On the (im)possibility of fairness

Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian

presented by Sarah Dean

Fairness in ML, September 2017

“Similar people should be treated similarly”

but similar in what sense?

On Monday, we considered the fairness constraint

$$D(f(x), f(x')) \leq d(x, x')$$

which amounts to a Lipschitz condition on the decision map $f : \mathcal{X} \to \mathcal{D}$

"Similar people should be treated similarly"

but similar in what sense?

On Monday, we considered the fairness constraint

$$D(f(x), f(x')) \leq d(x, x')$$

which amounts to a Lipschitz condition on the decision map $f : \mathcal{X} \to \mathcal{D}$

- Sensitivity to definition of d
- What is the space of individuals $\mathcal{X}$? The feature space?

Three spaces of the decision pipeline

We distinguish between the *construct space* $\mathcal{C}$, the *observed space* $\mathcal{O}$, and the *decision space* $\mathcal{D}$.

Decision space	Construct space	Observed space
College performance	intelligence	IQ
College performance	HS success	GPA
Recidivism	“criminality”	family history of crime
Recidivism	risk-averseness	age
Employee productivity	knowedge of job	years experience

Three spaces of the decision pipeline

We distinguish between the *construct space* $\mathcal{C}$, the *observed space* $\mathcal{O}$, and the *decision space* $\mathcal{D}$.

Decision space	Construct space	Observed space
College performance	intelligence	IQ
College performance	HS success	GPA
Recidivism	"criminality"	family history of crime
Recidivism	risk-averseness	age
Employee productivity	knowedge of job	years experience

- Imperfections in choice of construct vs. observed space?

Three spaces of the decision pipeline

We distinguish between the *construct space* $\mathcal{C}$, the *observed space* $\mathcal{O}$, and the *decision space* $\mathcal{D}$.

Decision space	Construct space	Observed space
College performance	intelligence	IQ
College performance	HS success	GPA
Recidivism	"criminality"	family history of crime
Recidivism	risk-averseness	age
Employee productivity	knowedge of job	years experience

- Imperfections in choice of construct vs. observed space?
- Should we distinguish between the decision space (label space) and the outcome space?

Decision pipeline as maps between spaces

We consider transformations between spaces,

- observation processes $g : \mathcal{C} \to \mathcal{O}$
- desired "ideal map" $f : \mathcal{C} \to \mathcal{D}$
- designed or learned map $\hat{f} : \mathcal{O} \to \mathcal{D}$

$$\begin{tikzcd} \mathcal{C} \arrow[r, "g"] \arrow[dr, "f"'] & \mathcal{O} \arrow[d, "\hat{f}"] \\ & \mathcal{D} \end{tikzcd}$$

Decision pipeline as maps between spaces

We consider transformations between spaces,

- observation processes $g : \mathcal{C} \rightarrow \mathcal{O}$
- desired "ideal map" $f : \mathcal{C} \rightarrow \mathcal{D}$
- designed or learned map $\hat{f} : \mathcal{O} \rightarrow \mathcal{D}$

$$
\begin{tikzcd}
\mathcal{C} \arrow[r, "g"] \arrow[dr, "f"'] & \mathcal{O} \arrow[d, "\hat{f}"] \\
 & \mathcal{D}
\end{tikzcd}
$$

To make up for lack of knowledge about $\mathcal{C}$, g, and f, we will have to make assumptions based on our "world view".

Decision pipeline as maps between spaces

We consider transformations between spaces,

- observation processes $g : \mathcal{C} \to \mathcal{O}$
- desired "ideal map" $f : \mathcal{C} \to \mathcal{D}$
- designed or learned map $\hat{f} : \mathcal{O} \to \mathcal{D}$

$$\begin{tikzcd} \mathcal{C} \arrow[r, "g"] \arrow[dr, "f"'] & \mathcal{O} \arrow[d, "\hat{f}"] \\ & \mathcal{D} \end{tikzcd}$$

To make up for lack of knowledge about $\mathcal{C}$, g, and f, we will have to make assumptions based on our "world view".

