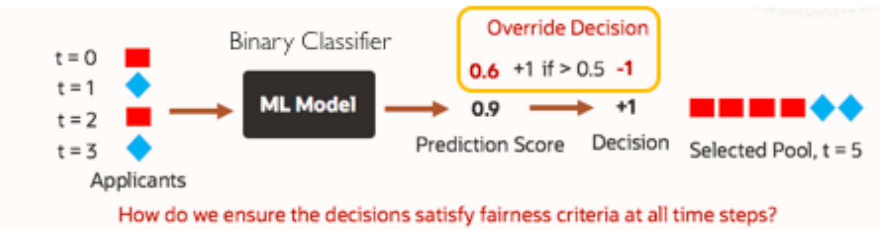
Fair Online Post-Processing For Black-Box ML Screening Systems

Oracle Labs

Swetasudha Panda, Ari Kobren, Jean-Baptiste Tristan and Michael Wick
Oracle Labs, Burlington, MA

Motivation

Addressing bias in decisions made by ML screening models (hiring/finance etc.).



Post-Processing Algorithms For ML Fairness

Learned classifier post-processed offline → *Derived* classifier is deployed. [Hardt et al. 2016]

Our experiments demonstrate that batch post-processing approaches are insufficient to mitigate fairness violations in the online setting.

Fair Online Post-Processing

a) Override classifier's decisions *at deployment time*; mitigate issues on the fly.
b) Sequential decision-making for continuous monitoring and audit.
c) Satisfy predefined fairness criteria at all time steps while maximizing long-term utility: *constrained optimization problem*.

Algorithmic Policies

Decide at each time step, whether to override classifier's decisions.

a) Deterministic greedy (**gbf**).
b) Randomized (**rpo, rpo-fl**).
c) Learned using imitation learning and learning to search (**il, l2s**).

Learning A Policy (l2s) With LOLS Variation [Chang et al. 2015]

Trained (offline) using a sequence of cost-sensitive examples.

a) **State** at t: statistics on data up to t, decisions up to t-1.
b) **Label**: max utility *roll-out* (accept/reject at t and *reference policy* afterwards).
c) **Weight**: difference between the two roll-out utilities.

Fairness Constraints

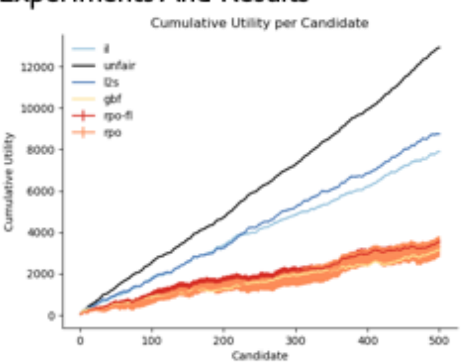
Can be general (predefined) group fairness constraints.
Demographic parity constraint in experiments.

Datasets

a) COMPAS
b) UCI Credit
c) UCI Income
d) Synthetic

Binary protected classes.

Experiments And Results



Cumulative utility (higher is better) for different policies (synthetic data, 500 time steps). *Unfair:* classifier without post-processing (max possible utility). **Learned policies (l2s, il) consistently outperform the rest in terms of both utility and fairness (across datasets).**

Analyzing l2s decisions: fairness audit score vs. classifier scores. Colors: two classes, circles: accept, 'x's: reject. **Learned policies trigger failsafe the least, operate further from the audit threshold and are able to learn soft thresholds for accepting per class (vs. gbf).**

Our work on generalization of online post-processing to ranking models [Gupta et al. WSDM 2021].