Demographic parity constraint for algorithmic fairness

a statistical perspective

Statlearn'23
Montpellier

Evgenii Chzhen
CNRS, Univ. Paris-Saclay

Today's plan

1. A biased intro to fairness and fairness zoology
2. Demographic Parity constraint and analogies
3. Regression with demographic parity constraint
4. Building estimators

Fairness in ML: a major societal concern

Mercer | mettl

PRODUCTS▼ CUSTOMERS▼ PRICING RESOURCES▼ REQUEST A DEMO

Talent Assessment | 16 Min Read

How AI-based HR Chatbots are Simplifying Pre-screening

Source https://www.mettl.com

Fairness in ML: a major societal concern

05-17-19

Schools are using software to help pick who gets in. What could go wrong?

Admissions officers are increasingly turning to automation and AI with the hope of streamlining the application process and leveling the playing field.

Source https://www.fastcompany.com

Fairness in ML: a major societal concern

SCIENCE ADVANCES | RESEARCH ARTICLE

RESEARCH METHODS

The accuracy, fairness, and limits of predicting recidivism

Julia Dressel and Hany Farid*

Algorithms for predicting recidivism are commonly used to assess a criminal defendant's likelihood of committing a crime. These predictions are used in pretrial, parole, and sentencing decisions. Proponents of these systems argue that big data and advanced machine learning make these analyses more accurate and less biased than humans. We show, however, that the widely used commercial risk assessment software COMPAS is no more accurate or fair than predictions made by people with little or no criminal justice expertise. In addition, despite COMPAS's collection of 137 features, the same accuracy can be achieved with a simple linear predictor with only two features.

EU regulation for AI

EUROPEAN COMMISSION

Brussels, 21.4.2021

COM(2021) 206 final

2021/0106(COD)

Proposal for a

REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS

Group fairness paradigm

Observations: $(\underbrace{\text{feature}}_{\boldsymbol{X}}, \underbrace{\text{sensitive attribute}}_{S}, \underbrace{\text{label}}_{Y}) \sim \mathbb{P}$ on $\mathcal{X} \times \mathcal{S} \times \mathcal{Y}$

Predictions: $f : \mathcal{Z} \to \mathcal{Y}$

- Fairness through awareness: $\mathcal{Z} = \mathcal{X} \times \mathcal{S}$ (disparate treatment)
- Fairness through unawareness: $\mathcal{Z} = \mathcal{X}$ (legal reasons: regulations)

Risk: $f \mapsto \mathcal{R}(f)$

- classification: $\mathcal{R}(f) = \mathbb{P}(Y \neq f(\boldsymbol{Z}))$
- regression: $\mathcal{R}(f) = \mathbb{E}(Y - f(\boldsymbol{Z}))^2$

Fairness criteria: dichotomy of prediction functions: which functions we call fair? There are a lot of definitions, maybe too many to parse.

Connections of ML fairness notions with political philosophy (Heidari et al., 2019)

Popular definitions of fair classifiers

- Demographic Parity (DP) (Calders, Kamiran, and Pechenizkiy, 2009)

$$\mathbb{P}(f(\boldsymbol{Z}) = 1 \mid S = 0) = \mathbb{P}(f(\boldsymbol{Z}) = 1 \mid S = 1)$$

  1. Prediction rate is the same for two groups
  2. Random variable $f(\boldsymbol{Z})$ is independent from S
  3. Only $\boldsymbol{X}|S$ matters
  4. Constant predictions satisfy DP

Popular definitions of fair classifiers

- Demographic Parity (DP) (Calders, Kamiran, and Pechenizkiy, 2009)

$$\mathbb{P}(f(\boldsymbol{Z}) = 1 \mid S = 0) = \mathbb{P}(f(\boldsymbol{Z}) = 1 \mid S = 1)$$

  1. Prediction rate is the same for two groups
  2. Random variable $f(\boldsymbol{Z})$ is independent from S
  3. Only $\boldsymbol{X}|S$ matters
  4. Constant predictions satisfy DP

- Equalized Odds (Hardt, Price, and Srebro, 2016)

$$\mathbb{P}(f(\boldsymbol{Z}) = y \mid Y = y, S = 0) = \mathbb{P}(f(\boldsymbol{Z}) = y \mid Y = y, S = 1) \quad \forall y \in \{0, 1\}$$

  1. Equal True Positive and True Negative rates
  2. Requires more knowledge about the distribution
  3. Constant predictions satisfy Equalized Odds

Popular definitions of fair classifiers

- Equal Opportunity (Hardt, Price, and Srebro, 2016)

$$\mathbb{P}(f(\boldsymbol{Z}) = 1 \mid Y = 1, S = 0) = \mathbb{P}(f(\boldsymbol{Z}) = 1 \mid Y = 1, S = 1)$$

  1. Equal True Positive rates
  2. If a person $\boldsymbol{Z}$ is qualified ($Y = 1$) then positive prediction ($f(\boldsymbol{Z}) = 1$) is given with the same probability for any sensitive attribute

Popular definitions of fair classifiers

- Equal Opportunity (Hardt, Price, and Srebro, 2016)

  $$\mathbb{P}(f(\boldsymbol{Z}) = 1 \mid Y = 1, S = 0) = \mathbb{P}(f(\boldsymbol{Z}) = 1 \mid Y = 1, S = 1)$$

