形式的公平性規準間の不可能性に関する考察

神嶌 敏弘 (https://www.kamishima.net)
産業技術総合研究所

2023年度人工知能学会全国大会（第37回）@ 熊本城ホール，2023-06-08

Motivation and Outline

Motivation

- Many fairness conditions for ML models are proposed
- We enumerate mathematically-possible conditions, and investigate the impossibility between these conditions
- **Outline**
  - bad discrimination
  - preliminary: formal fairness, independence, Markov network
  - enumeration
    - 2 variables: statistical parity
    - 3 variables, S, $\hat{Y}$, and Y: equalized odds, sufficiency
    - 3 variables, S, $\hat{Y}$, and $\mathbf{X}$: individual fairness, converse individual fairness
    - 4 variables
  - summary

Accounts of Discrimination

[Lippert-Rasmussen 2006]

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

    ⟷ 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

Baselines in Harm-based Account

[Lippert-Rasmussen 2006]

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

Formal Fairness

In fairness-aware data mining, 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 = observed / true, $\hat{Y}$ = predicted, Y° = fairized

- $Y=1$ advantageous decision / $Y=0$ disadvantageous decision

S **sensitive feature**

To ignore the influence to the sensitive feature from a target

Ex: socially sensitive information (gender, race), items' brand

- $S=1$ non-protected group / $S=0$ disadvantageous decision
- Specified by a user or an analyst depending on his/her purpose
- It may depend on a target or other features

$\mathbf{X}$ **non-sensitive feature vector**

All features other than a sensitive feature

Independence

(unconditional) independence

A pair sets of variables, $\mathbf{Y}$ and $\mathbf{S}$, are not influenced from each other

$$\mathbf{Y} \perp\!\!\!\perp \mathbf{S}$$

conditional independence

$\mathbf{Y}$ and $\mathbf{S}$ are independent, if conditional variables, $\mathbf{X}$, are fixed

$$\mathbf{Y} \perp\!\!\!\perp \mathbf{S} \mid \mathbf{X}$$

* **Conditional independence doesn't imply independence, and vice versa**

context-specific independence

$\mathbf{Y}$ and $\mathbf{S}$ are independent, if $\mathbf{X}$ are fixed to specific values, $\mathbf{x}$ [Boutiller+ 96]

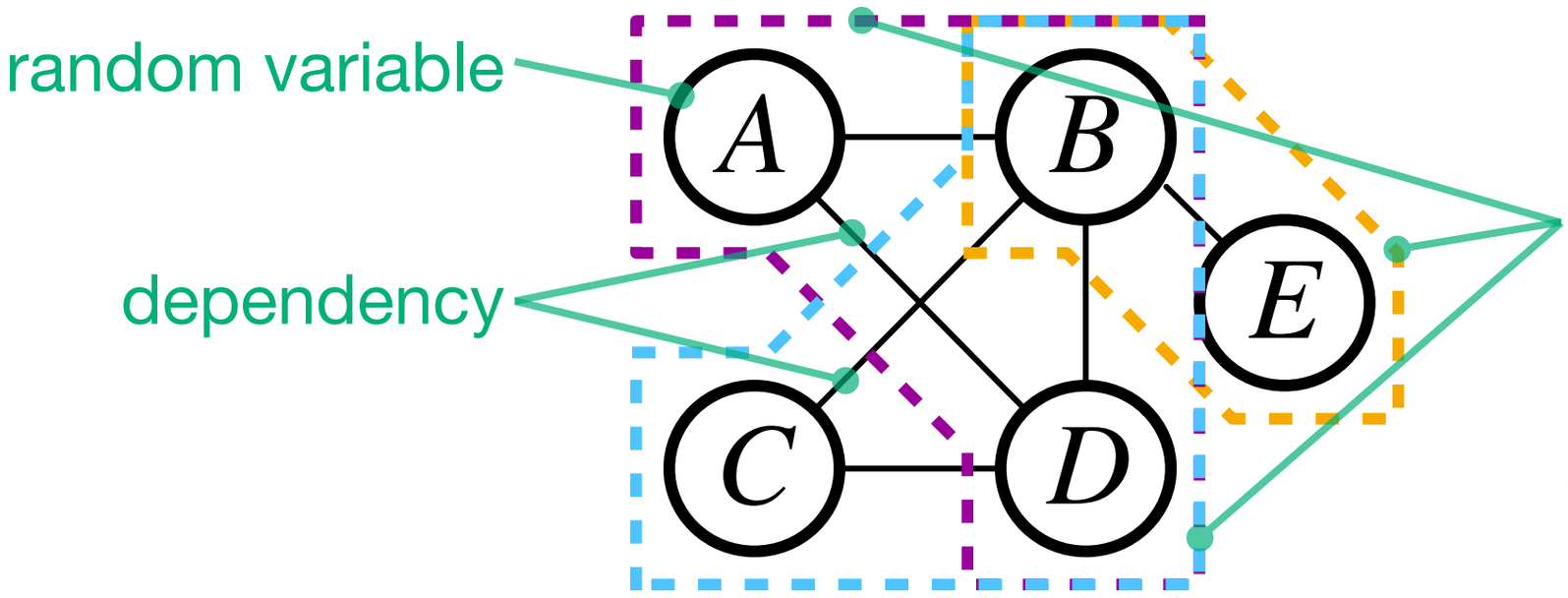
$$\mathbf{Y} \perp\!\!\!\perp \mathbf{S} \mid \mathbf{X}=\mathbf{x}$$

