Title: Bayesian Approach to Estimating Counterfactual Fairness Measures: A Case Study on COMPAS Recidivism Risk Score

Saeyoung Rho
Columbia University

Algorithmic fairness is receiving increasing attention due to the wide adoption of algorithms with significant socioeconomic implications, such as recidivism prediction, loan approval, and hiring decision making. Despite various definitions of algorithmic fairness proposed based on a counterfactual framework, the identification of counterfactual probabilities remains challenging without a concrete solution for measuring such quantities. In this paper, we propose a Bayesian approach to evaluate counterfactual fairness measures, given a graphical structure and data. To demonstrate the process, we analyze COMPAS recidivism risk score to identify the counterfactual fairness measure that signifies the effect of altering protected attributes (race, age, sex) on the algorithm's risk score output. Our results indicate that the degree of fairness/discrimination varies depending on the choice of the fairness definition.