  1. Equal True Positive rates
  2. If a person $\boldsymbol{Z}$ is qualified ($Y = 1$) then positive prediction ($f(\boldsymbol{Z}) = 1$) is given with the same probability for any sensitive attribute

- Test fairness (Chouldechova, 2017)

  $$\mathbb{P}(Y = 1 \mid S = 0, f(\boldsymbol{Z}) = 1) = \mathbb{P}(Y = 1 \mid S = 1, f(\boldsymbol{Z}) = 1)$$

  1. Y independent from S conditionally on $f(\boldsymbol{Z}) = 1$.
  2. Closely related to group-wise calibration.

Global view on group fairness constraints

Most of the definitions of fairness fall inside or try to reflect only 3 criteria

1. $f(\boldsymbol{Z}) \perp\!\!\!\perp S$ - independence (DP, Statistical Parity)

2. $(f(\boldsymbol{Z}) \perp\!\!\!\perp S) \mid Y$ - separation (Equal Odds, Equal Opportunity)

3. $(Y \perp\!\!\!\perp S) \mid f(\boldsymbol{Z})$ - sufficiency (Test fairness)

N.B. Sometimes we consider a score function $f(\boldsymbol{Z}) \in [0, 1]$.

Taken from Chapter 2 of (Barocas, Hardt, and Narayanan, 2019)

Impossibilities for score functions

1. $f(\boldsymbol{Z}) \perp\!\!\!\perp S$ - independence (DP, Statistical Parity)

2. $(f(\boldsymbol{Z}) \perp\!\!\!\perp S) \mid Y$ - separation (Equal Odds, Equal Opportunity)

3. $(Y \perp\!\!\!\perp S) \mid f(\boldsymbol{Z})$ - sufficiency (Test fairness)

- If S and Y are not independent, then sufficiency and independence cannot both hold.
- If $Y \in \{0, 1\}$, S and Y are not independent, $f(\boldsymbol{Z})$ is not independent from Y, then independence and separation cannot both hold.
- If S and Y are not independent, and $\mathbb{P}(Y = 1) \in (0, 1)$, then separation and sufficiency cannot both hold.

Taken from Chapter 2 of (Barocas, Hardt, and Narayanan, 2019)
propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Impossibilities for score functions

1. $f(\boldsymbol{Z}) \perp\!\!\!\perp S$ - independence (DP, Statistical Parity)

2. $(f(\boldsymbol{Z}) \perp\!\!\!\perp S) \mid Y$ - separation (Equal Odds, Equal Opportunity)

3. $(Y \perp\!\!\!\perp S) \mid f(\boldsymbol{Z})$ - sufficiency (Test fairness)

- If S and Y are not independent, then sufficiency and independence cannot both hold.
- If $Y \in \{0, 1\}$, S and Y are not independent, $f(\boldsymbol{Z})$ is not independent from Y, then independence and separation cannot both hold.
- If S and Y are not independent, and $\mathbb{P}(Y = 1) \in (0, 1)$, then separation and sufficiency cannot both hold.

A fact: famous example of COMPAS nearly satisfied sufficiency, but failed to satisfy separation. Due to the latter propublica published an article that extremely influenced the field of algorithmic fairness (Chouldechova, 2017).

Taken from Chapter 2 of (Barocas, Hardt, and Narayanan, 2019)
propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Three (rough) types of methods: pre-processing

Pre-processing – Fair representation

Find a feature representation $\boldsymbol{Z} \mapsto \hat{\varphi}(\boldsymbol{Z})$ such that

$$\hat{\varphi}(\boldsymbol{Z}) \perp\!\!\!\perp S$$

then use any method on this representation.
Typically, (unsupervised) optimal fair representation is defined as

$$\varphi^* \in \arg\min \{\mathbb{E}[d(\boldsymbol{X}, \varphi(\boldsymbol{Z}))] \,:\, \varphi(\boldsymbol{Z}) \perp\!\!\!\perp S\} \quad .$$

Three (rough) types of methods: pre-processing

Pre-processing – Fair representation

Find a feature representation $\boldsymbol{Z} \mapsto \hat{\varphi}(\boldsymbol{Z})$ such that

$$\hat{\varphi}(\boldsymbol{Z}) \perp\!\!\!\perp S$$

then use any method on this representation.
Typically, (unsupervised) optimal fair representation is defined as

$$\varphi^* \in \arg\min \left\{\mathbb{E}[d(\boldsymbol{X}, \varphi(\boldsymbol{Z}))] \, : \, \varphi(\boldsymbol{Z}) \perp\!\!\!\perp S\right\} \quad .$$

Methods

- Linear models (Zemel et al., 2013)
- Kernel methods (Grünewälder and Khaleghi, 2021)
- GANs (Xu et al., 2018)

Three (rough) types of methods: in-processing

Add the fairness constraint into training

$$f^*_{\mathcal{F}} \in \arg\min_{f \in \mathcal{F}} \{\mathcal{R}(f) \,:\, f(\boldsymbol{Z}) \perp\!\!\!\perp S\}$$