* Notation with a symbol '⫫' (Unicode 2AEB) is called Dawid's notation

Markov Network

[Bishop 2006]

Markov network: undirected graphical model for probabilistic distribution



potential function

Each corresponds to one clique

standardized constant or partition function

$$\Pr[A, B, C, D, E] = f(A, B, D) f(B, C, D) f(B, E) / Z$$

Variables, A and C, are separated by removing B and D

conditional independence: $A \perp\!\!\!\perp C \mid B, D$

Possible Models

Formal fairness is a model of the influence between S and $\hat{Y}$

⬇

At least, two variables, S and $\hat{Y}$, must be included into the model

To enumerate possible models

- Two variables, S and $\hat{Y}$ ➡ 2 cases
- Three variables, S, $\hat{Y}$, and $\mathbf{X}$ ➡ 6 cases
- Three variables, S, $\hat{Y}$, and Y ➡ 6 cases
- Four variables, S, $\hat{Y}$, Y, and $\mathbf{X}$ ➡ 62 cases

⬇

From these cases, models that is considered as unfair or unrealistic are removed

Conditioning

Conditioning by S is unfair

The information contained in $\mathbf{X}$ is pass to $\hat{Y}$ through only S

A fairness model conditioned by S, such as $\mathbf{X} \perp\!\!\!\perp \hat{Y} \mid S$, is unfair

Conditioning by $\mathbf{X}$ makes fairness individualize

Conditioning by $\mathbf{X}$

Fairness is judged individually

Individual Fairness = Treating like cases alike

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

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

Unrealistic Independence

Independence considered as unrealistic due to the following reasons

$$\hat{Y} \perp\!\!\!\perp Y$$

- Observed decisions are completely ignored when making predictions

$$\hat{Y} \perp\!\!\!\perp \mathbf{X} \text{ or } Y \perp\!\!\!\perp \mathbf{X}$$

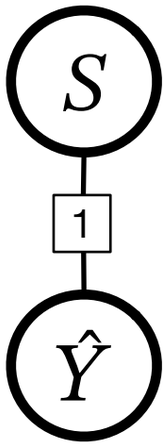
- Features of instances are completely ignored when making predictions or decisions

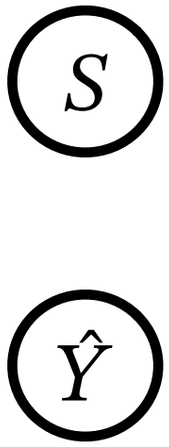
$$S \perp\!\!\!\perp \mathbf{X}$$

- Both sensitive and non-sensitive information are uncontrolable, and the probability that the independence is satisfied is almost zero

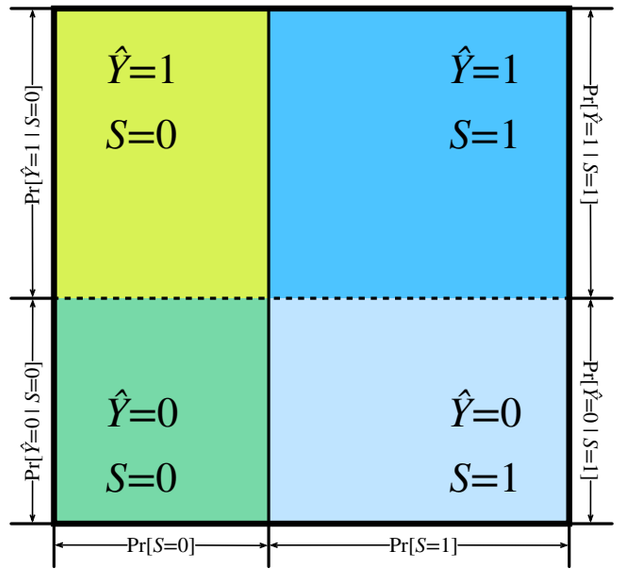
Two Variables

2 cases of two variables, S and $\hat{Y}$





$S \perp\!\!\!\perp \hat{Y}$



direct discrimination

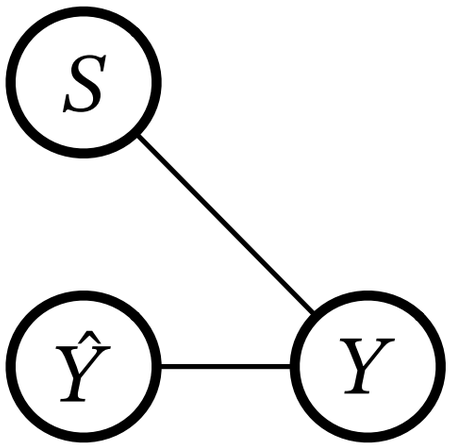
- Predictions directly depends on a sensitive feature, which is a unfair operation

statistical parity / demographic parity

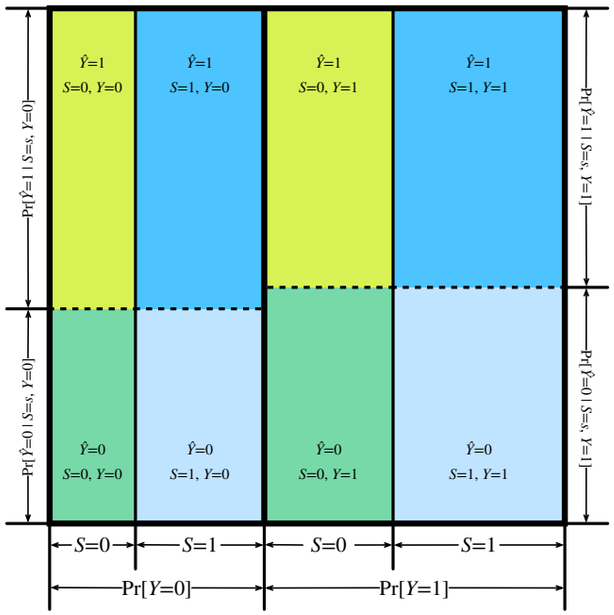
- Predictions are made so that the ratio of positives are the same between sensitive groups groups
- This condition corresponds to the ethical notion, distributive justice

