

COMP-SCI 396

Lecture 14: Algorithmic Bias

Do algorithms “discriminate”?

Algorithmic scoring guides decision-making in high stakes contexts:

- who should receive bail
- who should receive a medical treatment
- who should receive a loan
- who should be considered for a job

Much recent attention on how an algorithm's **errors** compare across social groups, and whether these errors are systematically borne by one group.

Example: Predicting Recidivism

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

- Highly publicized article in 2016 by ProPublica
- Algorithmic prediction tool mislabelled non-white defendants as future criminals twice as often as white defendants.

Example: Predicting Health Risks

COMPUTING

Racial Bias Found in a Major Health Care Risk Algorithm

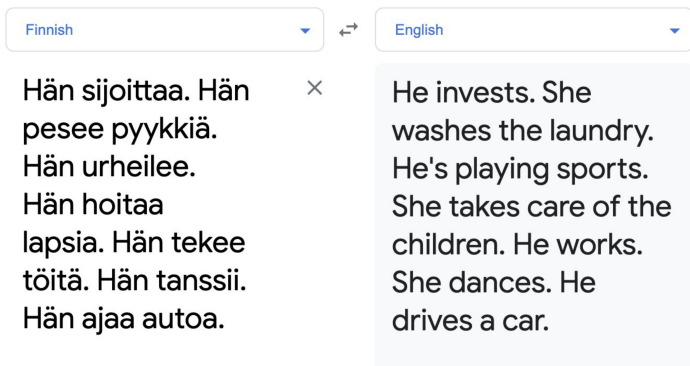
Black patients lose out on critical care when systems equate health needs with costs

By Starre Vartan on October 24, 2019

- Recent article in Science
- Among all patients classified as in need of high-risk health management/support, white individuals had 26.3% fewer chronic illnesses (i.e., were less sick)

Example: Google Translate

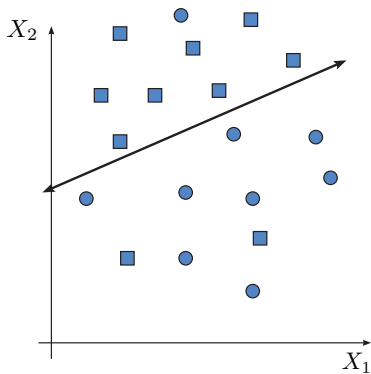
“Biased” translations from languages without gendered pronouns to languages with gendered pronouns.



The image shows a screenshot of the Google Translate interface. At the top, there are two dropdown menus for language selection: 'Finnish' on the left and 'English' on the right, with a double-headed arrow between them. Below the 'Finnish' menu, there is a list of Finnish sentences: 'Hän sijoittaa. Hän pesee pyykkiä.', 'Hän urheilee.', 'Hän hoitaa lapsia. Hän tekee töitä. Hän tanssii.', and 'Hän ajaa autoa.'. To the right of this list is a small 'x' icon. Below the 'English' menu, there is a light blue box containing the translated English text: 'He invests. She washes the laundry. He's playing sports. She takes care of the children. He works. She dances. He drives a car.'

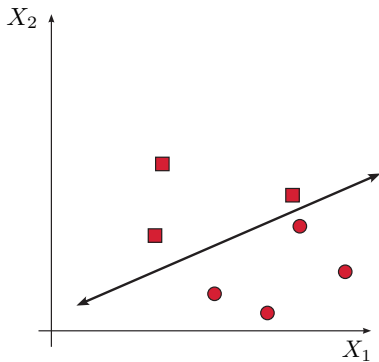
Why Might Algorithms Treat Groups Differently?

Different populations have different properties:



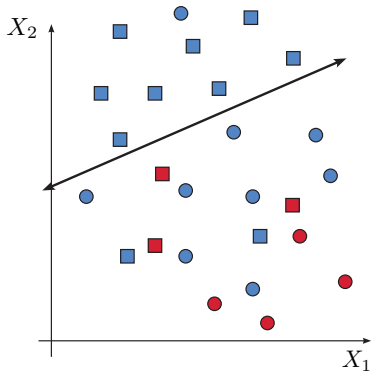
Why Might Algorithms Treat Groups Differently?

