Fairness of Learned Classifiers under Performative Effects

Sophia Günlük, António Góis, Simon Lacoste-Julien, Dhanya Sridhar

Mila, Université de Montréal

There is a growing interest in automating decision-making by using machine learning (ML) models to estimate scores that can be used to rank candidates [2, 4, 1]. Consider, as an example, loan application decisions. An institution might train an ML model from historical data to predict the probability that a candidate will default on their loan, based on various features in their application. For new applicants, the trained model predicts a score that the institution can use to approve or reject loan applications, or at least rank the applicants for further review.

In these ML-based decision-making contexts, the field of algorithmic fairness has developed a number of metrics to assess the disparity faced by different demographic groups [3], as well as by individuals [6]. However, algorithmic decision-making is often dynamic, with individuals responding to the deployment of ML models and their automated predictions. In the loan example, rejected applicants may adapt their applications to get a better outcome, potentially by gaming the classifier or making changes to become more loan-worthy candidates in the future. These dynamics, which we refer to as performative effects, lead to a feedback loop between the model's prediction and the true outcome (e.g., defaulting on a loan or not), which is hard to analyze and is often ignored by existing fairness metrics, which assume a static data-generating process. Recently the fields of strategic classification and performative prediction have begun to address this phenomenon in the context of machine learning [5], but most works ignore potential disparities between different segments of the population. In this emerging line of work, models often do not account for factors that may result in demographic groups adapting differently, which can lead to unfair decisions when applied to the different segments of the population.

The goal of our project is to analyze unfairness in performative ML-based decision-making settings, where individuals are incentivized by model predictions to change their attributes over time. More specifically, we are interested in understanding how, over time, the unfairness of the learned model can persist and perpetuate the systemic biases in settings where individuals put effort into improving their features after receiving an ML-based decision. We propose modeling such systems with four types of variables: protected attributes, causal attributes, spurious attributes, and outcomes of interest. In the loan example, protected attributes could be candidates' specified gender or race while the outcome of interest is defaulting on a loan or not. Causal attributes are those that when changed, would directly affect the probability of defaulting; for our loan example, a person's income. Spurious attributes are associated with the outcome but do not cause it directly, and are often also correlated with one's protected attributes. In the loan example, where someone attended university might be a spurious attribute. In our setting, we assume that an ML model is trained to predict the outcome, and candidates observe their predictions from the classifier. Crucially, we assume that rejected candidates put in effort to improve their attributes in order to receive a prediction value that gives them a positive decision in the future. Which features they change and by how much is defined by their knowledge of the true underlying structural model, the ML classifier, their cost for changing each feature, and a budget of the amount of effort they can put in.

The first key challenge we address is developing new measures of unfairness for this performative setting, where disparities in the ability of each group to improve their scores may compound any disparities at the start. We propose two metrics that capture these disparities in improvement across different segments of the population: i) "improvability" which measures, out of those who could have improved their true outcome, how many would truly improve when adapting according to the classifier, and ii) "gaming", which measures, out of those with a negative true outcome after adaptation, how many fooled the classifier into a positive prediction. Our preliminary experiments have highlighted models in which the learned classifier uses the spurious features significantly for prediction, resulting in a misalignment between the optimal adaptation according to the classifier and the optimal adaptation according to the data generation process. This can be a source of unfairness because if a bank's ML classifier puts a high weight on an applicant's spurious features, such as university degree, then an applicant from a privileged background may not need to put in as much effort into adaptation compared to a person who did not have the same educational opportunities, even with the same causal features. Using the developed metrics, we plan to tackle a second key challenge of understanding why the classifier misaligns with the true causal model. This will allow us to develop ML methods to train decision models on biased data that do not exploit spurious correlations between sensitive attributes and the outcome, thus reducing bias and promoting fairness.

References

[1] Amazon. *How Amazon leverages A.I. and M.L. to enhance the hiring experience for candidates*. URL: https://www.aboutamazon.com/news/workplace/how-amazon-leverages-ai-and-ml-to-enhance-the-hiring-experience-for-candidates.

[2] Edmund L. Andrews. *How Flawed Data Aggravates Inequality in Credit*. Accessed: 2023, September 4. 2020. URL: https://hai.stanford.edu/news/how-flawed-data-aggravates-inequality-credit.

[3] Ozgur Guldogan et al. *Equal Improvability: A New Fairness Notion Considering the Long-term Impact*. 2023. arXiv: 2210.06732.

[4] Jeff Larson et al. *How we analyzed the compas recidivism algorithm*. 2016. URL: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

[5] John Miller, Smitha Milli, and Moritz Hardt. *Strategic Classification is Causal Modeling in Disguise*. 2020. arXiv: 1910.10362.

[6] Victor Veitch et al. *Counterfactual Invariance to Spurious Correlations: Why and How to Pass Stress Tests*. 2021. arXiv: 2106.00545.