Towards Integrating Differential Privacy & Demographic Parity in Machine Learning

Keywords : Supervised learning, Algorithmic fairness, Privacy

Context : Machine learning is at the core of many data-oriented high-stake applications such as medicine, finance, heavy industry or recommendation algorithms on the internet. However, recent studies have identified several major flaws of machine learning and data analysis such as d*violation of data privacy* or *model bias*.

- **Violation of data privacy.** When the final model is publicly released, it may be exposed to membership inference attacks by external entities that could demonstrate the presence of some specific instances in the dataset. The risk of a model inversion attack, which infers sensitive attributes of the dataset by analyzing the final model, must also be seriously considered when deploying a model in real-world scenarios. Acordingly, several definitions have been introduced to characterize privacy preserving algorithms in the context of machine learning and data publishing. Among them, *differential privacy* has become the dominant standard to provide a formal and adaptive conception of privacy preserving data analysis.

- **Model Bias.** Besides, learning algorithms may inherit bias in the data during the training process, leading to undesired knock-on effect on future decisions. In particular, severe conflicts may arise with ethical criteria of the modern society using algorithms that only focus on prediction accuracy. *Algorithmic fairness*, which has been emerging in the last few years, try to give a solution to the problem of mitigating the bias in data. Among the existing definitions of algorithmic fairness, demographic parity is one of the simplest and most actionable notion so far. As such, it constitutes the starting point of many fairness-related machine learning studies.

These shortcomings raise questions about the legal liability of model providers and cause practitioners to reevaluate the trust they place in the systems they use. They also call for algorithms that learn accurate models while meeting strong privacy and fairness requirements. While these two crucial conditions have been extensively studied individually, their combination remains poorly understood. The objective of the internship is to propose and study algorithms that are able to handle both fairness and privacy constraints simultaneously, by focusing on the specific notions of demographic parity and differential privacy.

Objectives : We consider the supervised classification framework. As an introduction to the internship, the first step will involve reviewing the existing literature, with a particular focus on the following papers : [1, 2, 3].
The second step of the internship will focus on studying the problem within the Gaussian mixture model. This phase aims to formalize the problem's specific aspects and deepen the understanding of its nuances. A central question in this step will be comparing a fair predictor with another predictor that satisfies both fairness and differential privacy constraints.
Finally, the third step will concentrate on developing a general methodology to enforce fairness under privacy constraints for a given machine learning algorithm.
The objectives of this internship can thus be summarized as follows.

1. Review the relevant literature ;
2. Analyze the problem within the Gaussian mixture model ;

3. Design a classification algorithm that incorporates fairness under privacy constraints ;
4. Implement the procedure in `Python` ;
5. Study its theoretical and numerical properties.

Supervisors : Christophe Denis (SAMM, Université Paris1 Panthéon-Sorbonne), Rafael Pinot (LPSM, Sorbonne Université).

Required skills : M2 level trainee in statistics/machine learning/optimization. Python programming. Applicants should send a CV and transcripts of the last two years to christophe.denis1@univ-paris1.fr and pinot@lpsm.paris

Practical information :

- April-September 2025
- Location : LPSM, Sorbonne Université
- Grant from LPSM
- Possibility of getting a PhD position after the internship

References

[1] Mangold, Paul and Perrot, Michaël and Bellet, Aurélien and Tommasi, Marc Differential Privacy has Bounded Impact on Fairness in Classification, International Conference On Machine Learning 2023

[2] Dwork, Cynthia and Roth, Aaron The Algorithmic Foundations of Differential Privacy, Foundations and Trends in Theoretical Computer Science 2014

[3] Denis, Christophe and Elie, Romuald and Hebiri, Mohamed and Hu, François Fairness guarantees in multi-class classification with demographic parity, Journal of Machine Learning Research 2024