The **distortion** ρ_h of map $h : \mathcal{X} \to \mathcal{Y}$ is

$$\sup_{x,x' \in \mathcal{X}} |d_{\mathcal{X}}(x, x') - d_{\mathcal{Y}}(h(x), h(x'))|$$

and $\rho(\mathcal{X}, \mathcal{Y}) = \min_h \rho_h$

Individual fairness and what-you-see-is-what-you-get

A map $f : \mathcal{C} \to \mathcal{D}$ is (ϵ, ϵ')-**fair** if for all $x, x' \in \mathcal{C}$

$$d_{\mathcal{C}}(x, x') \leq \epsilon \implies d_{\mathcal{D}}(f(x), f(x')) \leq \epsilon'$$

Individual fairness and what-you-see-is-what-you-get

A map $f : \mathcal{C} \to \mathcal{D}$ is (ϵ, ϵ')-**fair** if for all $x, x' \in \mathcal{C}$

$$d_{\mathcal{C}}(x, x') \leq \epsilon \implies d_{\mathcal{D}}(f(x), f(x')) \leq \epsilon'$$

Axiom (WYSIWYG) The distortion between $\mathcal{C}$ and $\mathcal{O}$ is at most δ

Individual fairness and what-you-see-is-what-you-get

A map $f : \mathcal{C} \to \mathcal{D}$ is (ϵ, ϵ')-**fair** if for all $x, x' \in \mathcal{C}$

$$d_{\mathcal{C}}(x, x') \leq \epsilon \implies d_{\mathcal{D}}(f(x), f(x')) \leq \epsilon'$$

Axiom (WYSIWYG) The distortion between $\mathcal{C}$ and $\mathcal{O}$ is at most δ

An **individual fairness mechanism** (IFM_ϵ) is a nontrivial mapping $\hat{f} : \mathcal{O} \to \mathcal{D}$ with $\rho_{\hat{f}} \leq \epsilon$.

Fairness is possible!!!

Theorem

Under WYSIWYG, an IFM$_{\delta'}$ guarantees $(\epsilon, \delta + \delta')$-fairness.

$$\begin{tikzcd} \mathcal{C} \arrow[r, "g"] \arrow[dr, "f"'] & \mathcal{O} \arrow[d, "\hat{f}"] \\ & \mathcal{D} \end{tikzcd}$$

Fairness is possible!!!

Theorem

Under WYSIWYG, an $\mathrm{IFM}_{\delta'}$ guarantees $(\epsilon, \delta + \delta')$-fairness.

$$\begin{tikzcd} \mathcal{C} \arrow[r, "g"] \arrow[dr, "f"'] & \mathcal{O} \arrow[d, "\hat{f}"] \\ & \mathcal{D} \end{tikzcd}$$

Fairness is impossible :(

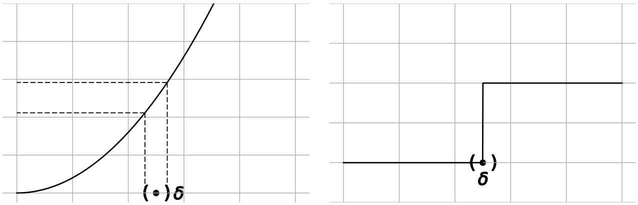
Theorem

Under WYSIWYG(ϵ), any nontrivial map $\hat{f} : \mathcal{O} \to \mathcal{D}$ is not $(\delta - \epsilon, \delta')$-fair for any $\delta, \delta' < 1$ if $\mathcal{D}$ is discrete

Fairness is impossible :(

Theorem

Under WYSIWYG(ϵ), any nontrivial map $\hat{f} : \mathcal{O} \to \mathcal{D}$ is not $(\delta - \epsilon, \delta')$-fair for any $\delta, \delta' < 1$ if $\mathcal{D}$ is discrete



- What about randomization?

Beyond individuals: structural bias

How might bias against certain groups manifest itself in this framework?

- Individuals belong to groups, partitioning the space $\mathcal{C} = X_1 \cup \ldots \cup X_k$, $\mathcal{O} = Y_1 \cup \ldots \cup Y_k$

$$\begin{tikzcd}
\mathcal{C} \arrow[r, "g"] \arrow[dr, "f"'] & \mathcal{O} \arrow[d, "\hat{f}"] \\
& \mathcal{D}
\end{tikzcd}$$

Beyond individuals: structural bias

How might bias against certain groups manifest itself in this framework?

