FairPFN: Transformers Can do Counterfactual Fairness

Jake Robertson [1 2] Noah Hollmann [3] Noor Awad [1] Frank Hutter [1 4]

Abstract

Machine Learning systems are increasingly prevalent across healthcare, law enforcement, and finance but often operate on historical data, which may carry biases against certain demographic groups. Causal and counterfactual fairness provides an intuitive way to define fairness that closely aligns with legal standards. Despite its theoretical benefits, counterfactual fairness comes with several practical limitations, largely related to the reliance on domain knowledge and approximate causal discovery techniques in constructing a causal model. In this study, we take a fresh perspective on counterfactually fair prediction, building upon recent work in in-context-learning (ICL) and prior-fitted networks (PFNs) to learn a transformer called FairPFN. This model is pretrained using synthetic fairness data to eliminate the causal effects of protected attributes directly from observational data, removing the requirement of access to the correct causal model in practice. In our experiments, we thoroughly assess the effectiveness of FairPFN in eliminating the causal impact of protected attributes on a series of synthetic case studies and real-world datasets. Our findings pave the way for a new and promising research area: transformers for causal and counterfactual fairness.

1. Introduction

Algorithmic bias is one of the most pressing AI-related risks, arising when ML-assisted decisions produce discriminatory outcomes towards historically underprivileged demographic groups (Angwin et al., 2016). Despite the topic of fairness receiving significant attention in the ML community, various critics from outside the fairness community argue that statistical measures of fairness and current methods to optimize them are largely misguided in terms of their context-dependence and transferability to effective legislation. Recent work in causal fairness has proposed the popular notion of counterfactual fairness, which provides the intuition that outcomes are the same in the real world as in the counterfactual world where *protected attributes* - such as gender, ethnicity, or sexual orientation - take on a different value. According to a recent review contrasting observational and causal fairness metrics (Castelnovo et al., 2022), the non-identifiability of causal models from observational data (Peters et al., 2012) presents a significant challenge in applying causal fairness in practice, as causal mechanisms are often complex due to the intricate nature of bias in real-world datasets. If causal model assumptions are incorrect - for example, when a covariate is assumed not to be influenced by a protected attribute when in fact it is - proposing the wrong causal graph can provide a false sense of security and trust (Ma et al., 2023).

In this study, we introduce a novel approach to counterfactual fairness based on the recently proposed TabPFN. Our transformer-based approach coined FairPFN, is pre-trained on a synthetic benchmark of causally generated data and learns to identify and remove the causal effect of protected attributes. In our experimental results across a series of synthetic case-studies and real-world datasets, we demonstrate the effectiveness, flexibility, and extensibility of transformers for causal and counterfactual fairness.

2. Background

Algorithmic Fairness Algorithmic bias occurs when past discrimination against a demographic group such as ethnicity or sex is reflected in the training data of an ML algorithm. In such cases, ML algorithms are well known to reproduce and even amplify this bias in their predictions (Barocas et al., 2023). Fairness as a topic of research concerns the measurement of algorithmic bias and the development of principled methods that produce non-discriminatory predicted outcomes.

Causal Fairness Analysis Causal ML is a new and emerging research field that aims to represent data-generating processes and prediction problems in the language of causality, offering support for causal modeling, mediation analysis, and counterfactual explanations. The Causal Fairness Anal-

[1] University of Freiburg, Freiburg, Germany [2] Zuse School ELIZA, Darmstadt, Germany [3] Charité, Berlin, Germany [4] ELLIS Institute Tübingen, Tübingen, Germany. Correspondence to: Jake Robertson <robertsj@cs.uni-freiburg.de>.

Next Gen AI Safety at the International Conference on Machine Learning 2024, Vienna, Austria. Copyright 2024 by the author(s).