Generalized Subgroup Fairness in Machine Learning

Julie E. Cestaro

NYU Gallatin School of Individualized Study

Master's Thesis Proposal

Advised by Professor Daniel B. Neill

15 June 2024

Introduction

In an effort to avoid the subjective and occasionally capricious nature of human decision making, predictive models have proliferated as they were hailed as ideal decision makers: both rational and objective. These models were, and continue to be, integrated into a variety of decision points across different high-impact industries including policing, healthcare, and finance (Al-Fuqaha et al., 2021; Astya et al., 2021; Rezende, 2020).

However, predictive models are not the ideal decision makers that they may have appeared to be initially, instead exhibiting their own biases in the decision making process. This was most prominently demonstrated in 2016 when Julia Angwin and Jeff Larson published an audit of the COMPAS recidivism algorithm, which spurred on further investigations into the distribution of errors in the outputs of predictive models and how they may be subject to their own biases (Angwin et al., 2016).

Due to these findings, among many others published in the following years (Bolukbasi et al., 2016; Buolamwini & Gebru, 2018), it is imperative that predictive models are rigorously audited for bias. Early methodologies for bias investigation would frequently involve pre-defining a set of groups for some protected attribute (like race or gender) and requiring parity of some fairness statistic across all of these groups. While this may appear to enforce fairness for all groups within the protected class, this methodology does not protect the subgroups of these groups. For example, if fairness is enforced by parity for all genders and for all age ranges, parity may not hold when looking at a group of young women and a group of young men (Angwin & Grassegger, 2017). Given this, it is not enough to simply enforce fairness at the group level. Fairness must also be maintained for the exponentially large set of subgroups within a population.

Related Work

In 2016, Neill and Zhang proposed Bias Scan, a fast subset scanning methodology to search an exponentially large set of subgroups and determine which of these subgroups have the most miscalibrated predictions. One of the strengths of Bias Scan is that it audits for outcome probabilities being either over- or under-estimated for some subgroup, which sets it apart from similar methods. However, it is limited by two key aspects: first, by relying only on calibration as the fairness statistic, and secondly by identifying the most miscalibrated subgroup of *all* the subgroups. This means that if a set of predictions exhibits a racial bias against Black females as compared to non-Black females, Bias scan would not identify this bias unless the subgroup of Black females was the most miscalibrated subgroup of the population (Neill & Zhang, 2016).

In 2018, two additional methodologies for enforcing this kind of subgroup fairness were published: GerryFair and Multiaccuracy Boost. GerryFair trains a set of linear regressions to estimate the extent to which the predictions for a subgroup have deviated from the baseline error rate. Unlike Bias Scan, GerryFair does not audit for calibration specifically, but instead examines equal opportunity[1] and statistical parity[2]. Multiaccuracy Boost is similar, but primarily focuses on correcting the bias for all subgroups. In other words, it does not guarantee the identification of the most biased subgroup, but instead makes iterative corrections to the predicted log odds of some subgroup (Kearns, et al., 2018; Kim, et al., 2023).

Methodology

The research for my thesis will focus on extending the Bias Scan methodology -- scanning all of the possible subgroups and identifying the subgroup with the most miscalibrated

[1] Kearns, et al. use the definition of equal opportunity established by Hardt, et. al to be the equality of false negative rates (Hardt, 2016).
[2] Kearns et al. use the standard definition of statistical parity as equality of test positive rates.

predictions as compared to the observed outcomes -- to include scans for both separation and sufficiency in addition to the original scan for calibration[3].

Sufficiency ensures that the outcome given the model's prediction is consistent across all subgroups. Mathematically, this means that the ground truth is independent of group membership given the score that the model predicts. Separation, on the other hand, is concerned with maintaining consistent log loss in predictions and consistent error rates in recommendations across subgroups. This means that the prediction is independent of group membership given the ground truth.

Incorporating these additional definitions into the Bias Scan methodology will involve allowing the scoring function to accept different definitions of fairness when performing the scan for deviations among subgroups. I will refer to the methodology outlined by Menghani, et. al for false positive rate scan for subgroups with a higher rate of false positive errors as an element of the generalized separation scan towards which my work will build (Menghani, et al., 2023).

