4-2022

The Impossible Theorem of Fairness

Man Nguyen '22

The Impossible Theorem of Fairness

Man Nguyen

Introduction

With the growth of machine learning, there has been an increase of machine biases that can cause wrongful discrimination. In the case of implementing "fairness," several conceptions of bias were created to target a fair system. However, statisticians have found that these conceptions contradict one another. Thus, we run into an impossible conundrum of fairness in machine learning. In cases that high risk, we want to investigate the best fairness measures if one is possible. Moreover, we would like to determine when these fairness measures fail or what conditions must be met for them to succeed.

Background Information

Impossible Theorem - states that no more than one of the three fairness metrics of demographic parity, predictive parity and equalized odds can hold at the same time for a well calibrated classifier and a sensitive attribute capable of introducing machine bias.

> **Theorem (Impossibility Result [26]).** Let h_1 and h_2 be classifiers for groups G_1 and G_2 with $\mu_1 \neq \mu_2$. h_1 and h_2 satisfy the Equalized Odds and calibration conditions if and only if h_1 and h_2 are perfect predictors.

Definitions:

Let $P \subset \mathbb{R}^k \times \{0, 1\}$ be the input space of a binary classification task. Assume there are two groups $G_1, G_2 \subset P$, which represent disjoint population subsets and that they have different base rates μ_1, or probabilities of belonging to the positive class:

$\mu_1 = P_{(x,y)\sim G1}[y = 1] \neq P_{(x,y)\sim G2}[y = 1] = \mu_2$. Let $h_1, h_2 : \mathbb{R}^k \to [0, 1]$ be binary classifiers, where h_1 classifies samples from G_1 and h_2 classifies samples from G_2.

Definition 1 (Kleinberg[1]). The generalized false-positive rate of classifier h_1 for group G_1 is $c_{fp}(h_1) = E_{(x,y)\sim Gt}[h_t(x) \mid y = 0]$. Similarly, the generalized false-negative rate of classifier h_t is $c_{fn}(h_1) = E_{(x,y)\sim Gt}[1 - h_t(x) \mid y = 0]$.

Definition 2 (Probabilistic Equalized Odds Kleinberg[2]). Classifiers h_1 and h_2 exhibit Equalized Odds for groups G_1 and G_2 if $c_{fp}(h_1) = c_{fp}(h_2)$ and $c_{fn}(h_1) = c_{fn}(h_2)$.

Definition 3 (Calibration Kleinberg[3]). A classifier h_t is perfectly calibrated if $\forall \rho \in [0, 1]$, $P_{(x,y)\sim Gt}[y = 1 \mid h_t(x) = \rho] = \rho$.