Different populations have different properties:



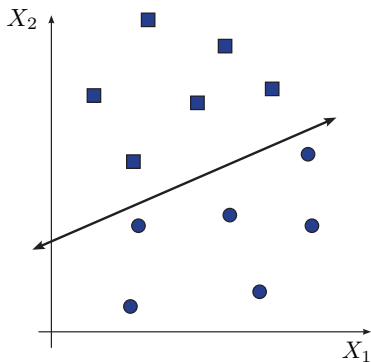
Why Might Algorithms Treat Groups Differently?

Different populations have different properties:



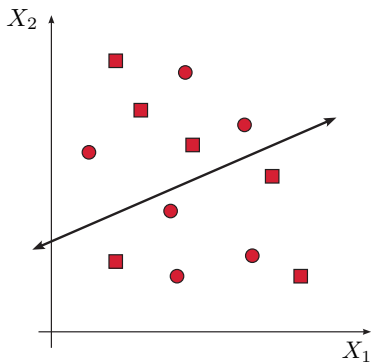
Why Might Algorithms Treat Groups Differently?

Different populations have different properties:



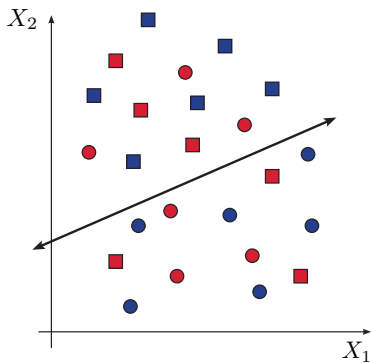
Why Might Algorithms Treat Groups Differently?

Different populations have different properties:



Why Might Algorithms Treat Groups Differently?

Different populations have different properties:



Why Might Algorithms Treat Groups Differently?

Endogeneity/selection:

- Training data includes historical biases, e.g.
 - label is not “committed a crime” but “was arrested”
 - potentially less data about minority groups
- Data collection feedback loops / selective data observation
 - only observe “committed a crime” if the subject was released
 - find more crimes at locations where more police officers are deployed

Other reasons?

Some Hard Questions

- 1 What does it mean for an algorithm to treat groups “unfairly?”

Some Hard Questions

- 1 What does it mean for an algorithm to treat groups “unfairly?”
- 2 What are the relevant groups over which fairness criteria should be defined?

Some Hard Questions

- 1 What does it mean for an algorithm to treat groups “unfairly?”
- 2 What are the relevant groups over which fairness criteria should be defined?
- 3 Is the “group” the right unit, or the individual?

Some Hard Questions

- 1 What does it mean for an algorithm to treat groups “unfairly?”
- 2 What are the relevant groups over which fairness criteria should be defined?
- 3 Is the “group” the right unit, or the individual?
- 4 How to trade off between fairness and accuracy?

Some Hard Questions

- 1 What does it mean for an algorithm to treat groups “unfairly?”
- 2 What are the relevant groups over which fairness criteria should be defined?
- 3 Is the “group” the right unit, or the individual?
- 4 How to trade off between fairness and accuracy?
- 5 What kind of features should we permit algorithms access to: should group labels be forbidden?

Framework

Each individual in the population is described by

- a feature vector $X = (X_1, \dots, X_n)$ belonging to \mathcal{X}
- a group membership $G \in \{b, r\}$
- a type $Y \in \{0, 1\}$ (will commit another crime versus won't)

Formally, $(X, G, Y) \sim \mathbb{P}$ is a random vector.

Example

Group (G)	Age (X_1)	Freq. Irreg. Heartbeats (X_2)	Arrythmia (Y)
r	50	13%	1
b	32	7%	0
b	80	20%	1
r	78	28 %	1
b	39	15%	0
b	50	24%	0

The Distribution \mathbb{P}

The distribution \mathbb{P} can reflect a prior belief, or the empirical distribution of observed data.

May be “asymmetries” depending on how the data was collected.

The Distribution \mathbb{P}

The distribution \mathbb{P} can reflect a prior belief, or the empirical distribution of observed data.

May be “asymmetries” depending on how the data was collected.

- e.g., could be that there is very little representation of group- r observations in the observed data

The Distribution \mathbb{P}

The distribution \mathbb{P} can reflect a prior belief, or the empirical distribution of observed data.

May be “asymmetries” depending on how the data was collected.

- e.g., could be that there is very little representation of group- r observations in the observed data
- or, could be that groups r and b are equally represented, but the observed data is of higher quality for group r (more noise in the data for group b)

The Distribution \mathbb{P}

The distribution \mathbb{P} can reflect a prior belief, or the empirical distribution of observed data.

