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

Type: Scientific
Advisor: Mário Sérgio Alvim

Artur Gaspar da Silva

05/04/2024

Abstract

Do que podemos falar aqui?

1 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 Explainability. 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[9]. 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 naturally-interpretable models (as explanations to more complex models) for auditing systems and checking if they are fair, although this might lead to problems[11]. 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[10], 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 2 discusses the already developed theoretical work on these areas, section 3 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 4 provides conclusions and possible future lines of work. The rest of this section is dedicated to introducing the concepts of Causality, Fairness, Privacy, Explainability in Machine Learning, and also Quantitative Information Flow (QIF).

1.1 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.

1.2 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.

1.3 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. All of these unfairness possibilities can be futther divided into other types of unfairness, as was done in [**?**].

Besides deliberate bias in the algorithm, such that the results of the algorithm do not reflect the data, and biased data, it is also possible to introduce bias because the algorithm might prioritize making correct predictions for the majority of the population, if it can't make correct predictions for both the majority and the minority. Another possibility is that the prediction might depend on past decisions of the algorithm, and we only know the result if the result provided is positive (for instance, we only know if someone will reincide if we release them). In this type of scenario, according to Learning Theory it's important to take suboptimal decisions to *explore* different options and gather more data [**?**], which might be considered unethical as it might have a big cost to society (releasing someone that's probably going to commit more crimes) or to the individual (not giving a life-saving drug to some patients as an experiment to see the survival rates for that specific group).

Many different notions of algorithmic fairness have been developed, and some are not compatible [**?**]. Initially, the notions of fairness could be grouped into two main types: statistical and individual definitions of fairness[**?**]. Statistical (group) notions of fairness require some statistical metric to be similar for certain demographic groups, and individual notions enforce constraints on pairs of individuals, for instance requiring similar individuals to be treated similarly. Many problems with statistical notions and why they, in general, don't provide good individual guarantees are presented in [**?**][**?**]. Some of these problems include: satisfying the constraints for two protected attributes individually but not to combinations of these attributes, . One problem with both individual and group notions is *composition*: it is not always the case that satisfying fairness constraints in individual, isolated, components of a system imply that fairness constraints will be satisfied for the whole system [**?**]. Finally, there are also causal approaches to fairness notions, which we will discuss more in Section 2.

The techniques developed to reduce unfairness in algorithmic decision making can be divided into *pre-processing*, *in-processing* and *post-processing*. Pre-processing techniques modify the training data to remove biases present there. In-processing techniques modifies the learning algorithm iteself, for instance by changing the objective learning function to include not only accuracy but also adding to it some statistical fairness metric, or including some constraint that it has to satisfy. Post-processing techniques act after the model is trained to reduce the unfairness in the decisions made by such a model.

We summarize below the ways in which unfairness might be introduced:

1. Algorithm results does not reflect the data.
   (a) The algorithm might optimize for the majority only, achieving good overall accuracy even though it's mostly wrong for minorities. This can be considered a type of Aggregation Bias.
   (b) Systematic errors in the algorithm, that leads to biased estimation.
2. The Data can be biased, not reflecting the reality.
   (a) We can have a structural biases in society, such that people in unpriviledged groups do not have the necessary oportunities, but if they were treated similary to the priviledged group by society, they would have similar results. For instance, an unpriviledged group that doesn't have good

education oportunities will have worse scores on exams because of historical discrimination, and although just looking at whether someone is in this group could lead to a good accuracy, it might be only perpetuate current unfair biases in society.

(b) The bias can also be introduced in a way that people in the unpriviledged group were misclassified before the data was colected, for instance maybe capable people in an unpriviledged group usually don't get a job even though they are actually as capable as the unpriviledged group.

(c) Data collection doesn't reflect the reality: Measurement bias (for instance, COMPASS used friend/family arrests as a proxy for a risk score present in the dataset), Ommited Variable bias (this violates assumptions of some learning models, for instance linear regression models usually assumes error terms uncorrelated with the parameters considered in the regression), Representation/Sampling Bias (biased sampling lacking the diversity of the population), Simpson's Paradox (if we don't have data on a confounder, correlations might be spurious [8]).

