Re-formalization of Individual Fairness

Toshihiro Kamishima
National Institute of Advanced Industrial Science and Technology (AIST), Japan

The 6th FAccTRec Workshop: Responsible Recommendation
in conjunction with RecSys2023 @ Singapore, Sep. 18, 2023

START

Outline

Re-formalization of Individual Fairness

Individual Fairness: the principle of “Treating Like Cases Alike”

Mapping similar individuals in an original space into similar positions in a fair space

↓

Conditioning fairness criterion by individuals

Outline

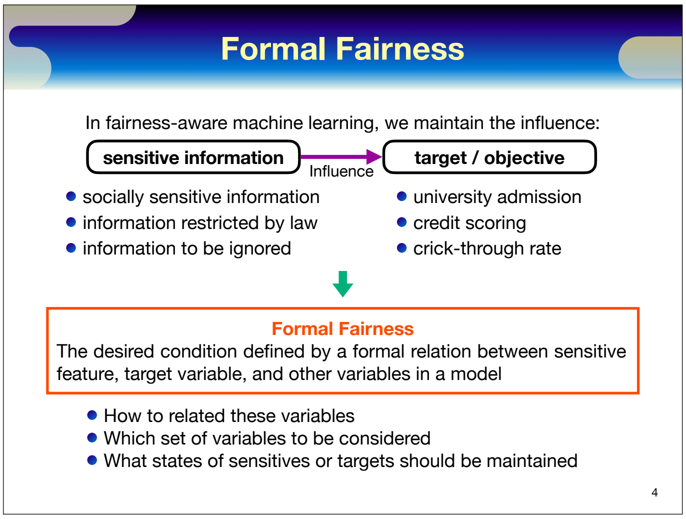
- Brief summary of formal fairness
- Our re-formalized individual fairness is compatible with that of Dwork et al.
- Extend equalized odds and sufficiency by applying our new re-formalized individual fairness

Individual fairness is the principle of "Treating Like Cases Alike", has been argued by Aristotle.
Dwork et al. formalized this principle as mapping similar individuals in an original space into similar positions in a fair subspace.
I this talk, we re-formalize this as conditioning fairness criterion by individuals.

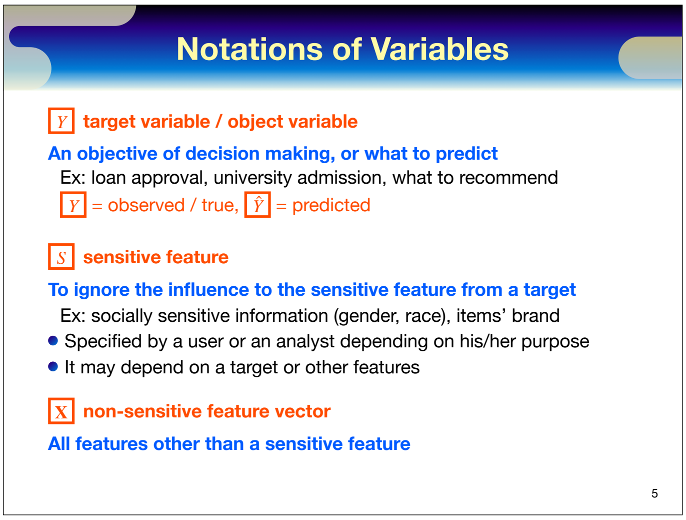
After showing brief summary of formal fairness, our re-formalized individual fairness is compatible with that of Dwork et al.
Then, we extend equalized odds and sufficiency