May be “asymmetries” depending on how the data was collected.

- e.g., could be that there is very little representation of group- r observations in the observed data
- or, could be that groups r and b are equally represented, but the observed data is of higher quality for group r (more noise in the data for group b)
- or, there may be certain correlations between group and other traits (e.g., if group r individuals are only observed if they have some other trait)

Algorithmic Scoring Rule

An algorithmic scoring rule is any function $S : \mathcal{X} \rightarrow \{0, 1\}$.

This maps observed features into a prediction of Y .

For example,

- we might map an individual's medical profile into an assessment of whether they are at high or low risk of a particular illness
- we might map a defendant's background and record into a prediction of whether they are at high or low risk of criminal reoffense
- we might map a job applicant's record and interview outcomes into a prediction of whether they are likely or unlikely to perform well on the job

False Positives and False Negatives

Every scoring rule S can be identified with a table like this:

	$S = 0$	$S = 1$
$Y = 0$	<i>TN</i>	<i>FP</i>
$Y = 1$	<i>FN</i>	<i>TP</i>

where:

- TN = True Negative
- FP = False Positive
- FN = False Negative
- TP = True Positive

What Makes a Scoring Rule “Fair”?

Statistical Parity

Definition

Say that S satisfies statistical parity if

$$\mathbb{P}(S = 1 \mid G = g) = \mathbb{P}(S = 1) \quad \text{for both groups } g$$

- This says that the probability of being scored 1 is the same for members of both groups
- Example violation: 40% of group r individuals receive score 1 and 80% of group b individuals receive score 1

Let's Discuss

- This fairness criterion doesn't take into account the true type Y at all
- So it makes more sense when everyone prefers the outcome (e.g., a job or a loan), rather than something when the value of the outcome depends on on your type (e.g., only want to get medical procedures that you actually need it)
- If 40% of group r individuals need a given medical treatment but no one in group b needs it, statistical parity is probably not desirable

Let's Discuss

- This fairness criterion doesn't take into account the true type Y at all
- So it makes more sense when everyone prefers the outcome (e.g., a job or a loan), rather than something when the value of the outcome depends on on your type (e.g., only want to get medical procedures that you actually need it)
- If 40% of group r individuals need a given medical treatment but no one in group b needs it, statistical parity is probably not desirable
- Also has the unfortunate property that it can be satisfied in a bunch of undesirable ways: e.g., put everyone in jail.

Let's Discuss

- This fairness criterion doesn't take into account the true type Y at all
- So it makes more sense when everyone prefers the outcome (e.g., a job or a loan), rather than something when the value of the outcome depends on on your type (e.g., only want to get medical procedures that you actually need it)
- If 40% of group r individuals need a given medical treatment but no one in group b needs it, statistical parity is probably not desirable
- Also has the unfortunate property that it can be satisfied in a bunch of undesirable ways: e.g., put everyone in jail.
- Incentive problems?

Conditional Statistical Parity

Definition

Say that S satisfies conditional statistical parity if

$$\mathbb{P}(S = 1 \mid G = g, X = x) = \mathbb{P}(S = 1 \mid X = x)$$

for any $g \in \{r, b\}$ and $x \in \mathcal{X}$.

- If two individuals have the same features but different group memberships, their probability of being scored 1 is the same
- Example violation: fixing a given resume, the probability of being hired is 40% if the individual belongs to group r and 80% if the individual belongs to group b

Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination

By MARIANNE BERTRAND AND SENDHIL MULLAINATHAN*

We study race in the labor market by sending fictitious resumes to help-wanted ads in Boston and Chicago newspapers. To manipulate perceived race, resumes are randomly assigned African-American- or White-sounding names. White names receive 50 percent more callbacks for interviews. Callbacks are also more responsive to resume quality for White names than for African-American ones. The racial gap is uniform across occupation, industry, and employer size. We also find little evidence that employers are inferring social class from the names. Differential treatment by race still appears to still be prominent in the U.S. labor market. (JEL J71, J64).

Apple Card Investigated After Gender Discrimination Complaints

A prominent software developer said on Twitter that the credit card was “sexist” against women applying for credit.