(d) If the data is collected on a group fundamentally distinct from the one where it will be used, for instance another population (Population Bias) or the same population but at another time (Temporal Bias), unfair bias might be introduced.

(e) Data that relies on people's opinion is prone to many biases: Social Bias (people do what others are doing), Self-Selection Bias (people think that everyone agrees with them), and many others.

3. Data might depend on the algorithm previous output: Presentation Bias (the user is presented to some selected advertisements, for instance), Ranking Bias (search engines ordering results in a biased way), Popularity Bias (more popular itens are shown more). This might strengthen biases through time.

4. Finally, the circunstances can change through time, either by influence of the algorithmt itself or other factors, which can worsen the quality of algorithms previously considered to provide good results (Emergent Bias).

1.4 Privacy

In the context of Machine Learning, a privacy-preserving algorithm is one that doesn't allow information considered private/sensitive to be obtained by unauthorized parties. This has been named *private ML*[**?**], and the private data to be protected can be the data using to train the model or the model parameters and structure itself. It is also possible to use Machine Learning to enhance privacy, or to serve as an attack tool. We focus on *private ML*.

We call **adversary** the agent that wants to discover the private information, and **secret** the private information itself. There are some possible goals of the adversary, she might wish to recover the model itself (Model Extraction Attack) by trying to approximate the function that represents the model, to recover some feature or statistical property of the dataset (Feature Estimation Attack), to discover whether some individual data point is present in the dataset (Membership Inference Attack), or even recover the exact values of individual samples in the datset (Model Memorization Attack). We distinguish between the **White-Box access** and **Black-Box access** scenarios as the situation in which the adversary has or does not have full access to the trained model and their parameters, respectively.

One approach to improving privacy in Machine Learning is by encrypting the data or the model in a way such that the computations can be done with the encrypted data/model, and just the result is decrypted. Secure Multi-Party Computation is also an option if there are multiple parties responsible for this computation. Other option is to obfuscate the data or the model, by introducing pertubation in a way that protects the private information in a way that (hopefully) doesn't worsen much the results. There is also the aggregation approach, that focuses on multy-party computations such that the data of one party remains private to other parties, and is used by aggregation learning, for instance. Finally, it's possible to develop *back-door* attacks: malicious modifications to the model might allow an adversary to infer private information, if she provides the specific inputs that are designed to trigger that. All of these are listed in more detail in [**?**].

1.5 Explainability

Explainability, loosely defined, concerns the ability to assign meaning to why some model took one or more decisions, in a way that can be interpreted by humans.

1.6 Causality

Most of our discussion of causality will be based on the works of Judea Pearl [8]. There are many alternative notions of causality, but most of them are already discussed in [8]. Judea Pearl divides causality into three levels of causation, according to the type of question we want to answer.

The first level deals with questions that can be answered by looking at the data. For instance, questions that can be answered by Bayesian networks, deep neural networks or other machine learning algorithms. According to Pearl, this kind of question can be written in terms of what we see: for instance, given that we see that the floor of an entire street is wet then it probably rained just before. Although the examples usually involve some form of conditional probability, Pearl states that anything that can be computed from the joint distribution belongs to the first level of causation. This includes basically all of the usual machine learning approaches. The notation used by Pearl is the usual probabilistic notation, $P(Y = y|X = x)$, abreviated as $P(y|x)$, this means that we observed $X = x$ and want to know how likely it is for Y to have the value y.

