corels/pycorels`
[4] `https://github.com/Trusted-AI/adversarial-robustness-toolbox`

Differential privacy in machine learning. The training of machine learning models raises important privacy risks, in particular with respect to the training data. The objective of privacy-preserving machine learning is to reconcile two antagonist purposes: extracting useful correlations from data without revealing private information about an individual user. Ensuring differential privacy when learning a model is a strong protection mechanism [JLE14, GXP+20], which theoretically bounds the amount of information that an adversary can infer from the access to the model. However, this privacy guarantee comes at a cost in terms of utility, in particular in terms of predictive accuracy. While methods were proposed to adapt deep learning frameworks to satisfy differential privacy [ACG+16], some recent works also considered interpretable models learning (*e.g.,* decision trees [FI19]).

Objective(s)

- The first part of the internship will consist in familiarizing with the different background notions, in particular;
  - understanding the basics and motivations for interpretability, rule list models and the CORELS algorithm,
  - learning the concept of differential privacy and the different algorithms that ensure such guarantee.
- The second and core part of the internship will consists in modifying the CORELS algorithm to ensure differential privacy guarantees, using one of the previously identified mechanisms. A theoretical study will also be conducted to prove that the modified algorithm satisfies differential privacy. In addition, an empirical evaluation will be performed to assess the trade-offs between DP budget and predictive accuracy (as well as eventually, sparsity or other interpretability notions). The adaptation of decision tree learning algorithms to comply with differential privacy guarantees [FI19] provides a good methodological example.
- Different differentially-private versions of the CORELS algorithm will also be explored, using different underlying DP-preserving mechanisms (*e.g.*, the Laplace mechanism, the Gaussian mechanism or the Exponential mechanism ...). The different versions can then be compared in terms of:
  1. (Theoretical) Differential privacy guarantees.
  2. (Empirical) Trade-offs between DP budget and learning accuracy (and, eventually, sparsity or other interpretability notions).
  3. (Empirical) Robustness against membership inference attacks [SSSS17] as compared to their non-DP counterparts.

References

[ACG+16] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

[AFM+13] Giuseppe Ateniese, Giovanni Felici, Luigi V Mancini, Angelo Spognardi, Antonio Villani, and Domenico Vitali. Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers. *arXiv:1306.4447*, 2013.

[ALSA+17] Elaine Angelino, Nicholas Larus-Stone, Daniel Alabi, Margo Seltzer, and Cynthia Rudin. Learning certifiably optimal rule lists. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, page 35–44. Association for Computing Machinery, 2017.

[ALSA+18] Elaine Angelino, Nicholas Larus-Stone, Daniel Alabi, Margo Seltzer, and Cynthia Rudin. Learning certifiably optimal rule lists for categorical data. *Journal of Machine Learning Research*, 18(234):1–78, 2018.

[C+21] Nicholas Carlini et al. Extracting training data from large language models. In *30th USENIX Security Symposium*, 2021.

[DR14] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4):211–407, aug 2014.

[FI19] Sam Fletcher and Md. Zahidul Islam. Decision tree classification with differential privacy. *ACM Computing Surveys (CSUR)*, 52:1 – 33, 2019.

[FJR15] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *ACM CCS*, 2015.

[GXP+20] Maoguo Gong, Yu Xie, Ke Pan, Kaiyuan Feng, and A. K. Qin. A survey on differentially private machine learning [review article]. *IEEE Computational Intelligence Magazine*, 15:49–64, 2020.

[JLE14] Zhanglong Ji, Zachary Chase Lipton, and Charles Elkan. Differential privacy and machine learning: a survey and review. *CoRR*, abs/1412.7584, 2014.

[Lip18] Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018.

[NA21] Joseph P. Near and Chiké Abuah. *Programming Differential Privacy*, volume 1. 2021.

[RG20] Maria Rigaki and Sebastian Garcia. A survey of privacy attacks in machine learning. *arXiv:2007.07646*, 2020.

[Riv87] Ronald L Rivest. Learning decision lists. *Machine learning*, 2(3):229–246, 1987.

[Rud19] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

[SSSS17] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017.

[TZJ+16] Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. Stealing machine learning models via prediction {APIs}. In *USENIX Security Symposium*, 2016.