In-processing type method: Given data $(\boldsymbol{X}_1, S_1, Y_1), \ldots, (\boldsymbol{X}_n, S_n, Y_n)$ build an estimator $\hat{f}$ as a solution

$$\min_{f \in \mathcal{F}} \left\{\hat{\mathcal{R}}(f) + \lambda_0 \cdot \Omega_{\text{compl}}(f) + \lambda_1 \cdot \Omega_{\text{UNfairness}}(f)\right\}$$

Three (rough) types of methods: in-processing

Add the fairness constraint into training

$$f^*_{\mathcal{F}} \in \arg\min_{f \in \mathcal{F}} \{\mathcal{R}(f) \, : \, f(\boldsymbol{Z}) \perp\!\!\!\perp S\}$$

In-processing type method: Given data $(\boldsymbol{X}_1, S_1, Y_1), \ldots, (\boldsymbol{X}_n, S_n, Y_n)$ build an estimator $\hat{f}$ as a solution

$$\min_{f \in \mathcal{F}} \left\{\hat{\mathcal{R}}(f) + \lambda_0 \cdot \Omega_{\text{compl}}(f) + \lambda_1 \cdot \Omega_{\text{UNfairness}}(f)\right\}$$

Methods

- Regularized ERM methods (Oneto, Donini, and Pontil, 2019)
- MWU-type methods for minmax games (Agarwal et al., 2018)

Three (rough) types of methods: post-processing

Given a base algorithm f, find a transformation

$$f \mapsto \hat{T}(f) \ ,$$

so that $\hat{T}(f)$ satisfies your fairness constraint

Three (rough) types of methods: post-processing

Given a base algorithm f, find a transformation

$$f \mapsto \hat{T}(f) \enspace ,$$

so that $\hat{T}(f)$ satisfies your fairness constraint
Typical algorithm construction is based on the connection between

$$f^*_{\text{fair}} \in \operatorname*{arg\,min}_{f : \mathcal{Z} \to \mathcal{Y}} \{\mathcal{R}(f) \,:\, f \text{ is fair}\} \quad \text{and} \quad f^*_{\text{Bayes}} \in \operatorname*{arg\,min}_{f : \mathcal{Z} \to \mathcal{Y}} \mathcal{R}(f)$$

Often we can show that

$$f^*_{\text{fair}} = T^*(f^*_{\text{Bayes}}) \enspace ,$$

treat the base algorithm f as if it were a Bayes and estimate T^*

Three (rough) types of methods: post-processing

Given a base algorithm f, find a transformation

$$f \mapsto \hat{T}(f) \enspace ,$$

so that $\hat{T}(f)$ satisfies your fairness constraint
Typical algorithm construction is based on the connection between

$$f^*_{\text{fair}} \in \mathop{\arg\min}_{f : \mathcal{Z} \to \mathcal{Y}} \{\mathcal{R}(f) \,:\, f \text{ is fair}\} \quad \text{and} \quad f^*_{\text{Bayes}} \in \mathop{\arg\min}_{f : \mathcal{Z} \to \mathcal{Y}} \mathcal{R}(f)$$

Often we can show that

$$f^*_{\text{fair}} = T^*(f^*_{\text{Bayes}}) \enspace ,$$

treat the base algorithm f as if it were a Bayes and estimate T^*

Methods

- Threshold adjustments (Hardt, Price, and Srebro, 2016; Menon and Williamson, 2018; C. et al., 2019)
- Optimal transport based (C. et al., 2020; Le Gouic, Loubes, and Rigollet, 2020)

What is the Demographic Parity constraint?

with C. Denis, S. Gaucher, M. Hebiri, L. Oneto, M. Pontil, and N. Schreuder

Learning with Demographic Parity

$$(\underbrace{\text{feature}}_{\boldsymbol{X}}, \underbrace{\text{sensitive attribute}}_{S}, \underbrace{\text{signal}}_{Y}) \sim \mathbb{P} \text{ on } \mathbb{R}^d \times \underbrace{\mathcal{S}}_{=\{1,\ldots,K\}} \times \mathcal{Y}$$

Prediction: $f : \mathbb{R}^d \times \mathcal{S} \to \mathcal{Y}$

Risk: $\mathcal{R}(f) = \mathbb{E}[(Y - f(\boldsymbol{X}, S))^2]$ or $\mathcal{R}(f) = \mathbb{P}(Y \neq f(\boldsymbol{X}, S))$

Demographic Parity fairness

$$f(\boldsymbol{X}, S) \perp\!\!\!\perp S$$

Optimal fair prediction:

$$f_0^* \in \arg\min \{\mathcal{R}(f) \ : \ f(\boldsymbol{X}, S) \perp\!\!\!\perp S\}$$

Our goals

1. Understand a relation between regression and classification under the Demographic Parity constraint
2. Understand a relation between constraint and unconstraint (Bayes optimal) problems
3. Try to explain the notion of Demographic Parity in a simple language
4. Figure out an estimation strategy and get some bounds on risk and unfairness

Classical classification-regression link

$$(\underbrace{\text{feature}}_{\boldsymbol{X}}, \underbrace{\text{sensitive attribute}}_{S}, \underbrace{\text{signal}}_{Y}) \sim \mathbb{P} \text{ on } \mathbb{R}^d \times \underbrace{\mathcal{S}}_{=\{1,\ldots,K\}} \times \{0,1\}$$

$$g^* \in \mathop{\arg\min}_{g:\mathcal{X}\times\mathcal{S}\to\{0,1\}} \mathbb{P}(Y \neq g(\boldsymbol{X}, S)) \qquad f^* \in \mathop{\arg\min}_{f:\mathcal{X}\times\mathcal{S}\to\mathbb{R}} \mathbb{E}[(Y - f(\boldsymbol{X}, S))^2]$$