- Individuals belong to groups, partitioning the space $\mathcal{C} = X_1 \cup \ldots \cup X_k$, $\mathcal{O} = Y_1 \cup \ldots \cup Y_k$

$$\begin{tikzcd} \mathcal{C} \arrow[r, "g"] \arrow[dr, "f"'] & \mathcal{O} \arrow[d, "\hat{f}"] \\ & \mathcal{D} \end{tikzcd}$$

Measure distance between subsets X, X' in the same space $\mathcal{X}$ with **Wasserstein distance**

$$\mathcal{W}_d(X, X') = \min_{\nu \in \mathcal{U}(X, X')} \int d_{\mathcal{X}}(x, x')\nu(x, x')$$

Beyond individuals: structural bias

How might bias against certain groups manifest itself in this framework?

- Individuals belong to groups, partitioning the space $\mathcal{C} = X_1 \cup \ldots \cup X_k$, $\mathcal{O} = Y_1 \cup \ldots \cup Y_k$

$$\begin{tikzcd} \mathcal{C} \arrow[r, "g"] \arrow[dr, "f"'] & \mathcal{O} \arrow[d, "\hat{f}"] \\ & \mathcal{D} \end{tikzcd}$$

Measure distance between subsets X, X' in the same space $\mathcal{X}$ with **Wasserstein distance**

$$\mathcal{W}_d(X, X') = \min_{\nu \in \mathcal{U}(X, X')} \int d_{\mathcal{X}}(x, x') \nu(x, x')$$

Measure distance between subsets X, Y in the different spaces with **Gromov-Wasserstein distance**

$$\mathcal{G}(X, Y) = \frac{1}{2} \inf_{\mu} \int |d_{\mathcal{X}}(x, x') - d_{\mathcal{Y}}(y, y')| d\mu_X \times d\mu_X d\mu_Y \times d\mu_Y$$

Structual bias and discrimination

The **between-groups** and **within-group distances** of $\mathcal{X} = \bigcup_{i=1}^{k} X_i$ and $\mathcal{Y} = \bigcup_{i=1}^{k} Y_i$ are respectively

$$\rho_b = \frac{1}{\binom{k}{2}} \mathcal{G}(\mathcal{X}, \mathcal{Y}), \quad \rho_w = \frac{1}{k} \sum_{i=1}^{k} \mathcal{G}(X_i, Y_i)\,,$$

and the **group skew** is

$$\sigma(\mathcal{X}, \mathcal{Y}) = \frac{\rho_b(\mathcal{X}, \mathcal{Y})}{\rho_w(\mathcal{X}, \mathcal{Y})}$$

Structual bias and discrimination

The **between-groups** and **within-group distances** of $\mathcal{X} = \bigcup_{i=1}^{k} X_i$ and $\mathcal{Y} = \bigcup_{i=1}^{k} Y_i$ are respectively

$$\rho_b = \frac{1}{\binom{k}{2}} \mathcal{G}(\mathcal{X}, \mathcal{Y}), \quad \rho_w = \frac{1}{k} \sum_{i=1}^{k} \mathcal{G}(X_i, Y_i) \,,$$

and the **group skew** is

$$\sigma(\mathcal{X}, \mathcal{Y}) = \frac{\rho_b(\mathcal{X}, \mathcal{Y})}{\rho_w(\mathcal{X}, \mathcal{Y})}$$

We have

- t-**structural bias**: $\sigma(\mathcal{C}, \mathcal{O}) > t$
- t-**direct discrimination**: $\sigma(\mathcal{O}, \mathcal{D}) > t$
- a t-**nondiscriminatory** mapping $f : \mathcal{C} \to \mathcal{D}$ if $\sigma(\mathcal{C}, \mathcal{D}) \leq t$

Structural bias and discrimination

How does this notion of structural bias compare with an intuitive one?

What does direct discrimination look like? Can affirmative action be direct discrimination?

