10 April 2020

Fairness with Multiple Protected Attributes A Causal Approach

Kailash Karthik S (ks3740) and Shiv Vidhyut (skv2109)

ABSTRACT

Most of the contemporary literature on fairness and bias using causal inference have focused on counterfactuals which have made it disjoint to the rest of the classical machine learning literature. This project aims to draw parallels between the two ideologies by formalizing causal analogues of classical machine learning fairness notions including the impossibility theorem and existing metrics like Equalized Odds, Predictive Parity and Demographic Parity. The project also aims to extend current literature on causal fairness to multiple protected attributes by (1) investigating the effects of intervention on one protected attribute on the others and (2) developing a debiasing algorithm in multi-attribute scenarios. We also aim to find bounds on fairness in situations where the evaluation of counterfactuals is not feasible to make the causal metrics usable under a wider range of scenarios.

Project Structure

The project goals have been structured such that (1) there exists a minimum set of achievable results so that a final report can be prepared; and (2) there exists an exploratory component to do introductory research in causal inference.

Fundamental Goals

1. While there have been statistical proofs for the impossibility theorem, the aid of causal diagrams yields simple a yet elegant proof of the theorem. It also helps to articulate specific "trivial" scenarios where the impossibility can be violated.

   We aim to construct this proof and conditions of the theorem's violation.

2. The causal equivalents of the existing metrics of fairness like demographic parity, equalized odds and predictive parity have been mentioned in passing in the causal inference literature.

   We aim to formalize these notions and interpret the relationship between the statistical and causal equivalents.

3. Much of the current literature on fairness focuses on a single protected attribute. An intersectional notion of fairness was introduced, but this is a statistical approach. We aim to investigate the interventional effects of protected attributes on each other by conducting an experimental analysis on the COMPAS and German Credit Datasets.

   The goals of the experiments are to infer causal relationships between the protected attributes and empirically validate/refute the intersectionality theory of fairness in these data settings.

Exploratory Goals

1. Devise a method to measure interaction of bias in the presence of multiple protected attributes. Here we will assume that the underlying causal mechanism and set of protected attributes are known.

2. Once the relationship between the attributes can be formally understood, devise a mechanism to debias the data. For this, we will base ourselves on current single attribute debiasing approaches in the current causal literature, possibly one that is based on propensity score.

3. A criticism with the notion of counterfactual fairness has been that it identifying counterfactuals demands certain strict conditions to be satisfied. Thus, when they are non-identifiable, a possible alternative would be to provide bounds on fairness using counterfactual bounds and we will formalize this.

References

Balke A, Pearl J. Counterfactual probabilities: computational methods, bounds and applications. In Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence, Seattle, WA, 1994; 46{54.

Bilal Qureshi, Faisal Kamiran, Asim Karim, Salvatore Ruggieri and Dino Pedreschi. Causal Inference for Social Discrimination Reasoning, 2016; arXiv:1608.03735.

James Foulds, Rashidul Islam, Kamrun Naher Keya and Shimei Pan. An Intersectional Definition of Fairness, 2018;

Jon Kleinberg, Sendhil Mullainathan and Manish Raghavan. Inherent Trade-Offs in the Fair Determination of Risk Scores, 2016; arXiv:1609.05807.

Junzhe Zhang and Elias Bareinboim. Fairness in decision-making{the causal explanation formula. In 32nd AAAI Conference on Artificial Intelligence, 2018.

Junzhe Zhang and Elias Bareinboim. Equality of opportunity in classification: A causal approach. In Advances in Neural Information Processing Systems.

Matt J. Kusner, Chris Russell, Joshua R. Loftus and Ricardo Silva. Causal Interventions for Fairness, 2018.

Matt J. Kusner, Joshua R. Loftus, Chris Russell and Ricardo Silva. Counterfactual Fairness, 2017; arXiv:1703.06856.

Niki Kilbertus et al. Avoiding Discrimination through Causal Reasoning, 2017, Advances in Neural Information Processing Systems 30, 2017, p. 656–666.

Silvia Chiappa and William S. Isaac. A Causal Bayesian Networks Viewpoint on Fairness, 2019, Springer, Cham, 2019;