Universidade Federal de Minas Gerais

Department of Computer Science

Undergraduate Thesis, Part II

Relations between Fairness, Privacy and Quantitative Information Flow in Machine Learning

Type: Scientific

Abstract

Recent years have witnessed an enormous advance in the area of Machine Learning, reflected by the popularity of Artificial Inteligence systems. For most of the history of machine learning research, the main goal was the development of machine learning algorithms that led to more accurate models, but it is now very clear that there are many other important areas to develop. We want models to be fair to unprivileged groups in society, to not reveal private information used in the model training, to provide comprehensible explanations to humans in order to help identifying causal relationships, among many relevant goals other than simply improving model accuracy. In this work, we explore possible new relationships between fairness, privacy and Quantitative Inforamtion Flow. The first exploration is an analysis of papers that explore the impact of privacy-enhancing mechanisms on Machine Learning fairness notions. Our second exploration is the possibility of dividing a fixed local differencial privacy budget between variables with varying degrees of sensitivity. Finally, we explore modeling both local differential privacy parameters within the Quantitative Information Flow framework.

Supervisor:

Prof. Mário Sérgio Alvim

Thesis written by:

Artur Gaspar da Silva

Academic Semester 2024/2

CONTENTS

Index Terms

Quantitative Information Flow, Differential Privacy, Fairness, Machine Learning.

I. Introduction

Recent resarch [1] [2] [3] [4] indicates numerous tensions and synergies between many concepts that surround the Machine Learning literature, including Fairness, Privacy, Accuracy and Interpretability. For instance, there is an inherent tradeoff between Fairness and Accuracy such that, depending on the data distribution, it might be impossible to develop a model that achieves acceptable values for both fairness and accuracy, if we considr some reasonable fairness metrics [5]. Also, there has been some work on introducing Causality concepts into the discussion, for instance, to develop better fairness metrics [6]. It has also been suggested to use interpretable models (as explanations to more complex models) for auditing systems and checking if they are fair, although this might lead to problems [7]. This area of research is especially relevant nowadays, given the importance that Machine Learning and Artificial Intelligence systems have: we now have computational systems that are part of processes of making decisions with big impacts on people's lives, for instance, recidivism prediction [8], loan approvals [9], hiring decisions [10], and others.

The goal of this Undergraduate Thesis is to review and reproduce results presented in the litearture, verify the viability of the connections between the aforementioned areas and Quantitative Information Flow, and, if possible, develop new theoretical results. This project is divided into two parts: POC I and POC II. In POC I, the specific goals were to research the literature for these concepts and focus on the connections that have been identified between them, so the expected result is a concise review of the literature on these topics. In this work (POC II), we verify possible connections between Privacy, Fairness and Quantitative Information Flow, and outline possible new theoretical results. We provide an in-depth theoretical analysis of the viability of Quantitative Information Flow approaches to these areas and the connections between privacy and fairness. We provide a formal exploration of the impact of privacy-enhancing obfuscation methods in fairness, based on important results in the literature reviewed in POC I, we explore how the privacy budget can be divided between many variables in the context of Local Differential Privacy, and, finally, we explore how viable is the application of the Quantitative Information Flow framework in Local Differential Privacy.

II. Theoretical Reference

Fairness in Machine Learning is concerned with measuring how unfair the results provided by Machine Learning models are to certain groups or individuals [11], and improving how fair the models are [12]. There are tensions between different fairness measures [13] [14]. Privacy is concerned with quantifying how much sensitive information leaks about individuals and methods to avoid this information leakage. In Machine Learning settings, the data collection might be hard for information that is considered very sensitive (for instance, whether or not a person regularly uses illegal drugs) and approaches such as Differential Privacy [15] might improve trust in the data collection. Also, the model itself might allow the identification of individuals and sensitive features, which is not desirable [16]. Quantitative Information Flow is a general theoretical framework for measuring amounts of information, with a focus on privacy applications but, in principle, a broader scope [17]. In recent research [1], the relationships between Fairness, Interpretability and Privacy have been extensively explored. Recent papers focuses on relationships between Privacy and Fairness [4], on the relationship between Privacy [3], and on the feasibility regions of Accuracy and Fairness metrics [5] [2].

More specifically to the relation between Differential Privacy and Quantitative Inforamtion Flow, there are important results in the literature. There are works discussing the relations between differential privacy and g-vulnerability, including bounds on g-leakage as a function of the ϵ parameter of differential privacy, and the fact that there is no bound on differential privacy as a function of the g-vulnerability [18]. Also, we have recent work [19] discussing how the ϵ parameter of Differential Privacy is related to max-case g-vulnerability: e^{ϵ} is exactly the multiplicative max-case channel capacity under a fixed prior. This work also discusses many other theoretical results relating g-vulnerability notions with differential privacy.

