Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness

Michael Kearns, Seth Neel, Aaron Roth and Zhiwei Steven Wu

Presenter: Wenxin Zhang

April 6, 2023

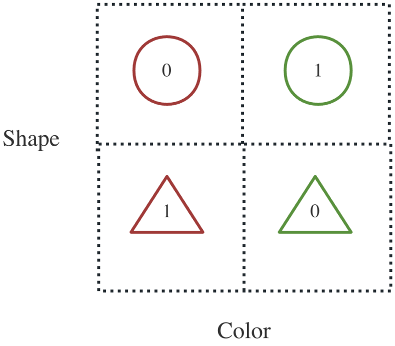
Fairness in Machine Learning

- Machine learning algorithms can amplify existing biases and unfairness in society
  - Example: COMPAS recidivism prediction algorithm, high false positive rate for Black defendants[1]
- Different approaches to fairness (e.g., group fairness, individual fairness, counterfactual fairness, etc.) [Friedler et al., 2019]
- Challenges in achieving fairness in machine learning
  - Trade-off between fairness and other objectives (e.g., accuracy, utility) [Kleinberg et al., 2016]
  - Lack of diversity in data and algorithms [Buolamwini and Gebru, 2018]
  - Need for transparency and accountability in algorithmic decision-making [Diakopoulos, 2018]

[1] `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`

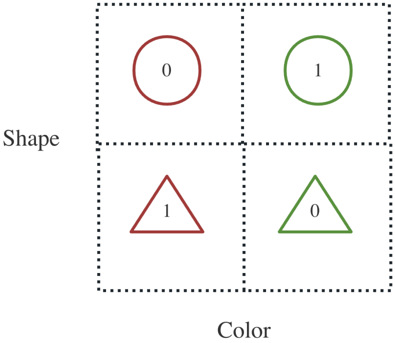
Fairness Gerrymandering

If we only look for unfairness over a small number of pre-defined groups:



Fairness Gerrymandering

If we only look for unfairness over a small number of pre-defined groups:



- Equitable with respect to single attributes
- Maximally violates statistical parity fairness for a red circle or green triangle

Roadmap

- To prevent Fairness Gerrymandering

Roadmap

- To prevent Fairness Gerrymandering $\rightarrow$
- Demand statistical notions of fairness across exponentially (or infinitely) many subgroups

Roadmap

- To prevent Fairness Gerrymandering $\rightarrow$
- Demand statistical notions of fairness across exponentially (or infinitely) many subgroups $\rightarrow$
- Computational Challenge

Roadmap

- To prevent Fairness Gerrymandering $\rightarrow$
- Demand statistical notions of fairness across exponentially (or infinitely) many subgroups $\rightarrow$
- Computational Challenge $\rightarrow$
- Show the equivalence between auditing subgroup fairness and weak agnostic learning

Roadmap

- To prevent Fairness Gerrymandering $\rightarrow$
- Demand statistical notions of fairness across exponentially (or infinitely) many subgroups $\rightarrow$
- Computational Challenge $\rightarrow$
- Show the equivalence between auditing subgroup fairness and weak agnostic learning $\rightarrow$
- Implications:
  - computationally hard in the worst case
  - common heuristics for learning can be applied successfully in practice

Roadmap

- To prevent Fairness Gerrymandering $\rightarrow$
- Demand statistical notions of fairness across exponentially (or infinitely) many subgroups $\rightarrow$
- Computational Challenge $\rightarrow$
- Show the equivalence between auditing subgroup fairness and weak agnostic learning $\rightarrow$
- Implications:
  - computationally hard in the worst case
  - common heuristics for learning can be applied successfully in practice
- Fictitious play in a two-player zero-sum game between a Learner and an Auditor

Model

- Individual: $(X, y) = ((x, x'), y)$, $x \in \mathcal{X}$: protected attributes; $x' \in \mathcal{X}'$: unprotected attributes; $y \in \{0, 1\}$: label
- (X, y): i.i.d. drawn from an unknown distribution $\mathcal{P}$
- D: decision making algorithm, $D(X) \in \{0, 1\}$
- $\mathcal{G}$: family of indicator functions, $g : \mathcal{X} \mapsto \{0, 1\}$, $g(x) = 1$ indicates than individual with x is in group g

