Fairness Considerations for Conformal Classification

Arlan Abzhanov ZCAKABZ@UCL.AC.UK and **Brieuc Lehmann** B.LEHMANN@UCL.AC.UK
Department of Statistical Science, University College London, United Kingdom

Editor: Simone Vantini, Matteo Fontana, Aldo Solari, Henrik Boström and Lars Carlsson

Abstract

Algorithmic fairness refers to the elimination of systematic and quantifiable disparities in statistical models' outputs for protected groups, such as gender or ethnicity. In recent years, algorithmic fairness has grown to become a popular and widely studied method in the machine learning community. Despite these advancements, the intersection of fairness and conformal prediction remains an underexplored area in the literature, with only a handful of papers covering the subject. In this poster, we provide a synthesis of recent developments in conformal prediction and algorithmic fairness. In particular, we explore the intersection of these topics, investigating how biases can be identified, measured and excluded in the context of conformal classification methods. Building on past research on group-balanced conditional coverage (Vovk, 2012; Barber et al., 2021) and Mondrian conformal predictors (Vovk et al., 2003), we develop a novel joint group-class conditional coverage (JGCC) framework, a special type of Mondrian conformal predictor that aims to satisfy balanced coverage for all protected groups within all classification output classes. Hence, this method aims to achieve equal coverage conditional on both x (protected groups) and y (output classes) by training separate nonconformity score functions for each group *within* each class. We test this method on a clinical dataset, MIMIC-III, that has been shown to exhibit bias against certain demographic groups, predicting in-hospital mortality. We build several conformal models, finding that our JGCC framework with group clustering ensures the most equal conditional coverage and set size metrics for all protected groups. Notably, all other methods tested, including group-balanced conformal prediction fail to mitigate underlying biases in the dataset, such as low mortality coverage for Black patients and high set sizes for Asian patients, showcasing our method's effectiveness in ensuring fairness across various metrics for conformal classification.

Keywords: Conformal Prediction, Classification, Uncertainty Quantification, Algorithmic Fairness, Clinical Prediction

References

R. Foygel Barber, E. J. Candes, A. Ramdas, and R. J. Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2):455–482, 2021.

V. Vovk. Conditional validity of inductive conformal predictors. In *Asian conference on machine learning*, pages 475–490. PMLR, 2012.

V. Vovk, D. Lindsay, I. Nouretdinov, and A. Gammerman. Mondrian confidence machine. *Technical Report*, 2003.