




UNFAIR

A hand is shown drawing the word 'UNFAIR' in blue chalk on a dark chalkboard. The 'UN' part of the word is crossed out with several white diagonal lines. A horizontal white line is drawn below the word, and a hand is holding a piece of white chalk, having just finished drawing the line.

ALGORITHMIC FAIRNESS IN MACHINE LEARNING

DR VIKTORIIA SHARMANSKA

Algorithmic fairness in machine learning

- ▶ **Fairness definitions** (20 Feb, 4pm-6pm)
- ▶ Fairness methods (26 Feb, 11am-1pm) *Extra!* 
- ▶ Practical session (27 Feb, 4pm-6pm, Huxley 225)
- ▶ Advancements in algorithmic fairness (5 Mar, 4pm-6pm)

OUTLINE TODAY

- ▶ Intro into algorithmic bias
- ▶ Fairness definitions
 - ▶ Group-based
 - ▶ Individual-based
- ▶ Running examples

**WHAT IS MACHINE LEARNING?
WHY ML COULD BE UNFAIR?**

MACHINE LEARNING

A study of computer programmes that improve their **performance** at some **task** with **experience** (data).

MACHINE LEARNING

A study of computer programmes that improve their **performance** at some **task** with **experience** (data).



Examples?

HUMAN-CENTRIC EXAMPLES

 REUTERS

BusinessMarketsWorldPoliticsTVMore

Amazon scraps secret AI recruiting tool that showed bias against women

TECHNOLOGY NEWS OCTOBER 10, 2018 / 4:12 AM / A YEAR AGO



SOCIAL CREDIT SYSTEM IN CHINA

Facebook Ads Can Still Discriminate Against Women and Older Workers, Despite a Civil Rights Settlement

New research and Facebook's own ad archive show that the company's new system to ensure diverse audiences for housing and employment ads has many of the same problems as its predecessor.

by Ava Kofman and Ariana Tobin, Dec. 13, 2019, 5 a.m. EST

ALGORITHM WATCH

Search...

NewsletterABOUT

story

At least 10 police forces use face recognition in the EU, AlgorithmWatch reveals

Read more >

HUMAN-CENTRIC MACHINE LEARNING

A study of computer programmes that improve their **performance** at some **task** with **experience** (data), **where automatic decisions are made about humans**.

HUMAN-CENTRIC MACHINE LEARNING

A study of computer programmes that improve their **performance** at some **task** with **experience** (data), **where automatic decisions are made about humans**.

PROBLEM



A risk of discrimination

HUMAN-CENTRIC MACHINE LEARNING

A study of computer programmes that improve their **performance** at some **task** with **experience** (data), **where automatic decisions are made about humans**.

PROBLEM

A risk of discrimination



Task

Data



Machine Learning
Model



Decision Maker

Fair Machine
Learning Model



Fair Decision Maker

e.g. $Accuracy_{male} = Accuracy_{female}$

QUESTIONS?

QUESTIONS?

IS SOFTWARE NOT NEUTRAL?
OR
WHY MACHINE LEARNING MODEL COULD BE UNFAIR?

EDUCATIONAL GAME

<https://www.survivalofthebestfit.com/>



Built by [Gabor Csapo](#), [Jihyun Kim](#), [Miha Klasinc](#), and [Alia ElKattan](#). Supported by the [Creative Media Award](#) from [Mozilla Foundation](#).

FAIRNESS DEFINITIONS

HOW TO DEFINE FAIRNESS?

A LEGAL PERSPECTIVE

HOW TO DEFINE FAIRNESS?

- ▶ Direct discrimination w.r.t. intent

If a decision making process is based on the subject's sensitive attribute

- ▶ Indirect discrimination w.r.t. consequences

If the outcomes disproportionately hurt (or benefit) people with certain sensitive attribute values

Sensitive attributes: gender, race, age, disability, religion etc.

HOW TO DEFINE FAIRNESS?

- ▶ Direct discrimination w.r.t. intent

Disparate
treatment in US

If a decision making process is based on the subject's sensitive attribute

- ▶ Indirect discrimination w.r.t. consequences

Disparate impact
in US

If the outcomes disproportionately hurt (or benefit) people with certain sensitive attribute values

Protected
characteristics

Sensitive attributes: gender, race, age, disability, religion etc.

HOW TO DEFINE FAIRNESS?

- ▶ Direct discrimination w.r.t. intent

If a decision making process is based on the subject's sensitive attribute

- ▶ Indirect discrimination w.r.t. consequences

If the outcomes disproportionately hurt (or benefit) people with certain sensitive attribute values

No consensus on the mathematical formulations of fairness!



HOW TO DEFINE FAIRNESS

QUANTITATIVELY SUCH THAT IT CAN BE USED IN ML SYSTEMS?

HOW TO DEFINE FAIRNESS

QUANTITATIVELY SUCH THAT IT CAN BE USED IN ML SYSTEMS?

DEFINITIONS

- ▶ Fairness through Unawareness

GROUP-BASED

- ▶ Demographic Parity
- ▶ Equality of Opportunity
- ▶ Equalized Odds
- ▶ Predictive Parity

INDIVIDUAL-BASED

- ▶ Individual Fairness
- ▶ Counterfactual fairness

EXAMPLE



PREDICTING IF HIRING AN APPLICANT:

$X \in R^d$ quantified features of the applicant (e.g. education, experience, college GPA, etc.)

$A \in \{0, 1\}$ a binary sensitive attribute (e.g. male/female)

$C(X, A)$ an ML predictor (e.g. hire/reject)

$Y \in \{0, 1\}$ target variable (e.g. if the candidate is truly capable of the position)

$X, A, Y \sim$ from an underlying distribution D .



PREDICTING IF HIRING AN APPLICANT:

$X \in R^d$ quantified features of the applicant (e.g. education, experience, college GPA, etc.)

$A \in \{0, 1\}$ a binary sensitive attribute (e.g. male/female)

$C(X, A)$ an ML predictor (e.g. hire/reject)

$Y \in \{0, 1\}$ target variable (e.g. if the candidate is truly capable of the position)

$X, A, Y \sim$ from an underlying distribution D .

FAIR MACHINE LEARNING:

Training the best $C(X, A)$ that is accurate and fair.

EXAMPLE



**WHAT IS THE EASIEST WAY OF
INTRODUCING FAIRNESS IN
DECISION MAKING?**

1. FAIRNESS THROUGH UNAWARENESS

Not including the sensitive attribute as a feature in the training data.

1. FAIRNESS THROUGH UNAWARENESS

Not including the sensitive attribute as a feature in the training data.

RECAP

$X \in \mathcal{R}^d$ features of the applicant;

$A \in \{0, 1\}$ sensitive attribute;

$C(X, A)$ predictor (hire/reject)

$Y \in \{0, 1\}$ target variable (truly capable)

$C(X)$ instead of $C(X, A)$

1. FAIRNESS THROUGH UNAWARENESS

Not including the sensitive attribute as a feature in the training data.

RECAP

$X \in \mathcal{R}^d$ features of the applicant;

$A \in \{0, 1\}$ sensitive attribute;

$C(X, A)$ predictor (hire/reject)

$Y \in \{0, 1\}$ target variable (truly capable)

$C(X)$ instead of $C(X, A)$

- ▶ This definition protects against direct discrimination.

1. FAIRNESS THROUGH UNAWARENESS

Not including the sensitive attribute as a feature in the training data.

RECAP

$X \in \mathcal{R}^d$ features of the applicant;

$A \in \{0, 1\}$ sensitive attribute;

$C(X, A)$ predictor (hire/reject)

$Y \in \{0, 1\}$ target variable (truly capable)

$C(X)$ instead of $C(X, A)$

- ▶ This definition protects against direct discrimination.

FLAWS

- ▶ there can be many highly correlated features in X (e.g. marital status, height) that are proxies of the sensitive attribute (e.g. gender).

2. DEMOGRAPHIC/STATISTICAL PARITY

The acceptance rates of the applicants from each of the groups must be equal.

2. DEMOGRAPHIC/STATISTICAL PARITY

The acceptance rates of the applicants from each of the groups must be equal.

C is independent of A :

$$P[C = 1 \mid A=0] = P[C = 1 \mid A=1]$$

RECAP

$X \in R^d$ features of the applicant;

$A \in \{0, 1\}$ sensitive attribute;

$C(X, A)$ predictor (hire/reject)

$Y \in \{0, 1\}$ target (truly capable)

2. DEMOGRAPHIC/STATISTICAL PARITY

The acceptance rates of the applicants from each of the groups must be equal.

RECAP

$X \in R^d$ features of the applicant;

$A \in \{0, 1\}$ sensitive attribute;

$C(X, A)$ predictor (hire/reject)

$Y \in \{0, 1\}$ target (truly capable)

C is independent of A :

$$P[C = 1 | A=0] = P[C = 1 | A=1]$$

A positive outcome is often the preferred decision, such as *getting to university, getting a loan or being shown the ad.*

2. DEMOGRAPHIC/STATISTICAL PARITY

The acceptance rates of the applicants from each of the groups must be equal.

RECAP

$X \in R^d$ features of the applicant;

$A \in \{0, 1\}$ sensitive attribute;

$C(X, A)$ predictor (hire/reject)

$Y \in \{0, 1\}$ target (truly capable)

C is independent of A :

$$P[C = 1 \mid A=0] = P[C = 1 \mid A=1]$$

Variations:

- ▶ The p% rule: $P[C=1 \mid A=0] / P[C=1 \mid A=1] \geq p/100$

2. DEMOGRAPHIC/STATISTICAL PARITY

The acceptance rates of the applicants from each of the groups must be equal.

RECAP

$X \in R^d$ features of the applicant;

$A \in \{0, 1\}$ sensitive attribute;

$C(X, A)$ predictor (hire/reject)

$Y \in \{0, 1\}$ target (truly capable)

C is independent of A :

$$P[C = 1 \mid A=0] = P[C = 1 \mid A=1]$$

Variations:

- ▶ The p% rule: $P[C=1 \mid A=0] / P[C=1 \mid A=1] \geq p/100$

Legal Support: “**four-fifth rule**” prescribes that a selection rate for any protected/disadvantaged group should be at least 80% (four-fifths) of the selection rate for the unprotected/advantaged group.

2. DEMOGRAPHIC/STATISTICAL PARITY

The acceptance rates of the applicants from each of the groups must be equal.

RECAP

$X \in R^d$ features of the applicant;

$A \in \{0, 1\}$ sensitive attribute;

$C(X, A)$ predictor (hire/reject)

$Y \in \{0, 1\}$ target (truly capable)

C is independent of A :

$$P[C = 1 \mid A=0] = P[C = 1 \mid A=1]$$

Variations:

- ▶ The p% rule: $P[C=1 \mid A=0] / P[C=1 \mid A=1] \geq p/100$
- ▶ $|P[C=1 \mid A=0] - P[C=1 \mid A=1]| \leq \epsilon$ where $\epsilon \in [0,1]$.

2. DEMOGRAPHIC/STATISTICAL PARITY

Whiteboard example.

2. DEMOGRAPHIC/STATISTICAL PARITY

The acceptance rates of the applicants from each of the groups must be equal.

RECAP

$X \in R^d$ features of the applicant;

$A \in \{0, 1\}$ sensitive attribute;

$C(X, A)$ predictor (hire/reject)

$Y \in \{0, 1\}$ target (truly capable)

C is independent of A :

$$P[C = 1 | A=0] = P[C = 1 | A=1]$$

QUIZZZ

- ▶ We have a perfect predictor, $C=Y$. Will it satisfy statistical parity?
- ▶ If we hire the qualified from one group and random people from the other group, we can still achieve demographic parity. How?

2. DEMOGRAPHIC/STATISTICAL PARITY

The acceptance rates of the applicants from each of the groups must be equal.

RECAP

$X \in R^d$ features of the applicant;

$A \in \{0, 1\}$ sensitive attribute;

$C(X, A)$ predictor (hire/reject)

$Y \in \{0, 1\}$ target (truly capable)

C is independent of A :

$$P[C = 1 | A=0] = P[C = 1 | A=1]$$

FLAWS

- ▶ This definition rules out a perfect predictor $C=Y$ when base rates are different (i.e. $P[Y=1 | A=0] \neq P[Y=1 | A=1]$).
- ▶ If we hire the qualified from one group and random people from the other group, we can still achieve demographic parity.

2. DEMOGRAPHIC/STATISTICAL PARITY

The acceptance rates of the applicants from each of the groups must be equal.

C is independent of A :

$$P[C = 1 \mid A=0] = P[C = 1 \mid A=1]$$

RECAP

$X \in R^d$ features of the applicant;

$A \in \{0, 1\}$ sensitive attribute;

$C(X, A)$ predictor (hire/reject)

$Y \in \{0, 1\}$ target (truly capable)



3. EQUALITY OF OPPORTUNITY

The acceptance rates of the **qualified** applicants from each of the groups must be equal.

3. EQUALITY OF OPPORTUNITY

The acceptance rates of the **qualified** applicants from each of the groups must be equal.

RECAP

$X \in \mathbb{R}^d$ features of the applicant;

$A \in \{0, 1\}$ sensitive attribute;

$C(X, A)$ predictor (hire/reject)

$Y \in \{0, 1\}$ target variable (truly capable)

C is independent of A conditioned on $Y=1$:

$$P[C = 1 \mid A=0, Y=1] = P[C = 1 \mid A=1, Y=1]$$

3. EQUALITY OF OPPORTUNITY

The acceptance rates of the **qualified** applicants from each of the groups must be equal.

RECAP

$X \in \mathbb{R}^d$ features of the applicant;

$A \in \{0, 1\}$ sensitive attribute;

$C(X, A)$ predictor (hire/reject)

$Y \in \{0, 1\}$ target variable (truly capable)

C is independent of A conditioned on $Y=1$:

$$P[C = 1 \mid A=0, Y=1] = P[C = 1 \mid A=1, Y=1]$$

- ▶ Allows a perfect predictor $C=Y$.
- ▶ Stimulates to reduce errors uniformly in all groups.

