Relations between Causality, Fairness, Privacy, Accuracy and Interpretability in Machine Learning

Type: Scientific

Advisor: Mário Sérgio Alvim

Artur Gaspar

05/04/2024

1 Introduction

Recent resarch[9][1][6][8] 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[16]. 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[5], loan approvals[22], hiring decisions[12], 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 will be divided into two parts: POC I and POC II. In POC I, the specific goals are 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 with Quantitative Information Flow, with the possibility of developing new theoretical results. The expected result is the reproduction of many results and an in-depth theoretical analysis of the viability of Quantitative Information Flow approaches to these areas. It's thus necessary to have a comprehensible review of the literature by the end of POC I.

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[15], and improving how fair the models are[11]. There are tensions between different fairness measures[10][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[7] might improve trust in the data collection. Also, the model itself might allow the identification of individuals and sensitive features, which is not desirable[13]. 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[9][19][6]. 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[4].

In [9], the relationships between Fairness, Interpretability and Privacy have been extensively explored. The paper [8] focuses on relationships between Privacy and Fairness, [6] on the relationship between Privacy, Fariness and Accuracy, [19] and [1] on the feasibility regions of Accuracy and Fairness metrics, [16] 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 ([3] explored the relations between Quantitative Information Flow and Fairness, but it is still possible to find relationships with the other topics mentioned).

3 Methodology

The methodology applied to this project consists, in general, of reading as many papers on the subject as possible, in order to gather what has been produced recently. Also, for developing the necessary theoretical background, part of the methodology is to read books on Causality[18][17], Quantitative Information Flow[4] and Information Theory[14]. Also, rigorous mathematical reasoning will be applied for any possible theoretical result, and computer simulations will be developed for the reproduction of results.

4 Expected Results

For the first part of the project (POC I), the expected result is 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 are the reproduction and verification of the viability of applying the theoretical framework of Quantitative Information Flow to these concepts, with the possibility of developing new theoretical results.

5 Steps and Cronogram

1. March 17, 2024 - March 23, 2024: Contacting advisor and preparing themes.
2. March 24, 2024 - March 30, 2024: Writing this proposal.
3. March 31, 2024 - April 06, 2024: Reading Causality book[18].
4. April 07, 2024 - April 13, 2024: Reading Causality book[18].
5. April 14, 2024 - April 20, 2024: Reading Causality book[18].
6. April 21, 2024 - April 27, 2024: Reading recent papers.
7. April 28, 2024 - May 04, 2024: Reading recent papers.
8. May 05, 2024 - May 11, 2024: Reading Causal Inference book[17] and preparing partial pitch.
9. May 12, 2024 - May 18, 2024: Reading Causal Inference book[17].
10. May 19, 2024 - May 25, 2024: Reading Causal Inference book[17].
11. May 26, 2024 - June 01, 2024: Reading Information Theory book[14].
12. June 02, 2024 - June 08, 2024: Reading Information Theory book[14].
13. June 09, 2024 - June 15, 2024: Reading Information Theory book[14] and preparing final pitch.
14. June 16, 2024 - June 22, 2024: Reading recent papers and writing final report.
15. June 23, 2024 - June 29, 2024: Reading recent papers and 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., 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

[4] 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

[5] 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)

[6] 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

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

[8] 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

[9] 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

[10] 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

[11] 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)

[12] 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)

[13] 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)

[14] MacKay, D.J.: Information Theory , Inference And Learning Algorithms. Cambridge University Press (2005), https://books.google.com.br/books?id=4WhiPgAACAAJ

[15] 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

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

[17] Pearl, J., Glymour, M., Jewell, N.: Causal Inference in Statistics: A Primer. Wiley (2016), https://books.google.com.br/books?id=L3G-CgAAQBAJ

[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