Abolfazl Asudeh
Fall 2020
10/06/2020

CS 594: In-process Interventions to Achieve Fairness

Classification

Reminder

- Finds the parameter θ that minimizes the loss function $L(f)$

$$\min_{\theta} L(f_\theta)$$

- For efficient learning, the loss function is designed to be convex
- Optimizing the loss function, without considering demographic groups may result in “unfair” models
- Changing the problem formulation to account for fairness

$$\min_{\theta} \quad L(f_\theta)$$
$$s.t. fairness$$

- Challenge: This is (often) not convex

Adding fairness makes the optimization non-convex

- e.g.:
  - $\min L(\theta)$
    - s.t. $P(f_\theta(X) = 1|S = 0) = P(f_\theta(X) = 1|S = 1)$ Demographic Parity
  - $\min L(\theta)$
    - s.t. $P(f_\theta(X) \neq y|S = 0) = P(f_\theta(X) \neq y|S = 1)$ Misclassification Partiy

Fairness constraints: Mechanisms for fair classification

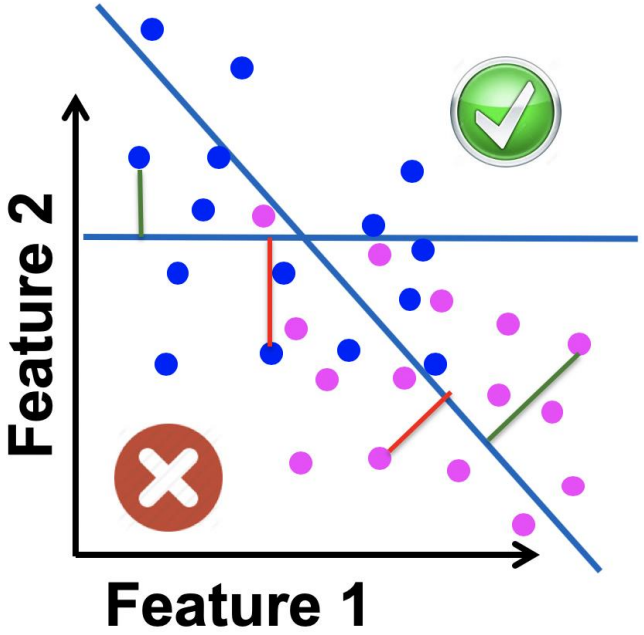
Muhammad Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P. Gummadi

Artificial Intelligence and Statistics, pp. 962-970. 2017.

- To resolve the non-convex optimization issue:
  - Proposes the (alternative) measure of “decision boundary (un)fairness” for convex margin-based classifiers such as SVM.

An alternative for disparate impact

- The difference between the strength of acceptance and rejection across different demographic groups.
- The covariance between demographic groups and their signed distance from classifier's decision boundary as the fairness measure



Decision-boundry fairness

$$cov\big(S, d_\theta(X)\big) = E[(S - \bar{S})d_\theta(X)] - E[(S - \bar{S})]E[d_\theta(X)]$$
$$\approx \frac{1}{n}\sum(S - \bar{S})d_\theta(X)$$

Considering the decision boundary at score zero: $\theta^\top X = 0$:

$$cov\big(S, d_\theta(X)\big) = \frac{1}{n}\sum_{i=1}^{n}(S_i - \bar{S})\theta^T X$$

Decision-boundry fairness:

$$\left|\frac{1}{n}\sum_{i=1}^{n}(S_i - \bar{S})\theta^T X\right| \leq \tau$$

Convex Optimization

- $\min L(\theta)$
- **s.t.**
  - $\frac{1}{n}\sum_{i=1}^{n}(S_i - \bar{S})\theta^T X \leq \tau$
  - $\frac{1}{n}\sum_{i=1}^{n}(\bar{S} - S_i)\theta^T X \geq -\tau$

Similar constraints can be applied for misclassification parity, false negative rate, and false positive rate parity

A reductions approach to fair classification

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach

ICML 2018

* This paper can handle multiple sensitive attributes and multiple fairness measure

1- How to handle different notions of fairness?

There is a cost associated with re-engineering the ML systems to satisfy fairness

→ This may be too much for many stakeholders

2- How to adopt the existing ML system?