Three Variables: S, $\hat{Y}$, and Y

Only known two cases

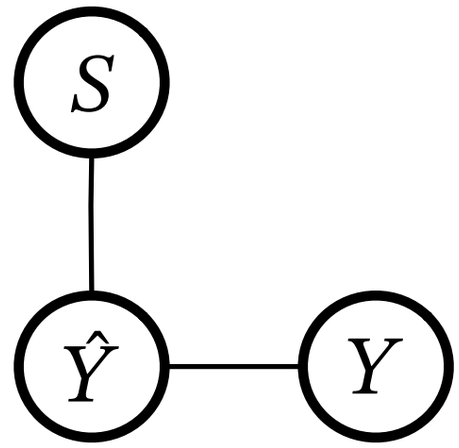


$S \perp\!\!\!\perp \hat{Y} \mid Y$

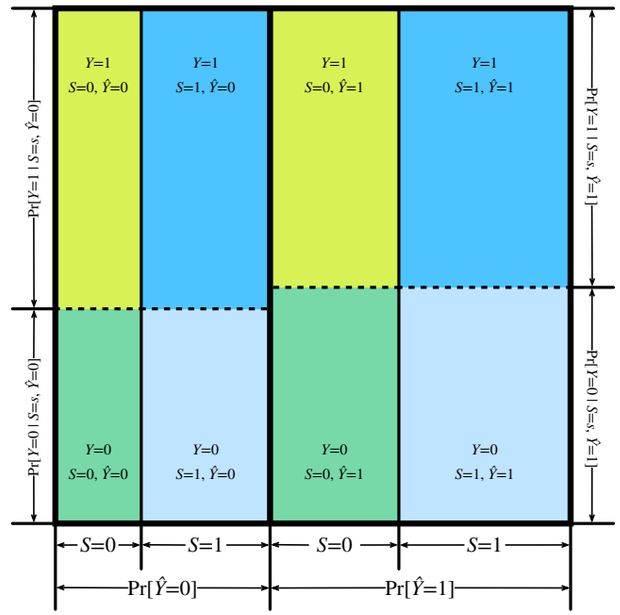


equalized odds / separation

- Prediction errors, FPR and FNR, are the same between sensitive groups
- ProPublica claimed that the COMPAS score doesn't satisfy the condition



$S \perp\!\!\!\perp Y \mid \hat{Y}$

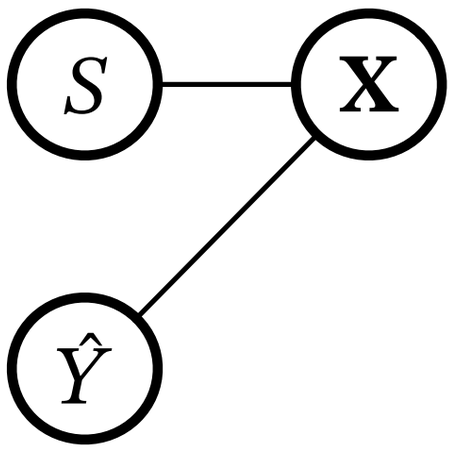


separation

- The predictions are equally likely tue between sensitive groups
- The US court countered that the COMPAS score satisfy the condition

Three Variables: S, $\hat{Y}$, and $\mathbf{X}$

a known case, individual fairness, and new one

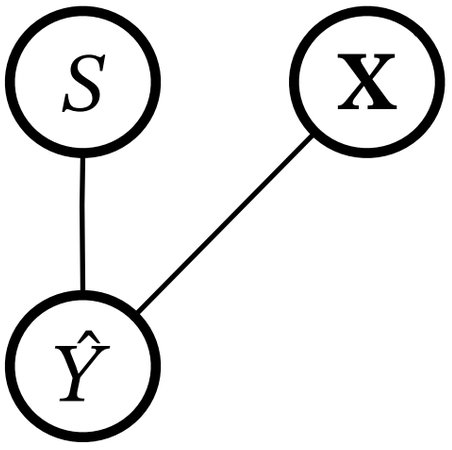


individual fairness = fairness through unawareness

Treat like cases alike

$$\Pr[\hat{Y} \mid S = 0, \mathbf{X} = \mathbf{x}] = \Pr[\hat{Y} \mid S = 1, \mathbf{X} = \mathbf{x}], \forall \mathbf{x} \in \mathrm{Dom}(\mathbf{X})$$

$$\hat{Y} \perp\!\!\!\perp S \mid \mathbf{X}$$



a new case

$$\mathbf{X} \perp\!\!\!\perp S \mid \hat{Y}$$

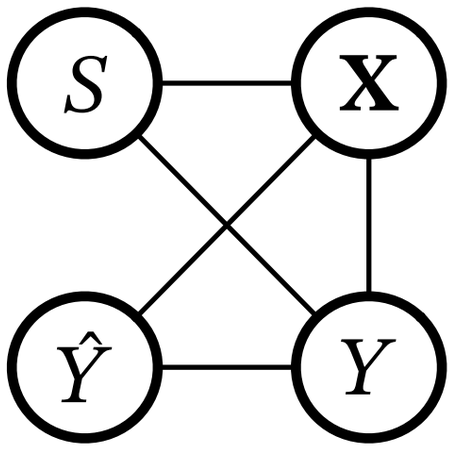
$$\Pr[\mathbf{X} \mid S = 0, \hat{Y} = \hat{y}] = \Pr[\mathbf{X} \mid S = 1, \hat{Y} = \hat{y}], \forall \hat{y} \in \mathrm{Dom}(Y)$$

Like treatment implies like cases

Considered as the **converse of individual fairness**

- In a practical use, it is hard to satisfy the independence from a high-dimensional variable, $\mathbf{X}$

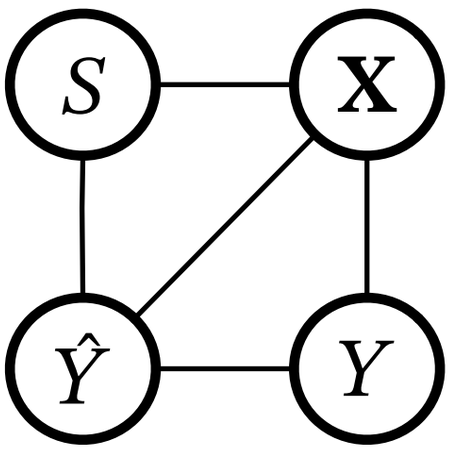
Four Variables (1)



$$\hat{Y} \perp\!\!\!\perp S \mid (\mathbf{X}, Y)$$

equalized odds + individual fairness

- If the condition of equalized odds is satisfied, that of individual fairness is satisfied in general

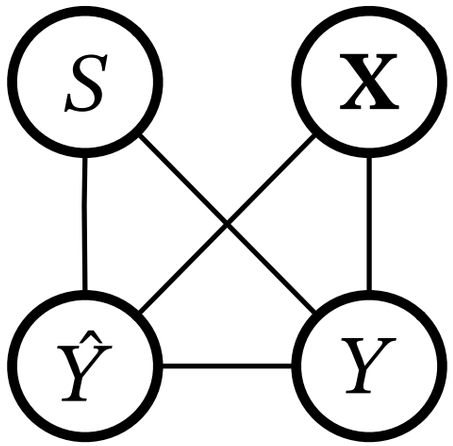


$$Y \perp\!\!\!\perp S \mid (\mathbf{X}, \hat{Y})$$

sufficiency + individual fairness

- If the condition of sufficiency is satisfied, that of individual fairness is satisfied in general

Four Variables (2)



$$\mathbf{X} \perp\!\!\!\perp S \mid (\hat{Y}, Y)$$

converse individual fairness conditioned by Y and $\hat{Y}$

- If both predicted and observed decision is alike, the cases are also alike
- This condition is more complicated than the converse individual fairness, and its utility in a real world would be low