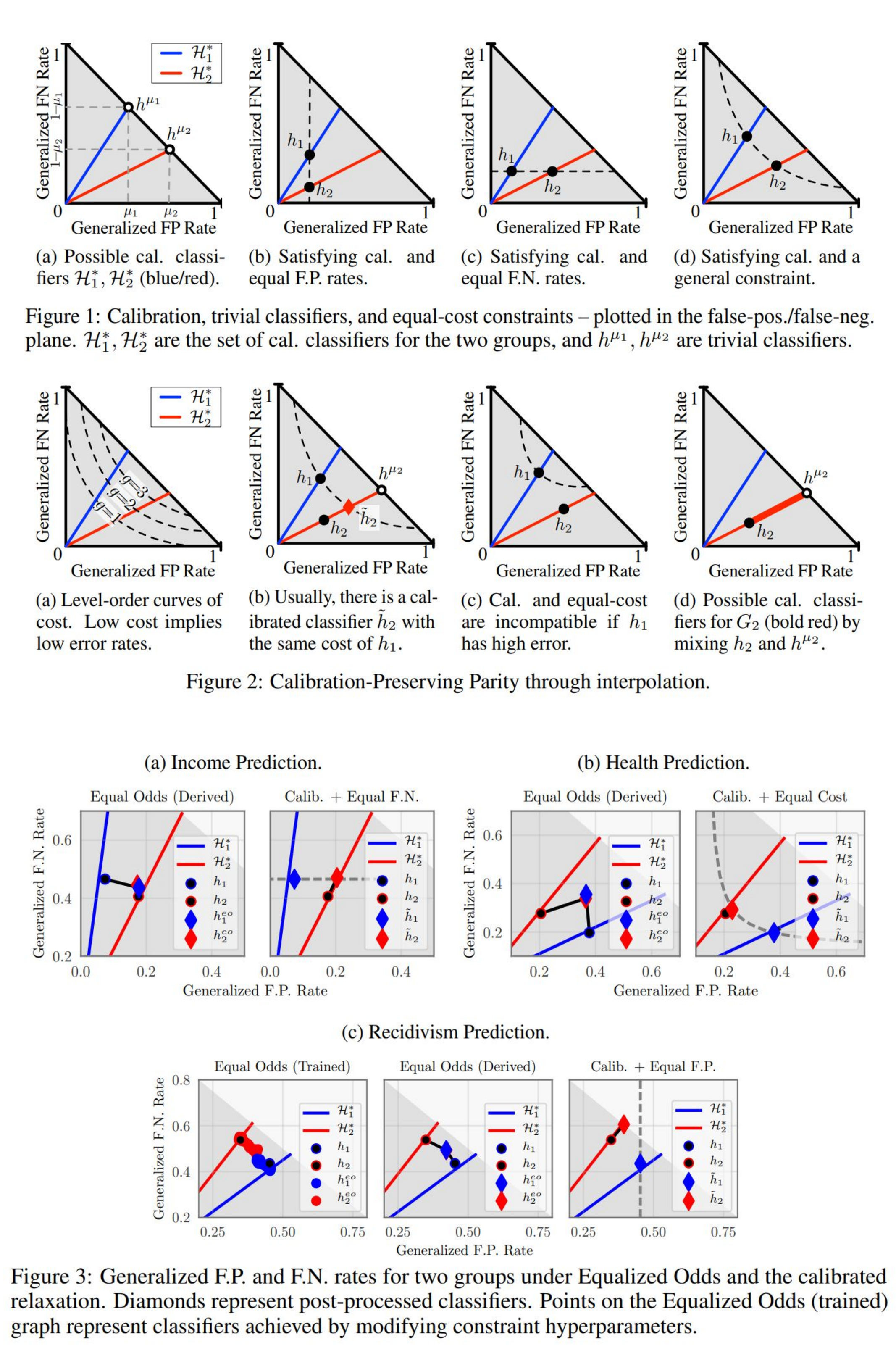

Figure I and Figure II. Calibration and error-rate constraints with simple geometric intuitions. *On Fairness and Calibration by Felix Wu, Jon Kleinberg, Kilian Q. Weinberger*

Criminal Recidivism Prediction

- Recidivism prediction instruments (RPI's) provide decision makers with an assessment of the likelihood that a criminal defendant will reoffend at a future point in time. Much of the controversy concerns potential discriminatory bias in the risk assessments that are produced. Many cases have reported false positives which ultimately hurts many innocent individuals.
- Calibration results in a tendency to disproportionately identify members of a certain population—that is, black people—as high risk.
  - COMPAS fails on both false positive and false negative error rate balance across the range of high-risk cutoffs.

In our model for criminal recidivism $(x, y) \sim P$ represents a person, with x representing the individual's history and y representing whether or not the person will commit another crime.