Classical classification-regression link

$$(\underbrace{\text{feature}}_{\boldsymbol{X}}, \underbrace{\text{sensitive attribute}}_{S}, \underbrace{\text{signal}}_{Y}) \sim \mathbb{P} \text{ on } \mathbb{R}^d \times \underbrace{\mathcal{S}}_{=\{1,\ldots,K\}} \times \{0,1\}$$

$$g^* \in \operatorname*{arg\,min}_{g:\mathcal{X}\times\mathcal{S}\to\{0,1\}} \mathbb{P}(Y \neq g(\boldsymbol{X}, S)) \qquad f^* \in \operatorname*{arg\,min}_{f:\mathcal{X}\times\mathcal{S}\to\mathbb{R}} \mathbb{E}[(Y - f(\boldsymbol{X}, S))^2]$$

A folklore result

$$f^*(\boldsymbol{X}, S) = \mathbb{E}[Y \mid \boldsymbol{X}, S] \qquad g^*(\boldsymbol{X}, S) = \mathbb{1}\{f^*(\boldsymbol{X}, S) \geq 1/2\}$$

present in every ML/Stat book

N.B. Simple to prove, but very useful in theory and in practice.

Classification-regression link under DP

$$(\underbrace{\text{feature}}_{\boldsymbol{X}}, \underbrace{\text{sensitive attribute}}_{S}, \underbrace{\text{signal}}_{Y}) \sim \mathbb{P} \text{ on } \mathbb{R}^d \times \underbrace{\mathcal{S}}_{=\{1,\ldots,K\}} \times \{0, 1\}$$

Can we expect the same result under the Demographic parity constraint?

There is really no reason for such a relation...

Classification-regression link under DP

$$(\underbrace{\text{feature}}_{\boldsymbol{X}}, \underbrace{\text{sensitive attribute}}_{S}, \underbrace{\text{signal}}_{Y}) \sim \mathbb{P} \text{ on } \mathbb{R}^d \times \underbrace{\mathcal{S}}_{=\{1,\ldots,K\}} \times \{0,1\}$$

Can we expect the same result under the Demographic parity constraint?

There is really no reason for such a relation... Indeed, if

$$g_0^* \in \operatorname*{arg\,min}_{g:\mathcal{X}\times\mathcal{S}\to\{0,1\}} \{\mathbb{P}(Y \neq g(\boldsymbol{X}, S)) \,:\, g(\boldsymbol{X}, S) \perp\!\!\!\perp S\}$$

$$f_0^* \in \operatorname*{arg\,min}_{f:\mathcal{X}\times\mathcal{S}\to\{0,1\}} \left\{\mathbb{E}[(Y - f(\boldsymbol{X}, S))^2] \,:\, f(\boldsymbol{X}, S) \perp\!\!\!\perp S\right\}$$

are such that

$$g_0^*(\boldsymbol{X}, S) = \mathbb{1}\{f_0^*(\boldsymbol{X}, S) \geq 1/2\}\ ,$$

then g_0^* is "much fairer" than we expect—f_0^* is fair at every threshold, while g_0^* needs to be fair only at one of them.

Classification-regression link under DP

$$(\underbrace{\text{feature}}_{\boldsymbol{X}}, \underbrace{\text{sensitive attribute}}_{S}, \underbrace{\text{signal}}_{Y}) \sim \mathbb{P} \text{ on } \mathbb{R}^d \times \underbrace{\mathcal{S}}_{=\{1,\dots,K\}} \times \{0,1\}$$

$$g_0^* \in \operatorname*{arg\,min}_{g:\mathcal{X}\times\mathcal{S}\to\{0,1\}} \{\mathbb{P}(Y \neq g(\boldsymbol{X}, S)) \,:\, g(\boldsymbol{X}, S) \perp\!\!\!\perp S\}$$

$$f^* \in \operatorname*{arg\,min}_{f:\mathcal{X}\times\mathcal{S}\to\mathbb{R}} \mathbb{E}[(Y - f(\boldsymbol{X}, S))^2]$$

Lemma

$$g_0^*(\boldsymbol{X}, S) = \mathbb{1}\left\{f^*(\boldsymbol{X}, S) \geq \frac{1}{2} + \frac{\lambda_s^*}{2w_s}\right\}$$

where $w_s = \mathbb{P}(S = s)$ and

$$(\lambda_1^*, \dots, \lambda_K^*) \in \operatorname*{arg\,min}_{(\lambda_1,\dots,\lambda_K)\in\mathbb{R}^K} \left\{\mathbb{E}\left|2f^*(\boldsymbol{X}, S) - 1 - \frac{\lambda_S}{w_S}\right| \,:\, \sum_{s\in\mathcal{S}} \lambda_s = 0\right\}$$

(Menon and Williamson, 2018; Gaucher, Schreuder, and C., 2023)

Classification-regression link under DP

$$(\underbrace{\text{feature}}_{\boldsymbol{X}}, \underbrace{\text{sensitive attribute}}_{S}, \underbrace{\text{signal}}_{Y}) \sim \mathbb{P} \text{ on } \mathbb{R}^d \times \underbrace{\mathcal{S}}_{=\{1,\ldots,K\}} \times \{0,1\}$$

Nevertheless

$$g_0^* \in \mathop{\arg\min}_{g : \mathcal{X}\times\mathcal{S}\to\{0,1\}} \left\{\mathbb{P}(Y \neq g(\boldsymbol{X}, S)) \ : \ g(\boldsymbol{X}, S) \perp\!\!\!\perp S\right\}$$

$$f_0^* \in \mathop{\arg\min}_{f : \mathcal{X}\times\mathcal{S}\to\mathbb{R}} \left\{\mathbb{E}[(Y - f(\boldsymbol{X}, S))^2] \ : \ f(\boldsymbol{X}, S) \perp\!\!\!\perp S\right\}$$

Lemma

$$g^*(\boldsymbol{X}, S) = \mathbb{1}\{f_0^*(\boldsymbol{X}, S) \geq 1/2\} \qquad \boxed{f_0^*(\boldsymbol{X}, S) = ??}$$

(Gaucher, Schreuder, and C., 2023)

N.B. It remains to understand the regression case

Regression + Demographic Parity

$$(\underbrace{\text{feature}}_{\boldsymbol{X}}, \underbrace{\text{sensitive attribute}}_{S}, \underbrace{\text{signal}}_{Y}) \sim \mathbb{P} \text{ on } \mathbb{R}^d \times \underbrace{\mathcal{S}}_{=\{1,\ldots,K\}} \times \mathbb{R}$$

Prediction: $f : \mathbb{R}^d \times \mathcal{S} \to \mathbb{R}$

Risk: $\mathcal{R}(f) = \mathbb{E}[(f^*(\boldsymbol{X}, S) - f(\boldsymbol{X}, S))^2]$ where $f^*(\boldsymbol{X}, S) = \mathbb{E}[Y \mid \boldsymbol{X}, S]$

Demographic Parity fairness

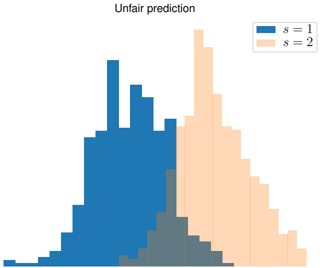
$$f(\boldsymbol{X}, S) \perp\!\!\!\perp S$$

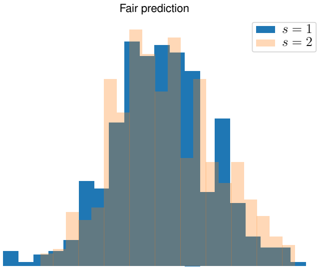
Optimal fair prediction:

$$f_0^* \in \arg\min \{\mathcal{R}(f) \ : \ f(\boldsymbol{X}, S) \perp\!\!\!\perp S\}$$

An illustration and main assumption

$$f(\boldsymbol{X}, S) \perp\!\!\!\perp S$$





Assumption (A)

The group-wise prediction distributions $\mathrm{Law}(f^*(\boldsymbol{X}, S) \mid S = s)$ have finite second moment and are non-atomic for any s in $\mathcal{S}$.

Optimal transport and the Wasserstein-2 metric

Define, for $\mu, \nu \in \mathcal{P}_2(\mathbb{R})$,

$$\mathsf{W}_2^2(\mu, \nu) \coloneqq \inf \left\{ \mathbb{E}_{(X,Y)}(\boldsymbol{X} - \boldsymbol{Y})^2 : \boldsymbol{X} \sim \mu, \boldsymbol{Y} \sim \nu \right\}.$$

- Metric on $\mathcal{P}_2(\mathbb{R}^d)$
- Optimal $T^*_{\mu \to \nu} \equiv F_\nu^{-1} \circ F_\mu$
- Nice interpretations

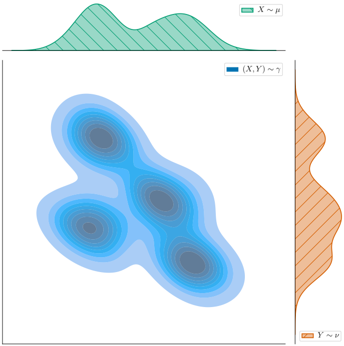


Figure: Transport plan illustration

Reminder: post-processing

Optional fair: $f_0^* \in \underset{f:\mathbb{R}^d \times \mathcal{S} \to \mathbb{R}}{\arg\min} \{\mathcal{R}(f) \,:\, f(\boldsymbol{X}, S) \perp\!\!\!\perp S\}$

Bayes optimal: $f^* \in \underset{f:\mathbb{R}^d \times \mathcal{S} \to \mathbb{R}}{\arg\min} \, \mathcal{R}(f)$

Question: is there a link between f_0^* and f^*?

More precisely, can we show that

$$f_0^* \equiv T \circ f^* \ ?$$

Main insight

Optimal fair: $f_0^* \in \operatorname*{arg\,min}_{f:\mathbb{R}^d \times \mathcal{S} \to \mathbb{R}} \{\mathcal{R}(f) : f(\boldsymbol{X}, S) \perp\!\!\!\perp S\}$

Bayes optimal: $f^* \in \operatorname*{arg\,min}_{f:\mathbb{R}^d \times \mathcal{S} \to \mathbb{R}} \mathcal{R}(f)$

Question: is there a link between f_0^* and f^*?

Theorem

Set $w_s = \mathbb{P}(S{=}s)$. Let Assumption (A) be satisfied, then

$$\mathrm{Law}(f_0^*(\boldsymbol{X}, S)) = \underbrace{\operatorname*{arg\,min}_{\nu \in \mathcal{P}_2(\mathbb{R})} \sum_{s \in \mathcal{S}} w_s \mathsf{W}_2^2\Big(\mathrm{Law}(f^*(\boldsymbol{X}, S) \mid S = s),\ \nu\Big)}_{\text{Wasserstein barycenter problem}} ,$$

$$f_0^*(\boldsymbol{x}, 1) = w_1 f^*(\boldsymbol{x}, 1) + w_2 T^*_{1\to 2} \circ f^*(\boldsymbol{x}, 1), \qquad \forall \boldsymbol{x} \in \mathbb{R}^d ,$$

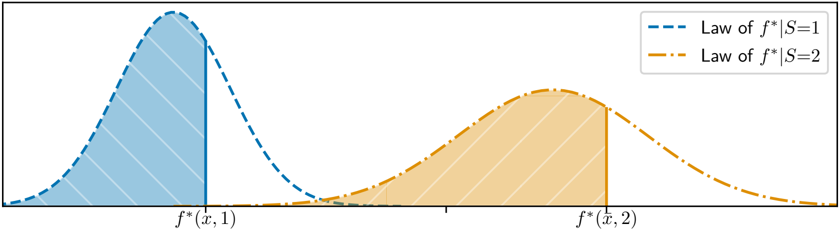
$T^*_{1\to 2}$ – optimal transport map from $\mathrm{Law}(f^* \mid S = 1)$ to $\mathrm{Law}(f^* \mid S = 2)$.

(C. et al., 2020)

Interpretation for $\mathcal{S} = \{1, 2\}$

Fair optimal: $f_0^*(\boldsymbol{x}, 1) = w_1 f^*(\boldsymbol{x}, 1) + w_2 F^{-1}_{f^*|S=2} \circ F_{f^*|S=1} \circ f^*(\boldsymbol{x}, 1)$

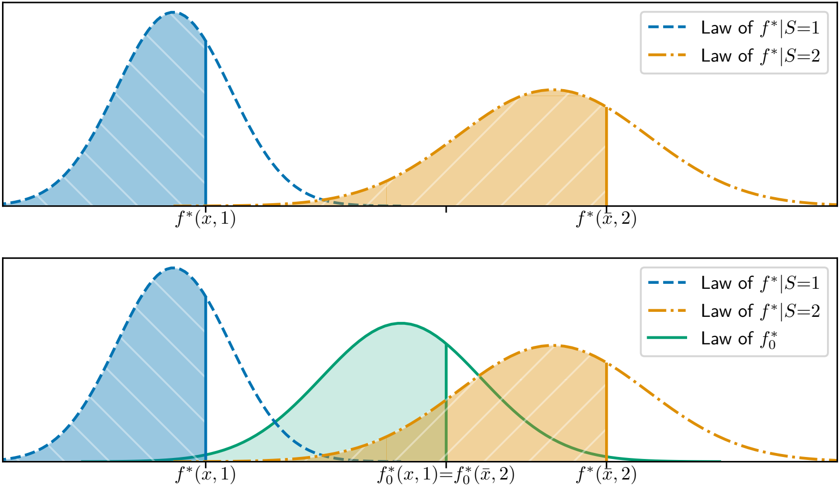
Fair optimal prediction f_0^* with $w_1 = 2/5$ and $w_2 = 3/5$



Interpretation for $\mathcal{S} = \{1, 2\}$

Fair optimal: $f_0^*(\boldsymbol{x}, 1) = w_1 f^*(\boldsymbol{x}, 1) + w_2 F^{-1}_{f^*|S=2} \circ F_{f^*|S=1} \circ f^*(\boldsymbol{x}, 1)$

Fair optimal prediction f_0^* with $w_1 = 2/5$ and $w_2 = 3/5$



Generic post-processing estimator ($\mathcal{S} = \{1, 2\}$)

Fair optimal: $f_0^*(\boldsymbol{x}, 1) = w_1 f^*(\boldsymbol{x}, 1) + w_2 T^*_{1\to 2} \circ f^*(\boldsymbol{x}, 1)$

- Base estimator: $\hat{f} : \mathbb{R}^d \times \{1, 2\} \to \mathbb{R}$ trained independently from the following data.
- Unlabeled data: $\forall s \in \mathcal{S}$ we observe $\boldsymbol{X}_1^s, \dots, \boldsymbol{X}_{N_s}^s \overset{i.i.d.}{\sim} \mathbb{P}_{\boldsymbol{X}|S=s}$

Meta algo:
1. estimate w_s if needed
2. estimate transport maps $T^*_{1\to 2}$ and $T^*_{2\to 1}$ using unlabeled data and base estimator

Generic post-processing estimator ($\mathcal{S} = \{1, 2\}$)