Definitions of Fairness

Definition (Statistical Parity (SP) Subgroup Fairness)

Fix any classifier D, distribution $\mathcal{P}$, collection of group indicators $\mathcal{G}$, and parameters $\alpha, \beta \in [0, 1]$. We say that D satisfies (α, β)-statistical parity (SP) Fairness with respect to $\mathcal{P}$ and $\mathcal{G}$ if for every $g \in \mathcal{G}$ such that $\min(\Pr[g(x) = 1], \Pr[g(x) = 0]) \geq \alpha$ we have:

$$|\Pr[D(X) = 1 \mid g(x) = 1] - \Pr[D(X) = 1]| \leq \beta$$

α: how small a fraction of population we are allowed to ignore
β: how much deviations positive classifications are allowed

Comparison of Concepts

Definition (Calibration [Hébert-Johnson et al., 2018])

For all but an α-fraction of a set S, the average of the true probabilities of the individuals receiving prediction v is α-close to v. Multicalibration requires $\alpha-$calibrated on all subsets of $\mathcal{C}$.

Comparison of Concepts

Definition (Calibration [Hébert-Johnson et al., 2018])

For all but an α-fraction of a set S, the average of the true probabilities of the individuals receiving prediction v is α-close to v. Multicalibration requires $\alpha-$calibrated on all subsets of $\mathcal{C}$.

- SP-fairness cares about the difference between the average prediction of groups.
- Calibration cares about the difference between the prediction accuracy within groups of same prediction.

Comparison of Concepts

Definition (Calibration [Hébert-Johnson et al., 2018])

For all but an α-fraction of a set S, the average of the true probabilities of the individuals receiving prediction v is α-close to v. Multicalibration requires $\alpha-$calibrated on all subsets of $\mathcal{C}$.

- SP-fairness cares about the difference between the average prediction of groups.
- Calibration cares about the difference between the prediction accuracy within groups of same prediction.
- SP-fairness can be seen as constraints on learning a good predictor
- Calibration aligns with learning a good predictor

Auditing

Theorem (Informal)

Auditing for an arbitrary D w.r.t. $\mathcal{G}$ is computationally equivalent to weak agnostic learning of $\mathcal{G}$ under the marginal distribution on $(x, D(X))$.

Auditing

Theorem (Informal)

Auditing for an arbitrary D w.r.t. $\mathcal{G}$ is computationally equivalent to weak agnostic learning of $\mathcal{G}$ under the marginal distribution on $(x, D(X))$.

Definition (Auditing (in English))

Given access to samples $(x, x', y, D(X))$, can we decide if D is SP fair, or output a violated g?

Definition (Weak Agnostic Learning (in English))

Learn patterns purely from the training data with no assumptions about the underlying data distribution of the data; 'Weak' in the sense that model can make errors in its predictions, but still needs to perform better than random guessing.

Auditing

Intuition: For $\mathbb{P}(D(X) = g(x))$ to be better than random guess, the group should be imbalanced.

$$\mathbb{P}(D(X) = 1|g(x) = 1)\mathbb{P}(g(x) = 1) + \mathbb{P}(D(X) = 1|g(x) = 0)\mathbb{P}(g(x) = 0)$$

- If g is violated, then g or $\neg g$ predict the decisions made by algorithm D better than random guess
- If g predicts the decisions made by the algorithm D better then random guess, then g or $\neg g$ is violated

Learning

Theorem (Worst-case intractability of auditing (informal))

Even for $\mathcal{G}$ with simple structure such as conjunctions of Boolean attributes, there exist distributions $\mathcal{P}$ such that the auditing problem cannot be solved in polynomial time.