Summary of Fairness Conditions

	individual fairness $S \perp\!\!\!\perp \hat{Y} \mid \mathbf{X}$	**converse individual fairnes** $S \perp\!\!\!\perp \mathbf{X} \mid \hat{Y}$	**equalized odds** $S \perp\!\!\!\perp \hat{Y} \mid Y$	**sufficiency** $S \perp\!\!\!\perp Y \mid \hat{Y}$	**statistical parity** $S \perp\!\!\!\perp \hat{Y}$
unit	individual	group	group (individual)	group (individual)	group
awareness	unaware	aware	aware	aware	aware
worldview	WAE	WAE	WYSIWYG	WYSIWYG	WAE

Worldview and Bias

[Friedler+ 21]

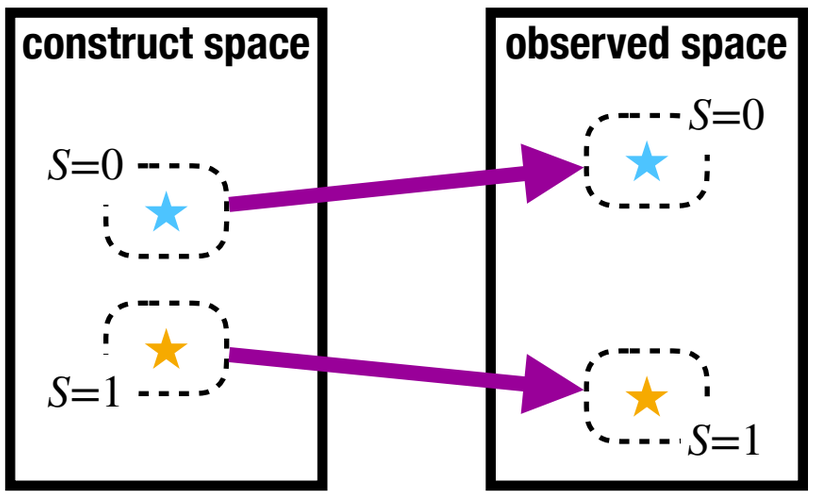
Worldview is an assumption about mapping from construct space to observed space

- **construct space**: underlying ideal features and decisions
- **observed space**: observed features and decisions

Structural bias Worldview

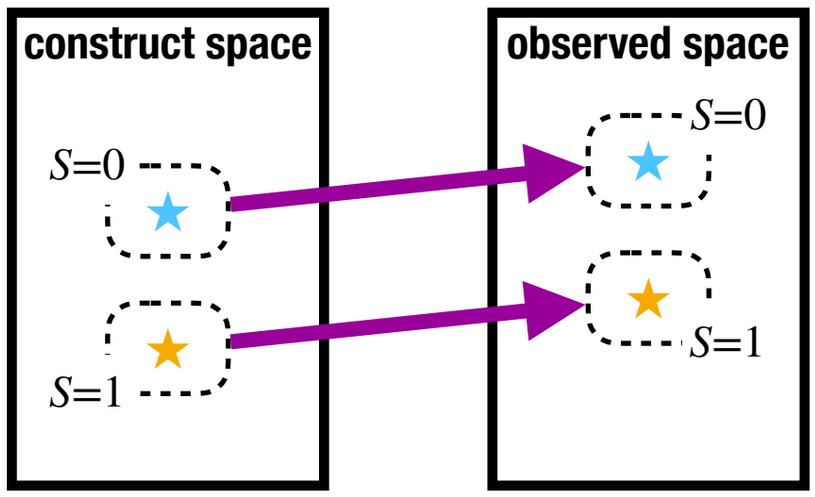
Instances in different groups are mapped differently

data bias



What You See Is What You Get Worldview

Mapping while keeping relative positions between groups



Conclusion

Conclusion

- We enumerated all mathematically-possible fairness conditions.
- While almost valid conditions are existing, we discovered some new conditions
  - converse individual fairness
  - individual equalized odds, individual sufficiency
  - converse individual fairness conditioned by Y and $\hat{Y}$

Future work

- We will further dig into the newly-found fairness conditions
  - Utilities of new conditions
  - Algorithms satisfying new conditions