Formal Fairness

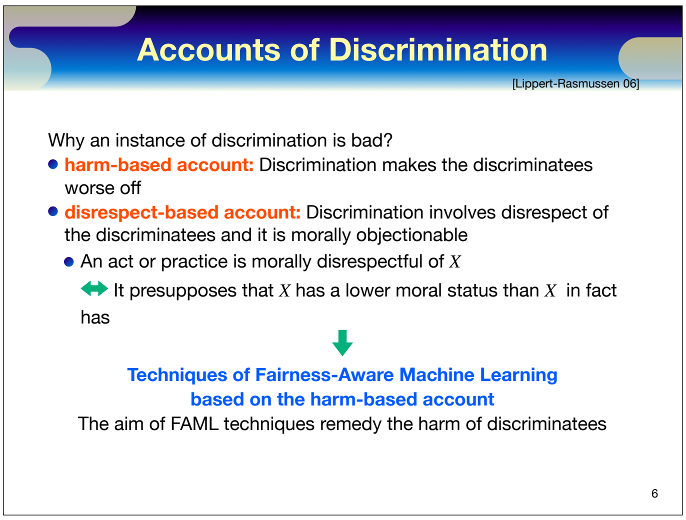
We begin with a brief summary of formal fairness



In fairness-aware machine learning, we maintain the influence of sensitive information to an objective.
For this purpose, we have to satisfy formal fairness, which is the desired condition defined by a formal relation between sensitive feature, target variable, and other variables in a model.



We define some notations.
An objective variable Y represents an objective of decision making.
Vanilla Y indicates an observed label, and $\hat{Y}$ indicates a predicted label.
A sensitive feature S represents socially sensitive information to ignore.
All features other than a sensitive feature consist of non-sensitive feature vector, X.



There are two major accounts why an instance of discrimination is bad.
In a harm-based account, discrimination makes the discriminatees worse off.
In a disrespect-based account, discrimination involves disrespect of the discriminatees and it is morally objectionable.
A harm-based accounts relates to the Mill's utilitarianism, and a disrespect-based account relates to Kantian deontology.
Techniques of Fairness-Aware Machine Learning based on the harm-based account.

Judgements Related to Formal Fairness

[Ishiguro+ 14, Bareinboim+ 21, Pearl+ 18]

Hazelwood School District v. United States, 433 U.S. 299 (1977)

- Where **gross statistical disparities** can be shown, they alone may, in a proper case, constitute *prima facie* proof

Gross Statistical Disparity: Discrimination in employment is determined whether the ratio of protected and non-protected groups of employees is diverged from the corresponding ratio in general population

Jack Gross, Petitioner, v. FBL Financial Services, US Supreme Court (2008)

- To establish a disparate-treatment claim under this plain language, a plaintiff must prove that age was **the but-for cause** of the employer's adverse decision
- A plaintiff must prove by a preponderance of the evidence (which may be direct or circumstantial), that age was **the but-for cause** of the challenged employer decision

7

This is mainly due to these judgements.

Baselines in Harm-based Account

[Lippert-Rasmussen 06]

A harm-based account requests a baseline for determining whether the discriminatees have been made worse off

⬇

- **Ideal outcome:** the discriminatees are in just, or the morally best
  ➡ **association-based fairness:** letting predictors get ideal outcomes
- **Counterfactual:** the discriminatees had not been subjected to the discrimination
  ➡ **counterfactual fairness:** comparing with the counterfactuals that a status of a sensitive feature was different

8

Further, a harm-based account requests a baseline for determining whether the discriminatees have been made worse off. Association-based fairness uses an ideal outcome as a baseline, and counterfactual fairness uses counterfactuals.

Individual Fairness

9

Next, we show our individual fairness and its compatibility.