1- multiple fairness measures

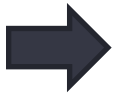
- Define generic fairness constraints
- Each fairness constrains is defined as
- $\mu_j(\theta) = E\left[g_j\left(X, S, Y, f_\theta(X)\right) \middle| \varepsilon_j\right], \forall j \in$ demographic groups
  - ε_j does not depend on h → does not support measures based on sufficiency

- Example:
  - DP
  - EO

1- multiple fairness measures

- Define generic fairness constraints
- Each fairness constrains is defined as
- $\mu_j(\theta) = E\left[g_j\left(X, S, Y, f_\theta(X)\right) \middle| \varepsilon_j\right], \forall j \in$ demographic groups
  - ε_j does not depend on h → does not support measures based on sufficiency

- Example:
  - DP: $g_j\left(X, S, Y, f_\theta(X)\right) = f_\theta(x)$ and $\varepsilon_j = \{S = S_j\}$, $\varepsilon_* = true$
  - EO: $g_j\left(X, S, Y, f_\theta(X)\right) = f_\theta(x)$ and $\varepsilon_j = \{S = S_j, Y = y\}$, $\varepsilon_* = \{Y = y\}$
- Constraints:
  - $\mu_j(\theta) - \mu_*(\theta) \leq \tau$
  - $-\mu_j(\theta) + \mu_*(\theta) \leq \tau$



$$M\mu(\theta) \leq \tau$$

2- adopt the existing ML system

- Solution: build a wrapper around the existing learning system that ensures fairness
  - Key idea: reduce fair classification to a sequence of cost-sensitive classification problems, whose solutions yield a (randomized) classifier with the lowest (empirical) error subject to the desired constraints

→ The fairness component can seamlessly integrate to the system

- Find the classifier *f* that
  1. Minimizes the loss (classification error)
  2. Satisfies fairness constraints

- Iteratively call the black-box learner and apply reweighting and (possibly) relabeling the data

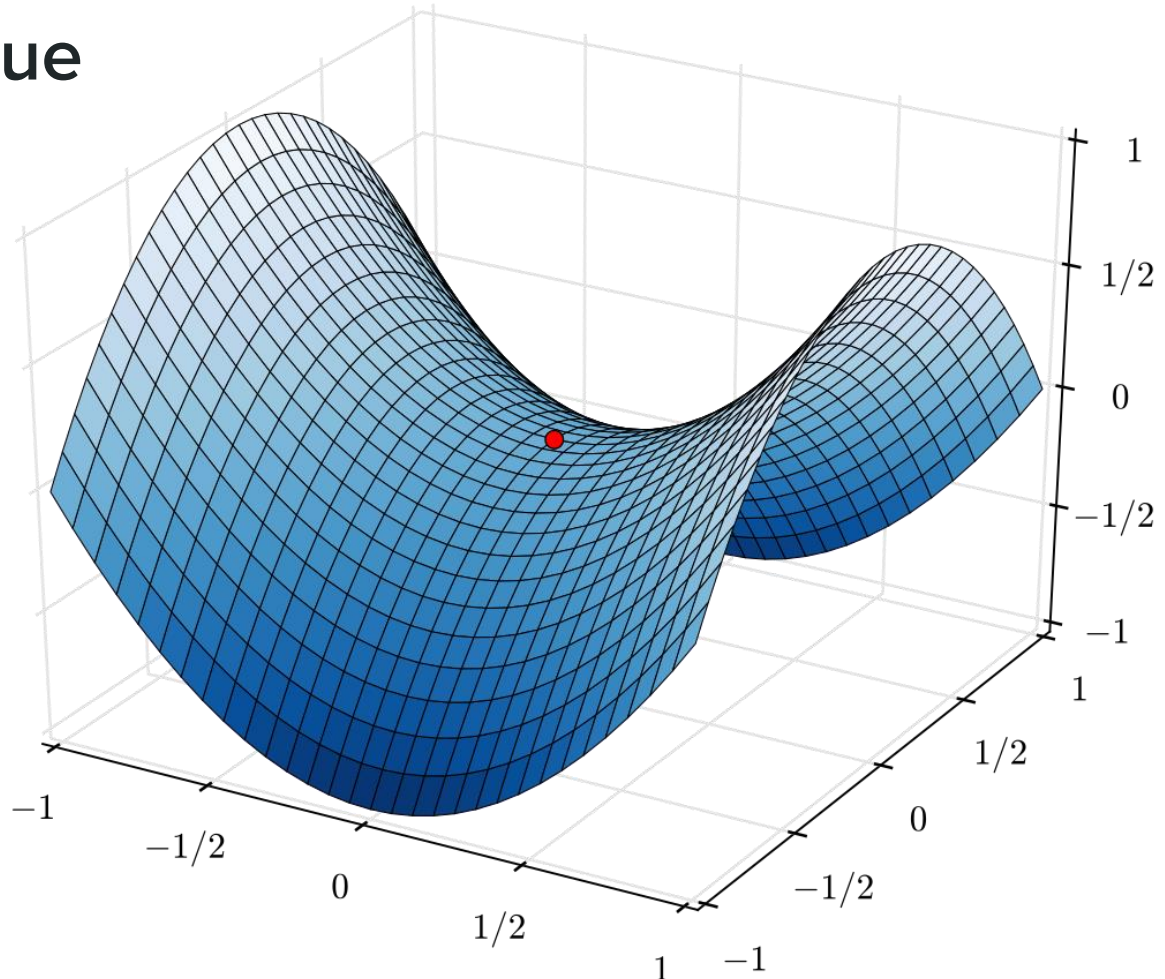
- It guarantees to find the most accurate fair classifier in not too many iterations (~5 in experiments)

