Botar data de entreg da proposta dpsRelations between Fairness, Privacy and Quantitative Information Flow in Machine Learning

Type: Scientific
Advisor: Mário Sérgio Alvim

Artur Gaspar da Silva

23/09/2024

1 Introduction

Recent resarch[11][1][7][9] 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[19]. Also, there has been some work on introducing Causality concepts into the discussion, for instance, to develop better fairness metrics[17]. 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[23]. 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[6], loan approvals[22], hiring decisions[14], 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 POC II, the specific goals are to reproduce the results and verify possible connections between Privacy, Fairness and Quantitative Information Flow, with the possibility of developing new theoretical results. The expected result is an in-depth theoretical analysis of the viability of Quantitative Information Flow approaches to these areas and the connections between privacy and fairness. It's thus necessary to have a comprehensible review of the literature by the end of POC I. By the end of POC2, the expected result is to have a formal exploration of the impact of privacy-enhancing obfuscation methods in fairness, to explore how the privacy budget can be divided between many variables in the context of Local Differential Privacy, and explore how viable is the application of Quantitative Information Flow tools for the development of privacy and fairness research.

2 Theoretical Reference

Causality refers to the study of causal relationships between variables, and how to model and infer causal relationships from the combination of domain knowledge and data[18]. This area of research has matured a lot in the last 50 years, with many different approaches still being developed. Fairness in Machine Learning is concerned with measuring how unfair the results provided by Machine Learning models are to certain groups or individuals[16], and improving how fair the models are[13]. There are tensions between different fairness measures[12][2]. 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[8] might improve trust in the data collection. Also, the model itself might allow the identification of individuals and sensitive features, which is not desirable[15]. Accuracy is a metric of how many mistakes the Machine Learning model makes, and there are trade-offs between Accuracy and the other concepts presented[11][19][7]. The area of Interpretability focus on developing Machine Learning models that have human-comprehensible decisions (either directly or to explain the decisions of more complex models), which might be useful when developing these models[21] and also to help experts with domain knowledge decide when to trust the results presented by the models[20]. 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[5].

In [11], the relationships between Fairness, Interpretability and Privacy have been extensively explored. The paper [9] focuses on relationships between Privacy and Fairness, [7] on the relationship between Privacy, Fariness and Accuracy, [19] and [1] on the feasibility regions of Accuracy and Fairness metrics, [17] on Causality-Aware fairness metrics. One of the goals of the first part of this project is to increase this list of references with the added analysis of the possibility of approaches based on Quantitative Information Flow ([4] explored the relations between Quantitative Information Flow and Fairness, but it is still possible to find relationships with the other topics mentioned).

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 [3]. Also, we have recent work [10] 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.

3 Methodology

The methodology applied to this project consists, in general, of reading many papers on the relevant subjects, in order to gather what has been produced recently. Also, for developing the necessary theoretical background, part of the methodology is to read the main book on Causality[18], by Judea Pearl. Also, rigorous mathematical reasoning will be applied for any possible theoretical result, and computer simulations will be developed for the reproduction of relevant results.

4 Expected Results

For the first part of the project (POC I), the expected result was an extensive review of the literature on Causality, Fairness, Privacy, Accuracy and Interpretability in Machine Learning, and the relationships between these concepts. For the second part (POC II), the expected results is to have a formal exploration of the impact of privacy-enhancing obfuscation methods in fairness, to explore how the privacy budget can be divided between many variables in the context of Local Differential Privacy, and explore how viable is the application of Quantitative Information Flow tools for the development of privacy and fairness research.

5 Steps and Cronogram

1. Why should we try to reverse the noise after LDP. We can show that we lose info (utility, accuracy, etc.), or that we worsen the Paretto Front of the fairness-accuracy trade-off.
2. How to better distribute the privacy budget if we have only a few sensitive attributes.
3. How to model $(\epsilon, \delta) - DP$ with QIF. Maybe δ-max case, the best that has at least δ probability of hapennning.