3. EQUALITY OF OPPORTUNITY

Whiteboard example.

3. EQUALITY OF OPPORTUNITY

The acceptance rates of the **qualified** applicants from each of the groups must be equal.

RECAP

$X \in R^d$ features of the applicant;

$A \in \{0, 1\}$ sensitive attribute;

$C(X, A)$ predictor (hire/reject)

$Y \in \{0, 1\}$ target variable (truly capable)

C is independent of A conditioned on $Y=1$:

$$P[C = 1 \mid A=0, Y=1] = P[C = 1 \mid A=1, Y=1]$$

QUIZZZ

Group $A=0$ has 100 applicants and 58 of them are qualified while group $A=1$ also have 100 applicants but only 2 of them are qualified.

If the company decides to accept 30 applicants and satisfies equality of opportunity,

? offers will be conferred to group $A=0$ and

? offers will be conferred to group $A=1$.

3. EQUALITY OF OPPORTUNITY

The acceptance rates of the **qualified** applicants from each of the groups must be equal.

FLAWS

The gap between the groups has tendency to grow over time.

If the job is a well-paid job, group $A=0$ tends to have a better living condition and affords better education for their kids, and thus enable them to be qualified for such well-paid jobs when they grow up. The gap between group $A=0$ and group $A=1$ will tend to be enlarged over time.

4. EQUALIZED ODDS

The acceptance/rejection rates of the qualified/unqualified applicants from each of the groups must be equal.

4. EQUALIZED ODDS

The acceptance/rejection rates of the qualified/unqualified applicants from each of the groups must be equal.

RECAP

$X \in \mathbb{R}^d$ features of the applicant;

$A \in \{0, 1\}$ sensitive attribute;

$C(X, A)$ predictor (hire/reject)

$Y \in \{0, 1\}$ target variable (truly capable)

C is independent of A conditioned on Y :

$$P[C = 1 \mid A=0, Y=1] = P[C = 1 \mid A=1, Y=1]$$

$$P[C = 0 \mid A=0, Y=0] = P[C = 0 \mid A=1, Y=0]$$

4. EQUALIZED ODDS

The acceptance/rejection rates of the qualified/unqualified applicants from each of the groups must be equal.

RECAP

$X \in \mathbb{R}^d$ features of the applicant;

$A \in \{0, 1\}$ sensitive attribute;

$C(X, A)$ predictor (hire/reject)

$Y \in \{0, 1\}$ target variable (truly capable)

C is independent of A conditioned on Y :

$$P[C = 1 \mid A=0, Y=1] = P[C = 1 \mid A=1, Y=1]$$



TPR

$$P[C = 0 \mid A=0, Y=0] = P[C = 0 \mid A=1, Y=0]$$



TNR

- ▶ How about false positive and false negative rates?

4. EQUALIZED ODDS

The acceptance/rejection rates of the qualified/unqualified applicants from each of the groups must be equal.

RECAP

$X \in \mathbb{R}^d$ features of the applicant;

$A \in \{0, 1\}$ sensitive attribute;

$C(X, A)$ predictor (hire/reject)

$Y \in \{0, 1\}$ target variable (truly capable)

C is independent of A conditioned on Y :

$$P[C = 1 \mid A=0, Y=1] = P[C = 1 \mid A=1, Y=1]$$



TPR

$$P[C = 0 \mid A=0, Y=0] = P[C = 0 \mid A=1, Y=0]$$



TNR

- ▶ How about false positive and false negative rates? $FPR=1-TNR$

4. EQUALIZED ODDS

The acceptance/rejection rates of the qualified/unqualified applicants from each of the groups must be equal.

RECAP

$X \in \mathbb{R}^d$ features of the applicant;

$A \in \{0, 1\}$ sensitive attribute;

$C(X, A)$ predictor (hire/reject)

$Y \in \{0, 1\}$ target variable (truly capable)

C is independent of A conditioned on Y :

$$P[C = 1 | A=0, Y=1] = P[C = 1 | A=1, Y=1]$$



TPR

$$P[C = 0 | A=0, Y=0] = P[C = 0 | A=1, Y=0]$$



TNR

- ▶ How about false positive and false negative rates?
- ▶ Can we do $P[C \neq Y | A=0] = P[C \neq Y | A=1]$ instead?

Accuracy parity



4. EQUALIZED ODDS

The acceptance (rejection) rates of the qualified (unqualified) applicants from each of the groups must be equal.

RECAP

$X \in R^d$ features of the applicant;

$A \in \{0, 1\}$ sensitive attribute;

$C(X, A)$ predictor (hire/reject)

$Y \in \{0, 1\}$ target variable (truly capable)

C is independent of A conditioned on Y :

$$P[C = 1 | A=0, Y=1] = P[C = 1 | A=1, Y=1]$$



TPR

$$P[C = 0 | A=0, Y=0] = P[C = 0 | A=1, Y=0]$$



TNR

- ▶ How about false positive and false negative rates?
- ▶ Can we do $P[C \neq Y | A=0] = P[C \neq Y | A=1]$ instead?

Accuracy parity



Could lead to a tradeoff: rejecting ($C=0$) qualified applicants ($Y=1$) from one group ($A=0$) for accepting ($C=1$) unqualified people ($Y=0$) from another group ($A=1$).

5. PREDICTIVE PARITY

Given acceptance (rejection), there is an equal chance of being qualified (unqualified) for each of the groups.

5. PREDICTIVE PARITY

Given acceptance (rejection), there is an equal chance of being qualified (unqualified) for each of the groups.

RECAP

$X \in \mathbb{R}^d$ features of the applicant;

$A \in \{0, 1\}$ sensitive attribute;

$C(X, A)$ predictor (hire/reject)

$Y \in \{0, 1\}$ target variable (truly capable)

Y is independent of A conditioned on C :

$$P[Y = 1 \mid A=0, C=1] = P[Y = 1 \mid A=1, C=1],$$

$$P[Y = 0 \mid A=0, C=0] = P[Y = 0 \mid A=1, C=0].$$

5. PREDICTIVE PARITY

Given acceptance (rejection), there is an equal chance of being qualified (unqualified) for each of the groups.

RECAP

$X \in R^d$ features of the applicant;

$A \in \{0, 1\}$ sensitive attribute;

$C(X, A)$ predictor (hire/reject)

$Y \in \{0, 1\}$ target variable (truly capable)

Y is independent of A conditioned on C :

$$P[Y = 1 \mid A=0, C=1] = P[Y = 1 \mid A=1, C=1],$$

$$P[Y = 0 \mid A=0, C=0] = P[Y = 0 \mid A=1, C=0].$$

- ▶ Allows a perfect predictor $C=Y$.
- ▶ Predictor C reflects the candidate's real capability of doing the job.

5. PREDICTIVE PARITY

Given acceptance (rejection), there is an equal chance of being qualified (unqualified) for each of the groups.

RECAP

$X \in R^d$ features of the applicant;

$A \in \{0, 1\}$ sensitive attribute;

$C(X, A)$ predictor (hire/reject)

$Y \in \{0, 1\}$ target variable (truly capable)

Y is independent of A conditioned on C :

$$P[Y = 1 \mid A=0, C=1] = P[Y = 1 \mid A=1, C=1],$$

$$P[Y = 0 \mid A=0, C=0] = P[Y = 0 \mid A=1, C=0].$$

- ▶ Allows a perfect predictor $C=Y$.
- ▶ Predictor C reflects the candidate's real capability of doing the job.

GROUP FAIRNESS DEFINITIONS

Let's have all fairness metrics together, then we are surely fair. right?

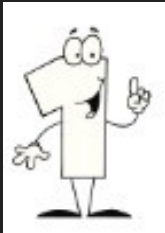
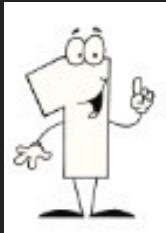
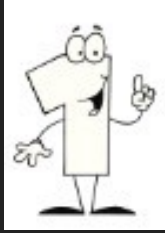











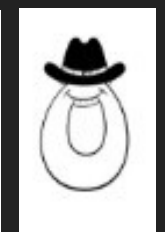
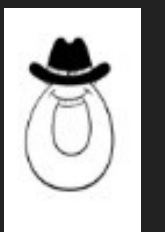
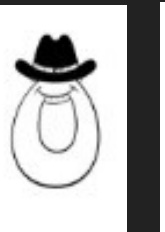
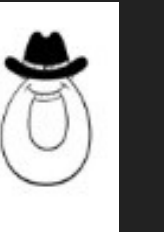
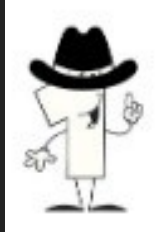
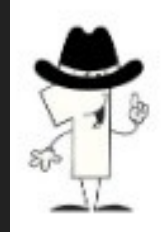

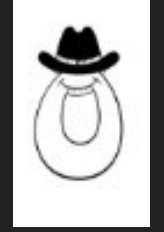
Let's have all fairness metrics together, then we are surely fair. right?

THE IMPOSSIBILITY THEOREM

Any two of the three criteria (demographic parity, equalised odds, predictive parity) are mutually exclusive except in non-degenerate cases.

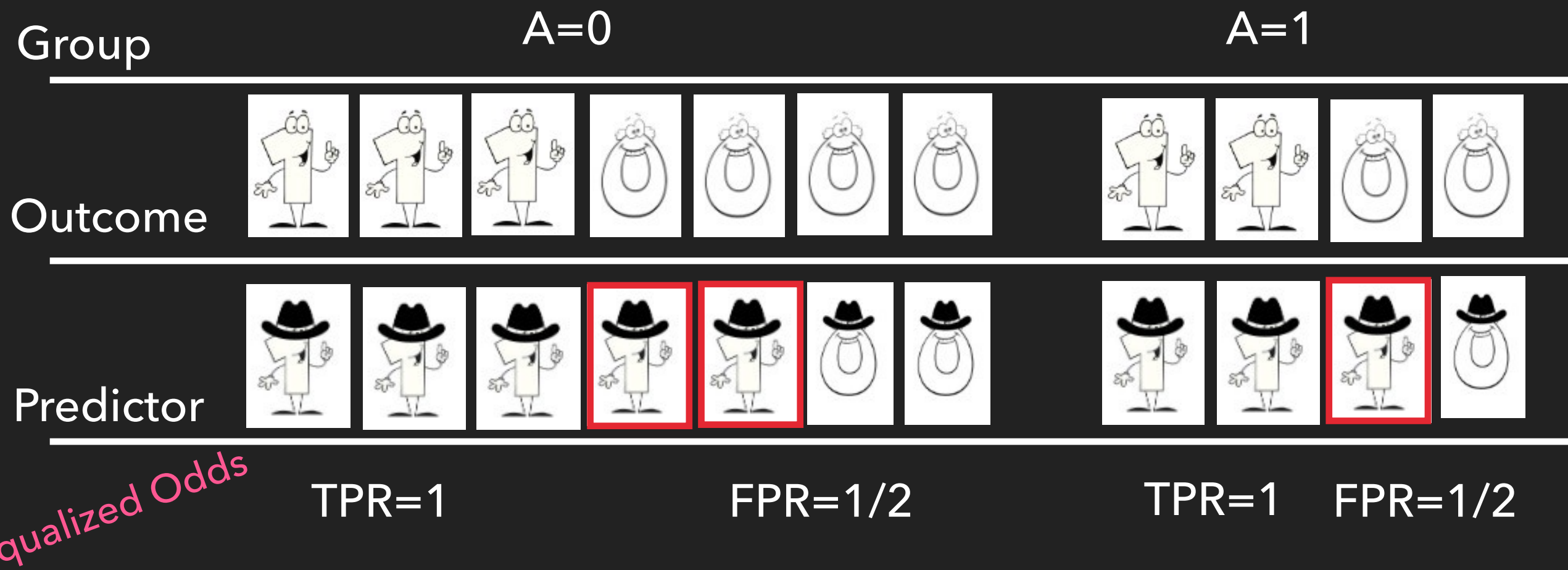
THE IMPOSSIBILITY THEOREM BY EXAMPLE

Equalized odds versus Predictive Parity

Group	A=0							A=1			
Outcome											
Predictor											
	TPR=1			FPR=0				TPR=1	FPR=1/2		

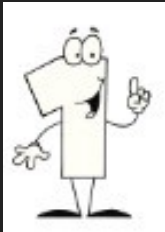
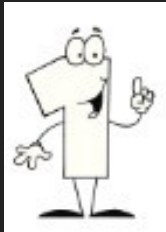
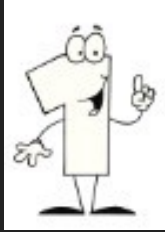















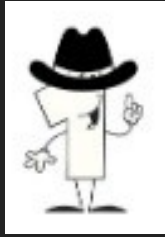



THE IMPOSSIBILITY THEOREM BY EXAMPLE

Equalized odds versus Predictive Parity



THE IMPOSSIBILITY THEOREM BY EXAMPLE

Equalized odds versus Predictive Parity

Group	A=0							A=1			
Outcome											
Predictor											
Equalized Odds				TPR=1				TPR=1			
Predictive parity does not hold!				FPR=1/2				FPR=1/2			
PPP=3/5				NPP=1				PPP=2/3			
								NPP=1			

THE IMPOSSIBILITY THEOREM BY BAYES' RULE

For each group, $A=0$ and $A=1$, we compute:

$$P[Y = 1|C = 1] = \frac{P[C = 1|Y = 1] P[Y = 1]}{P[C = 1|Y = 1] P[Y = 1] + P[C = 1|Y = 0] (1 - P[Y = 1])}$$

THE IMPOSSIBILITY THEOREM BY BAYES' RULE

For each group, $A=0$ and $A=1$, we compute:

$$P[Y = 1|C = 1] = \frac{\overset{\text{TPR (true positive rate)}}{P[C = 1|Y = 1]} \overset{\text{Base rate}}{P[Y = 1]}}{\overset{\text{FPR (false positive rate)}}{P[C = 1|Y = 0]} (1 - P[Y = 1]) + \overset{\text{TPR (true positive rate)}}{P[C = 1|Y = 1]} P[Y = 1]}$$

(Positive) prediction parity

- ▶ Suppose we have equalized odds (TPR and FPR rates are equal for $A=0$ and $A=1$), can we have (positive) prediction parity?