We're all equal

If structural bias is suspected, WYSIWYG doesn't hold. How can we get around our lack of knowledge about the construct space?

$$\begin{tikzcd} \mathcal{C} \arrow[r, "g"] \arrow[dr, "f"'] & \mathcal{O} \arrow[d, "\hat{f}"] \\ & \mathcal{D} \end{tikzcd}$$

We're all equal

If structural bias is suspected, WYSIWYG doesn't hold. How can we get around our lack of knowledge about the construct space?

$$\begin{tikzcd} \mathcal{C} \arrow[r, "g"] \arrow[dr, "f"'] & \mathcal{O} \arrow[d, "\hat{f}"] \\ & \mathcal{D} \end{tikzcd}$$

> **Axiom** (WAE) For $\mathcal{C} = X_1 \cup \ldots \cup X_k$,
>
> $$\mathcal{W}_{d_\mathcal{C}}(X_i, X_j) < \epsilon \text{ for all } 1 \leq i, j \leq k$$

We're all equal

If structural bias is suspected, WYSIWYG doesn't hold. How can we get around our lack of knowledge about the construct space?

$$\begin{tikzcd} \mathcal{C} \arrow[r, "g"] \arrow[dr, "f"'] & \mathcal{O} \arrow[d, "\hat{f}"] \\ & \mathcal{D} \end{tikzcd}$$

> **Axiom** (WAE) For $\mathcal{C} = X_1 \cup \ldots \cup X_k$,
>
> $$\mathcal{W}_{d_\mathcal{C}}(X_i, X_j) < \epsilon \text{ for all } 1 \leq i, j \leq k$$

A **group fairness mechanism** (GFM_ϵ) $f : \mathcal{O} \to \mathcal{D}$ with $\mathcal{O} = Y_1 \cup \ldots \cup Y_k$ satisfies $\mathcal{W}_{d_\mathcal{O}}(f(Y_i), f(Y_j)) \leq \epsilon$

Nondiscrimination is possible!

Theorem

Under WAE, a GFM$_{\epsilon'}$ guarantees $\frac{\max(\epsilon,\epsilon')}{\delta}$-nondiscrimination

$$\begin{tikzcd} \mathcal{C} \arrow[r, "g"] \arrow[dr, "f"'] & \mathcal{O} \arrow[d, "\hat{f}"] \\ & \mathcal{D} \end{tikzcd}$$

Nondiscrimination is possible!

Theorem

Under WAE, a $GFM_{\epsilon'}$ guarantees $\frac{\max(\epsilon,\epsilon')}{\delta}$-nondiscrimination

$$\begin{tikzcd} \mathcal{C} \arrow[r, "g"] \arrow[dr, "f"'] & \mathcal{O} \arrow[d, "\hat{f}"] \\ & \mathcal{D} \end{tikzcd}$$

Nondiscrimination is possible!

Theorem

Under WAE, a $GFM_{\epsilon'}$ guarantees $\frac{\max(\epsilon, \epsilon')}{\delta}$-nondiscrimination

$$\begin{tikzcd} \mathcal{C} \arrow[r, "g"] \arrow[dr, "f"'] & \mathcal{O} \arrow[d, "\hat{f}"] \\ & \mathcal{D} \end{tikzcd}$$

Could achieving nondiscrimination in this setting require direct discrimination?

Worldview comparison

WYSIWYG vs. WAE

Each axiom induces fairness mechanism (group v. individual) to achieve fairness or nondiscrimination. Are they always incompatible?

Worldview comparison

WYSIWYG vs. WAE

Each axiom induces fairness mechanism (group v. individual) to achieve fairness or nondiscrimination. Are they always incompatible?

How can we understand the following as part of this framework?

- Observational measures:
  - demographic parity (equalized odds)
  - accuracy parity
  - true positive parity (equal opportunity)
  - predictive value parity
- Credit scores?
- COMPAS: Kristian Lum's approach

Beyond conceptual design?

$$\begin{tikzcd} \mathcal{C} \arrow[r, "g"] \arrow[dr, "f"'] & \mathcal{O} \arrow[d, "\hat{f}"] \\ & \mathcal{D} \end{tikzcd}$$

Assume good observations

$$\begin{tikzcd} \mathcal{C} \arrow[r, "g"] \arrow[dr, "f"'] & \mathcal{O} \arrow[d, "\hat{f}"] \\ & \mathcal{D} \end{tikzcd}$$

Assume inherent equality

Framework allows for conceptual exploration and justification of a type of fairness mechanism.

- How do we model these spaces? How to explicitly encode structural bias at the modeling level?
- More sophisticated axioms?