Theoretical Guarantee of Efficiently Satisfying Subgroup Fairness in Multi-class Classification Settings

Fahim Tajwar (tajwar93@stanford.edu), Matthew King (mcking@stanford.edu)
Department of Computer Science, Department of Mathematics

Abstract
We build on recent theory of binary classification which treats arbitrary combinations of protected classes fairly and prove analogous results in the multi-class setting.

Introduction

- **Background**: Fairness in machine learning is increasingly topical as machine learning algorithms are leveraged to predict convict recidivism, future ability to pay loans, and many other predictions which correct or not have the ability to influence individuals’ lives for decades afterward.

- **What Fairness Means**: Fairness in machine learning generally deals with some predefined groups (based on age, gender etc.) in the dataset, and ensures no protected groups suffer drastic differences/discriminations in terms of some statistical measure (for example, the false positive rate).

- **Problem**: This often does not work as there can be combinations and intersections of protected groups (called subgroups) which can be discriminated against, still maintaining fairness in the protected group level.

Definitions

**Statistical Parity Subgroup Fairness**

Fix any classifier $D$, distribution $P$, group indicator $g$, threshold $y \in [0,1]$, 

$$a_{sp}(g, P) = \Pr_{x \sim P}(g(x) = 1)$$

$$\beta_{sp}(g, D, P) = |SP(D) - SP(D, g)|$$

where $SP(D) = \Pr_{x \sim P}(D(X) = 1)$ and $SP(D, g) = \Pr_{x \sim P}(D(X = 1 | g(x) = 1)$

We say $D$ satisfies $y$-statistical parity subgroup fairness with respect to $P$ and $g$ if we have,

$$a_{sp}(g, P)\beta_{sp}(g, D, P) \leq y$$

We say $D$ satisfies $y$-statistical parity subgroup fairness with respect to $P$ if it satisfies the above condition for $P$ and all $g \in G$.

**Previous Work**

Kearns et al., 2018 provides an algorithm that can efficiently maintain fairness for any number of combinatorial subgroups of the originally protected subgroups, without hurting the classification error too much. They prove that:

**Theorem**: Fix any $\nu, \delta \in [0,1]$. Then given an input of $n$ data points and accuracy parameters $\nu, \delta$, and access to Oracles $\text{CSC}(H)$ and $\text{CSC}(G)$, there exists an algorithm that runs in polynomial time, and with probability at least $1 - \delta$, outputs a randomized classifier $D'$ such that $err(D', P) < \text{OPTIMUM}(\nu) + \nu$, and for any group indicator $g$, the fairness constraint below is satisfied:

$$a_{sp}(g, P)\beta_{sp}(g, D', P) \leq y + 0(\nu)$$

where we use false positive rate as our measure of statistical parity.

Our Approach

We take the results from Kearns et al., and extend them from binary classification to multi-class classification settings, giving novel theoretical results.

- We define subgroup fairness in the multi-class classification settings.
- We prove similar bounds as the binary classification settings but for algorithms that satisfy this definition of multi-class fairness.
- We prove runtime bounds on our algorithm.

**Definition of Fairness in Multi-Class Settings**

Consider $k$-class classification. For any $j \in \{1, \ldots, k-1\}$ and any classifier $D$, distribution $P$, group indicator $g$, threshold $y \in [0,1]$, we define,

$$a_{\mu}(j, P) = \Pr_{x \sim P}(j(x) = 1)$$

$$\beta_{\mu}(j, D, P, g) = |\text{SP}(D(j)) - \text{SP}(D(j, g))|$$

where $\text{SP}(D(j)) = \Pr_{x \sim P}(D(j) = 1)$ and $\text{SP}(D(j, g)) = \Pr_{x \sim P}(D(j, g) = 1)$.

We say $D$ satisfies $y$-statistical parity subgroup fairness with respect to $P$ if we have,

$$a_{\mu}(j, P)\beta_{\mu}(j, D, P, g) \leq y$$

for all $j \in \{1, \ldots, k-1\}$ and all group indicators $g$.

**Theorem**: Given $n$ points to classify into $k$ classes, accuracy parameters $\nu, \delta$, and access to Oracles $\text{CSC}(H)$ and $\text{CSC}(G)$ for multi-class classification, there exists an algorithm that runs in time polynomial in $n$, $\delta$ and $k$, and with probability $1 - \delta$ outputs a randomized classifier $D'$ such that $err(D', P) < \text{OPTIMUM}(y) + \nu$, and is $y$ fair, according to the multi-class definition of fairness.

Figure: A toy example showing fairness gerrymandering. (Kearns et al., 2018)

Discussion: Supplementary results proven

- Uniform convergence: we prove that sufficiently large datasets yield empirical errors within any epsilon of the true error, and that for fixed $\epsilon$, this dataset grows only according to $k \log(k)$.

- Statistical parity uniform convergence: we prove that for each class $A$, $a_{\mu}(A, P)\beta_{\mu}(A, D, P, j)$ from the multi-class statistical parity definition can be taken arbitrarily close to true and empirical distributions, and the required dataset size grows analogously to the same uniform convergence difference.

- Worst-case intractability: it follows from earlier results in the field that no deterministic polynomial time algorithm can be guaranteed to find the optimal-accuracy classifier which is some $\gamma$-fair.

Future Work

- Empirical validation: We plan to apply our results to multi-class prediction on a fairness-sensitive dataset for which multi-class classification is more appropriate than binary, such as criminal risk-score generation.

- Extension to “fair” continuous output: Fairness concerns apply equally to continuous predictions, but is much harder to evaluate fairness. We propose the objective of minimizing a function of this form, for $n$ examples:

$$C(b) = \sum_{i=1}^{n} h_i^2(c_i)$$

where $h_i$ gives a predicted probability density of output $c_i$, and $c_i$ is integrable, mapping predictions to costs, and $C$ gives the cost of $h$, the set of $h_i$ which represents a single element of the hypothesis class.

Acknowledgements

We would like to thank CS 229 Course Staff for all the assistance provided throughout the quarter. We would like to thank Jingbo Yang specifically for mentoring us on this project.

Citations

1. Michael Kearns, Seth Neel, Aaron Roth and Zhiwei Wu, Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness