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
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
Formal Fairness
In fairness-aware machine learning, we maintain the influence:

- socially sensitive information
- information restricted by law
- information to be ignored

- university admission
- credit scoring
- crick-through rate

**Formal Fairness**

The desired condition defined by a formal relation between sensitive feature, target variable, and other variables in a model

- How to related these variables
- Which set of variables to be considered
- What states of sensitives or targets should be maintained
Notations of Variables

- $Y$: target variable / object variable

  An objective of decision making, or what to predict
  - Ex: loan approval, university admission, what to recommend
  - $Y = \text{observed / true}$, $\hat{Y} = \text{predicted}$

- $S$: sensitive feature

  To ignore the influence to the sensitive feature from a target
  - Ex: socially sensitive information (gender, race), items’ brand
  - Specified by a user or an analyst depending on his/her purpose
  - It may depend on a target or other features

- $X$: non-sensitive feature vector

  All features other than a sensitive feature
Accounts of Discrimination

Why an instance of discrimination is bad?

- **harm-based account:** Discrimination makes the discriminatees worse off
- **disrespect-based account:** Discrimination involves disrespect of the discriminatees and it is morally objectionable
  - An act or practice is morally disrespectful of $X$
    - $\iff$ It presupposes that $X$ has a lower moral status than $X$ in fact has

Techniques of Fairness-Aware Machine Learning based on the harm-based account

The aim of FAML techniques remedy the harm of discriminatees

- 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


- 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
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

[Lippert-Rasmussen 06]
Individual Fairness
Individual Fairness: the principle of “Treating Like Cases Alike”

We re-formalize individual fairness as conditioning a fairness criterion by $X$

\[\hat{Y} \perp S \quad \text{conditioned by } X \quad \hat{Y} \perp S \mid X\]

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
Dwork’s Individual Fairness

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
**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.

- **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, $\Pr[ \hat{Y} \mid X, S ]$, is replaced with a fair model, $\Pr[ \hat{Y} \mid X ]$.

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

**Fairness through Unawareness**: $\hat{Y} \perp \text{S} \mid X$
Distributions of a target variable are equal for all possible sensitive groups given a specific non-sensitive values:

\[
\Pr[\hat{Y} \mid S, X=x] = \Pr[\hat{Y} \mid X=x], \quad \forall x \in \text{Dom}(X) \Rightarrow \hat{Y} \perp S \mid X
\]

We re-formalize Individual fairness as conditioning fairness criteria by \(X\):

This formula, \(\hat{Y} \perp S \mid X\), is coincident with Fairness through Unawareness.

Our re-formalized individual fairness is compatible with Dwork's Individual Fairness.
Extended Individual Fairness
Equalized Odds and Sufficiency

Fairness in errors of predictions to mitigate an inductive bias

Equalized Odds

\[ \hat{Y} \perp S \mid Y \]

Matching false positive ratio (FPR) and true positive ratio (TPR), if \( Y \) is binary

Sufficiency

\[ Y \perp S \mid \hat{Y} \]

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]
Extended Individual Fairness

Conditioning by $X$ can convert Equalized Odds and Sufficiency to individual versions of them.

- **Statistical Parity**
  \[ Y \perp S \mid \hat{Y} \]

- **Equalized Odds**
  \[ \hat{Y} \perp S \mid Y \]

- **Sufficiency**
  \[ Y \perp S \mid \hat{Y} \]

 conditioned by $X$

- **Individual Statistical Parity**
  \[ Y \perp S \mid X \]

- **Individual Equalized Odds**
  \[ \hat{Y} \perp S \mid (Y, X) \]

- **Individual Sufficiency**
  \[ Y \perp S \mid (\hat{Y}, X) \]

The phrase, “treating alike,” means predicting in similar error rate
Summary of Formal Association-based Fairness

### Mitigating Data Biases

<table>
<thead>
<tr>
<th>Statistical Parity</th>
<th>Equalized Odds</th>
<th>Sufficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{Y} \perp S$</td>
<td>$\hat{Y} \perp S \mid Y$</td>
<td>$Y \perp S \mid \hat{Y}$</td>
</tr>
</tbody>
</table>

### Mitigating Inductive Biases

<table>
<thead>
<tr>
<th>Group Fairness</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{Y} \perp S \mid (Y, X)$</td>
</tr>
</tbody>
</table>

### Individual Fairness

<table>
<thead>
<tr>
<th>Individual Statistical Parity</th>
<th>Individual Equalized Odds</th>
<th>Individual Sufficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{Y} \perp S \mid X$</td>
<td>$\hat{Y} \perp S \mid (Y, X)$</td>
<td>$Y \perp S \mid (\hat{Y}, X)$</td>
</tr>
</tbody>
</table>

Fairness through Unawareness
Conclusion

- We re-formalize the notion of individual fairness by conditioning by 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-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