Dasha Asienga

STATISTICS HONORS THESIS PROPOSAL

Proposed Title: Statistical Notions of Fairness and Algorithmic Bias: A Focus on Support Vector Machines.
Proposed Related Course: COSC 247: Machine Learning (Spring '23).

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. Outside healthcare, credit-scoring algorithms predict outcomes based on income, which disadvantages low-income groups that have no economic mobility. Policing algorithms result in increased scrutiny of black neighborhoods, and hiring algorithms, which predict employment decisions, are affected by race and gender biases. We have some ideas of societal notions of fairness regarding disparities within certain demographic identities, but are there any such statistical notions of fairness and bias within classification and prediction algorithms?

There are many different types of bias, including but not limited to omitted variable bias, representation bias, sampling bias, aggregation bias, self-selection bias, and Simpson's paradox. However, the primary focus of this proposed thesis is algorithmic bias, which refers to the different ways in which prediction and classification algorithms replicate and even amplify human biases. Biased algorithms can have huge societal consequences through decisions that have a collective, disparate impact on certain groups of people. For example, an algorithm that automates bail, parole, and sentencing decisions, if not thoroughly audited for bias, can generate incorrect conclusions that target certain groups.

Statistical fairness, on the other hand, refers to classifiers behaving appropriately equally on average across protected groups according to some metric. However, a model that appears fair with respect to individual groups may actually discriminate over specific intersections of those groups. A preliminary step in this thesis work would be understanding ways of identifying and measuring algorithmic bias. Some existing statistical methods include defining protected demographic groups, statistical and demographic parity, raw positive classification rate, equalized odds, false positive and false negative rates, and positive predictive value, most of which are probabilistic definitions.

The next step would then be understanding different ways to correct such bias. Some existing techniques include transfer learning, which involves learning each group separately, multi-task learning, fair regression, and synthetic minority over-sampling technique (SMOTE). A particular method of interest to explore in greater detail would be fair representation learning, which is a data debiasing process that produces transformations of the original data that retain as much relevant information as possible while removing information about sensitive or protected attributes. This method is particularly applicable to support vector machines, which are a set of supervised learning methods that are highly effective in high-dimensional spaces, even when the number of dimensions is higher than the number of sample observations. To gain a deeper understanding of support vector machines, the thesis work would involve studying their objective/ loss function, regularization of bias and variance, the use of kernel functions to classify data in higher dimensions, the potential for algorithmic bias, the potential for correcting this bias using fair representation learning, and the corresponding trade-offs between fairness, accuracy, variance, overfitting, and other metrics in a simulation study, as well as the quantifying long-term impact for different populations. Another method of interest is the Seldonian algorithm which is based on the premise that if 'unfair' or 'unsafe' behavior can be defined mathematically, then it should be possible to create algorithms that can learn from the data on how to avoid these unwanted outcomes with high confidence.

The proposed thesis would be concluded with an application of the learnings to a real-world biased data set of interest. There are many application areas and biased data sets available for use in studying the topic of bias. The COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) dataset is a popular commercial dataset fed into an algorithm used by judges to score a criminal defender's likelihood of recidivism (reoffending) and to aid in bail and parole decisions.

Ultimately, there are numerous potential future applications of this study in healthcare, college admissions, social justice, facial recognition software, employment, retail, advertisement, and credit institutions, to name a few. Further research in this area can have significant implications in various fields, and while it is not possible to have a singular definition nor to satisfy all definitions of statistical fairness, understanding and addressing algorithmic bias is crucial in promoting fairness in data-driven decision-making.

References

Chouldechova, A., & Roth, A. (2020, May 1). *A Snapshot of the Frontiers of Fairness in Machine Learning*. Communications of the ACM. Retrieved March 30, 2023, from https://cacm.acm.org/magazines/2020/5/244336-a-snapshot-of-the-frontiers-of-fairness-in-machine-learning/fulltext

La Cava, W., & Moore, J. H. (2020, April 28). *Genetic Programming Approaches to Learning Fair Classifiers*. arXiv.org. Retrieved March 31, 2023, from https://arxiv.org/abs/2004.13282

Lee, N. T., Resnick, P., & Barton, G. (2022, March 9). *Algorithmic Bias Detection and Mitigation: Best Practices and Policies to Reduce Consumer Harms*. Brookings. Retrieved March 31, 2023, from https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2022, January 25). *A Survey on Bias and Fairness in Machine Learning*. arXiv.org. Retrieved March 31, 2023, from https://arxiv.org/abs/1908.09635

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019, October 25). *Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations*. Science. Retrieved March 31, 2023, from https://www.science.org/doi/10.1126/science.aax2342

Thomas, P. S., Castro da Silva, B., Barto, A. G., Giguere, S., Brun, Y., & Brunskill, E. (2019, November 22). *Preventing Undesirable Behavior of Intelligent Machines*. Science. Retrieved March 31, 2023, from https://www.science.org/doi/10.1126/science.aag3311