Stable and Fair Classification

Nisheeth Vishnoi

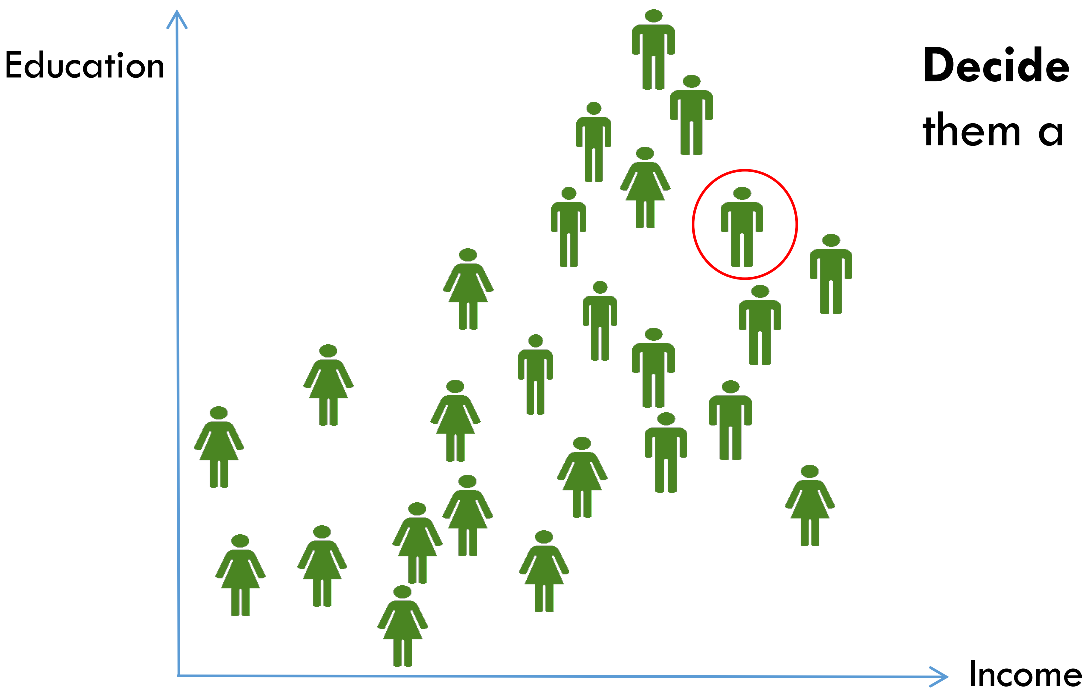
Yale

@NisheethVishnoi

Joint work with Lingxiao Huang (EPFL)

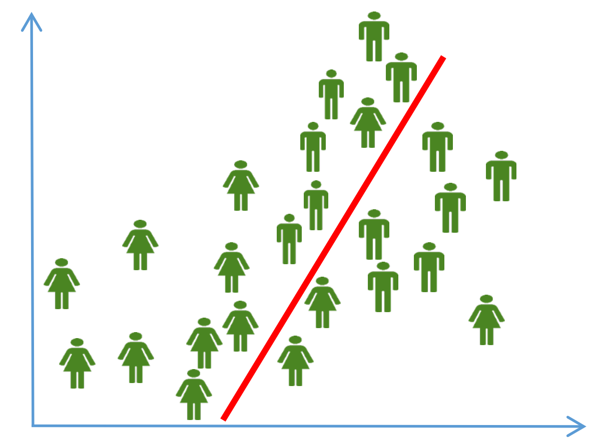
Poster 130, Pacific Ballroom, 6:30-9:00 pm, Thursday, June 13

Classification

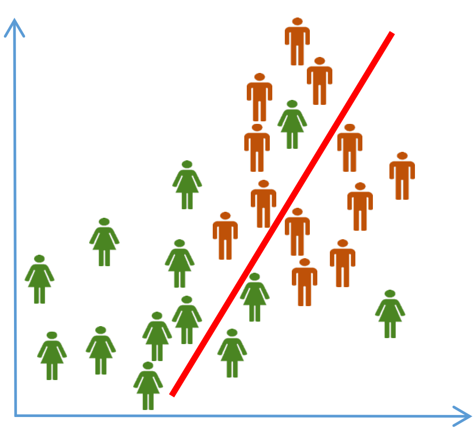


Given the data of an individual

Decide whether to recommend them a high-salary job or not?