The second level of causation concerns questions of the type: "what's the consequence of some specific action". For instance: "if we make a street wet by throwing water manually into it, what is the probability that it rained?". The difference between seeing and doing is the fundamental distinction between Pearl's first and second levels of causation. Not only is the meaning completelly different, it's also impossible to answer questions at the second level of causation by only looking at the data, extra assumptions are necessary. In our example, if we look only at the data what we see is a strong correlation between raining and the floor getting wet, and in principle we have no idea which one causes which. There are also scenarios in which two variables are strongly correlated but neither one causes the other. For instance, in some places the number of ice cream sales is strongly correlated with the number of deaths by shark attacks, but clearly that's because the times of the year when more people go to the beach are the same of when people buy more ice creams. Pearl proclaims that any method of solving questions at the second level of causation must rely on assumptions beyond the data. One way of structuring many of the relevant assumptions is by using a directed acyclic graph in which each vertex is a variable and each edge represents the causal directions between two variables. There are many other types of assumptions that can be made, Pearl discusses them in detail at [8]. The notation used by Pearl is $P(Y = y|do(X = x))$, which can be abreviated as $P(y|do(x))$ or $P(y|\hat{x})$, which can be interpreted as how likely Y is to have value y when we set the value of X to x manually, and when we say "manually", we mean by an external intervention, that is not causally affected by the variables in the system.

The third level of causality regards questions of the type: "what would have happened had something been different?". For instance, "what would be the average temperature of Earth had the Industrial Revolution never happened?". Note that we make an observation about something that happened, then we think what would have happened if we changed something that already happened. This is what Pearl calls a counterfactual question: it requires imagining alternative worlds. To deal with this kind of question, Pearl proposes Structured Causal Models (SCMs). Also, in the way the Pearl defines counterfactuals, there is a fundamental distinction between counterfactuals and actions with observations: in his notation, when we write the probability of some variable reaching some value given that there was an action and an observation, we interpret this as if the observation came *after* the action. This is different from the counterfactual notation, in which the observation comes before the action. The notation used by Pearl for the counterfactual notion is $P(Y_{x'} = y|X = x)$, which can be interpreted as how likely it is for Y to have value y in the world that we manually set X to x', given that we observed that the value of X is x. This is different from $P(Y = y|do(X = x'), X = x)$, as this should be interpreted as the probability that Y has a value y when we manually set X to x' and then observe that X has value x, which doesn't make sense if $x \neq x'$ (we shouldn't be able to condition on impossible scenarios). This difference of interpretation can lead to confusion, and was mentioned in the second part of the "question to author" at subsection 11.7.2 of [8].

1.7 Quantitative Information Flow

2 Theoretical Reference

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

2.1 Accuracy × Fairness

Carlos' paper, Rachel's paper, Microsoft's paper, and maybe others.

2.2 Accuracy × Privacy

Differential privacy literature should have something

2.3 Accuracy × Explainability

Probably can find it in the explainability literature

2.4 Accuracy × Causality

Maybe we can say that accuracy lies in the first ladder of causation, and usually we want to answer questions on other levels?

2.5 Accuracy × QIF

Accuracy can be seen as a form of utility, maybe this is usefull for statistical disclosure control.

2.6 Fairness × Privacy

Rachel's paper, fairness through awareness maybe, and others. Awareness is about being fair by a generalization of differential privacy if I recall correctly, at least in individual fairness, the relations with group fairness and other stuff.

2.7 Fairness × Explainability

Some people try to judge fairness based on the explanation, we can mention the How to Justify Almost Anything aqui, que pra auditar pode não ser uma boa ideia, só isso mesmo talvez.

2.8 Fairness × Causality

Citar paper Karima sobre noções causais de fairness.

2.9 Fairness × QIF

Citar trabalho do Bruno sobre QIF ao contrário como noção de fairness talvez

2.10 Privacy × Explainability

No idea!

2.11 Privacy × Causality

No idea! Maybe be private to causal discovery? Maybe be private but allow causal discovery? I think that there is a paper by Sylvia about this...

2.12 Privacy × QIF

QIF was desined to work with privacy, quantifying how much the sensitive information is leaking is in a way quantifying privacy.

2.13 Explainability × Causality

Causality is inheretely easier to explain.

2.14 Explainability × QIF

No idea! Maybe a way to see which variables leak more information?

2.15 Causality × QIF

I don't think that the paper I saw was published, but maybe we can mention that.

3 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.

4 Conclusions and future work

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

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

[9] 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

[10] 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)

[11] 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