M2 Internship — 5 to 6 months, starting between Feb. and Apr. 2022

Fairness vs Privacy Tradeoff in Machine Learning algorithms

Nesrine KAANICHE ans Sophie CHABRIDON
SAMOVAR lab, Télécom SudParis
Évry and Palaiseau, France
Contacts: kaaniche.nesrine [at] telecom-sudparis.eu, Sophie.Chabridon [at] telecom-sudparis.eu

Context. Machine learning algorithms are gaining an expanding interest in different sectors, for their promising results and expected performances and efficiency. This includes decision-aided algorithms implemented in recommendation and scoring applications, or autonomous algorithms embedded in smart settings, such as autonomous vehicles.

However, these results are increasingly discussed and questioned by the scientific community [Coo21]. In particular, they are accused of being black boxes that process a massive amount of personal data, leading thus to (i) privacy attacks and (ii) discriminatory practices linked, for instance, to gender or ethnic origin [VR18] Indeed, intelligent algorithms are vulnerable to privacy attacks, i.e., inference and reconstruction attacks, and suffer from increased biases due to the imbalanced distribution of data. Several recent surveys have confirmed that privacy attacks under pre-trained models are not uniform across all classes, they are showing around 20% more success of identification for minority classes [SSSS17, NSH19].

Objective The objective of this project is to study the different definitions of fairness and to evaluate so-called "pre-processing" methods to reduce the indesirable discriminatory impact, and their impact on privacy protection.

The following tasks are foreseen:

- Implement a "pre-processing" method, that consists of selecting the appropriate "subsets" of attributes/features to be considered during the classification phase. This process will be repeated until reaching an acceptable level of fairness;
- Evaluate the impact of the implemented pre-processing method in relation to the equity objectives.
- Evaluate the privacy attacks with respect to the proposed algorithm.

Deliverables

- Report
- Demonstrator and source code of the implemented method

To apply Contact the supervisors Nesrine Kaaniche and Sophie Chabridon with motivation letter, résumé, M1 and M2 academic transcripts, and an example of a report or article written in English or French.

References

[Coo21] Alexis Cook. Ai fairness. https://www.kaggle.com/alexisbcook/ai-fairness, 2021.

[NSH19] Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *2019 IEEE Symposium on Security and Privacy, SP 2019, San Francisco, CA, USA, May 19-23, 2019*, pages 739–753. IEEE, 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 2017, San Jose, CA, USA, May 22-26, 2017*, pages 3–18. IEEE Computer Society, 2017.

[VR18] Sahil Verma and Julia Rubin. Fairness definitions explained. In Yuriy Brun, Brittany Johnson, and Alexandra Meliou, editors, *Proceedings of the International Workshop on Software Fairness, FairWare@ICSE 2018, Gothenburg, Sweden, May 29, 2018*, pages 1–7. ACM, 2018.