Fair Wrapping for Black-box Predictions

Objectives

Exploit the properties (im)proper loss functions to develop a black-box classification post-processing method for algorithmic fairness.

We also want an approach with: ↪ flexibility to fairness criteria ↪ limited distortion of black-box ↪ algorithmic guarantees ↪ interpretability of correction

Black-box Post-Processing Setting

- $\mathcal{X}$ a domain of observations;
- $\mathcal{Y} \in \{-1, +1\}$ labels to classify;
- S sensitive modalities partitioning $\mathcal{X}$ such as age, race, *etc.*;
- η_u an accurate but unfair black-box classifier.

The goal of post-processing is to find a correction

$$\eta_u \mapsto \eta_f \tag{1}$$

which can be fitted to make η_f fairer than η_u.

Limitation on Distortion

Theorem 1. Suppose black-box η_u is bounded from extremes, s.t.

$$\mathrm{Im}(\eta_u) \subseteq \left[(1+\exp(\mathrm{B}))^{-1}, (1+\exp(-\mathrm{B}))^{-1}\right], \quad 0 < \mathrm{B} \leq 3 \quad \text{(Assum. 1)}$$

$$\text{and} \quad |\alpha(\boldsymbol{x}) - 1| \leq 1/B \quad \text{(a.s.)},$$

$$\text{Then for any } \mathrm{M}: \quad \mathrm{KL}(\eta_u, \eta_f; \mathrm{M}) \leq \frac{\pi^2}{6 \cdot (2 + \exp(B) + \exp(-B))}.$$

Twist-Properness to Corrections

Class probability estimation loss functions define what the "best responses" of η is via

$$t_\ell(\eta) = \arg \inf_{\hat{\eta} \in [0,1]} L(\hat{\eta}, \eta). \tag{2}$$

The family of *improper* α*-loss* provides an (twist-proper) implementation of (1) through t_ℓ (2):

$$\eta_f(\boldsymbol{x}) = \frac{\eta_u(\boldsymbol{x})^{\alpha(\boldsymbol{x})}}{\eta_u(\boldsymbol{x})^{\alpha(\boldsymbol{x})} + (1 - \eta_u(\boldsymbol{x}))^{\alpha(\boldsymbol{x})}}. \tag{3}$$

TopDown to Learn Alpha-Trees

With (3), our goal is to learn $\alpha(.)$ to improve fairness.

To learn $\alpha(.)$, we utilize a TopDown tree induction algorithm for general risk minimization:

$$\eta_f \overset{\text{TopDown}}{\longleftarrow} \min_{\alpha\text{-tree}} \mathbb{E}_{X \sim \mathrm{M}_t}\left[L(\eta_f(X), \eta_t(X))\right] \tag{4}$$

- $\alpha(.)$ is learnt as a *tree* (see Fig ↙);
- L is chosen to be the log-loss;
- M_t, η_t can be chosen for various fairness metrics.

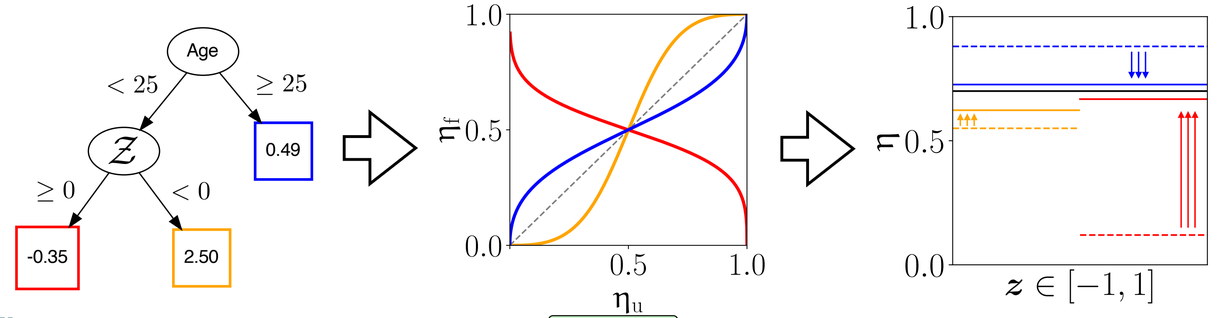

Figure: Toy Hiring Dataset: An example of an α-tree improving the fairness of a black-box (dotted colored) through a corrected posterior (solid colored).

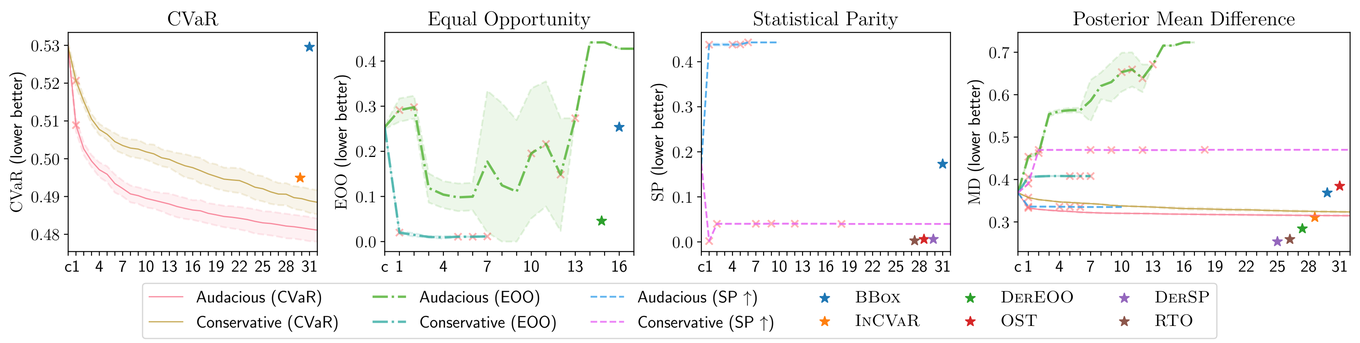

Figure: Improving fairness for a random forest black-box over a CA Income dataset for various criteria.

Algorithmic Convergence

Theorem 2. Suppose (Assum. 1) and a *Weak Learning Assumption* (with γ-witness) holds for all splits, then for any $\varepsilon > 0$, with TopDown we have $\mathbb{E}_{X \sim \mathrm{M}_t}\left[L(\eta_f(X), \eta_t(X))\right] \leq \varepsilon$ if

$$\text{(“\# Leaves in } \alpha\text{-tree”)} \quad |\Lambda(\alpha)| \geq (1/\varepsilon)^{c \log(1/\varepsilon)/\gamma^2}$$

- As this holds for any M_t, η_t, careful selection of targets allows for fairness guarantees for different criteria.

Alexander Soen[†], Ibrahim Alabdulmohsin[‡], Sanmi Koyejo[*‡], Yishay Mansour[‡◇], Nyalleng Moorosi[‡], Richard Nock[‡†], Ke Sun[†○], Lexing Xie[†]

Australian National University[†] Google Research[‡] Stanford University[*] Tel Aviv University[◇] CSIRO's Data61[○]