I will also add an option to condition subgroup separation and sufficiency on a specific protected class. This will compare the performance of the model for each subgroup of the protected class to the same subgroup outside of the protected class. This methodology is already outlined in Auditing Predictive Models for Intersectional Biases and will be incorporated as a specific functionality of this more generalized framework (Boxer, et al., 2023).

The primary data sets used in this research will be the COMPAS dataset or the German Credit Dataset, both of which have become benchmarks in the field, and both of which have categorical features with binary outcomes, making them suitable for the scope of this particular problem. While they are the standard datasets used in this type of work, they are not without fault. Therefore, while I expect to benchmark on one or both of these datasets, an element of my research will also involve finding and processing an additional dataset to use in this work.

---

[3] Note that while calibration and sufficiency are similar, a predictive model may demonstrate sufficiency while being equally miscalibrated for all groups. Therefore, it is necessary to view each as separate and necessary fairness criteria.

Evaluation of my methodology will occur in two parts. First, I will inject bias into a dataset by shifting the base rate of a randomly selected subgroup of the population represented by the dataset. Then, I will perform the new bias scanning methodology and evaluate the extent to which the injected bias is detected. I will perform this analysis for each measure of bias that I expect the new methodology to identify. Secondly, after verifying that synthetic bias can be detected by the methodology, I will perform discovery work on an unaltered dataset to identify any previously undetected biases. I will benchmark my findings against two similar methods that also focus on subgroup fairness: GerryFair and MultiAccuracy Boost (Kearns, et al., 2018; Kim, et al., 2023).

Limitations

A key limitation of this work, and the work of applying the notion of “fairness” to machine learning systems in general, is how fairness itself is defined. By the nature of auditing a model given specific definitions of fairness, we become limited to understanding fairness only as far as these definitions. The goal of this work of generalizing the original Bias Scan methodology is to make it less limited in detecting violations of fairness, but this set of definitions remains finite and may not generalize as new definitions of fairness are discovered or introduced.

Further, by nature of writing an auditing methodology for a predictive model, the implication is that the solution to bias in this area is to satisfy separation and sufficiency across all subgroups. This work does not account for instances where a predictive model should not be used in the first place or where a new dataset is needed to achieve results more reflective of reality.

Conclusion

Society is not likely to slow in its application of machine learning and ML-backed artificial intelligence systems to various problem spaces. In fact, the trend seems to be toward outsourcing ever more high impact decisions to machine learning. As we willfully progress (or are involuntarily yanked) down this path, it becomes imperative to recognize the importance of the deliberate pursuit of fairness and equity in machine learning along the entirety of the machine learning pipeline, which encompasses everything from initial data collection and the training of the model to predictions and the impact of those predictions.

References

Al-Fuqaha, A., Bilal, M., Qadir, J. & Qayyum, A. (2021). Secure and Robust Machine Learning for Healthcare: A Survey. *IEEE Reviews in Biomedical Engineering, (14)*, 156-180. https://doi.org/10.1109/RBME.2020.3013489.

Al-Fuqaha, et. al review recent uses of machine learning and deep learning in healthcare settings, focusing on cardiovascular procedures and imaging to explore security and privacy concerns related to these use cases and present novel methods to mitigate the identified challenges. This paper was published at the onset of 2020, just as a new phase of fairness research was on the precipice.

Angwin, J. & Grassegger, H. (2017, June 28). *Facebook's Secret Censorship Rules Protect White Men From Hate Speech But Not Black Children*. ProPublica. https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms.

Julia Angwin and Hannes Grassegger's investigation of Facebook's censorship policy in the wake of a terrorist attack in London and recent Black Lives Matter activism demonstrate the need for subgroup fairness. Their key finding is that while Facebook's content moderation promises to protect all *groups* equally, it does not protect all *subgroups* equally.

Angwin, J., Kirchner, L., Larson, J. & Mattu, S. (2016, May 23). *Machine Bias*. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Angwin, et. al's 2016 investigative report uncovers the fallibility of more than a decade's worth of use of software called COMPAS by New York State to predict recidivism rates of defendants to help judges determine criminal sentencing. This story became the staple example of human bias perpetuated by machine learning, finding that the formula was

particularly likely to falsely flag Black defendants as future reoffenders while White defendants were often incorrectly labeled as unlikely to reoffend.

