Identify and Correct Unfairness in Machine Learning Models
A Practical Approach

Davide Scandella

January 31, 2022

Machine learning uses experience-based computer algorithms trained on historical data in order to make predictions or decisions without the need to explicitly program them to achieve such goal. Being able to choose the appropriate input data is crucial for a machine learning developer, both for the creation of a well-performing algorithm, and for the need to keep the decision process fair. What is the best way to address the issue of algorithm fairness in every day practice? Can machine learning models address the issue of fairness and correct potential unfairness in existing data?

Algorithms are a set of mathematical functions aimed at a specific purpose. Once algorithms are trained on data and used to make decisions or predictions, a machine learning model is obtained. A biased model derives from use of biased data during its training phase. Usually, the more complex the algorithm, the less interpretable it is. Model interpretability is often correlated to the issue of fairness, as being able to inspect the process of decision making of a model also helps understand the fairness of such decision.

There have been many attempts to define the concept of algorithm fairness, depending on the context of its application. Pessach and Shmueli [1] provide a definition of fairness in a legal domain. As a *disparate treatment*, that is the "intention of treating an individual based on his/her membership to a protected class", and as a *disparate impact*, that is "affecting member of a protected class more than others even if by a seemingly neutral policy". Friedler et al. [2] provide a more general definition of fairness metric. According to them, a decision-making model relies on the mapping between a "construct space", consisting of the true input parameters intended to be used for the decision, a "decision space", consisting of the actual decision parameters. However, what the model is trained with is an indirect representation of the construct space, called "observed space", not necessarily close to the construct space it intends to represent. If the distance between subjects of decision in construct and observed spaces depends on the membership to groupings like gender, race, religion, algorithmic unfairness needs to be addressed and corrected.

Literature has indicated different possible causes of unfairness [3, 4]:

- Biases that are already part of the datasets used for learning; examples of such biases could be human decisions, erroneous reports.
- Missing data; it causes bias as datasets are not representative of the full target population.
- Bias induced by algorithm targets; models inherently benefit majority groups over minorities in order to minimise prediction errors.

- Bias caused by "proxy attributes for sensitive attributes". Sensitive attributes differentiate privileged and unpriviliged groups, such as race, gender and age, and are not suitable for use in decision-making models. This concept is analogous to the definition given above.

Pessach and Shmueli [1] investigated different mechanisms to reduce unfairness in decision making for machine learning and divided them in three classes: Pre-, In- and Post-Process mechanisms depending on whether the unfairness reduction occurs before, during or after the training of the model.
They also research in literature regarding the emerging research on algorithmic fairness and reported many topics. Here the most valuable findings are listed:

- Fair Sequential Learning: online learning models rely on the fact the decisions taken at a certain step in time may influence the outcome of future decisions. In such systems, it is very important to consider fairness at each step, making sure that algorithmic decisions are consistent over time.

- Fair Adversarial Learning: many strategies have been proposed in the context of using adversarial learning to identify and contrast unfairness presence in machine learning models. One approach suggested to use GANs in feedback loops to assess whether a trained model was fair or not and update it accordingly [5]. Other approaches propose the use of GANs to generate fair syntethic data to be used for training or, most importantly, propose adversarial learning in its domain transfer capability to learn fairly to predict decision outcomes for unprivileged groups.

- Fair Causal Learning: casual learning aims at creating models based on the additional knowledge deriving from cause and effect relationship. Understanding causes and effects in the data, the model may help facing the problem of fairness by assessing which type of discrimination should be allowed and which not. Casual models can also help in case of missing data or dataset containing sample or selection bias [6].

How can a data scientist identify, evaluate and correct models fairness and bias in practice? Several open source libraries are available. Their scope is to help the developers and users to identify and correct fairness issues in datasets and models by using bias mitigations algorithms described previously and various fairness metric. Examples of such libraries are Fairness Measures, FairML, FairTest and Aequitas. However, the most comprehensive open-source library available to date for use is AI Fairness 360 [7]. It is designed to "translate algorithmic research from the lab into the actual practice" and answer to common questions such as "Should the data be debiased?", "Should we create new classifiers that learn unbiased models?", "Is it better to correct predictions from the model?". AIF360 is a valid tool for both fairness researchers and machine learning developers. For the former, it allows experiments and comparison between existing bias detection and mitigation algorithms, contribution and benchmarking of new algorithms and datasets. For the latter, it guides through the choice of metrics and algorithms for unfairness reduction and provides a Python package for detecting and mitigating bias from a practical point of view.

In conclusion, the current essay aimed at illustrating how to deal with algorithmic unfairness in practice. A definition of algorithm fairness and bias was given first. Secondly possible causes of model unfairness were listed. Subsequently, focus was put on current literature trends on unfairness reduction in decision-making models, distinguishing between Pre-, In- and Post-Process mechanism. At last, AI Fairness 360 was introduced and proposed as the most comprehensive library to deal with fairness in practical machine learning problems.

References

[1] D. Pessach and E. Shmueli, "Algorithmic fairness," 2020.

[2] S. A. Friedler, C. Scheidegger, and S. Venkatasubramanian, "On the (im)possibility of fairness," 2016.

[3] A. Chouldechova and A. Roth, "The frontiers of fairness in machine learning," 2018.

[4] F. Martínez-Plumed, C. Ferri, D. Nieves, and J. Hernández-Orallo, "Fairness and missing values," 2019.

[5] L. E. Celis and V. Keswani, "Improved adversarial learning for fair classification," 2019.

[6] E. Bareinboim and J. Pearl, "Causal inference and the data-fusion problem," *Proceedings of the National Academy of Sciences*, vol. 113, no. 27, pp. 7345–7352, 2016.

[7] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, S. Nagar, K. N. Ramamurthy, J. T. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang, "AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias," *CoRR*, vol. abs/1810.01943, 2018.

[8] R. Binns, "What can political philosophy teach us about algorithmic fairness?," *IEEE Security Privacy*, vol. 16, no. 3, pp. 73–80, 2018.