III. Contributions

A. The impact of Local Differential Privacy mechanisms on fairness metrics

B. How to distribute the Local Differential Privacy budget between variables with varying levels of importance

C. Modeling the δ parameter of (ϵ, δ)-LDP within the Quantitative Information Flow framework

IV. Conclusions

References

[1] J. Ferry, U. Aïvodji, S. Gambs, M.-J. Huguet, and M. Siala, "Sok: Taming the triangle - on the interplays between fairness, interpretability and privacy in machine learning," *ArXiv*, vol. abs/2312.16191, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:266573131

[2] A. Agarwal, A. Beygelzimer, M. Dudik, J. Langford, and H. Wallach, "A reductions approach to fair classification," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. PMLR, 10–15 Jul 2018, pp. 60–69. [Online]. Available: https://proceedings.mlr.press/v80/agarwal18a.html

[3] R. Cummings, V. Gupta, D. Kimpara, and J. Morgenstern, "On the compatibility of privacy and fairness," in *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*, ser. UMAP'19 Adjunct. New York, NY, USA: Association for Computing Machinery, 2019, p. 309–315. [Online]. Available: https://doi.org/10.1145/3314183.3323847

[4] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ser. ITCS '12. New York, NY, USA: Association for Computing Machinery, 2012, p. 214–226. [Online]. Available: https://doi.org/10.1145/2090236.2090255

[5] C. Pinzón, C. Palamidessi, P. Piantanida, and F. Valencia, "On the incompatibility of accuracy and equal opportunity," *Machine Learning*, May 2023. [Online]. Available: https://doi.org/10.1007/s10994-023-06331-y

[6] K. Makhlouf, S. Zhioua, and C. Palamidessi, "Survey on causal-based machine learning fairness notions," 2022.

[7] J. Zhou and T. Joachims, "How to explain and justify almost any decision: Potential pitfalls for accountability in ai decision-making," in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 12–21. [Online]. Available: https://doi.org/10.1145/3593013.3593972

[8] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks." *URL https://www. propublica. org/article/machine-bias-risk-assessments-in-criminal-sentencing*, 2019.

[9] M. A. Sheikh, A. K. Goel, and T. Kumar, "An approach for prediction of loan approval using machine learning algorithm," in *2020 international conference on electronics and sustainable communication systems (ICESC)*. IEEE, 2020, pp. 490–494.

[10] L. Li, T. Lassiter, J. Oh, and M. K. Lee, "Algorithmic hiring in practice: Recruiter and hr professional's perspectives on ai use in hiring," in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 2021, pp. 166–176.

[11] K. Makhlouf, S. Zhioua, and C. Palamidessi, "On the applicability of machine learning fairness notions," *SIGKDD Explor. Newsl.*, vol. 23, no. 1, p. 14–23, may 2021. [Online]. Available: https://doi.org/10.1145/3468507.3468511

[12] K. Holstein, J. Wortman Vaughan, H. Daumé III, M. Dudik, and H. Wallach, "Improving fairness in machine learning systems: What do industry practitioners need?" in *Proceedings of the 2019 CHI conference on human factors in computing systems*, 2019, pp. 1–16.

[13] S. A. Friedler, C. Scheidegger, and S. Venkatasubramanian, "The (im)possibility of fairness: different value systems require different mechanisms for fair decision making," *Commun. ACM*, vol. 64, no. 4, p. 136–143, mar 2021. [Online]. Available: https://doi.org/10.1145/3433949

[14] G. Alves, F. Bernier, M. Couceiro, K. Makhlouf, C. Palamidessi, and S. Zhioua, "Survey on fairness notions and related tensions," *EURO Journal on Decision Processes*, vol. 11, p. 100033, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2193943823000067

[15] C. Dwork, "Differential privacy," in *International colloquium on automata, languages, and programming*. Springer, 2006, pp. 1–12.

[16] B. Liu, M. Ding, S. Shaham, W. Rahayu, F. Farokhi, and Z. Lin, "When machine learning meets privacy: A survey and outlook," *ACM Computing Surveys (CSUR)*, vol. 54, no. 2, pp. 1–36, 2021.

[17] M. Alvim, K. Chatzikokolakis, A. McIver, C. Morgan, C. Palamidessi, and G. Smith, *The Science of Quantitative Information Flow*, ser. Information Security and Cryptography. Springer International Publishing, 2020. [Online]. Available: https://books.google.com.br/books?id=jJH-DwAAQBAJ

[18] M. S. Alvim, M. E. Andrés, K. Chatzikokolakis, P. Degano, and C. Palamidessi, "On the information leakage of differentially-private mechanisms," *Journal of Computer Security*, vol. 23, no. 4, pp. 427–469, 2015.

[19] N. Fernandes, A. McIver, and P. Sadeghi, "Explaining epsilon in local differential privacy through the lens of quantitative information flow," *arXiv preprint arXiv:2210.12916*, 2022.