The dates are in month/day format. Obviamente, refazer as datas né

Cronogram							
	03/17	03/31	04/14	04/28	05/12	05/26	06/09 - 06/27
Contacting advisor and preparing themes	▌						
Writing this proposal	▐						
Reading recent papers		█	▌				
Exploring LDP noise reversal			▐	█			
Preparing Partial Pitch				█			
Exploring privacy bidget distribution				▐	█		
Exploring $(\epsilon, \delta) - DP$ modeling with QIF						█	▌
Preparing Final Pitch						▐	▌
Writing final report							█

6 References

References

[1] Agarwal, A., Beygelzimer, A., Dudik, M., Langford, J., Wallach, H.: A reductions approach to fair classification. In: Dy, J., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 80, pp. 60–69. PMLR (10–15 Jul 2018), `https://proceedings.mlr.press/v80/agarwal18a.html`

[2] Alves, G., Bernier, F., Couceiro, M., Makhlouf, K., Palamidessi, C., Zhioua, S.: Survey on fairness notions and related tensions. EURO Journal on Decision Processes **11**, 100033 (2023). https://doi.org/https://doi.org/10.1016/j.ejdp.2023.100033, `https://www.sciencedirect.com/science/article/pii/S2193943823000067`

[3] Alvim, M.S., Andrés, M.E., Chatzikokolakis, K., Degano, P., Palamidessi, C.: On the information leakage of differentially-private mechanisms. Journal of Computer Security **23**(4), 427–469 (2015)

[4] Alvim, M., Fernandes, N., Nogueira, B., Palamidessi, C., Silva, T.: On the duality of privacy and fairness (extended abstract). In: International Conference on AI and the Digital Economy (CADE 2023). Institution of Engineering and Technology, United Kingdom (2023). https://doi.org/10.1049/icp.2023.2563, 9th International Conference on AI and the Digital Economy, CADE 2023 ; Conference date: 26-06-2023 Through 28-06-2023

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

[6] Angwin, J., Larson, J., Mattu, S., Kirchner, L.: 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)

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

[8] Dwork, C.: Differential privacy. In: International colloquium on automata, languages, and programming. pp. 1–12. Springer (2006)

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

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

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

[12] Friedler, S.A., Scheidegger, C., Venkatasubramanian, S.: The (im)possibility of fairness: different value systems require different mechanisms for fair decision making. Commun. ACM **64**(4), 136–143 (mar 2021). https://doi.org/10.1145/3433949, https://doi.org/10.1145/3433949

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

[14] Li, L., Lassiter, T., Oh, J., Lee, M.K.: 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. pp. 166–176 (2021)

[15] Liu, B., Ding, M., Shaham, S., Rahayu, W., Farokhi, F., Lin, Z.: When machine learning meets privacy: A survey and outlook. ACM Computing Surveys (CSUR) **54**(2), 1–36 (2021)

[16] Makhlouf, K., Zhioua, S., Palamidessi, C.: On the applicability of machine learning fairness notions. SIGKDD Explor. Newsl. **23**(1), 14–23 (may 2021). https://doi.org/10.1145/3468507.3468511, https://doi.org/10.1145/3468507.3468511

[17] Makhlouf, K., Zhioua, S., Palamidessi, C.: Survey on causal-based machine learning fairness notions (2022)

[18] Pearl, J.: Causality: Models, Reasoning and Inference. Cambridge University Press, USA, 2nd edn. (2009)

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

[20] Roscher, R., Bohn, B., Duarte, M.F., Garcke, J.: Explainable machine learning for scientific insights and discoveries. Ieee Access **8**, 42200–42216 (2020)

[21] Santos, G., Figueiredo, E., Veloso, A., Viggiato, M., Ziviani, N.: Predicting software defects with explainable machine learning. In: Proceedings of the XIX Brazilian Symposium on Software Quality. pp. 1–10 (2020)

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

[23] Zhou, J., Joachims, T.: 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. p. 12–21. FAccT '23, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3593013.3593972, https://doi.org/10.1145/3593013.3593972