Fair optimal: $f_0^*(\boldsymbol{x}, 1) = w_1 f^*(\boldsymbol{x}, 1) + w_2 T^*_{1\to 2} \circ f^*(\boldsymbol{x}, 1)$

- Base estimator: $\hat{f} : \mathbb{R}^d \times \{1, 2\} \to \mathbb{R}$ trained independently from the following data.
- Unlabeled data: $\forall s \in \mathcal{S}$ we observe $\boldsymbol{X}_1^s, \ldots, \boldsymbol{X}_{N_s}^s \overset{i.i.d.}{\sim} \mathbb{P}_{\boldsymbol{X}|S=s}$

Meta algo:
1. estimate w_s if needed
2. estimate transport maps $T^*_{1\to 2}$ and $T^*_{2\to 1}$ using unlabeled data and base estimator

Put together:
3. $\hat{f}_0(\boldsymbol{x}, 1) = w_1 \hat{f}(\boldsymbol{x}, 1) + w_2 \hat{T}_{1\to 2} \circ \hat{f}(\boldsymbol{x}, 1)$

Theoretical guarantees

Theorem

For any joint distribution $\mathbb{P}$ of $(\boldsymbol{X}, S, Y)$, any base estimator $\hat{f}$ it holds that

$$\boxed{\hat{f}_0(\boldsymbol{X}, S) \perp\!\!\!\perp S}$$

Under additional assumptions on $\mathbb{P}$ we have

$$\mathbf{E}\|\hat{f}_0 - f_0^*\|_1 \lesssim \underbrace{\mathbf{E}\|\hat{f} - f^*\|_1}_{\text{quality of base estimator}} \bigvee \underbrace{\sum_{s \in \mathcal{S}} w_s N_s^{-1/2}}_{\text{transport estimation}}$$

(C. and Schreuder, 2022)

Additional assumptions: $(f^*(\boldsymbol{X}, S) \mid S = s)$ admits density which is upper and lower bounded

N_s – number of *unlabeled* samples from $\mathbb{P}_{\boldsymbol{X}|S=s}$ and $\mathbb{P}_{\boldsymbol{X}|S=2}$

How did we get exact independence and a cute lemma from conformal prediction theory

Lemma for "smoothed ranks"

Let $\boldsymbol{V} = (V, V_1, \dots, V_n)$ be *i.i.d.* real valued random variables and let U be distributed uniformly on $(0,1)$ and independent of $\boldsymbol{V}$. Let

$$F(U, V_1, \dots, V_n, x) = \frac{1}{n+1}\left(\sum_{i=1}^{n} \mathbb{1}\{V_i < x\} + U \cdot \left(1 + \sum_{i=1}^{n} \mathbb{1}\{V_i = x\}\right)\right) \quad .$$

Then, $F(U, V_1, \dots, V_n, V)$ is distributed uniformly on $(0,1)$.

V. Vovk and A. Gammerman

N.B. No assumptions on the distribution of the data, to compare with rank statistics.

How did we get exact independence and a cute lemma from conformal prediction theory

Lemma for "smoothed ranks"

Let $\boldsymbol{V} = (V, V_1, \ldots, V_n)$ be *i.i.d.* real valued random variables and let U be distributed uniformly on $(0, 1)$ and independent of $\boldsymbol{V}$.

$$F(U, V_1, \ldots, V_n, V) \sim \mathrm{Unif}(0, 1)$$

The optimal fair prediction can be expressed as

$$f_0^*(\boldsymbol{x}, s) = Q \circ (F_s(f^*(\boldsymbol{x}, s))) \quad ,$$

where Q is a monotone and F_s is the CDF of $\mathrm{Law}(f^*(\boldsymbol{X}, S) \mid S = s)$.

Idea. Use the above lemma for estimation of $F_s(f^*(\boldsymbol{x}, s))$ as it always produces uniform distributions on $(0, 1)$ (conditionally on $S = s$)

Conclusions

1. Group fairness – enforce some independence criterion

$$f(\boldsymbol{Z}) \perp\!\!\!\perp S, \qquad (f(\boldsymbol{Z}) \perp\!\!\!\perp S) \mid Y, \qquad (Y \perp\!\!\!\perp S) \mid f(\boldsymbol{Z})$$

2. Demographic parity preserves classical classification-regression

$$g_0^* = \mathbb{1}\{f_0^* \geq 1/2\}$$

3. Regression with demographic parity ($f(\boldsymbol{Z}) \perp\!\!\!\perp S$) can be characterized by Wasserstein barycenter problem

4. Demographic parity simply matches ranks of individuals from different groups

Thank you for your attention! Questions?

PROHIBITED ARTIFICIAL INTELLIGENCE PRACTICES

Article 5

1. The following artificial intelligence practices shall be prohibited:

(a) the placing on the market, putting into service or use of an AI system that deploys subliminal techniques beyond a person's consciousness in order to materially distort a person's behaviour in a manner that causes or is likely to cause that person or another person physical or psychological harm;

(b) the placing on the market, putting into service or use of an AI system that exploits any of the vulnerabilities of a specific group of persons due to their age, physical or mental disability, in order to materially distort the behaviour of a person pertaining to that group in a manner that causes or is likely to cause that person or another person physical or psychological harm;

(c) the placing on the market, putting into service or use of AI systems by public authorities or on their behalf for the evaluation or classification of the trustworthiness of natural persons over a certain period of time based on their social behaviour or known or predicted personal or personality characteristics, with the social score leading to either or both of the following:

(i) detrimental or unfavourable treatment of certain natural persons or whole groups thereof in social contexts which are unrelated to the contexts in which the data was originally generated or collected;

(ii) detrimental or unfavourable treatment of certain natural persons or whole groups thereof that is unjustified or disproportionate to their social behaviour or its gravity;

Bibliography I

Agarwal, A. et al. (2018). “A reductions approach to fair classification”. In: *arXiv preprint arXiv:1803.02453*.

