Algorithmic Fairness as a Missing Data Problem: A Causal Perspective

Amit Sharma[1]

[1] Microsoft Research, Bangalore, India

Abstract:

Training datasets for machine learning often have some form of missingness. For example, to learn a model for deciding whom to give a loan, the available training data includes individuals who were given a loan in the past, but not those who were not. This missingness, if ignored, nullifies any fair ness guarantee of the training procedure when the model is deployed. Using causal graphs, we characterize the missingness mechanisms in different real-world scenarios. We show conditions under which various distributions, used in popular fairness algorithms, can or cannot be recovered from the training data. Our theoretical results imply that many of these algorithms cannot guarantee fairness in practice. Modelling missingness also helps to identify correct design principles for fair algorithms. For example, in multi-stage settings where decisions are made in multiple screening rounds, we use our framework to derive the minimal distributions required to design a fair algorithm. Our proposed algorithm decentralizes the decision-making process and still achieves similar performance to the optimal algorithm that requires centralization and non-recoverable distributions.

Keywords:

Causality, algorithmic fairness, missing data, machine learning