- $\min_{\forall\theta} \; L(\theta) \quad \text{s.t.} \quad M\mu(\theta) \leq \tau$

- Lagrangian dual form:

$$L(\theta, \lambda) = L(\theta) + \lambda(M\mu(\theta) - \tau)$$

Saddle point issue



Theorem: After $O(n^2 \log \#constraints)$ iterations, finds the classifier with probability $(1-\delta)$

- $\min_{\forall\theta} \; L(\theta) \quad \text{s.t.} \quad M\mu(\theta) \leq \tau$

- **Lagrangian dual form:** $L(\theta, \lambda) = L(\theta) + \lambda(M\mu(\theta) - \tau)$
- **Solve for Saddle point:**

$$\max_{\lambda} \min_{\theta} L(\theta, \lambda)$$

Existing ML system

Iterate while reweighting examples

Classification with fairness constraints: A meta-algorithm with provable guarantees

Elisa Celis, Lingxiao Huang, Vijay Keswani,
and Nisheeth K. Vishnoi

FAT* 2019

- Proposes a meta-algorithm for a general class of fairness constraints with respect to multiple non-disjoint and multi-valued sensitive attributes

- Can handle non-convex linear fractional constraints, including predictive parity

Generalization of fairness functions: group performance function

- At a high-level, fairness requires equal “performance” of a classifier f for different demographic groups.

- For a classifier f, the group performance of group S_i is defined as

$$q_i(f) = P[\varepsilon | S_i, \varepsilon']$$

- Example:
  - Accuracy rate: $\varepsilon := (f = y), \varepsilon' := \emptyset$
  - False negative rate: $\varepsilon := (f = 0), \varepsilon' := (y = 1)$

The family of classifications with linear constraints

		$q_i(f)$		$Q_{\text{lin}}/Q_{\text{linf}}$
		$\mathcal{E}$	$\mathcal{E}'$	
fairness metrics	statistical	$f = 1$	$\emptyset$	Q_{lin}
	conditional statistical	$f = 1$	$X \in S$	Q_{lin}
	false positive	$f = 1$	$Y = 0$	Q_{lin}
	false negative	$f = 0$	$Y = 1$	Q_{lin}
	true positive	$f = 1$	$Y = 1$	Q_{lin}
	true negative	$f = 0$	$Y = 0$	Q_{lin}
	accuracy	$f = Y$	$\emptyset$	Q_{lin}
	false discovery	$Y = 0$	$f = 1$	Q_{linf}
	false omission	$Y = 1$	$f = 0$	Q_{linf}
	positive predictive	$Y = 1$	$f = 1$	Q_{linf}
	negative predictive	$Y = 0$	$f = 0$	Q_{linf}

ρ-Fair formulation

Nonconvex

- $\min_{\forall\theta} \; L(\theta)$ Loss term
- s.t.
  - $\rho_{q^{(i)}}(f_\theta) = \frac{\min q_j^{(i)}}{\max q_j^{(i)}} \geq \tau$ Fairness Constraint

*: $q^{(1)}$ … $q^{(m)}$ are the performance functions

Group-fair formulation

$$\min_{\forall\theta} \; L(\theta)$$

Loss term

s.t.

$$\ell_j^i \le q_j^{(i)}(f_\theta) \le u_j^i, \forall\, i \in [m], j \in [p]$$

Fairness Constraint

- Fairness constraints are linear → Convex

For any feasible classifier f of Group-Fair and any $i \in [m]$, f satisfies ρ-fair rule for:

$$\rho = \frac{\min \ell_j^{(i)}}{\max u_j^{(i)}}$$