Individual Fairness

Individual Fairness: the principle of “Treating Like Cases Alike”

We re-formalize individual fairness as conditioning a fairness criterion by $\mathbf{X}$

$$\underset{\textbf{statistical parity}}{\hat{Y} \perp\!\!\!\perp S} \xrightarrow{\text{conditioned by } \mathbf{X}} \underset{\textbf{individual fairness}}{\hat{Y} \perp\!\!\!\perp S \mid \mathbf{X}}$$

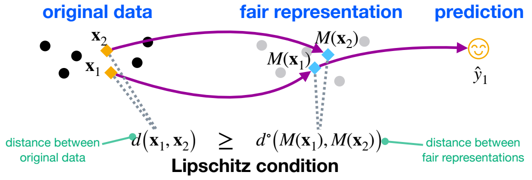
(individual statistical parity)

1. This formulation is compatible with the one proposed by Dwork et al.
2. This newly formalized criterion can be used for in- or post- process methods as well as pre-process methods of fairness
3. This formalization can be applied to fairness criteria, equalized odds or sufficiency

Individual Fairness is the principle of “Treating Like Cases Alike.”
We re-formalize individual fairness as conditioning a fairness criterion by X.
For example, statistical parity, independence between Y hat and S, is conditioned by X, and we got conditional independence between Y hat and S given X.
This formulation is compatible with the one proposed by Dwork.

Dwork’s Individual Fairness

[Dwork+ 12]



$$d(\mathbf{x}_1, \mathbf{x}_2) \geq d'(M(\mathbf{x}_1), M(\mathbf{x}_2))$$

Lipschitz condition

To formalize the principle “Treating Like Cases Alike,”

1. Similar original data are mapped to similar fair representations
2. Predictors make similar predictions for similar representations

No sensitive information in fair representations

⬇

The predictions satisfy a Fairness through Unawareness condition

We explain the Dwork's original formalization.
First, similar original data are mapped to similar fair representations, and predictors make similar predictions for similar representations.
In this formulation, there is no sensitive information in fair representations.
This implys that the predictions satisfy a "Fairness through Unawareness" condition.

Fairness through Unawareness

Fairness through Unawareness: Prohibiting to access individuals' sensitive information during the process of learning and inference
This is a kind of procedural fairness, in which a decision is fair, if it is made by following pre-specified procedure

$\Pr[\ \hat{Y} \mid \mathbf{X}, S\ ]$		$\Pr[\ \hat{Y} \mid \mathbf{X}\ ]$
A **unfair model** is trained from a dataset including sensitive and non-sensitive information	➡	A **fair model** is trained from a dataset eliminating sensitive information

A unfair model, $\Pr[\ \hat{Y} \mid \mathbf{X}, S]$, is replaced with a fair model, $\Pr[\ \hat{Y} \mid \mathbf{X}\ ]$

$$\Pr[\ \hat{Y}, \mathbf{X}, S\ ] = \Pr[\ \hat{Y} \mid \mathbf{X}, S]\ \Pr[\ S \mid \mathbf{X}\ ]\ \Pr[\ \mathbf{X}\ ] \Rightarrow \Pr[\ \hat{Y} \mid \mathbf{X}\ ]\ \Pr[\ S \mid \mathbf{X}\ ]\ \Pr[\ \mathbf{X}\ ]$$

⬇

Fairness through Unawareness: $\hat{Y} \perp\!\!\!\perp S \mid \mathbf{X}$

Fairness through Unawareness is prohibiting to access individuals' sensitive information during the process of learning and inference.
This means that a predictive model of Y hat given X and S is replaced with the model Y hat given X.
This leads the conditional independence between Y hat and S given X.

Re-formalization of Individual Fairness

Distributions of a target variable are equal for all possible sensitive groups given a specific non-sensitive values

$\Pr[\hat{Y} \mid S, \mathbf{X}=\mathbf{x}] = \Pr[\hat{Y} \mid \mathbf{X}=\mathbf{x}], \forall \mathbf{x} \in \mathrm{Dom}(X) \Rightarrow \hat{Y} \perp\!\!\!\perp S \mid \mathbf{X}$

=

We re-formalize Individual fairness as conditioning fairness criteria by $\mathbf{X}$

⬇

This formula, $\hat{Y} \perp\!\!\!\perp S \mid \mathbf{X}$, is coincident with Fairness through Unawareness

Dwork's Individual Fairness ↔ Fairness through Unawareness ↔ Our re-formalized individual fairness

Our re-formalized individual fairness is compatible with Dwork's

13

Fortunately, our re-formalized individual fairness is coincident with Fairness through Unawareness. From the fact that Dwork's individual fairness is also compatible with fairness through unawareness, our formulation is compatible with that of Dwork et al.