Astya, R., Sinha, J., Tripathi, K., Verma, A. & Verma, M. (2021). Machine Learning based Loan Allocation Prediction System for Banking Sector. *2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*. https://doi.org/10.1109/ICAC3N53548.2021.9725599.

This paper proposes a new model to determine whether potential loan customers are reliable to pay their loan amount back in response to the backlog in India's banking sector. If proven useful, this model would alleviate a significant burden on the country's financial sector.

Bolukbasi, T., Chang, K., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.* 1-25. https://doi.org/10.48550/arXiv.1607.06520.

Bolukbasi et al. define metrics to quantify both direct and indirect gender biases in embeddings, and develop algorithms to "debias" the embedding. They empirically demonstrate that our algorithms significantly reduce gender bias in embeddings while preserving its useful properties such as the ability to cluster related concepts and to solve analogy tasks. The resulting embeddings can be used in applications without amplifying gender bias.

Boxer, K., McFowland, E., & Neill, B. (2023). Auditing Predictive Models for Intersectional Biases. https://doi.org/10.48550/arXiv.2306.13064.

Boxer, et. al, introduce a method for subgroup fairness that allows the auditor to investigate the subgroups of a given protected class, and compares that subgroup to the analogous subgroup outside of the protected class.

Buolamwini, J. & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. 77-91. https://proceedings.mlr.press/v81/buolamwini18a.html.

Buolamwini and Gebru demonstrate that the primary datasets used to identify faces are overwhelmingly composed of light-skinned subjects, resulting in substantial disparities in classifying individuals of different skin tones -- primarily dark-skinned women.

Hardt, M., Price, E., & Srebro, N., (2016). Equality of Opportunity in Supervised Learning. https://doi.org/10.48550/arXiv.1610.02413.

Hardt, et. al introduce formal definitions for equalized odds and equal opportunity as an interpretable measure for detecting discrimination, and further offer a framework for constructing classifiers that meet these definitions of fairness. The definitions for equalized odds and equal opportunity put forth in this paper are still relevant to fairness research today.

Kearns, M., Neel, S., Roth, A., & Wu, Z., (2018). Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness*. International Conference on Machine Learning.* https://doi.org/10.48550/arXiv.1711.05144.

This group of researchers present an alternative in fairness to statistical definitions, instead proposing to demand statistical notions of fairness across infinitely many subgroups, as statistical definitions appear to be fair on each individual group but often fail to be fair on one or more subgroup.

Kim, M., Ghorbani, A. & Zou, J. (2023). Multiaccuracy: Black-Box Post-Processing for Fairness

in Classification. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*.

https://doi.org/10.48550/arXiv.1805.12317.

This group introduces multiaccuracy auditing and post-processing to ensure accurate predictions. This proposal uses a black-box framework that allows for improved fairness and accountability even when the predictor is minimally transparent.

Menghani, N., McFowland, E., & Neill, D. (2016). Insufficiently Justified Disparate Impact: A New Criterion for Subgroup Fairness.

https://doi.org/10.48550/arXiv.2306.11181.

In this paper, Menghani, et. al introduce a new fairness criterion: insufficiently justified disparate impact. They extend the Bias Scan methodology put forth by Neill & Zhang (2016) to identify significant error rate imbalances, as well as to scan for their new fairness criterion.

Neill, D. & Zhang, Z. (2016). Identifying Significant Predictive Bias in Classifiers. *NIPS 2016 Workshop on Interpretable Machine Learning in Complex Systems*.

https://doi.org/10.48550/arXiv.1611.08292.

This paper introduces a subgroup scanning methodology that scans over the exponentially large set of subgroups of a population to find the subgroup with the most miscalibrated predictions. The methodology introduced here will be the foundation for my proposed generalized bias scan.

Rezende, I. N. (2020). Facial recognition in police hands: Assessing the 'Clearview case' from a European perspective. *New Journal of European Criminal Law, (11)*3,

https://doi.org/10.1177/2032284420948161.

A history of law enforcement use of Clearview AI's facial recognition technology and review in light of European Union (EU) legal framework on privacy and data protection, assessing data scraping practices, the transfer and handling of said data, and the app's

compliance with lawful processing of biometric data and use of the strict necessity tests. This assessment will suggest that Clearview's app is highly problematic in criminal proceedings.