Project Summary

Measuring Sources of Bias and Unfairness in Machine Learning

Jamie Morgenstern

PROJECT DESCRIPTION

1 Introduction

Machine learning's explosive growth and integration into our everyday lives, changing the way our healthcare 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.

2 Background

Should we regularize? Should we gather additional data? Should we consider more or less complex models? Should we use two models for different po

3 Proposed Research

Very little work has actually attempted to measure the relative weight of these potential sources of unfairness in our

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

1. Classifying types of possible unfairness sources
1. Methodology of measuring

Question 2. *Which of these sources of unfairness are most significant in different domains?*

Question 3. *How should we treat these most significant sources?* 1. Constraints in learning? 1. Change in design of data collection process? 1. early collection measurement guiding what should be gathered 1. Change in our description of a dataset/model's documentation to acknowledge the limitations of the dataset/ML model

Question 4. *What part of the ML "pipeline" need the most work in terms of guaranteeing fairness?*

Intellectual Merit

Goal: measure where bias in ML comes from

Not having demographic info

Underfitting: too little data for some populations?

Underimportant in optimizations?

Less realizable for the models being used?

Less mutual information in the set of features that were collected for the problem at hand?

Subset over/underfitting

How much of this is more of a problem for intersectional definitions of minorities? Text ....

4 Summary: Significance of proposed work

4.1 Intellectual Merit

We summarize the main points of intellectual merit of this work.

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.

4.2 Broader Impacts

Broader impacts

Data science team : collect and analyze data- undergrads or even high schoolers

Develop tools for measuring these for different datasets, making this easier for different groups to know/modify/describe the bias in their data science

Develop a course on the foundations of fairness in ML

share the benefits of datascience/ML with the broader world