“My wife and I filed joint tax returns, live in a community-property state, and have been married for a long time,” Mr. Hansson wrote Thursday on Twitter. “Yet Apple’s black box algorithm thinks I deserve 20x the credit limit she does.”

Let's Discuss

- What this definition means depends a lot on what the features X are.
- If we have a lot of fine-grained information on individuals, then each unique realization of X might identify a single person, making this fairness definition vacuous.
- Could generalize to a selected subset of features, or put a similarity metric on X and require individuals with similar features to have similar scores.

Let's Discuss

- But could also be that (say, because of omitted features) the same realization of X carries different meanings for individuals in the two groups.
- e.g., suppose that the groups are low- and high- income categories, X is frequency of hospital visits, Y is whether you have a serious chronic illness.
- Could be that low-income individuals have a harder time getting to sick leave to visit the hospital, so conditional on the same Y they visit the hospital less.
- In this case, might not want to insist on similar treatments for individuals with the same X 's across the two groups.

Let's Discuss

- Previous definition asked for similar treatment of individuals with the same feature vectors X
- Another idea is to ask for similar treatment of individuals with the same type Y

Equality of False Positive Rates

	$S = 0$	$S = 1$
$Y = 0$	TN	FP
$Y = 1$	FN	TP

Definition

Say that S has equal false positive rates if

$$P(S = 1 \mid Y = 0, G = b) = P(S = 1 \mid Y = 0, G = r)$$

- This says that the probability that an individual with $Y = 0$ is wrongly classified as $S = 1$ is the same within each group
- Example violation: 40% of innocent group- b defendants are wrongly scored as high-risk of criminal offense, while 80% of innocent group- r defendants are wrongly scored in this way.

Equality of TNR and FPR

	$S = 0$	$S = 1$
$Y = 0$	TN	FP
$Y = 1$	FN	TP

Since this implies (and is implied by) equality of the true negative rate, we can equivalently say

$$S \perp\!\!\!\perp G | Y = 0$$

e.g., conditional on **not** committing a crime in the future, probability of being assessed as high-risk of committing a crime is the same regardless of group membership.

Equality in False Negatives

Definition

Say that S has equal false negative rates if

$$P(S = 0 \mid Y = 1, G = b) = P(S = 0 \mid Y = 1, G = r)$$

- This says that the probability that an individual with $Y = 1$ is wrongly classified as $S = 0$ is the same within each group
- Example violation: 40% of group- b people who need intensive medical care are wrongly scored as “low-risk” while 80% of group- b individuals who need intensive medical care are wrongly scored in this way.

Equality of TPR and FNR

	$S = 0$	$S = 1$
$Y = 0$	TN	FP
$Y = 1$	FN	TP

Since this implies (and is implied by) equality of the true positive rate, we can equivalently say

$$S \perp\!\!\!\perp G | Y = 1$$

e.g., conditional on repaying the loan, then the probability of being granted the loan is the same regardless of group membership.

Equalized Odds

	$S = 0$	$S = 1$
$Y = 0$	TN	FP
$Y = 1$	FN	TP

We could also ask for all of the four cells to be the same.

Definition

Say that S satisfies equalized odds if

$$\mathbb{E}_Y \left[\underbrace{\mathbb{E}_S(S \mid G = r, Y) - \mathbb{E}_S(S \mid G = b, Y)}_{\text{difference with which type } Y \text{ individuals are treated across groups}} \right] = 0$$

Let's Discuss

- Two groups can have equal TNR, FPR, TPR, FNR, and still have rather different overall accuracy rates
- e.g., suppose in both groups we have

$$\begin{array}{cc} & S = 0 & S = 1 \\ Y = 0 & 1/2 & 1/2 \\ Y = 1 & 0 & 1 \end{array}$$

but 90% of individuals in group r have $Y = 1$ while 10% of individuals in group b have $Y = 0$

Let's Discuss

- Two groups can have equal TNR, FPR, TPR, FNR, and still have rather different overall accuracy rates
- e.g., suppose in both groups we have

$$\begin{array}{cc} & S = 0 & S = 1 \\ Y = 0 & 1/2 & 1/2 \\ Y = 1 & 0 & 1 \end{array}$$

but 90% of individuals in group r have $Y = 1$ while 10% of individuals in group b have $Y = 0$

