Statistical and Algorithmic Fairness and the Seldonian Algorithm (temp)

Dasha Asienga

2023-11-06

Introduction

The public and private sector are increasingly turning to data-driven methods to automate and to guide simple and complex decision-making. However, this trend raises an important question of bias. There is a lot of misinterpretation when it comes to the collection of data in many application areas, and there is a major concern for data-driven methods to further introduce and perpetuate discriminatory practices, or to otherwise be unfair because of the social and historical processes that operate to the disadvantage of certain groups.

For example, within healthcare, using mortality or readmission rates to measure hospital performance penalizes hospitals serving poor or non-White populations as those inherently have higher mortality and readmission rates due to confounding societal factors. Outside healthcare, credit-scoring algorithms predict outcomes based on income, which disadvantages low-income groups further perpetuating economic immobility. Policing algorithms result in increased scrutiny of black neighborhoods because of the bias against black people that is already present in the U.S. policing system, and hiring algorithms, which predict employment decisions, are affected by historical race and gender biases.

Yet, these algorithms are regarded as ground truth and free of human limitations because they are based on mathematics, statistics, and computer science – otherwise regarded as objective disciplines. In theory, this should lead to greater fairness. However, left unregulated, these mathematical models privilege majority groups and discriminate against minority groups because they often learn from inherently biased data. If the data used to train models contains bias, then the resulting algorithms will learn the bias and reflect it into their predictions. In some cases, this can be detrimental.

While there are widely-accepted, though sometimes disputed, social and societal notions of fairness, one key question emerges: are there any established statistical notions of fairness and bias? Is it possible to mathematically and statistically define algorithmic bias and unfairness, thereby paving a way for addressing the challenges they pose? This thesis paper aims to explore and answer precisely this question.

Algorithmic Bias

There are multiple different types and sources of bias in the realm of statistics. In particular, algorithmic bias arises when an algorithm's decisions are skewed towards a particular group of people, either positively or negatively (Mehrabi, 2021). The danger with biased algorithmic outcomes is that they can generate a feedback loop. Take, for example, a hiring algorithm that discriminates against female applicants for a specific job. In the long run, this can perpetuate, and even amplify, existing gender biases by further widening the gender-based class imbalance.

One such key example of algorithmic bias often cited in literature is regarding the broad use of the COMPAS – or the Correctional Offender Management Profiling for Alternative Sanctions – tool to predict a defendant's risk of recidivism – committing another crime – within two years.

talk about COMPAS, statistical results from different studies, and the detrimental effects of the bias that arises

Conclude that we want to capture this bias/ unfairness mathematically to segway into the next section on fairness definitions.

Fairness Definitions

Talk about group v individual and also classification v regression. Preface the focus on group definitions.

Group (classification):

Independence

Sufficiency

Separation

Group (regression):

Fairness Conflicts

Addressing Algorithmic Bias