Data has sensitive types



Classifier's performance/accuracy can vary with the sensitive type

Group Fairness/Statistical parity
False positive
Calibration
Negative predictive parity
False discovery
False omission …

Optimization for Fair Classification

Fair Classification

[Kamishima et al. '12]
[Hardt et al. '16]
[Zafar et al. '17b]
[Krishna Menon et al. '18]
[Woodworth et al. '17]
[Goel et al. '18]
[Krasanakis et al. '18]
[Celis et al. '19] …

→ **Constrained Optimization**

- $\mathcal{F}$: reproducing kernel Hilbert space, e.g., $\{\langle \beta, \cdot \rangle\}$
- $L(\cdot,\cdot)$: loss function
- N Samples: $s_i = (x_i, z_i, y_i) \in X\ (fea) \times [p]\ (type) \times \{0,1\}$ (label)
- $\Omega(f)$: fairness constraints

$$min_{f\in\mathcal{F}} \frac{1}{N} \sum_{i\in[N]} L(f, s_i)$$
$$\text{s.t.} \quad \Omega(f) \le 0$$

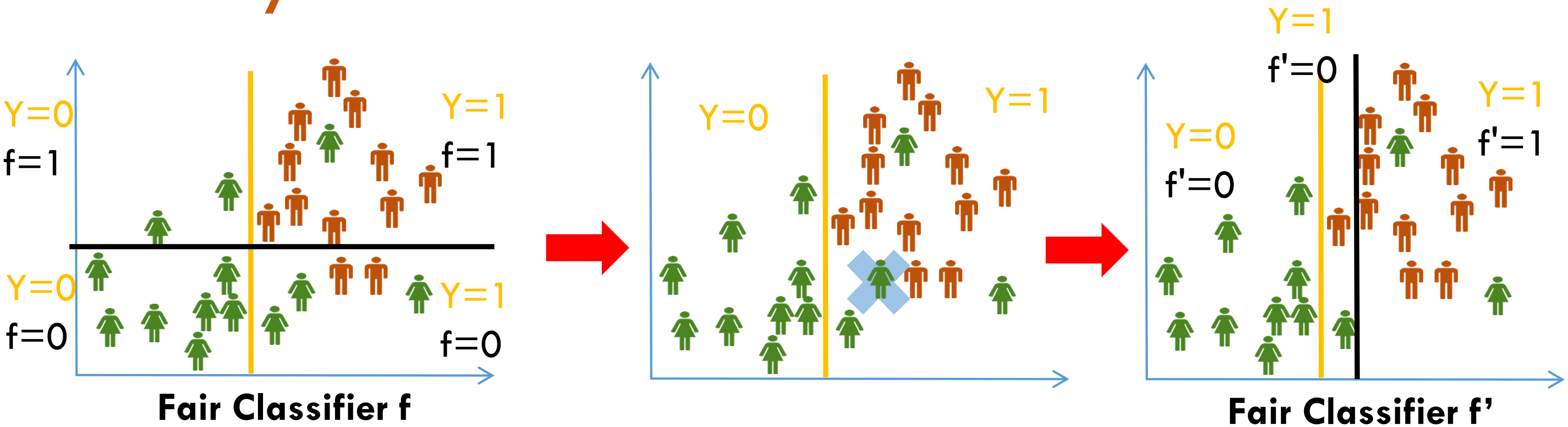
Example (logistic regression loss function + statistical parity/80%-rule)

$$min_{\alpha\in\mathbb{R}^N} \frac{1}{N} \sum_{i\in[N]} ln\left(1 + y_i \cdot e^{-\sum_{j\in[N]} \alpha_j k(x_i, x_j)}\right), s.t.,$$
$$0.8 - \frac{min_{i\in[p]} Pr_D[f = 1 | Z = i]}{max_{i\in[p]} Pr_D[f = 1 | Z = i]} \le 0$$

← statistical rate

- $k(\cdot,\cdot)$: kernel function
- D: empirical distribution over the training dataset S

Stability Problem in Fair Classification



Definition (β-Uniform stability [Bousquet & Elisseeff '02])

The maximum l_∞-distance between the risks of two classifiers learned from two training sets that differ in a single sample is upper bounded by β, i.e., $||L(f,\cdot) - L(f',\cdot)||_\infty \le \beta$

