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

Type: Scientific
Advisor: Mário Sérgio Alvim

Artur Gaspar da Silva

05/04/2024

Abstract

Do que podemos falar aqui?

1 Interpretabilidade ou explicabilidade?

2 Introduction

Recent resarch[5][1][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 consider some reasonable fairness metrics[8]. Also, there has been some work on introducing Causality concepts into the discussion, for instance, to develop better fairness metrics[7]. 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[10]. 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[2], loan approvals[9], hiring decisions[6], and others.

In this first part of the Undergraduate Thesis, we provide a concise review of the literature on the topics presented. The main goal is to provide a solid basis for future work on these topics and identify the known connections among them. Section 3 discusses the already developed theoretical work on these areas, section 4 provides the contributions of this analysis in the form of a higher-level discussion of what we can gather from the literature and the meaning of these results, and section 5 provides conclusions and possible future lines of work. The rest of this section is dedicated to introducing the concepts of Causality, Fairness, Privacy, Interpretability in Machine Learning, and also Quantitative Information Flow (QIF).

2.1 Machine Learning

Machine Learning is the field of study that focuses on developing methods of learning patterns from data in a way that generalizes to situations not present in the data. In recent years, significative advances have been observed in Machine Learning research, and also the popularity and applications of some methods increased significantly. One example is the improvement of Neural Network Architectures[**?**], and the popularity of Generative Models[**?**]. Also, some recent research is focused on the theory behind such machine learning methods[**?**], and statistical learning in general [**?**]. Part of the goal of the formalization of such theory is to provide more qualitative guarantees, for instance, in regard not only to accuracy, but also fairness, privacy and interpretability. We discuss these different goals in the next four subsections.

In general, we consider **supervised learning** problems: a machine learning model is an algorithm that receives many data points, which we call **training data**, and outputs a model. This model is itself an algorithm that receives a data point without the target value and outputs a prediction of the target value, which represents one or more variable of interest. This is callled supervised learning because the algorithm has acces to the target variable during the training process.

2.2 Accuracy

Accuracy is the notion of how close some estimate is to the true value we are estimating. In the context of Machine Learning, it represents how close the predictions of a given model are to the real value of the variable the model aims to predict. For binary classification (the scenario in which the target value has only two possible values), accuracy is defined in machine learning as $\frac{TP+TN}{TP+TN+FP+FN}$, where:

1. TP is the number of True Positives: how many predictions were labeled as True and were really True.
2. TN is the number of True Positives: how many predictions were labeled as True but were actually False.
3. FP is the number of True Positives: how many predictions were labeled as False and were really False.
4. FN is the number of True Positives: how many predictions were labeled as False but were actually True.

So, the usual notion of accuracy is the proportion of the predictions from the model that were correct. This generalizes to multiclass classification problems (in which the target variable has a finite number of possible values) by considering the proportion of times that the model's prediction was correct. Regression problems are the ones in which the target variable has an infinite number of possible values but can be codified as a vector of numbers, for instance, the value of some building at two different times. We can access how accurate a regression model is in many ways, for instance the square difference between the prediction value and the real value, added through all training data points.

One important point to consider is that it is better to evaluate how accurate the system is with data different from the data used for training. This is because during the training phase the Machine Learning algorithm usually has the goal of providing the model that provides the best possible accuracy in the training data, among the possible models supported. If there is enough freedom among the possible models, then it might be possible to obtain a model that has a very good accuracy, but if we try to use this same model on other data, this same model performs very badly. For instance, if all functions from the training data to the output are allowed, then a model that simply memorizes the training data and outputs the correct result by looking at the memorized data and outputs a random answer if the input is not in the training data will have one hundred percent accuracy in the training data, but we can obviously provide no guarantee for data points not present. This problem is called **overfitting**, and is usually avoided by limiting how powerfull the model can be, in conjuction with verifying real accuracy values with data other than the training data, and we call this the **testing data**. Also, notice that to evaluate the model, the test data should also have the target value of each data point.

The classical goal of Machine Learning is to provide models with good *test accuracy*, and an Machine Learning algorithm that produces models such that good results on the training data reflet on good results on testing data are said to **generalize** well. However, with the groth of Machine Learning applications in real-life scenarios and the impact on society as a whole, there has been a crescent focus on aspects other than simply maximizing the accuracy, in a similar way that we usually are worried only about how fast a cryptographic algorithm is but also about how safe it is.

2.3 Fairness

In the context of Machine Learning, fairness refers to the reduction, as much as possible, of **algorithmic bias**. Algorithm Bias is the bias introduced by algorithmic decisions. This bias might have a big social impact, as this can further existing unfair discrimination in society, as machine learning algorithms are being used to make more and more important decisions. One famouse example is the COMPAS recidivism

algorithm, that has been used by the United States courts to estimate how likely someone is to reoffend in the future. It was revealed [2] that this tool was heavily biased against black people.

We will say that the result is **positive** for a data point if it benefits the person represented by that data point, and **negative** otherwise. We will say that the **unpriviledged group** is the group of people affected negatively by the bias, and the **priviledged group** is the other group of people.

Such biases can happen because of many factors. The algorithm itself might be introducing bias, or the data might be biased. The data might have been collected in a biased way (in the compass example, this would be the case if reincidivism data was collected more for black reincidivists than for white), or the data might be simply reinforcing some bias of the society.

Also, the bias in society might be such that the data is in disagreement with reality (the unpriviledged group true values for the target variable would affect them in the same way as the priviledged group), or it is in agreement with reality because of structural biases in society. For instance, if the prediction of the algorithm is whether or not someone will have good grades if accepted to some university, people in the unpriviledged group might not have had as good oportunities in life as people in the priviledged group, so the data is correct when it says that those people will have worse grades. Even though, the results might still be considered unfair: this depends on the notion of fairness we consider.

Many different notions of algorithmic fairness have been developed, and some are not compatible [**?**].

2.4 Privacy

2.5 Interpretability

2.6 Causality

2.7 Quantitative Information Flow

3 Theoretical Reference

Começar mencionando artigos ou livros que falam sobre esses assuntos individualmente, mencionando a importância deles e tal.

3.1 Accuracy × Fairness

3.2 Accuracy × Privacy

3.3 Accuracy × Interpretability

3.4 Accuracy × Causality

3.5 Accuracy × QIF

Utility

3.6 Fairness × Privacy

3.7 Fairness × Interpretability

3.8 Fairness × Causality

3.9 Fairness × QIF

3.10 Privacy × Interpretability

3.11 Privacy × Causality

3.12 Privacy × QIF

3.13 Interpretability × Causality

3.14 Interpretability × QIF

3.15 Causality × QIF

4 Contributions

Citar artigos lidos e o livro de causalidade, falando que as atividades consistiram principalmente de entender um pouco mais do assunto, e dar uma discutida resumida das coisas que foram vistas.

5 Conclusions and future work

Aqui dá pra falar sobre quais dos caminhos parecem mais promissores pra fazer uma pesquisa futura...

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] 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)

[3] 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

[4] 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

[5] 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

[6] 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)

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

[8] 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

[9] 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)

[10] 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