A Feature-Level Adjuster to Debias Proxy Features

Nirek Sharma*
Upstart, Inc.

2026-04-20

Abstract

In the context of machine learning fairness, proxy features are defined as features that encode membership in some protected demographic group. Typically, these features are simply removed from the model because using them would constitute disparate treatment and would be illegal under various regulations. Here, we introduce a general methodology for debiasing proxy features as an alternative to outright removal. We learn a transformation of the feature that reduces proxiness while preserving useful task-relevant information. We show that learning this per-observation adjustment via a trainable model is effective at practically reducing any arbitrary statistical definition of proxy risk while preserving predictive signal. We demonstrate this effectiveness on a wide range of synthetic and real-world datasets.

*nirek.sharma@upstart.com