- ***Existing fair classification algorithms may not be stable*** [Friedler et al. '19]
  - Study the standard deviation of the statistical rate γ over ten random training-testing splits with respect to race/sex attribute over the **Adult** dataset
  - The standard deviation of γ is 2.4% for [Kamishima et al. '12] with respect to the race attribute, and is 4.1% for [Zafar et al. '17b] with respect to the sex attribute

Question: *Can we design stable and fair classification algorithms?*

Our results

We provide an extended algorithmic framework to constrained-optimization based fair classification algorithms that ensures both stability and fairness

1. Provable guarantees: our framework provides a uniform stability guarantee $\left(\frac{\sigma^2\kappa^2}{\lambda N}\right)$ and an empirical risk guarantee $\left(\frac{\sigma^2\kappa^2}{\lambda N} + \lambda B^2\right)$
2. Empirical risk guarantee can be used to inform the selection of the regularization parameter $\lambda \left(\frac{\sigma\kappa}{B\sqrt{N}}\right)$
3. The resulting optimization problem is polynomial time solvable

We introduce a stability-focused regularization term $\lambda\|f\|_k^2$ where $k(\cdot,\cdot)$ is the kernel function

$$min_{f\in\mathcal{F}} \frac{1}{N}\sum_{i\in[N]} L(f, s_i) + \lambda\|f\|_k^2$$
$$s.t.\ \Omega(f) \le 0$$

Assumptions:

- For all x, $k(x,x) \le \kappa^2$;
- $L(f,s)$ is σ-Lipschitz w.r.t f;
- $\Omega(f)$ is convex;
- For all $f \in \mathcal{F}$, $\|f\|_k \le B$

			λ					
			0	0.01	0.02	0.03	0.04	0.05
ZVRG-St	Race	Acc.	0.844(0.001)	0.842(0.001)	0.841(0.001)	0.840(0.001)	0.838(0.001)	0.838(0.001)
		γ	0.577(0.031)	0.667(0.020)	0.686(0.015)	0.711(0.016)	0.743(0.013)	0.761(0.012)
	Sex	Acc.	0.844(0.001)	0.840(0.001)	0.838(0.001)	0.838(0.001)	0.837(0.001)	0.836(0.001)
		γ	0.331(0.041)	0.501(0.011)	0.495(0.009)	0.478(0.009)	0.463(0.009)	0.469(0.009)
KAAS-St	Race	Acc.	0.850(0.001)	0.844(0.001)	0.843(0.001)	0.839(0.001)	0.837(0.001)	0.835(0.001)
		γ	0.571(0.019)	0.359(0.024)	0.302(0.011)	0.301(0.011)	0.300(0.015)	0.298(0.015)
	Sex	Acc.	0.850(0.002)	0.848(0.001)	0.844(0.001)	0.839(0.001)	0.837(0.001)	0.835(0.001)
		γ	0.266(0.011)	0.226(0.011)	0.165(0.008)	0.136(0.007)	0.128(0.006)	0.128(0.005)
GYF-St	Race	Acc.	0.849(0.001)	0.845(0.001)	0.844(0.001)	0.842(0.001)	0.840(0.001)	0.835(0.001)
		γ	0.558(0.020)	0.679(0.013)	0.690(0.017)	0.710(0.018)	0.740(0.014)	0.753(0.013)
	Sex	Acc.	0.850(0.002)	0.845(0.001)	0.844(0.001)	0.842(0.001)	0.840(0.001)	0.839(0.001)
		γ	0.275(0.010)	0.245(0.004)	0.242(0.004)	0.241(0.005)	0.245(0.005)	0.234(0.008)

Adult dataset. γ: statistical rate; "-St": our extended framework on fair classification algorithms; ZVRG [Zafar et al. '17b], KAAS [Kamishima et al. '12], GYF [Goel et al. '18]

More stable with slight loss in accuracy

- As λ increases, the average accuracy slightly decreases, by at most 1.5%
- The standard deviation of γ decreases from 4.1% to around 1% as λ increases → more stable

Conclusion and Future Work

- We propose an extended framework that for the first time combines stability and fairness in classification
- Our framework comes with a stability guarantee and we also provide an analysis of the resulting accuracy
- There exist other fair classification algorithms that are not formulated as optimization problems. Can we investigate and improve the stability guarantee of those algorithms?
- Combine stability and fairness for other automated decision-making tasks?

Thank you!

Poster 130, Pacific Ballroom, 6:30-9:00 pm, Thursday, June 13