Learning

Effective heuristics on specific (non-worst case) distributions:

- Formulate as a two-player repeated zero-sum game
- Given oracles to solve agnostic learning problem and auditing problem
- Learner objective: minimize error subject to fairness w.r.t. $\mathcal{G}$
- Learner: propose a classifier $h \in \mathcal{H}$
- Auditor: find a group that is being discriminated against most
- Provably convergent learning algorithm: theoretical convergence rate quite slow, but in practice converges quickly

Summary

- Statistical notions of fairness across exponentially (or infinitely) many subgroups
- Computational problem of auditing subgroup fairness is equivalent to the problem of weak agnostic learning
- Formulation of subgroup fairness as fictitious play in a two-player zero-sum game between a Learner and an Auditor

Auditing

Definition

Fix a notion of fairness (either statistical parity or false-positive fairness), a collection of group indicators $\mathcal{G}$ over the protected features, and any $\alpha, \beta, \alpha', \beta' \in (0, 1]$ such that $\alpha' \leq \alpha$ and $\beta' \leq \beta$. A collection of classifiers $\mathcal{H}$ is $(\alpha, \beta, \alpha', \beta')$-(efficiently) auditable under distribution $\mathcal{P}$ for groups $\mathcal{G}$ if there exists an auditing algorithm A such that for every classifier $D \in \mathcal{H}$, when given access the distribution $P_{\text{audit}}(D)$, A runs in time $\text{poly}\,(1/\alpha, 1/\alpha', 1/\beta, 1/\beta', 1/\delta)$, and with probability $(1 - \delta)$, outputs an (α', β')-unfair certificate for D whenever D is (α, β)-unfair with respect to $\mathcal{P}$ and $\mathcal{G}$.

Weak Agnostic Learning

Definition ([Kalai et al., 2008])

Let Q be a distribution over $\mathcal{X} \times \{0,1\}$ and let $\varepsilon, \gamma \in (0, 1/2)$ such that $\varepsilon \geq \gamma$. We say that the function class $\mathcal{G}$ is (ε, γ)-weakly agnostically learnable under distribution Q if there exists an algorithm L such that when given sample access to Q, L runs in time poly $(1/\gamma, 1/\delta)$, and with probability $1 - \delta$, outputs a hypothesis $h \in \mathcal{G}$ such that

$$\min_{f \in \mathcal{G}} \mathrm{err}(f, Q) \leq 1/2 - \varepsilon \implies \mathrm{err}(h, Q) \leq 1/2 - \gamma.$$

where $\mathrm{err}(h, Q) = \Pr_{(x,y)\sim Q}[h(x) \neq y]$.

Two Fairness Notions

Statistical

- group-level outcomes: the outcomes for different groups are not too different.
- e.g. equal false positive or negative rates across groups (equal opportunity); equality of classification rates (statistical parity)
- can be obtained and checked without making any assumptions about the underlying population

Individual

- individual-level outcomes: treating similar individuals similarly, regardless of group membership
- more difficult to achieve: require more assumptions on the setting
- similarity measures between individuals include k-nearest neighbors or kernel density estimation

References I

Buolamwini, J. and Gebru, T. (2018).
Gender shades: Intersectional accuracy disparities in commercial gender classification.
Conference on Fairness, Accountability and Transparency.

Diakopoulos, N. (2018).
Algorithmic accountability reporting: On the investigation of black boxes.
Journalism, 19(1):1–17.

Friedler, S. A., Scheidegger, C. E., and Venkatasubramanian, S. (2019).
A comparative study of fairness-enhancing interventions in machine learning.
In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 329–338.

Hébert-Johnson, U., Kim, M., Reingold, O., and Rothblum, G. (2018).
Multicalibration: Calibration for the (computationally-identifiable) masses.
In *International Conference on Machine Learning*, pages 1939–1948. PMLR.

Kalai, A. T., Mansour, Y., and Verbin, E. (2008).
On agnostic boosting and parity learning.
In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 629–638.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. (2016).
Inherent trade-offs in the fair determination of risk scores.
In *Conference on Innovations in Theoretical Computer Science*, pages 43–52. ACM.