THE IMPOSSIBILITY THEOREM BY BAYES' RULE

For each group, $A=0$ and $A=1$, we compute:

$$P[Y = 1|C = 1] = \frac{\overset{\text{TPR (true positive rate)}}{P[C = 1|Y = 1]} \overset{\text{Base rate}}{P[Y = 1]}}{\underset{\text{FPR (false positive rate)}}{P[C = 1|Y = 0]} (1 - P[Y = 1]) + \overset{\text{TPR (true positive rate)}}{P[C = 1|Y = 1]} \overset{\text{Base rate}}{P[Y = 1]}}$$

(Positive) prediction parity

- ▶ Suppose we have equalized odds (TPR and FPR rates are equal for $A=0$ and $A=1$), can we have (positive) prediction parity?
- ▶ YES! But only if we had a perfect dataset (base rates are equal)

THE IMPOSSIBILITY THEOREM BY BAYES' RULE

For each group, $A=0$ and $A=1$, we compute:

$$P[Y = 1|C = 1] = \frac{\overset{\text{TPR (true positive rate)}}{P[C = 1|Y = 1]} \overset{\text{Base rate}}{P[Y = 1]}}{\overset{\text{FPR (false positive rate)}}{P[C = 1|Y = 0]} (1 - P[Y = 1]) + \overset{\text{TPR (true positive rate)}}{P[C = 1|Y = 1]} P[Y = 1]}$$

(Positive) prediction parity

- ▶ Suppose we have equalized odds (TPR and FPR rates are equal for $A=0$ and $A=1$), can we have (positive) prediction parity?
- ▶ YES! But only if we had a **perfect dataset** (base rates are equal) or a **perfect predictor** (TPR=1 and FPR=0 for $A=0$ and $A=1$).

DEFINITIONS

- ▶ Fairness through Unawareness

GROUP-BASED

- ▶ Demographic Parity
- ▶ Equality of Opportunity
- ▶ Equalized Odds
- ▶ Predictive Parity

INDIVIDUAL-BASED

- ▶ Individual Fairness
- ▶ Counterfactual fairness

EXAMPLE



5. INDIVIDUAL FAIRNESS

Similar individuals should be treated similarly.

5. INDIVIDUAL FAIRNESS

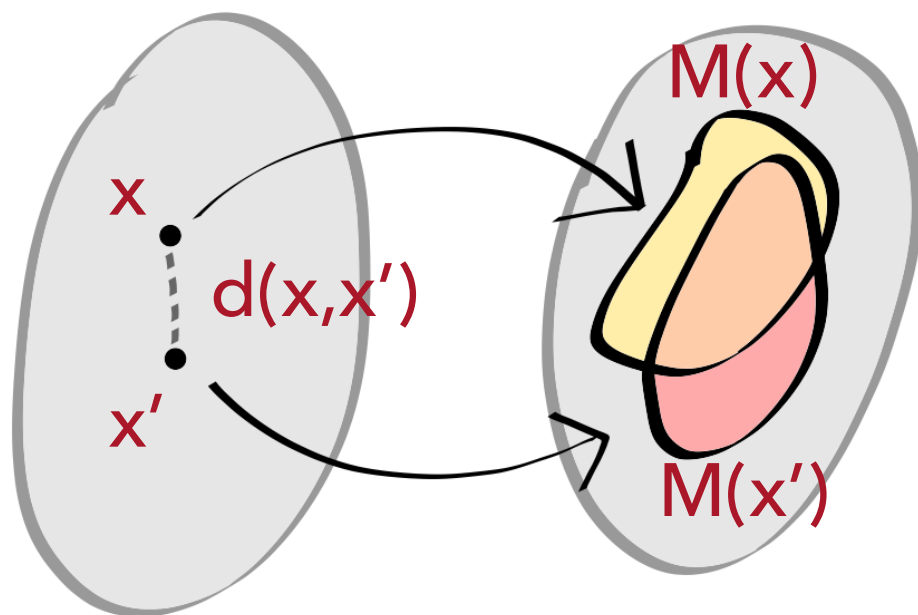
Similar individuals should be treated similarly.

- ▶ How to define a similarity metric?

5. INDIVIDUAL FAIRNESS

Similar individuals should be treated similarly.

- ▶ How to define a similarity metric?



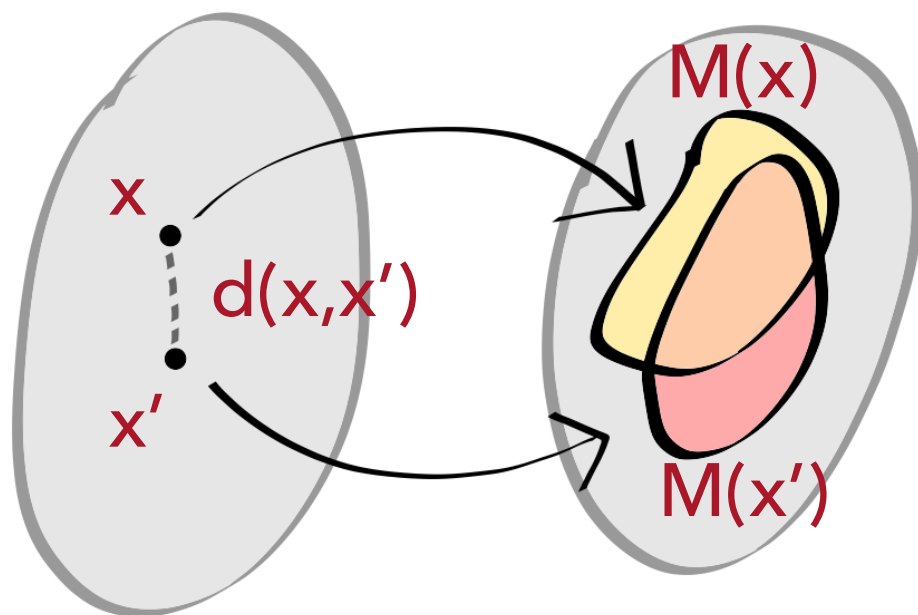
The Lipschitz condition:

any two individuals X, X' that are at distance $d(X, X') \in [0, 1]$ map to distributions $M(X)$ and $M(X')$, respectively, such that the statistical distance between $M(X)$ and $M(X')$ is at most $d(X, X')$.

5. INDIVIDUAL FAIRNESS

Similar individuals should be treated similarly.

► How to define a similarity metric?



Or:

The distributions over outcomes observed by X and X' are indistinguishable up to their distance $d(X, X')$.

5. INDIVIDUAL FAIRNESS

Similar individuals should be treated similarly.

- ▶ How to define a similarity metric?

EXAMPLE

Imagine three job applicants, A, B and C.

A has a bachelor degree and 1 year related work experience.

B has a master degree and 1 year related work experience.

C has a master degree but no related work experience.

Is A closer to B than C? If so, by how much?

5. INDIVIDUAL FAIRNESS

Similar individuals should be treated similarly.

- ▶ How to define a similarity metric?

EXAMPLE

Imagine three job applicants, A, B and C.

A has a bachelor degree and 1 year related work experience.

B has a master degree and 1 year related work experience.

C has a master degree but no related work experience.

Is A closer to B than C? If so, by how much?

- ▶ How about the sensitive attribute?
- ▶ How to count for the difference of group membership in the metric?

6. COUNTERFACTUAL FAIRNESS

If we intervene the sensitive feature, the prediction should not change.

6. COUNTERFACTUAL FAIRNESS

If we intervene the sensitive feature, the prediction should not change.

$$P[C\{A \leftarrow 0\}=1 \mid X, A=a] = P[C\{A \leftarrow 1\}=1 \mid X, A=a]$$

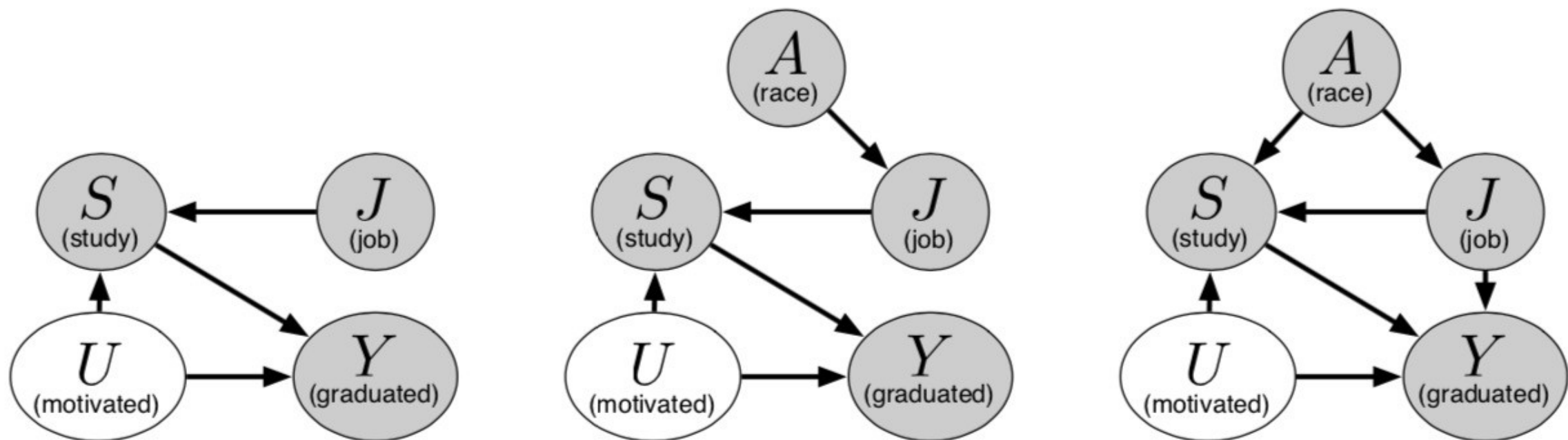
$$P[C\{A \leftarrow 0\}=0 \mid X, A=a] = P[C\{A \leftarrow 1\}=0 \mid X, A=a]$$

6. COUNTERFACTUAL FAIRNESS

If we intervene the sensitive feature, the prediction should not change.

$$P[C\{A \leftarrow 0\}=1 \mid X, A=a] = P[C\{A \leftarrow 1\}=1 \mid X, A=a]$$

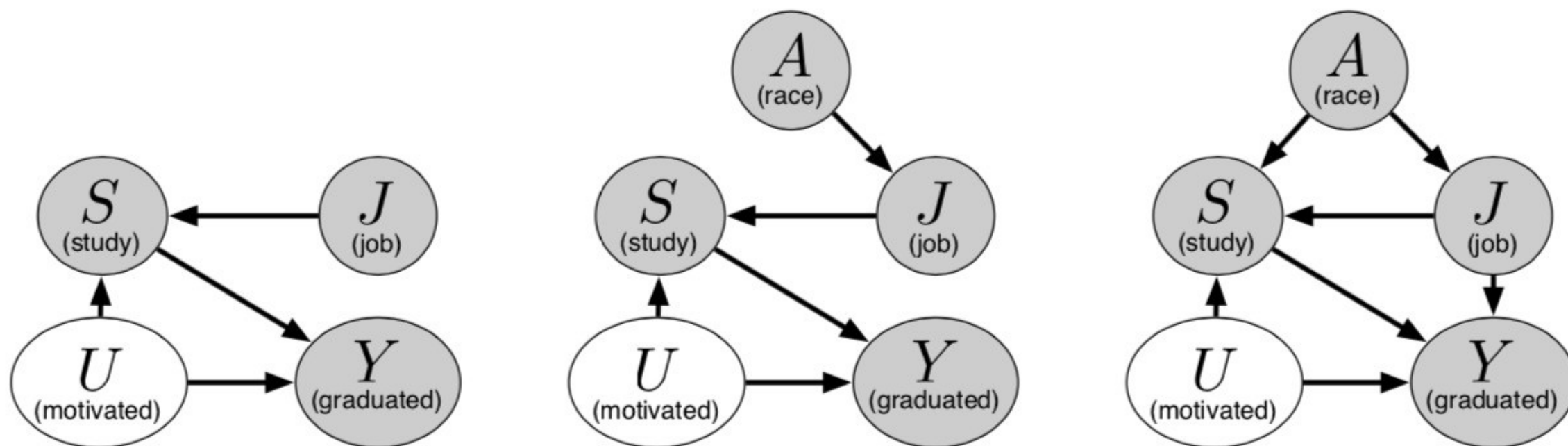
$$P[C\{A \leftarrow 0\}=0 \mid X, A=a] = P[C\{A \leftarrow 1\}=0 \mid X, A=a]$$



Causal graphs (scenario: applying to college)

6. COUNTERFACTUAL FAIRNESS

If we intervene the sensitive feature, the prediction should not change.

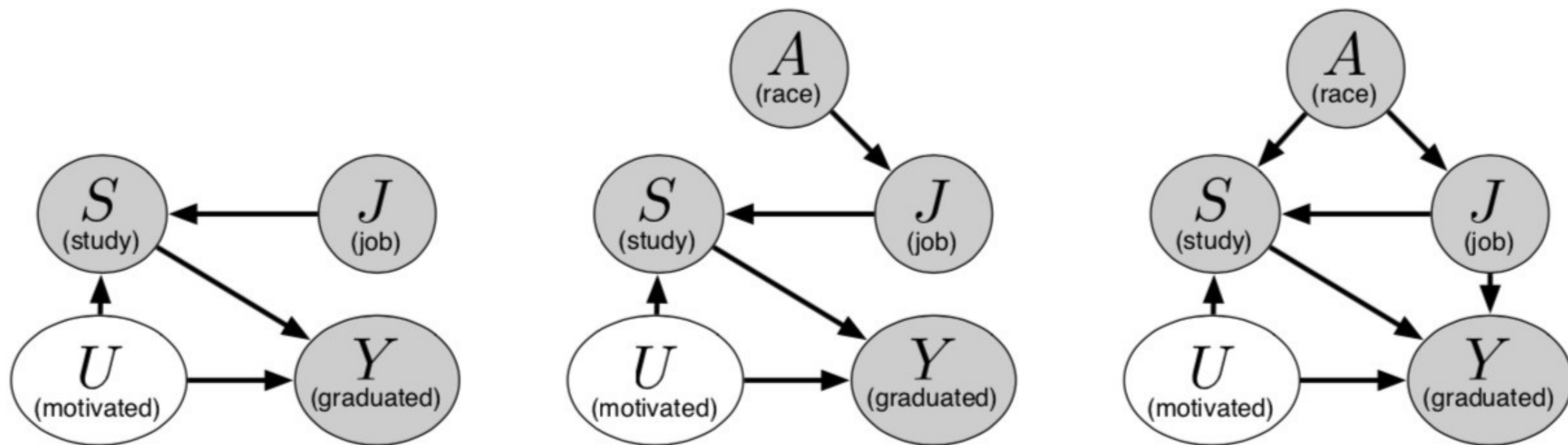


Causal graphs (scenario: applying to college)

We can compute what (the distribution of) any of the variables would have been *had certain other variables been different, or being equal*.

6. COUNTERFACTUAL FAIRNESS

If we intervene the sensitive feature, the prediction should not change.



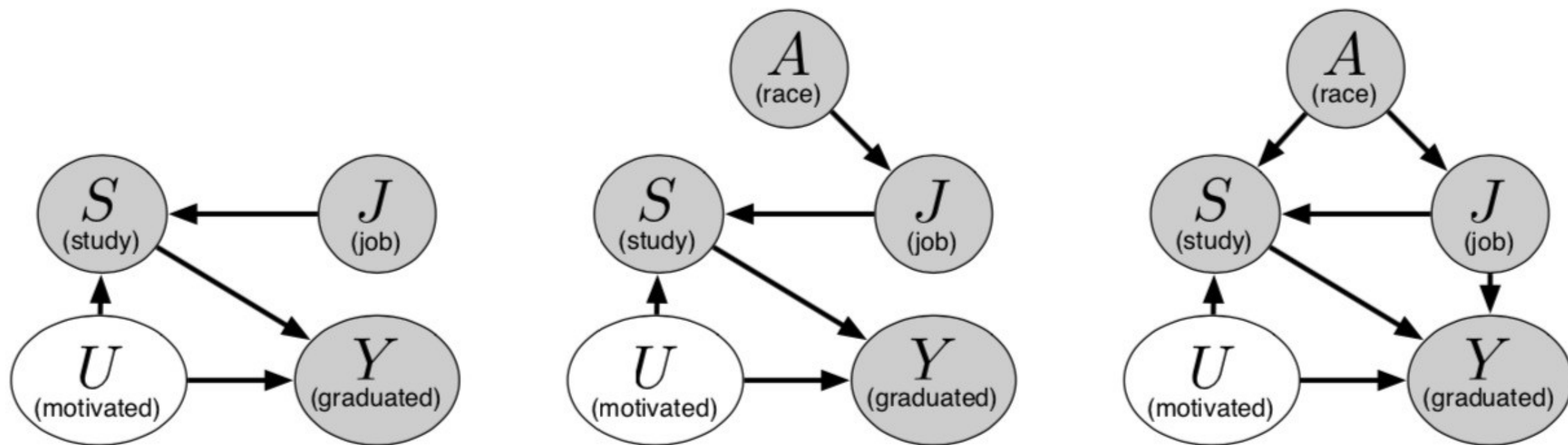
Causal graphs (scenario: applying to college)

We can compute what (the distribution of) any of the variables would have been *had certain other variables been different, or being equal*.

What would Y have been (would the student graduate, $Y=1$) if he/she hadn't had a job?

6. COUNTERFACTUAL FAIRNESS

If we intervene the sensitive feature, the prediction should not change.



Causal graphs (scenario: applying to college)

- ▶ Allows to check the possible impact of replacing only the sensitive attribute;
- ▶ In practice: what the causal graph should look like? How to decide which features to use even if we have such a graph (we may suffer large loss on accuracy if we eliminate all the correlated features).

CONCLUSIONS

FAIRNESS DEFINITIONS

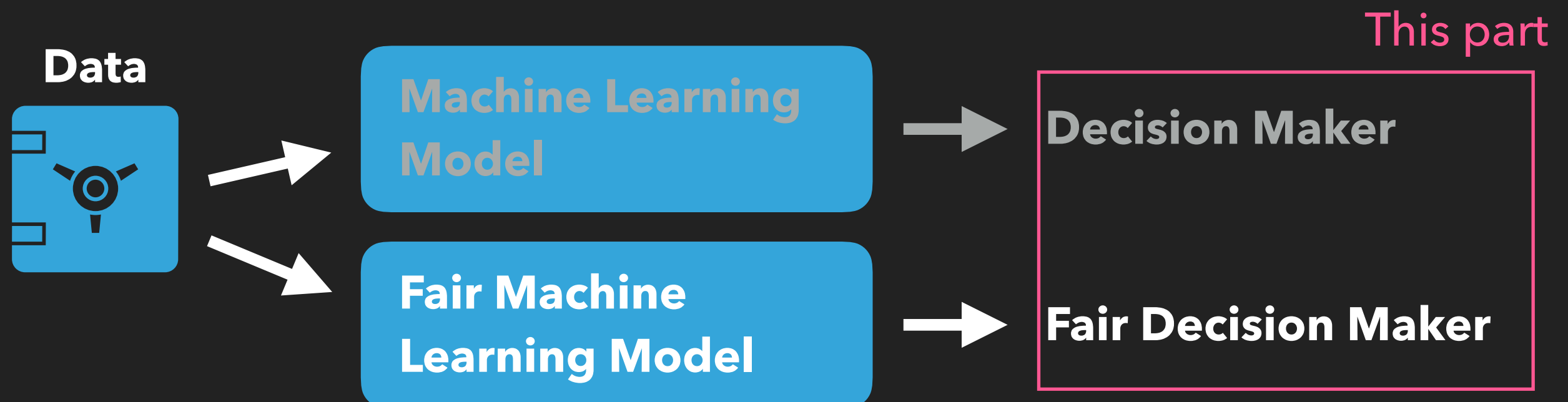
- ▶ Fairness through Unawareness

GROUP-BASED

- ▶ Demographic Parity
- ▶ Equality of Opportunity
- ▶ Equalized Odds
- ▶ Predictive Parity

INDIVIDUAL-BASED

- ▶ Individual Fairness
- ▶ Counterfactual fairness



FAIRNESS DEFINITIONS

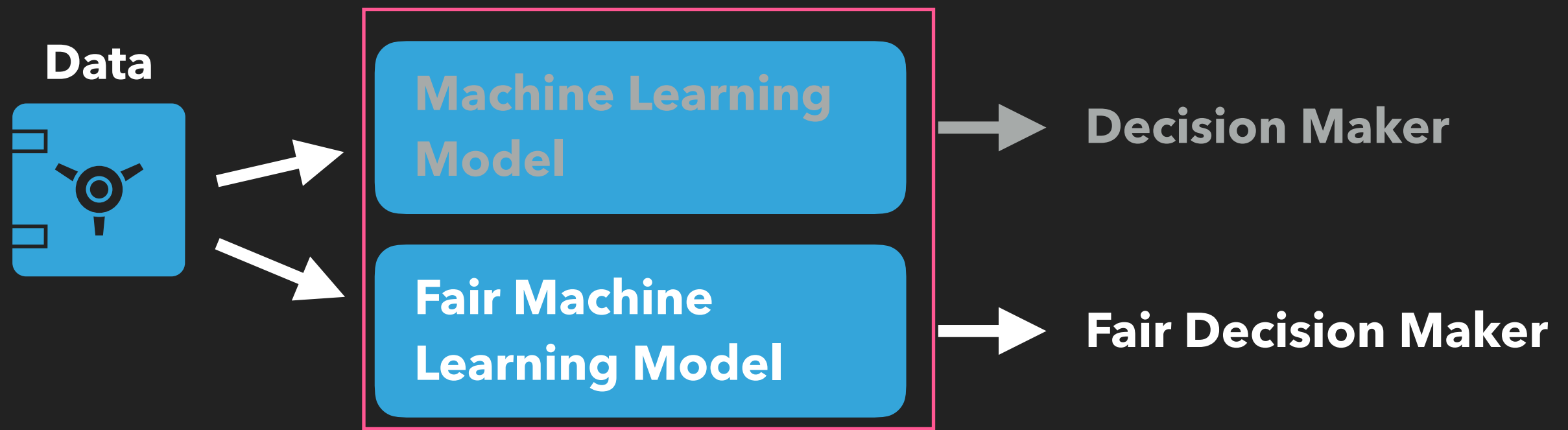
- ▶ Fairness through Unawareness

GROUP-BASED

- ▶ Demographic Parity
- ▶ Equality of Opportunity
- ▶ Equalized Odds
- ▶ Predictive Parity

INDIVIDUAL-BASED

- ▶ Individual Fairness
- ▶ Counterfactual fairness




PRO PUBLICA Graphic

Trump Administration Education Criminal Justice

Technology

SERIES



MACHINE BIAS
Investigating Algorithmic Injustice

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

<https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm>

<https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

Algorithmic fairness in machine learning

- ▶ Fairness definitions (20 Feb, 4pm-6pm)
- ▶ **Fairness methods** (26 Feb, 11am-1pm)
- ▶ Practical session (27 Feb, 4pm-6pm, Huxley 225)
- ▶ Advancements in algorithmic fairness (5 Mar, 4pm-6pm)

OUTLINE

- ▶ Fairness methods
 - ▶ Pre-processing
 - ▶ In-processing
 - ▶ Post-processing
- ▶ Running examples

RECAP

STATISTICAL PARITY

The acceptance rates of the applicants from each of the groups must be equal.

$$P[C = 1 \mid A=0] = P[C = 1 \mid A=1]$$

EQUALITY OF OPPORTUNITY

The acceptance rates of the **qualified** applicants from each of the groups must be equal.

$$P[C = 1 \mid A=0, Y=1] = P[C = 1 \mid A=1, Y=1]$$

RECAP

$X \in R^d$ features of the applicant;

$A \in \{0, 1\}$ sensitive attribute;

$C(X, A)$ predictor (hire/reject)

$Y \in \{0, 1\}$ target (truly capable)

RECAP

EQUALISED ODDS

The acceptance (rejection) rates of the qualified (unqualified) applicants from each of the groups must be equal.

$$P[C = 1 \mid A=0, Y=1] = P[C = 1 \mid A=1, Y=1]$$

$$P[C = 0 \mid A=0, Y=0] = P[C = 0 \mid A=1, Y=0]$$

PREDICTIVE PARITY

Given acceptance (rejection), there is an equal chance of being qualified (unqualified) for each of the groups.

$$P[Y = 1 \mid A=0, C=1] = P[Y = 1 \mid A=1, C=1],$$

$$P[Y = 0 \mid A=0, C=0] = P[Y = 0 \mid A=1, C=0].$$

RECAP

$X \in R^d$ features of the applicant;

$A \in \{0, 1\}$ sensitive attribute;

$C(X, A)$ predictor (hire/reject)

$Y \in \{0, 1\}$ target (truly capable)

CLASSIFICATION

RECAP

$X \in R^d$ features of the applicant;

$A \in \{0, 1\}$ sensitive attribute;

$C(X, A)$ predictor (hire/reject)

~~$Y \in \{0, 1\}$ target (truly capable)~~



$Y \in \{0, 1\}$ hired/not hired

$Y \in \{+1, -1\}$ hired/not hired

Classification

- Given
 - $\mathcal{X} = \mathbb{R}^d$ – input features, e.g. (experience, ML grade) $\in \mathcal{X}$,
 - $\mathcal{Y} = \{ \text{hired, not hired} \}$ – target labels,
- Learn $C : \mathcal{X} \rightarrow \mathcal{Y}$.

Classification

- Given
 - $\mathcal{X} = \mathbb{R}^d$ – input features, e.g. (experience, ML grade) $\in \mathcal{X}$,
 - $\mathcal{Y} = \{ \text{hired, not hired} \}$ – target labels,Learn $C : \mathcal{X} \rightarrow \mathcal{Y}$.
- Conditional probability $p(y|x)$

$$p(y = \text{hired} | \text{experience} = 24, \text{ML grade} = 90)$$

Classification

- Given
 - $\mathcal{X} = \mathbb{R}^d$ – input features, e.g. (experience, ML grade) $\in \mathcal{X}$,
 - $\mathcal{Y} = \{ \text{hired, not hired} \}$ – target labels,Learn $C : \mathcal{X} \rightarrow \mathcal{Y}$.

- Conditional probability $p(y|x)$

$$p(y = \text{hired} | \text{experience} = 24, \text{ML grade} = 90)$$

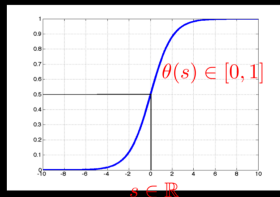
- Given $p(y|\mathbf{x})$, how should you classify?
 - Bayes optimal classifier:

$$C(x) = \arg \max_{y \in \{\text{hired, not hired}\}} p(y|\mathbf{x})$$

EXAMPLE: Logistic regression

- Conditional probability $p(y|x)$ is modeled via a logistic function:

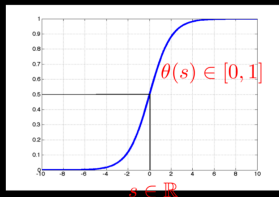
$$\sigma(s) = \frac{\exp(s)}{1 + \exp(s)} = \frac{1}{1 + \exp(-s)}.$$



EXAMPLE: Logistic regression

- Conditional probability $p(y|x)$ is modeled via a logistic function:

$$\sigma(s) = \frac{\exp(s)}{1 + \exp(s)} = \frac{1}{1 + \exp(-s)}.$$



- Now

$$p(y|\mathbf{x}) = \sigma(y\mathbf{w}^\top \mathbf{x})$$

where $y \in \{+1, -1\}$

EXAMPLE: Logistic regression

Given training data $\mathcal{D} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N)\}$, e.g.

$$\mathcal{D} = \{(88.7, 90, \text{hired}), (85.7, 87.2, \text{hired}), (50.1, 62.0, \text{not hired}) \dots\},$$

the objective minimises the regularised logistic loss:

$$\underset{\mathbf{w}}{\text{minimise}} \quad \lambda \|\mathbf{w}\|^2 + \sum_{i=1}^N -\log \left(1 + \exp(-y_i \mathbf{w}^\top \mathbf{x}_i) \right)^{-1}.$$

Fairness methods

- **pre-processing**
- in-processing
- post-processing

Pre-processing

Aim: to remove discrimination before a classifier is learned.

Pre-processing

Aim: to remove discrimination before a classifier is learned.

- The simplest pre-processing is to reweight the training data: those with higher weight are used more often and vice versa with lower weight.

Pre-processing

Aim: to remove discrimination before a classifier is learned.

- The simplest pre-processing is to reweight the training data: those with higher weight are used more often and vice versa with lower weight.
- E.g. for a data point with $Y = 1$ and $A = 0$ the weight is:

$$W_{(Y=1,A=0)} = \frac{P(Y=1)P(A=0)}{P(Y=1,A=0)} = \frac{\#(Y=1)\#(A=0)}{\#(Y=1,A=0)N}$$

Pre-processing

Aim: to remove discrimination before a classifier is learned.

- The simplest pre-processing is to reweight the training data: those with higher weight are used more often and vice versa with lower weight.
- E.g. for a data point with $Y = 1$ and $A = 0$ the weight is:

$$W_{(Y=1,A=0)} = \frac{P(Y=1)P(A=0)}{P(Y=1,A=0)} = \frac{\#(Y=1)\#(A=0)}{\#(Y=1,A=0)N}$$

- To make Y discrimination-free w.r.t. A in the reweighted dataset.

Pre-processing:reweighing

Whiteboard example

Gender	Ethnicity	Highest degree	Job type	Cl.
M	Native	H. school	Board	+
M	Native	Univ.	Board	+
M	Native	H. school	Board	+
M	Non-nat.	H. school	Healthcare	+
M	Non-nat.	Univ.	Healthcare	—
F	Non-nat.	Univ.	Education	—
F	Native	H. school	Education	—
F	Native	None	Healthcare	+
F	Non-nat.	Univ.	Education	—
F	Native	H. school	Board	+

Kamiran and Calders, Data preprocessing techniques for classification without discrimination, KAIS 2012.

Pre-processing:reweighing

Whiteboard example

Gender	Ethnicity	Highest degree	Job type	Cl.	Weight
M	Native	H. school	Board	+	0.75
M	Native	Univ.	Board	+	0.75
M	Native	H. school	Board	+	0.75
M	Non-nat.	H. school	Healthcare	+	0.75
M	Non-nat.	Univ.	Healthcare	=	2
F	Non-nat.	Univ.	Education	-	0.67
F	Native	H. school	Education	-	0.67
F	Native	None	Healthcare	+	1.5
F	Non-nat.	Univ.	Education	-	0.67
F	Native	H. school	Board	+	1.5

$$W(Y = 0, A = 0) = 2, W(Y = 0, A = 1) = 0.67, W(Y = 1, A = 0) = 0.75, \\ W(Y = 1, A = 1) = 1.5$$

Classification with reweighing

Input data (\mathbf{x}_i, y_i, a_i) , $i = 1 \dots N$

Compute weights $W(Y = y, A = a)$ for all combinations of Y, A :

$$W_{(Y=y, A=a)} = \frac{P(Y = y)P(A = a)}{P(Y = y, A = a)}$$

Train a classifier C on reweighed data

$\mathbf{w} \leftarrow \text{solve } \lambda \|\mathbf{w}\|^2 - \sum_{i=1}^N W_{(Y=y_i, A=a_i)} \log(1 + \exp(-y_i \mathbf{w}^\top \mathbf{x}_i))^{-1}$

Output classifier $C = (1 + \exp(-\mathbf{w}^\top \mathbf{x}))^{-1}$.

Classification with reweighing

- Logistic regression (logistic loss with reweighing):

$$\underset{\mathbf{w}}{\text{minimise}} \lambda \|\mathbf{w}\|^2 - \sum_{i=1}^N W_{(Y=y_i, A=a_i)} \log \left(1 + \exp(-y_i \mathbf{w}^\top \mathbf{x}_i) \right)^{-1}.$$

Classification with reweighing

- Logistic regression (logistic loss with reweighing):

$$\underset{\mathbf{w}}{\text{minimise}} \lambda \|\mathbf{w}\|^2 - \sum_{i=1}^N W_{(Y=y_i, A=a_i)} \log \left(1 + \exp(-y_i \mathbf{w}^\top \mathbf{x}_i) \right)^{-1}.$$

- Other examples of ML methods that allow reweighing:
 - SVMs (hinge loss with reweighing)

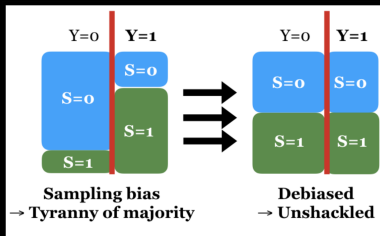
$$\underset{\mathbf{w}}{\text{minimise}} \lambda \|\mathbf{w}\|^2 + \sum_{i=1}^N W_{(Y=y_i, A=a_i)} \max(0, 1 - y_i \mathbf{w}^\top \mathbf{x}_i).$$

- Neural networks: `torch.nn.BCELoss(weight)`.

Pre-processing: resampling

Aim: to remove discrimination before a classifier is learned.

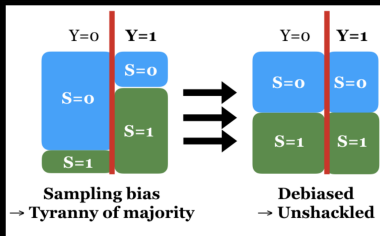
- From reweighing to resampling: sample data points with replacement according to weights.



Pre-processing: resampling

Aim: to remove discrimination before a classifier is learned.

- From reweighing to resampling: sample data points with replacement according to weights.



- E.g. from the group with $Y = 1$ and $A = 0$ sample

$$\frac{P(Y = 1)P(A = 0)}{P(Y = 1, A = 0)} \#(Y = 1, A = 0).$$

Kamiran and Calders, Data preprocessing techniques for classification without discrimination, KAIS 2012.

Classification with resampling

Input data (\mathbf{x}_i, y_i, a_i) , $i = 1 \dots N$

Compute weights $W(Y = y, A = a)$ for all combinations of Y, A :

$$W_{(Y=y, A=a)} = \frac{P(Y = y)P(A = a)}{P(Y = y, A = a)}$$

Sample uniformly $W_{(Y=y, A=a)} \times |(Y = y, A = a)|$ instances for all groups accordingly.

Output classifier C trained on resampled data.

Pre-processing: resampling

Whiteboard example

Gender	Ethnicity	Highest degree	Job type	Cl.	Weight
M	Native	H. school	Board	+	0.75
M	Native	Univ.	Board	+	0.75
M	Native	H. school	Board	+	0.75
M	Non-nat.	H. school	Healthcare	+	0.75
M	Non-nat.	Univ.	Healthcare	=	2
F	Non-nat.	Univ.	Education	-	0.67
F	Native	H. school	Education	-	0.67
F	Native	None	Healthcare	+	1.5
F	Non-nat.	Univ.	Education	-	0.67
F	Native	H. school	Board	+	1.5

2x1 samples ($Y = 0, A = 0$), 3x0.67=2 samples ($Y = 0, A = 1$), 0.75x4=3 samples ($Y = 1, A = 0$), 1.5x2=3 samples ($Y = 1, A = 1$).

Pre-processing: Representation learning

- Another popular approach is to produce a fair representation.
- Consider that we have 2 roles, a data vendor, who is charge of collecting the data and preparing it.
- Our other role is a data user, someone who will be making predictions based on our data.
- The data vendor is concerned that the data user may be using their data to make unfair decisions. So the data vendor decides to learn a new, fair representation.

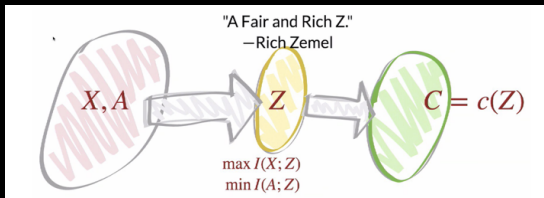
Pre-processing: Representation learning

Aim: to remove discrimination before a classifier is learned.

Representation learning

Aim: to remove discrimination before a classifier is learned.

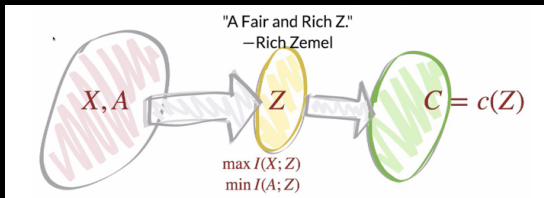
- Learn a new representation Z such that it removes the information about the sensitive attribute A and preserves the information of X .



Representation learning

Aim: to remove discrimination before a classifier is learned.

- Learn a new representation Z such that it removes the information about the sensitive attribute A and preserves the information of X .
- Train a classifier C on discrimination-free Z .



Representation learning

Discriminative clustering approach

- cluster training data into K clusters with \mathbf{v}_k centers $\mathbf{v}_k \in \mathcal{X}$, such that the probability of being assigned to the cluster k is independent of the sensitive feature A :

$$P[Z = k | \mathbf{x}_{|A=0}] = P[Z = k | \mathbf{x}_{|A=1}] \quad \forall k = 1 \dots K,$$

Representation learning

Discriminative clustering approach

- cluster training data into K clusters with \mathbf{v}_k centers $\mathbf{v}_k \in \mathcal{X}$, such that the probability of being assigned to the cluster k is independent of the sensitive feature A :

$$P[Z = k | \mathbf{x}_{|A=0}] = P[Z = k | \mathbf{x}_{|A=1}] \quad \forall k = 1 \dots K,$$

- where

$$P[Z = k | \mathbf{x}] = \frac{\exp(-\text{dist}(\mathbf{x}, \mathbf{v}_k))}{\sum_{i=1}^K \exp(-\text{dist}(\mathbf{x}, \mathbf{v}_i))}.$$

Representation learning

- To enforce fairness

$$P[Z = k | \mathbf{x}_{|A=0}] = P[Z = k | \mathbf{x}_{|A=1}] \quad \forall k = 1 \dots K,$$

we minimise the fair representation loss across all K clusters:

$$L_z = \sum_{k=1}^K |\mathbb{E}_{\mathbf{x}_{|A=0}} P[Z = k | \mathbf{x}] - \mathbb{E}_{\mathbf{x}_{|A=1}} P[Z = k | \mathbf{x}]|.$$

Representation learning

- To ensure the mapping to Z space retains information in X , we minimise the reconstruction loss on the training data:

$$L_x = \sum_{i=1}^N (\mathbf{x}_i - \hat{\mathbf{x}}_i)^2, \quad \hat{\mathbf{x}}_i = \sum_{k=1}^K P[Z = k | \mathbf{x}_i] \mathbf{v}_k.$$

Representation learning

- To ensure the mapping to Z space retains information in X , we minimise the reconstruction loss on the training data:

$$L_x = \sum_{i=1}^N (\mathbf{x}_i - \hat{\mathbf{x}}_i)^2, \quad \hat{\mathbf{x}}_i = \sum_{k=1}^K P[Z = k | \mathbf{x}_i] \mathbf{v}_k.$$

- To ensure accurate predictions of the induced mapping from X to Y (by first mapping probabilistically to Z -space, and then mapping Z to Y), we minimise the classification error:

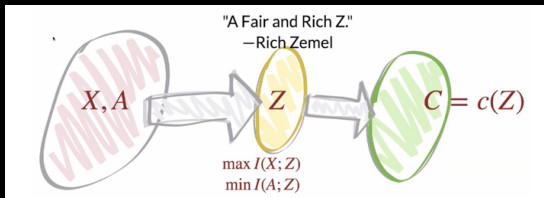
$$L_y = \sum_{i=1}^N -y_i \log(\hat{y}_i) - (1 - y_i) \log(1 - \hat{y}_i),$$

where $\hat{y}_i = \sum_{k=1}^K P[Z = k | \mathbf{x}_i] w_k$, $w_k \in [0, 1]$. Here $y_i \in \{0, 1\}$.

Representation learning

minimise $\alpha L_z + \beta L_x + \gamma L_y$, where \mathbf{v}_k, \mathbf{w}

- L_z is fair representation loss,
- L_x is reconstruction loss,
- L_y is classification loss,
- \mathbf{v}_k are the prototype locations, \mathbf{w} are the classifier weight parameters.



Questions?

Zemel et al, Learning fair representations, ICML 2013.

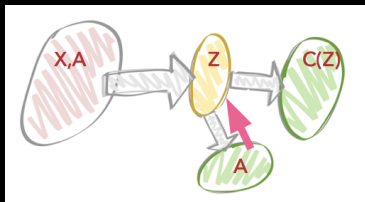
Pre-processing: Representation learning

Aim: to remove discrimination before a classifier is learned.

- Which fairness metric is enforced?
- Can we enforce a different one?
- Is Z interpretable?

Adversarial representation learning

- Many adversarial-based fair representation learning approaches.



Edwards and Storkey, Censoring representations with an adversary, ICLR 2016

Ganin et al., Domain-adversarial training of neural networks, JMLR 2016

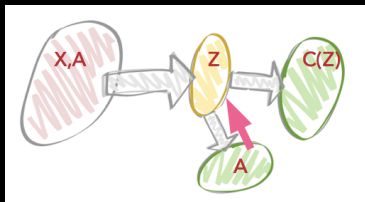
Beutel et al., Data decisions and theoretical implications when adversarially learning fair representations, Jul 2017

Madras et al., Learning adversarially fair and transferable representations, ICML 2018.

Adversarial representation learning

- Many adversarial-based fair representation learning approaches.
- For an encoder $X \rightarrow Z$ with parameters θ and an adversary classifier $R : Z \rightarrow [0, 1]$ with parameters ϕ :

$$\underset{\theta}{\text{minimise}} \underset{\phi}{\text{maximise}} \mathbb{E}_{X,A} A \log(R(Z)) + (1 - A) \log(1 - R(Z)).$$



Edwards and Storkey, Censoring representations with an adversary, ICLR 2016

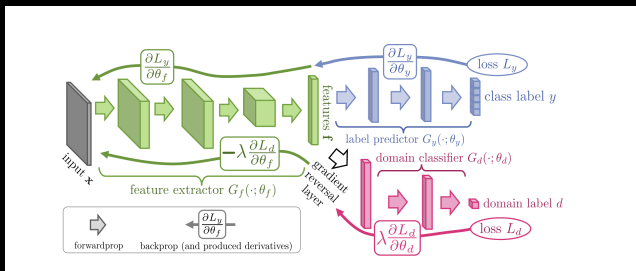
Ganin et al., Domain-adversarial training of neural networks, JMLR 2016

Beutel et al., Data decisions and theoretical implications when adversarially learning fair representations, Jul 2017

Madras et al., Learning adversarially fair and transferable representations, ICML 2018.

Adversarial representation learning

- Many adversarial-based fair representation learning approaches, e.g. using a **Gradient-Reversal Layer**.



Ganin et al., Domain-adversarial training of neural networks, JMLR 2016

Edwards and Storkey, Censoring representations with an adversary, ICLR 2016

Beutel et al., Data decisions and theoretical implications when adversarially learning fair representations, Jul 2017

Madras et al., Learning adversarially fair and transferable representations, ICML 2018.

Adversarial fair representation learning

- Does this representation really hide A ?

Adversarial fair representation learning

- Does this representation really hide A ?
- A work by Elazar and Goldberg shows that adversarially trained latent embeddings still retain sensitive attribute information when a post-hoc classifier is trained on them.

Elazar and Goldberg, Adversarial removal of demographic attributes from text data, EMNLP 2018.

Fairness methods

- pre-processing
- **in-processing**
- post-processing

In-processing

Aim: to constraint learning with fairness metrics.

In-processing

Aim: to constraint learning with fairness metrics.

- Given we have a loss function, $L(\mathbf{w})$,

In-processing

Aim: to constraint learning with fairness metrics.

- Given we have a loss function, $L(\mathbf{w})$,
- In an unconstrained classifier, we would expect to see:

$\underset{\mathbf{w}}{\text{minimise}} L(\mathbf{w})$.

In-processing

Aim: to constraint learning with fairness metrics.

- Given we have a loss function, $L(\mathbf{w})$,
- In an unconstrained classifier, we would expect to see:

$$\underset{\mathbf{w}}{\text{minimise}} L(\mathbf{w}).$$

- To enforce a **statistical parity** metric, a constraint is added:

$$\underset{\mathbf{w}}{\text{minimise}} L(\mathbf{w}) \text{ s.t. } P[\hat{Y} = 1|A = 0] = P[\hat{Y} = 1|A = 1].$$

In-processing

Aim: to constraint learning with fairness metrics.

- Given we have a loss function, $L(\mathbf{w})$,
- In an unconstrained classifier, we would expect to see:

$$\underset{\mathbf{w}}{\text{minimise}} L(\mathbf{w}).$$

- To enforce a **statistical parity** metric, a constraint is added:

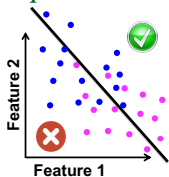
$$\underset{\mathbf{w}}{\text{minimise}} L(\mathbf{w}) \text{ s.t. } P[\hat{Y} = 1|A = 0] = P[\hat{Y} = 1|A = 1].$$

- Problem: The formulation is not convex!

In-processing

Need to find a better way to specify the constraints

Disparate impact constraints: Intuition



$$P(\hat{y} = 1 | z = 0) = P(\hat{y} = 1 | z = 1)$$

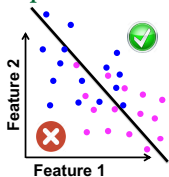
Limit the differences in the acceptance (or rejection) ratios across members of different sensitive groups

Zafar et al., Fairness constraints: Mechanisms for fair classification, AISTATS 2017
Zafar et al., Fairness Beyond Disparate Treatment and Disparate Impact: Learning Classification without Disparate Mistreatment, WWW 2017

In-processing

Need to find a better way to specify the constraints

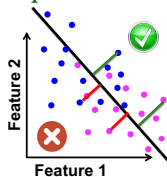
Disparate impact constraints: Intuition



$$P(\hat{y} = 1 | z = 0) = P(\hat{y} = 1 | z = 1)$$

Limit the differences in the acceptance (or rejection) ratios across members of different sensitive groups

Disparate impact constraints: Intuition



A **proxy** measure for $P(\hat{y} = 1 | z = 0) = P(\hat{y} = 1 | z = 1)$

Limit the differences in the average strength of acceptance and rejection across members of different sensitive groups

In-processing

- Instead of $P[\hat{Y} = 1|A = 0] = P[\hat{Y} = 1|A = 1]$

In-processing

- Instead of $P[\hat{Y} = 1|A = 0] = P[\hat{Y} = 1|A = 1]$
- Bound covariance between items sensitive feature values and their signed distance from classifier's decision boundary to less than a threshold:

$$\left| \frac{1}{N} \sum_{i=1}^N (a_i - \bar{a}) \mathbf{w}^T \mathbf{x}_i \right| \leq \epsilon$$

In-processing

- Instead of $P[\hat{Y} = 1|A = 0] = P[\hat{Y} = 1|A = 1]$
- **Bound covariance** between items **sensitive feature values** and **their signed distance from classifier's decision boundary** to less than a threshold:

$$\left| \frac{1}{N} \sum_{i=1}^N (a_i - \bar{a}) \mathbf{w}^T \mathbf{x}_i \right| \leq \epsilon$$

$$\begin{aligned} \text{Cov}(A - \bar{A}, f) &= \mathbb{E}[(A - \bar{A})f(X)] - \mathbb{E}[(A - \bar{A})]\mathbb{E}[f(X)] \\ &= \frac{1}{N} \sum_{i=1}^N (a_i - \bar{a}_i) f_i \end{aligned}$$

In-processing

- Instead of $P[\hat{Y} = 1|A = 0] = P[\hat{Y} = 1|A = 1]$
- **Bound covariance** between items **sensitive feature values** and **their signed distance from classifier's decision boundary** to less than a threshold:

$$\left| \frac{1}{N} \sum_{i=1}^N (a_i - \bar{a}) \mathbf{w}^T \mathbf{x}_i \right| \leq \epsilon$$

•

$$\begin{aligned} \text{Cov}(A - \bar{A}, f) &= \mathbb{E}[(A - \bar{A})f(X)] - \mathbb{E}[(A - \bar{A})]\mathbb{E}[f(X)] \\ &= \frac{1}{N} \sum_{i=1}^N (a_i - \bar{a}_i) f_i \end{aligned}$$

- From $\text{Cov}(A - \bar{A}, f) = 0 \rightarrow$ equal positive rates across groups
 \rightarrow statistical parity

Zafar et al., Fairness constraints: Mechanisms for fair classification, AISTATS 2017

Zafar et al., Fairness Beyond Disparate Treatment and Disparate Impact: Learning Classification without Disparate Mistreatment, WWW 2017

In-processing

Aim: to constraint learning with fairness metrics.

- Given we have a loss function, $L(\mathbf{w})$, e.g.
$$L(\mathbf{w}) = - \sum_{i=1}^N \log (1 + \exp(-y_i \mathbf{w}^\top \mathbf{x}_i))^{-1}$$

Zafar et al., Fairness constraints: Mechanisms for fair classification, AISTATS 2017
Zafar et al., Fairness Beyond Disparate Treatment and Disparate Impact: Learning Classification without Disparate Mistreatment, WWW 2017

In-processing

Aim: to constraint learning with fairness metrics.

- Given we have a loss function, $L(\mathbf{w})$, e.g.
$$L(\mathbf{w}) = - \sum_{i=1}^N \log \left(1 + \exp(-y_i \mathbf{w}^\top \mathbf{x}_i) \right)^{-1}$$
- the formulation is easy to solve!

$$\underset{\mathbf{w}}{\text{minimise}} L(\mathbf{w}) \text{ s.t. } \left| \frac{1}{N} \sum_{i=1}^N (a_i - \bar{a}) \mathbf{w}^\top \mathbf{x}_i \right| \leq \epsilon.$$

Zafar et al., Fairness constraints: Mechanisms for fair classification, AISTATS 2017
Zafar et al., Fairness Beyond Disparate Treatment and Disparate Impact: Learning Classification without Disparate Mistreatment, WWW 2017

Fairness methods

- pre-processing
- in-processing
- **post-processing**

Post-processing

Aim: to fix the trained model with fairness metrics.

Post-processing

- Train two separate models: one for all datapoints with $A=0$ and another one for $A=1$.

Calders and Verwer: Three naive Bayes approaches for discrimination-free classification, Data Mining and Knowledge Discovery 2010.

Hardt et al, Equality of Opportunity in Supervised Learning, NeurIPS 2016.

Post-processing

- Train two separate models: one for all datapoints with $A=0$ and another one for $A=1$.
- The thresholds of the model are then tweaked until they produce the same positive rate, i.e. $P[\hat{Y} = 1|A = 0] = P[\hat{Y} = 1|A = 1]$.

Calders and Verwer: Three naive Bayes approaches for discrimination-free classification, Data Mining and Knowledge Discovery 2010.

Hardt et al, Equality of Opportunity in Supervised Learning, NeurIPS 2016.

Post-processing

- Train two separate models: one for all datapoints with $A=0$ and another one for $A=1$.
- The thresholds of the model are then tweaked until they produce the same positive rate, i.e. $P[\hat{Y} = 1|A = 0] = P[\hat{Y} = 1|A = 1]$.
- Disadvantage: A has to be known for making predictions in order to choose the correct model.

Calders and Verwer: Three naive Bayes approaches for discrimination-free classification, Data Mining and Knowledge Discovery 2010.

Hardt et al, Equality of Opportunity in Supervised Learning, NeurIPS 2016.

Summary

- pre-processing

Aim: to remove discrimination before a classifier is learned.

- in-processing

Aim: to constraint learning with fairness metrics.

- post-processing

Aim: to fix the trained model with fairness metrics.

Algorithmic fairness in machine learning

- ▶ Fairness definitions (20 Feb, 4pm-6pm)
- ▶ Fairness methods (26 Feb, 11am-1pm)
- ▶ Practical session (27 Feb, 4pm-6pm, Huxley 225)
- ▶ **Advancements in algorithmic fairness** (5 Mar, 4pm-6pm)

OUTLINE TODAY

- ▶ Fairness methods
 - ▶ Post-processing
- ▶ Sources of unfairness
 - ▶ Bias from the data
 - ▶ Bias from the models
- ▶ Transparency in algorithmic fairness

Post-processing

- Given
 - $\mathcal{X} = \mathbb{R}^d$ – input features, e.g. (experience, ML grade) $\in \mathcal{X}$,
 - $\mathcal{Y} = \{ \text{hired, not hired} \}$ – target labels,
 - $\mathcal{A} = \{ 0, 1 \}$ – sensitive feature,

Post-processing

- Given
 - $\mathcal{X} = \mathbb{R}^d$ – input features, e.g. (experience, ML grade) $\in \mathcal{X}$,
 - $\mathcal{Y} = \{ \text{hired, not hired} \}$ – target labels,
 - $\mathcal{A} = \{ 0, 1 \}$ – sensitive feature,
- Learn $C : \mathcal{X} \rightarrow \mathcal{Y}$.

Post-processing

- Given
 - $\mathcal{X} = \mathbb{R}^d$ – input features, e.g. (experience, ML grade) $\in \mathcal{X}$,
 - $\mathcal{Y} = \{ \text{hired, not hired} \}$ – target labels,
 - $\mathcal{A} = \{ 0, 1 \}$ – sensitive feature,
- Learn $C : \mathcal{X} \rightarrow \mathcal{Y}$.
- Post-processing:

Aim: to fix the trained model with fairness metrics.

Adjust C to satisfy fairness, e.g.

$P[C = 1|A = 0] = P[C = 1|A = 1]$ parity

$P[C = 1|A = 0, Y = 1] = P[C = 1|A = 1, Y = 1]$ equality of opportunity.

Classifier

Learn $C : \mathcal{X} \rightarrow \mathcal{Y}$, e.g. by

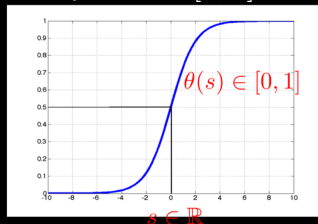
- modeling $p(y = 1|x)$ and using Bayes optimal classifier:
$$C(x) = \arg \max_{y \in \{\text{hired, not hired}\}} p(y|\mathbf{x})$$

Classifier

Learn $C : \mathcal{X} \rightarrow \mathcal{Y}$, e.g. by

- modeling $p(y = 1|x)$ and using Bayes optimal classifier:
$$C(x) = \arg \max_{y \in \{\text{hired, not hired}\}} p(y|\mathbf{x})$$
- thresholding a logistic regression classifier $f : \mathcal{X} \rightarrow [0, 1]$ at 0.5.

$$\theta(s) = \frac{1}{1 + \exp(-s)}$$



$$C(\mathbf{x}) = \mathbb{I}[f(\mathbf{x}) \geq 0.5], \quad f(x) = \theta(\mathbf{w}^\top \mathbf{x}).$$

Post-processing

- Skewed, unbalanced training data;
- A feature set that supports accurate predictions for the majority group may not for a minority group.

Train two separate models:

Calders and Verwer: Three naive Bayes approaches for discrimination-free classification, Data Mining and Knowledge Discovery 2010.

Post-processing

- Skewed, unbalanced training data;
- A feature set that supports accurate predictions for the majority group may not for a minority group.

Train two separate models:

- one, f_0 , for all datapoints with $A=0$ and another one, f_1 , for $A=1$.

Calders and Verwer: Three naive Bayes approaches for discrimination-free classification, Data Mining and Knowledge Discovery 2010.

Post-processing

- Skewed, unbalanced training data;
- A feature set that supports accurate predictions for the majority group may not for a minority group.

Train two separate models:

- one, f_0 , for all datapoints with $A=0$ and another one, f_1 , for $A=1$.
- Find the thresholds of the models f_0, f_1 to satisfy fairness (with the lowest loss in accuracy).

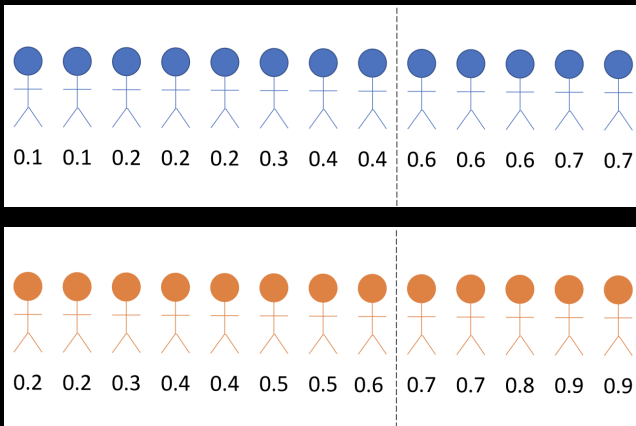
$$C(\mathbf{x}|A = 0) = \mathbb{I}[f_0 \geq \theta_0],$$

$$C(\mathbf{x}|A = 1) = \mathbb{I}[f_1 \geq \theta_1],$$

$$P[C = 1|A = 0] = P[C = 1|A = 1].$$

Calders and Verwer: Three naive Bayes approaches for discrimination-free classification, Data Mining and Knowledge Discovery 2010.

Post-processing



Post-processing

Aim: to fix the trained model with fairness metrics.

Post-processing

Aim: to fix the trained model with fairness metrics.

Given the binary predictor $C : \mathcal{X} \rightarrow \{0, 1\}$,

how to adjust its predictions

Post-processing

Aim: to fix the trained model with fairness metrics.

Given the binary predictor $C : \mathcal{X} \rightarrow \{0, 1\}$,

how to adjust its predictions

- to retain accuracy, i.e. minimize the expected loss $\mathbb{E}l(\hat{Y}, Y)$,

Post-processing

Aim: to fix the trained model with fairness metrics.

Given the binary predictor $C : \mathcal{X} \rightarrow \{0, 1\}$,

how to adjust its predictions

- to retain accuracy, i.e. minimize the expected loss $\mathbb{E} l(\hat{Y}, Y)$,
- adhere to fairness metric,

Post-processing

Aim: to fix the trained model with fairness metrics.

Given the binary predictor $C : \mathcal{X} \rightarrow \{0, 1\}$,

how to adjust its predictions

- to retain accuracy, i.e. minimize the expected loss $\mathbb{E} l(\hat{Y}, Y)$,
- adhere to fairness metric,
- without retraining/changes to C ?

Post-processing

- TPR/FRP coordinate system for \hat{Y} predictor:

$$\gamma_{A=0} = (P[\hat{Y} = 1|A = 0, Y = 0], P[\hat{Y} = 1|A = 0, Y = 1])$$

Similarly for $\gamma_{A=1}$.

Post-processing

- TPR/FRP coordinate system for \hat{Y} predictor:

$$\gamma_{A=0} = (P[\hat{Y} = 1|A = 0, Y = 0], P[\hat{Y} = 1|A = 0, Y = 1])$$

Similarly for $\gamma_{A=1}$.

- A predictor \hat{Y} satisfies equalized odds, if and only if $\gamma_{A=0} = \gamma_{A=1}$.
Second coordinate for equality of opportunity.

Post-processing

- TPR/FRP coordinate system for \hat{Y} predictor:

$$\gamma_{A=0} = (P[\hat{Y} = 1|A = 0, Y = 0], P[\hat{Y} = 1|A = 0, Y = 1])$$

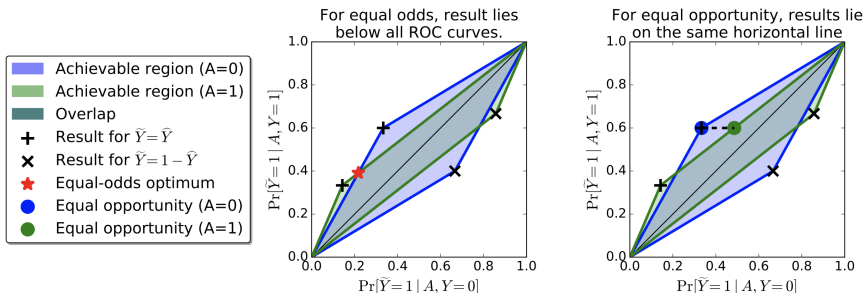
Similarly for $\gamma_{A=1}$.

- A predictor \hat{Y} satisfies equalized odds, if and only if $\gamma_{A=0} = \gamma_{A=1}$.
Second coordinate for equality of opportunity.
- Convex polytops:

$$P_{A=0} = \text{convhull}\{(0, 0), \gamma_{A=0}(\hat{Y}), \gamma_{A=0}(1 - \hat{Y}), (1, 1)\}$$

Similarly for $P_{A=1}$.

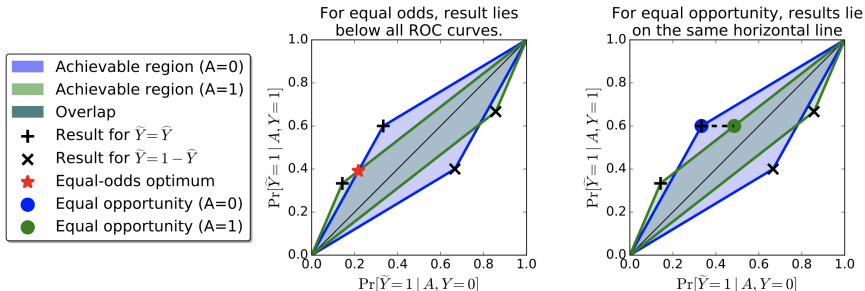
Post-processing



$$P_{A=0} = \text{convhull}\{(0,0), \gamma_{A=0}(\hat{Y}), \gamma_{A=0}(1 - \hat{Y}), (1,1)\}$$

$$P_{A=1} = \text{convhull}\{(0,0), \gamma_{A=1}(\hat{Y}), \gamma_{A=1}(1 - \hat{Y}), (1,1)\}$$

Post-processing

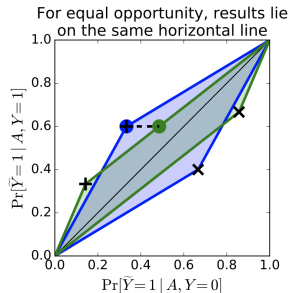
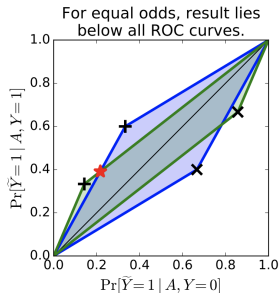
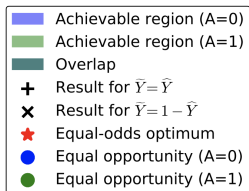


$$P_{A=0} = \text{convhull}\{(0,0), \gamma_{A=0}(\hat{Y}), \gamma_{A=0}(1 - \hat{Y}), (1,1)\}$$

$$P_{A=1} = \text{convhull}\{(0,0), \gamma_{A=1}(\hat{Y}), \gamma_{A=1}(1 - \hat{Y}), (1,1)\}$$

$P_{A=0}$ and $P_{A=1}$ characterize the trade-offs between false positives and true positives that we can achieve with any derived classifier.

Post-processing



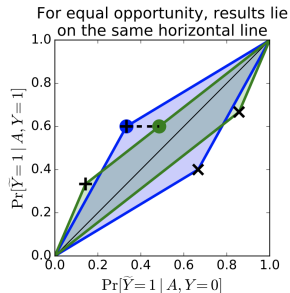
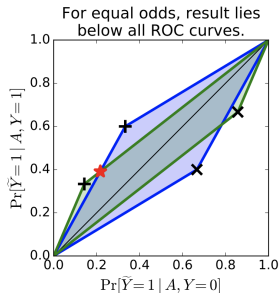
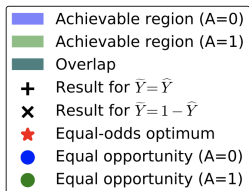
$$\underset{\bar{Y}}{\text{minimise}} \mathbb{E} l(\bar{Y}, Y)$$

$$s.t. \forall A \in \{0, 1\} : \gamma_A(\bar{Y}) \in P_A(\hat{Y})$$

$$\gamma_{A=0}(\bar{Y}) = \gamma_{A=1}(\bar{Y})$$

- a linear program

Post-processing



$$\underset{\bar{Y}}{\text{minimise}} \mathbb{E} l(\bar{Y}, Y)$$

$$s.t. \forall A \in \{0, 1\} : \gamma_A(\bar{Y}) \in P_A(\hat{Y})$$

$$\gamma_{A=0}(\bar{Y}) = \gamma_{A=1}(\bar{Y})$$

- a linear program
- solution is an optimal equalized odds predictor derived from \hat{Y} and A .

Sources of unfairness

Sources of unfairness

The problem can be divided into **two categories** (both types of bias can appear together):

Bias stemming from biased training data

Bias stemming from the algorithms themselves

Sources of unfairness

Bias stemming from biased training data

- **Sampling bias**: the data sample on which the algorithm is trained for is not representative of the overall population.

Chouldechova and Roth: The frontiers of fairness in machine learning, Oct 2018

Tolan: Fair and unbiased algorithmic decision making: current state and future challenges, Dec 2018

Sources of unfairness

Bias stemming from biased training data

- **Sampling bias**: the data sample on which the algorithm is trained for is not representative of the overall population.
E.g. training data contains most applicants from a certain region but the model is applied to the whole population.

Chouldechova and Roth: The frontiers of fairness in machine learning, Oct 2018

Tolan: Fair and unbiased algorithmic decision making: current state and future challenges, Dec 2018

Sources of unfairness

Bias stemming from biased training data

- **Sampling bias**: the data sample on which the algorithm is trained for is not representative of the overall population.
- **Selective labels**: only observe the outcome of one side of the decision.

Chouldechova and Roth: The frontiers of fairness in machine learning, Oct 2018

Tolan: Fair and unbiased algorithmic decision making: current state and future challenges, Dec 2018

Sources of unfairness

Bias stemming from biased training data

- **Sampling bias**: the data sample on which the algorithm is trained for is not representative of the overall population.
- **Selective labels**: only observe the outcome of one side of the decision.
E.g. in university admissions, we do not have data on performance of the applicants who were not admitted.

Chouldechova and Roth: The frontiers of fairness in machine learning, Oct 2018

Tolan: Fair and unbiased algorithmic decision making: current state and future challenges, Dec 2018

Sources of unfairness

Bias stemming from biased training data

- **Sampling bias**: the data sample on which the algorithm is trained for is not representative of the overall population.
- **Selective labels**: only observe the outcome of one side of the decision.
- **Proxy labels**

Sources of unfairness

Bias stemming from biased training data

- **Sampling bias**: the data sample on which the algorithm is trained for is not representative of the overall population.
- **Selective labels**: only observe the outcome of one side of the decision.
- **Proxy labels**
E.g. for predictive policing, we do not have data on who commits crimes, and only have data on who is arrested.

Chouldechova and Roth: The frontiers of fairness in machine learning, Oct 2018

Tolan: Fair and unbiased algorithmic decision making: current state and future challenges, Dec 2018

Sources of unfairness

Bias stemming from biased training data

Typical setup

X features of an individual

A sensitive attribute (race, gender, ...)

$C = C(X, A)$ classifier mapping X and A to some prediction

Y actual outcome

Sources of unfairness

Bias stemming from biased training data

Typical setup

X features of an individual

A sensitive attribute (race, gender, ...)

$C = C(X, A)$ classifier mapping X and A to some prediction

Y actual outcome

All of this is a lie

X incorporates all sorts of measurement biases

A often not even known, ill-defined, misreported, inferred

C often not well defined, e.g., large production ML system

Y often poor proxy of actual variable of interest

Sources of unfairness

Bias stemming from the algorithms themselves

- **Tyranny** of the majority

Sources of unfairness

Bias stemming from the algorithms themselves

- **Tyranny** of the majority

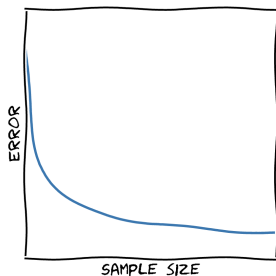
It is simpler to fit to the majority groups than to the minority groups because of generalization.

Sources of unfairness

Bias stemming from the algorithms themselves

- **Tyranny** of the majority

It is simpler to fit to the majority groups than to the minority groups because of generalization.



Generally, more data means smaller error

By definition, less data on minority groups.

Can lead to higher error rates on minority.

Sources of unfairness

Bias stemming from the algorithms themselves

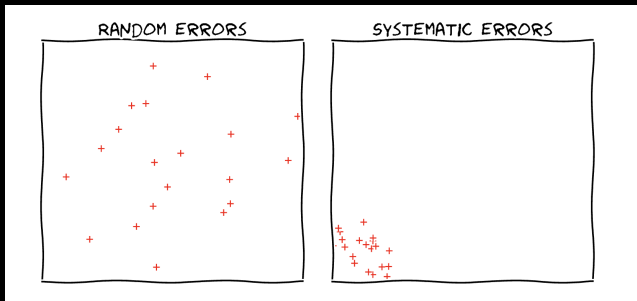
- **Tyranny** of the majority
- The meaning of **low error**

Sources of unfairness

Bias stemming from the algorithms themselves

- **Tyranny** of the majority
- The meaning of **low error**

Two classifiers with 5% average error:



Sources of unfairness

Bias stemming from the algorithms themselves

- **Tyranny** of the majority
- The meaning of **low error**
- **Feedback effects**

Chouldechova and Roth: The frontiers of fairness in machine learning, Oct 2018
Tolan: Fair and unbiased algorithmic decision making: current state and future challenges, Dec 2018

Sources of unfairness

Bias stemming from the algorithms themselves

- **Tyranny** of the majority
- The meaning of **low error**
- **Feedback effects**

Model at time $t + 1$ has to consider training data plus decisions of the model at time t .

Chouldechova and Roth: The frontiers of fairness in machine learning, Oct 2018
Tolan: Fair and unbiased algorithmic decision making: current state and future challenges, Dec 2018

Feedback effects

Model at time $t + 1$ has to consider training data plus decisions of the model at time t

Consider a bank's lending decision.

The outcome is not simply reject or accept the applicant for a loan. In fact, there are multiple effects of this decision.

If the applicant receives a loan, then goes on to successfully pay it back, then not only will the bank make a profit, but then the applicant's credit score will increase.

This will make future loan decisions more favorable to the applicant.

Feedback effects

Model at time $t + 1$ has to consider training data plus decisions of the model at time t

We could think of this in terms of loss functions.

The bank's objective was to maximize profit. After deciding to give the loan, the loan was repaid (with interest). The bank made money.

To maximize the objective (minimize the negative of the loss) the bank wants to give out as many loans that are likely to be repaid as possible.

Feedback effects

Model at time $t + 1$ has to consider training data plus decisions of the model at time t

We could think of this in terms of loss functions.

The bank's objective was to maximize profit. After deciding to give the loan, the loan was repaid (with interest). The bank made money.

To maximize the objective (minimize the negative of the loss) the bank wants to give out as many loans that are likely to be repaid as possible.

Our concern is that if the number of people who receive loans from one group outweigh the other, then as a whole, the credit rating will become a euphemism for race.

Examples

Sampling bias → the tyranny

In the **imSitu** situation recognition **dataset**, the activity cooking is over 33% more likely to involve females than males in a training set, and a trained algorithm further **amplifies** the disparity to 68%.

Zhao et al.: Men also like shopping, EMNLP 2017

The figure displays five examples of cooking scenes, each with a corresponding table of roles and values, and a probability distribution represented by colored squares and connecting lines.

- Example 1:** A woman in a green apron cooking in a kitchen.

COOKING	
ROLE	VALUE
AGENT	WOMAN
FOOD	PASTA
HEAT	STOVE
TOOL	SPATULA
PLACE	KITCHEN
- Example 2:** A woman in a blue shirt preparing food in a kitchen.

COOKING	
ROLE	VALUE
AGENT	WOMAN
FOOD	FRUIT
HEAT	∅
TOOL	KNIFE
PLACE	KITCHEN
- Example 3:** A woman in a blue tank top cooking outdoors.

COOKING	
ROLE	VALUE
AGENT	WOMAN
FOOD	MEAT
HEAT	STOVE
TOOL	SPATULA
PLACE	OUTSIDE
- Example 4:** A man in a white shirt and orange apron cooking in a kitchen.

COOKING	
ROLE	VALUE
AGENT	WOMAN
FOOD	∅
HEAT	STOVE
TOOL	SPATULA
PLACE	KITCHEN
- Example 5:** A man in a black shirt and apron cooking in a kitchen.

COOKING	
ROLE	VALUE
AGENT	MAN
FOOD	∅
HEAT	STOVE
TOOL	SPATULA
PLACE	KITCHEN

Examples

Sampling bias → the tyranny

The reason is: the algorithm predicts the gender from the activity and not from looking at the person.

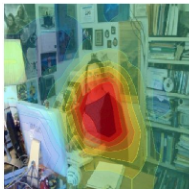
Anne Hendricks et al.: Women also snowboard, ECCV 2018

Wrong



Baseline:
*A **man** sitting at a desk with a laptop computer.*

Right for the Right Reasons



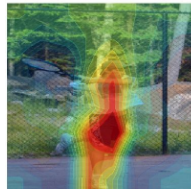
Our Model:
*A **woman** sitting in front of a laptop computer.*

Right for the Wrong Reasons



Baseline:
*A **man** holding a tennis racquet on a tennis court.*

Right for the Right Reasons



Our Model:
*A **man** holding a tennis racquet on a tennis court.*

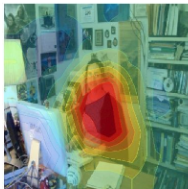
Examples

Wrong



Baseline:
*A **man** sitting at a desk with
a laptop computer.*

Right for the Right
Reasons



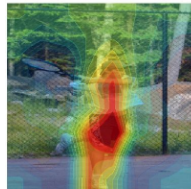
Our Model:
*A **woman** sitting in front of a
laptop computer.*

Right for the Wrong
Reasons



Baseline:
*A **man** holding a tennis
racquet on a tennis court.*

Right for the Right
Reasons



Our Model:
*A **man** holding a tennis
racquet on a tennis court.*

Anne Hendricks et al.: Women also snowboard, ECCV 2018

<https://github.com/kayburns/women-snowboard/tree/master/research/im2txt>

Examples

Sampling bias → the tyranny

- In the UCI Adult Income dataset, 30% of the male individuals earn more than 50K per year (high income), however of the female individuals only 11% have a high income.
- If an algorithm is trained on this data, the skewness ratio is amplified from 3:1 to 5:1.
- Simply removing sensitive attribute gender from the training data is not sufficient.

Examples

Skewed sample → feedback loop

- Future observations of crime confirm predictions
- Fewer opportunities to observe crime that contradicts predictions
- Initial bias may compound over time

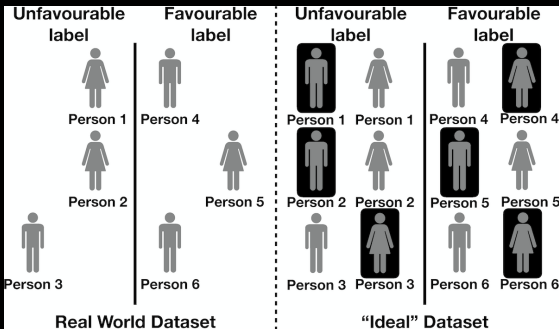
Transparency in fairness

Can we provide an individual-level explanation of fair systems without the difficult learning of fair (e.g. *genderless*) representations?

Transparency in fairness

Pre-processing with contrastive examples

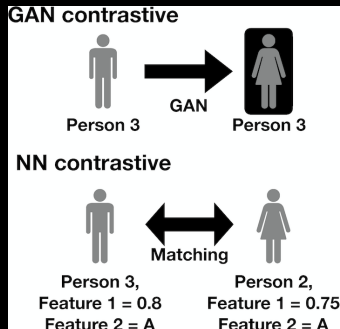
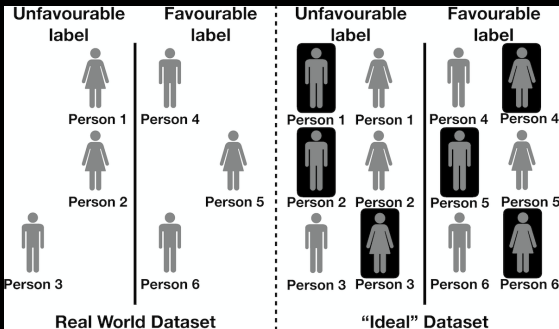
The ideal dataset contains an imaginary data point for each person, i.e. the one inside the black box, whereby we intervene and set the gender attribute to the opposite that is in real life.



Transparency in fairness

Pre-processing with contrastive examples

The ideal dataset contains an imaginary data point for each person, i.e. the one inside the black box, whereby we intervene and set the gender attribute to the opposite that is in real life.



Contrastive examples

- All previous work with adversarial learning try to **remove** protected attributes from data
- Instead, we use adversarial learning to **generate** data points with pre-specified protected attributes (contrastive examples)
- Contrastive examples "can be easily interpreted"

Real

GAN contrastive

NN contrastive

Male



Contrastive examples: StarGAN model objective

- A standard adversarial loss:

minimise $\underset{G}{\mathcal{L}_{adv}}$ maximise $\underset{D}{\mathcal{L}_{adv}}$

$$\mathcal{L}_{adv} = \mathbb{E}_x[\log D(x)] + \mathbb{E}_{x, \bar{s}}[\log(1 - D(G(x, \bar{s})))]$$

Contrastive examples: StarGAN model objective

- A standard adversarial loss:

$$\underset{G}{\text{minimise}} \underset{D}{\text{maximise}} \mathcal{L}_{adv}$$

$$\mathcal{L}_{adv} = \mathbb{E}_x[\log D(x)] + \mathbb{E}_{x, \bar{s}}[\log(1 - D(G(x, \bar{s})))]$$

- An auxiliary classifier D_{cls} to predict the correct attributes of the real samples:

$$\mathcal{L}_{cls}^{real} = \mathbb{E}_{x, s}[-\log D_{cls}(x, s)].$$

and to guide the generator to produce contrastive examples with correct attributes \bar{s} :

$$\mathcal{L}_{cls}^{contrastive} = \mathbb{E}_{x, \bar{s}}[-\log D_{cls}(G(x, \bar{s}), \bar{s})],$$

Contrastive examples: StarGAN model objective

- A standard adversarial loss:

$$\underset{G}{\text{minimise}} \underset{D}{\text{maximise}} \mathcal{L}_{adv}$$

$$\mathcal{L}_{adv} = \mathbb{E}_x[\log D(x)] + \mathbb{E}_{x,\bar{s}}[\log(1 - D(G(x, \bar{s})))]$$

- An auxiliary classifier D_{cls} to predict the correct attributes of the real samples:

$$\mathcal{L}_{cls}^{real} = \mathbb{E}_{x,s}[-\log D_{cls}(x, s)].$$

and to guide the generator to produce contrastive examples with correct attributes \bar{s} :

$$\mathcal{L}_{cls}^{contrastive} = \mathbb{E}_{x,\bar{s}}[-\log D_{cls}(G(x, \bar{s}), \bar{s})],$$

- A cycle consistency loss:

$$\mathcal{L}_{cyc} = \mathbb{E}_{x,\bar{s},s}[||G(G(x, \bar{s}), s) - x||_1].$$

Contrastive examples: StarGAN model

The full objective for training the StarGAN model for generating contrastive examples:

$$\mathcal{L}_G = \mathcal{L}_{adv} + \lambda_{cls} \mathcal{L}_{cls}^{contrastive} + \lambda_{cyc} \mathcal{L}_{cyc}$$

$$\mathcal{L}_D = -\mathcal{L}_{adv} + \lambda_{cls} \mathcal{L}_{cls}^{real}$$

where λ_{cls} and λ_{cyc} are hyper-parameters.

Choi et al, StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation, CVPR2018.

Code: <https://github.com/yunjey/stargan>

Contrastive examples: CelebA dataset

- We use **gender and age** as the two protected attributes.
- We use **smiling** as the classification task.

method	Acc.	TPR Diff.	FPR Diff.
logistic regression (original)	89.71	6.69	6.40
logistic regression (original and GAN contrastive)	88.94	3.50	2.79
logistic regression (original and NN contrastive)	88.78	3.32	3.53
‡ logistic regression (original and GAN contrastive with output consistency)	94.15	3.51	2.18

‡: **Rejection learning** -- classifier only makes a prediction if there is an agreement between original and contrastive examples (occurs in 17,237 out of 20,000 test examples, i.e. 86.185%).

What is fairness ML about?

Mentimeter



8

<https://www.mentimeter.com>

Acknowledgements

Some of the materials and results are taken from:

- Ziyuan Zhong: A Tutorial on Fairness in Machine Learning
- Novi Quadrianto: Deep Learning and Bayesian Methods Summer School 2019, University of Sussex, UK
- Krishna Gummadi, Manuel Gomez Rodriguez: Human-Centered Machine Learning course, Saarland University, MPI
- Moritz Hardt: Fairness in Machine Learning seminar, CS 294 UC Berkeley.

Thank you!