Barocas, Solon, Moritz Hardt, and Arvind Narayanan (2019). *Fairness and Machine Learning*. `http://www.fairmlbook.org`. fairmlbook.org.

C., E. and N. Schreuder (2022). “A minimax framework for quantifying risk-fairness trade-off in regression”. In: *Annals of Statistics*.

C., E et al. (2019). “Leveraging Labeled and Unlabeled Data for Consistent Fair Binary Classification”. Submitted to NeurIPS19.

– (2020). “Fair Regression with Wasserstein Barycenters”. In: *NeurIPS 2020*.

Calders, T., F. Kamiran, and M. Pechenizkiy (2009). “Building classifiers with independency constraints”. In: *IEEE international conference on Data mining*.

Chouldechova, Alexandra (2017). “Fair prediction with disparate impact: A study of bias in recidivism prediction instruments”. In: *Big data* 5.2, pp. 153–163.

Gaucher, Solenne, Nicolas Schreuder, and Evgenii C. (2023). “Fair learning with Wasserstein barycenters for non-decomposable performance measures”. In: *AISTATS*.

Bibliography II

Grünewälder, Steffen and Azadeh Khaleghi (2021). "Oblivious Data for Fairness with Kernels". In: *Journal of Machine Learning Research* 22.208, pp. 1–36.

Hardt, M., E. Price, and N. Srebro (2016). "Equality of opportunity in supervised learning". In: *Neural Information Processing Systems*.

Heidari, Hoda et al. (2019). "A moral framework for understanding fair ML through economic models of equality of opportunity". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 181–190.

Le Gouic, Thibaut, Jean-Michel Loubes, and Philippe Rigollet (2020). "Projection to fairness in statistical learning". In: *arXiv e-prints*, arXiv–2005.

Menon, A. and R. C. Williamson (2018). "The cost of fairness in binary classification". In: *Conference on Fairness, Accountability and Transparency*.

Oneto, L., M. Donini, and M. Pontil (2019). "General Fair Empirical Risk Minimization". In: *arXiv preprint arXiv:1901.10080*.

Bibliography III

Xu, Depeng et al. (2018). "Fairgan: Fairness-aware generative adversarial networks". In: *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, pp. 570–575.

Zemel, R. et al. (2013). "Learning fair representations". In: *International Conference on Machine Learning*.

Agarwal, A. et al. (2018). "A reductions approach to fair classification". In: *arXiv preprint arXiv:1803.02453*.

Barocas, Solon, Moritz Hardt, and Arvind Narayanan (2019). *Fairness and Machine Learning*. `http://www.fairmlbook.org`. fairmlbook.org.

C., E. and N. Schreuder (2022). "A minimax framework for quantifying risk-fairness trade-off in regression". In: *Annals of Statistics*.

C., E et al. (2019). "Leveraging Labeled and Unlabeled Data for Consistent Fair Binary Classification". Submitted to NeurIPS19.

– (2020). "Fair Regression with Wasserstein Barycenters". In: *NeurIPS 2020*.

Calders, T., F. Kamiran, and M. Pechenizkiy (2009). "Building classifiers with independency constraints". In: *IEEE international conference on Data mining*.

Bibliography IV

Chouldechova, Alexandra (2017). "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments". In: *Big data* 5.2, pp. 153–163.

Gaucher, Solenne, Nicolas Schreuder, and Evgenii C. (2023). "Fair learning with Wasserstein barycenters for non-decomposable performance measures". In: *AISTATS*.

Grünewälder, Steffen and Azadeh Khaleghi (2021). "Oblivious Data for Fairness with Kernels". In: *Journal of Machine Learning Research* 22.208, pp. 1–36.

Hardt, M., E. Price, and N. Srebro (2016). "Equality of opportunity in supervised learning". In: *Neural Information Processing Systems.*

Heidari, Hoda et al. (2019). "A moral framework for understanding fair ML through economic models of equality of opportunity". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 181–190.

Le Gouic, Thibaut, Jean-Michel Loubes, and Philippe Rigollet (2020). "Projection to fairness in statistical learning". In: *arXiv e-prints*, arXiv–2005.

Bibliography V

Menon, A. and R. C. Williamson (2018). “The cost of fairness in binary classification”. In: *Conference on Fairness, Accountability and Transparency.*

Oneto, L., M. Donini, and M. Pontil (2019). “General Fair Empirical Risk Minimization”. In: *arXiv preprint arXiv:1901.10080.*

Xu, Depeng et al. (2018). “Fairgan: Fairness-aware generative adversarial networks”. In: *2018 IEEE International Conference on Big Data (Big Data).* IEEE, pp. 570–575.

Zemel, R. et al. (2013). “Learning fair representations”. In: *International Conference on Machine Learning.*