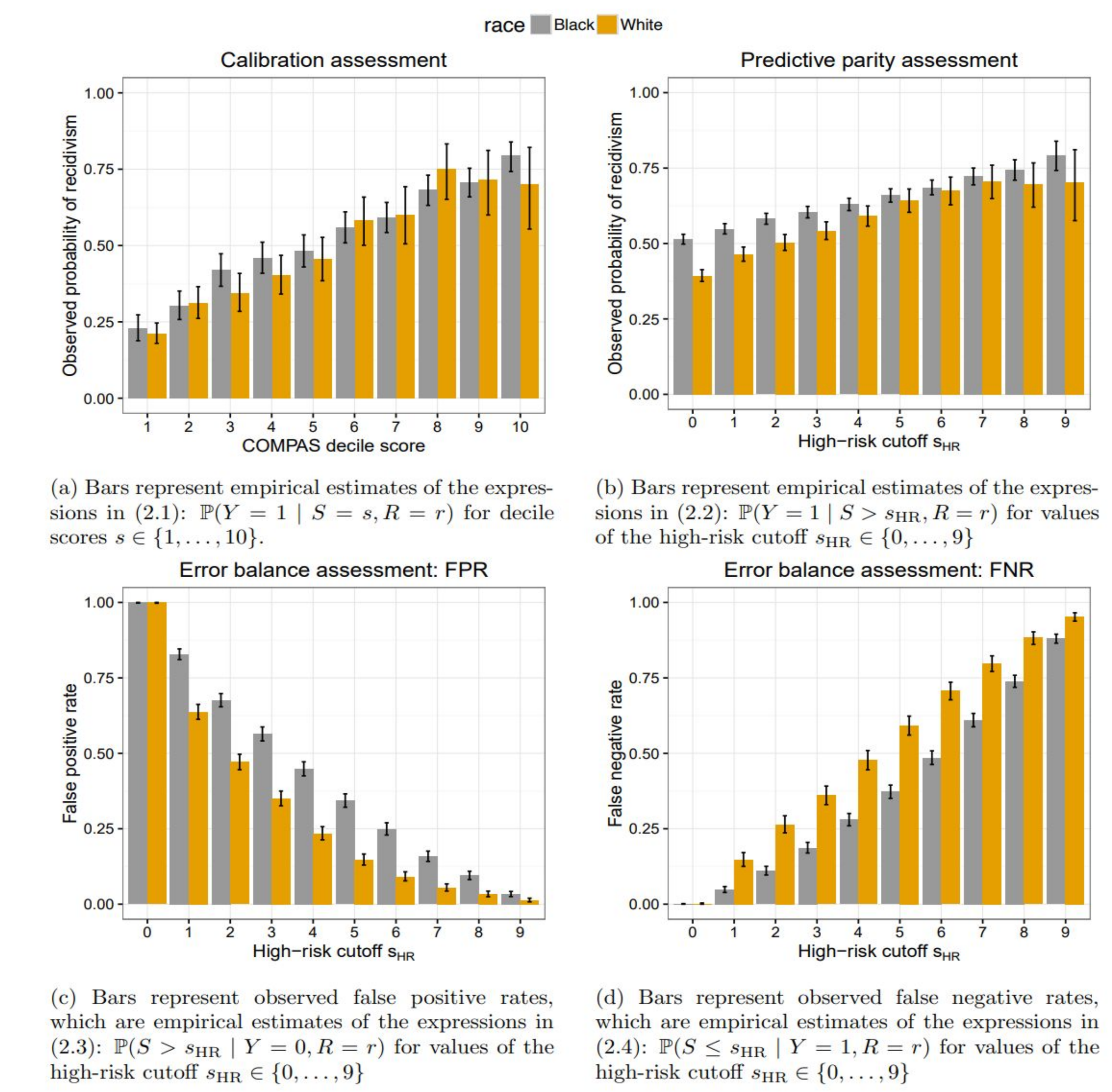

Figure III. Empirical assessment of the COMPAS RPI according to three of the fairness criteria presented in Section 2.1. Error bars represent 95% confidence intervals. These Figures confirm that COMPAS is (approximately) well-calibrated, satises predictive parity for high-risk cutoff values of 4 or higher, but fails to have error rate balance (*Fair prediction with disparate impact: A study of bias in recidivism prediction instruments by Alexandra Chouldechova*).

Concluding Thoughts

- The impossibility of fairness not only applies to recidivism but many other life altering prediction methods.
- Maintaining cost parity and calibration is desirable yet often difficult in practice because we need to find perfect classifiers.
- For recidivism prediction, calibration is completely incompatible with any error-rate constraints.
  - The most meaningful change in such a setting would be an improvement to the classifier for African Americans.
- The penalty of equalizing cost is amplified if the base rates between groups differ significantly.

Sources

[1] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments.
[2] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q. Weinberger. On fairness and calibration, 2017.
[3] Kailash Karthik Saravanakumar. The impossibility theorem of machine fairness – a causal perspective.
[4] Clinton Castro. What's wrong with machine bias. Ergo: An Open Access Journal of Philosophy, 6, 2019.