Project Summary

Measuring Sources of Bias and Unfairness in Machine Learning

Jamie Morgenstern

Machine learning's explosive growth and integration into our everyday lives, changing the way our health care providers provide us with treatment alternatives, the financial instruments provided to us by different lending agencies, to the the news stories and advertisements we are shown online. The constant interaction of these models with people. their lives and decisions, has been accompanied by many examples of these models behaving in unexpected and alarming ways in less common scenarios. These examples show that machine learning models capture and amplify biases evident in their training data. Without extreme care, gathered data will reflect societal biases.

The study of fairness and bias in machine learning, still in its infancy, has focused primarily upon creating mathematical formalizations of fairness, and proving those constraints degrade the performance of many standard batch learning algorithms. Much of this work suggests that classic models have wildly different performance on different populations due to imperfections or imbalances in our training datasets: if we have 10 or 100 or 1000 times as much data about a majority population as a minority, one would expect models trained on that data to perform better on tasks for the majority as the minority. This could stem from the fact that our models have less statistical information about the minority, or that the training algorithm emphasizes performance on larger populations over smaller ones. These two sources of unfairness (less statistical power and less importance in training) are only two of many which might exist in any particular context. It might separately be that the features collected for this task are more useful in the predictive task for one population than another, or that the set of models the algorithms choose amongst are a better fit for one than another.

We cannot hope to "fix" the unfair treatment of different groups by machine learning models until we understand the most pressing sources of this differing treatment. The sources of bias in datasets and models likely differ in different contexts, as will the appropriate and effective solutions to these biases.

1 Background

The work on fairness in machine learning, in the broadest terms, has mostly focused on two classes of fairness measures, group and individual notions of fairness, in either one-shot settings or settings with feedback loops. Group fairness definitions (including statistical parity, equality of false positives and negatives) capture the idea that an ML model should treat pre-specified groups similarly, on average over those groups. Individual fairness notions, on the other hand, initially popularized in TCS and ML by **?**, constrains a ML model to treat *each particular individual*. Much of this work has focused on the the degradation of overall accuracy when forcing a model to improve its fairness, or displaying the violation of a particular definition of fairness of a particular model or dataset.

The full proposal will give more precise notions of these technical definitions of fairness, and additionally review the relationship between fairness in machine learning and fair division, fair scheduling and other resource allocation, and draw connections between the technical aspects of fairness and the study of fairness in philosophy.

2 Proposed Research

Very little work has actually attempted to measure the relative weight of these potential sources of unfairness of machine learning datasets and models. Our work will focus on answering the following questions.

Question 1. *How do we measure these different sources of unfairness for a variety of datasets and pre-trained machine learning models? Which of these sources are most significant in different domains?*

This will require identifying different causes of different groups being treated differently by machine learning models. Some initially identified sources of unfairness in different domains include less accurate labels for different groups, less importance in the optimization done over the data, and the possibility that less data for a population will correlate with lower accuracy when a model is deployed. These are only a few of many additional imagined and identified sources of unfairness in machine learning. Here, we suggest several others which we hope to quantify in terms of their impact. Are our ML systems using models which are too low-dimensional to accurately capture nuanced correlations for different populations? Alternatively, is the model being used overfitting on subpopulations due to its complexity relative to the amount of data present representing that population? Are the features present in a given dataset more strongly correlated with a prediction task for some populations than others? Are there fundamentally different models with much better accuracy for different populations (e.g., is treating two heterogeneous populations as homogeneous degrading the model's ability to better predict for the minority)? How much of these issues are amplified when considering intersectional groups, e.g. minority women or physically challenged people originating from rural areas? We aim to carefully designing statistical tests to identify these and other sources of bias in datasets and models.

Question 2. *How should we treat these most significant sources?* The answers to Question 1 will spur us towards designing solutions for the most pressing of these sources of unfairness. Some will be more amenable to retraining a similar model subject to some regularization or constraints on the same dataset, others will require collection of additional data or features, and others still will require us to consider different classes models entirely. Our methods developed for question 1 should be simple enough that users of ML can test for these early on in their generation of data and models. This way, fewer resources will be wasted gathering datasets and training models with fundamental flaws in terms of its disparate treatment of different populations.

Question 3. *What part of the ML "pipeline" need the most work in terms of guaranteeing fairness?*

Much of the work combating bias of ML models focuses on different parts of the process of selecting an ML model in isolation, and in particular the area has focused on (1) preprocessing datasets by adding noise to labels or reweighting/subsampling, (2) modifying the training procedure using constraints or a modified objective, or (3) postprocessing a black-box model to be fair. Understanding how these pieces work together, and how "meta-questions" such as choice of features to gather or provide as input to an algorithm, or the choice of models to learn with respect to, will give a richer set of tools to improve ML's ability to fit technology to improve the lives of everyone rather than just those mist similar to the majority.

3 Summary: Significance of proposed work

3.1 Intellectual Merit

We summarize the main points of intellectual merit of this work. This work will provide a careful way to measure different sources of bias of ML in different domains. It will provide helpful guidelines for data collection, feature selection, model training and measurement, when the organization involved wants to prioritize their system treating different populations equitably.

Finally, the proposed work brings together work from both the machine learning, theoretical computer science, and datamining communities. This will allow us to draw insights from each of these communities, and make new connections.

3.2 Broader Impacts

The PI plans to take this line of research and integrate them with educational goals as follows:

- While developing a methodology for measuring sources of unfairness in machine learning, the PI proposes to release tools for automating these measurements with publicly available code, which will help non-experts pinpoint certain sources of bias in their datasets or models, so that earlier on in the development of a machine learning system these imbalances can be idenitifed and addressed.

- Involving undergraduates who currently participate in the Georgia Tech Data Science Team in gathering and analyzing new human-centric datasets. This group has a large amount of interest in machine learning, and are in a prime position to think of unique settings in which people are generating data. Many of these students are on the cusp of deciding between working in industry or applying for graduate school. This will expose these students to research in machine learning and data measurement. This will additionally generate new datasets to test our methodology of measuring unfairness and the sources of unfairness in new domains.

- Additionally, each group of students which generates one of these datasets will be encouraged (by offering an NSF REU position) to analyze the datasets using the methodology developed by the PI and her graduate students. The results of this analysis, and careful documentation of the data gathering process, will be made publicly available in an easily digestible format, as the PI advocates for in ongoing related work (**?**). The PI will encourage the students to record a nontechnical podcast describing their dataset, their inspiration for gathering the data and its potential use-cases, as well as describing any discrepancies in the representation of different groups in the dataset. Making these resources accessible to the less technical to the public can encourage young people to think about the world through the lens of machine learning and data science. Such reorientation will have two critical effects: thinking carefully about the data one is generating for organizations, and how that might be used by the organization, and second, drawing younger people into thinking about machine learning as a future career path.