The Impossible Theorem of Fairness

Man Nguyen '22
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 [24]]. Let \( h_1 \) and \( h_2 \) be classifiers for groups \( G_1 \) and \( G_2 \), \( \mu \neq \mu_1, \mu_2 \). \( h_1, h_2 \) satisfy the Equalized Odds and calibration conditions if and only if \( h_1 \) and \( h_2 \) are perfect predictors.

Definitions:

Let \( P \subset \mathbb{R}^n \times \{0, 1\} \) 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 \( \mu_1, \mu_2 \) of belonging to the positive class.

\[ \mu_i = \frac{|P(h_i = 1)\}}{|P|} \neq \mu \neq \mu_1, \mu_2 \]. Let \( h_1, h_2 : \mathbb{R}^n \rightarrow \{0, 1\} \) 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 \( G_1 \), \( c_{FP}(h_1) = \frac{|P(h_1 = 1) \cap G_1|}{|P|} \) if \( \forall x \in G_1 \). Similarly, the generalized false-negative rate of classifier \( h_2 \) is \( c_{FP}(h_2) = \frac{|P(h_2 = 1) \cap G_1|}{|P|} \).

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 \( h_1 \) is perfectly calibrated if \( \forall y \in \{0, 1\}, P(h_1 = y) = \mu \).

Criminal Recidivism Prediction

- Recidivism prediction instruments (RPIs) 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.

  o 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.

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.

  o 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