Extended Individual Fairness

14

We than apply our formalization to equalized odds and sufficiency.

Equalized Odds and Sufficiency

Fairness in errors of predictions to mitigate an inductive bias

Equalized Odds	Sufficiency
$\hat{Y} \perp\!\!\!\perp S \mid Y$	$Y \perp\!\!\!\perp S \mid \hat{Y}$
Matching false positive ratio (FPR) and true positive ratio (TPR), if Y is binary	Matching positive and negative predictive values (PPV & NPV), if Y is binary

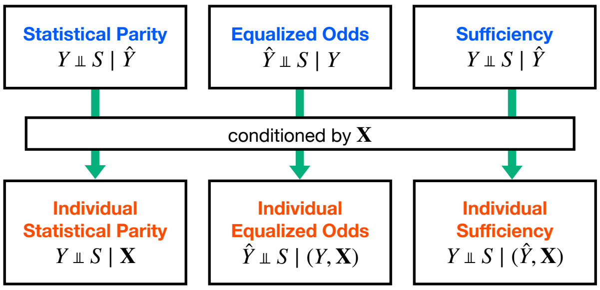
- The ProPublica pointed out the recidivism score, the COMPAS, does not satisfy equalized odds [Angwin+ 2016]
- The US Court refuted that the score is designed to satisfy a sufficiency condition [Flores+ 2016]

15

There are two types of fairness criteria in errors of predictions to mitigate an inductive bias: equalized odds and sufficiency.
The ProPublica pointed out the recidivism score, the COMPAS, does not satisfy equalized odds.
The US Court refuted that the score is designed to satisfy a sufficiency condition.

Extended Individual Fairness

Conditioning by $\mathbf{X}$ can convert Equalized Odds and Sufficiency to individual versions of them

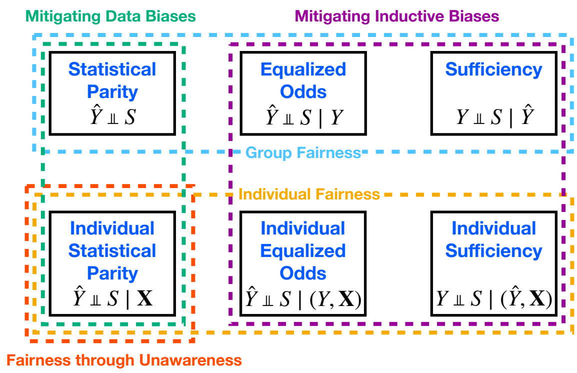


The phrase, “treating alike,“ means predicting in similar error rate

16

Conditioning by X can convert Equalized Odds and Sufficiency to individual versions of them: individual equalized odds and individual sufficiency.
In these criteria, the phrase, “treating alike,“ means predicting in similar error rate.

Summary of Formal Association-based Fairness



17

By adding our new criteria, formal association-based fairness can be summarized as this figure.
In my humble opinion, fairness criteria are well-organized.

Conclusion

Conclusion

- We re-formalize the notion of individual fairness by conditioning by $\mathbf{X}$
  - Compatible with that of Dwork et al.
  - Equalized odds or sufficiency can be extensible to their corresponding individual versions
  - Our individual fairness can be used in in in-process or post-process approaches as well as pre-process approaches

Future work

- One of the limitation is an interpretation of the term, *like*
  - if non-sensitive features take exactly the same values, two assumptive individuals are considered as *like*
  - To relax the limitation, the introduction of similarities between individuals would be required

My FAML tutorial slide: https://www.kamishima.net/archive/faml.pdf

18