- Then the overall error rate for group r is

$$0.1 \cdot 1/2 + 0.9 \cdot 0 = 0.05$$

while the overall error rate for group b is

$$0.9 \cdot 1/2 + 0.1 \cdot 0 = 0.45.$$

Let's Discuss

- Two groups can have equal TNR, FPR, TPR, FNR, and still have rather different overall accuracy rates
- e.g., suppose in both groups we have

$$\begin{array}{rcc} & S = 0 & S = 1 \\ Y = 0 & 1/2 & 1/2 \\ Y = 1 & 0 & 1 \end{array}$$

but 90% of individuals in group r have $Y = 1$ while 10% of individuals in group b have $Y = 0$

- Then the overall error rate for group r is

$$0.1 \cdot 1/2 + 0.9 \cdot 0 = 0.05$$

while the overall error rate for group b is

$$0.9 \cdot 1/2 + 0.1 \cdot 0 = 0.45.$$

- The base rate of Y in the two groups is not a factor.

Let's Discuss

- Next set of fairness criteria instead condition on the score S
- Instead of asking for the distribution of S to be similar for individuals with similar Y , ask that the distribution of Y is similar for individuals with similar S

Those Scored $S = 1$

Definition

S satisfies equality of positive predictive values if

$$P(Y = 1 \mid S = 1, G = r) = P(Y = 1 \mid S = 1, G = b)$$

Example violation: 40% of group- b loan applicants who are scored as “creditworthy” in fact pay back their loan, while 80% of group- r loan applicants scored as “creditworthy” pay back their loan.

Those Scored $S = 1$

Definition

S satisfies equality of positive predictive values if

$$P(Y = 1 \mid S = 1, G = r) = P(Y = 1 \mid S = 1, G = b)$$

Example violation: 40% of group- b loan applicants who are scored as “creditworthy” in fact pay back their loan, while 80% of group- r loan applicants scored as “creditworthy” pay back their loan.

This implies:

Definition

S satisfies equality of false discovery rates if

$$P(Y = 0 \mid S = 1, G = r) = P(Y = 0 \mid S = 1, G = b)$$

Example violation: 40% of group- b job applicants who are hired perform poorly at the job, while 80% of group- r job applicants who are hired perform poorly at the job.

Those Scored $S = 0$

Definition

S satisfies equality of negative predictive values if

$$P(Y = 0 \mid S = 0, G = r) = P(Y = 0 \mid S = 0, G = b)$$

Example violation: 40% of group- b loan applicants who are scored as “not creditworthy” in fact don’t pay back their loan, while 80% of group- r loan applicants scored as “not creditworthy” don’t.

Those Scored $S = 0$

Definition

S satisfies equality of negative predictive values if

$$P(Y = 0 \mid S = 0, G = r) = P(Y = 0 \mid S = 0, G = b)$$

Example violation: 40% of group- b loan applicants who are scored as “not creditworthy” in fact don’t pay back their loan, while 80% of group- r loan applicants scored as “not creditworthy” don’t.

This implies:

Definition

S satisfies equality of false omission rates if

$$P(Y = 1 \mid S = 0, G = r) = P(Y = 1 \mid S = 0, G = b)$$

Example violation: 40% of group- b job applicants who are not hired would have performed well at the job, while 80% of group- r job applicants who are not hired would have performed well.

Calibration

Combining these...

Definition (Calibrated)

A score S is calibrated if for each $s \in \{0, 1\}$,

$$P(Y = 1 \mid S = s, G = b) = P(Y = 1 \mid S = s, G = r)$$

Fixing an outcome (e.g., being hired or not being hired), the two groups look the same in distribution.

Let's Discuss

- This is a lot of criteria, and I won't blame you if your head started hurting a while back.

Let's Discuss

- This is a lot of criteria, and I won't blame you if your head started hurting a while back.
- Why does this matter? These all sound reasonable and kind of similar, don't many of these criteria imply one another?

Let's Discuss

- This is a lot of criteria, and I won't blame you if your head started hurting a while back.
- Why does this matter? These all sound reasonable and kind of similar, don't many of these criteria imply one another?
- It's actually the opposite.

Recall This Incident

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

- ProPublica showed that the algorithmic scoring rule had a false positive rate that was twice as high for one group as for another, and thus wasn't fair.

Recall This Incident

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

- ProPublica showed that the algorithmic scoring rule had a false positive rate that was twice as high for one group as for another, and thus wasn't fair.
- The company that made the algorithmic scoring responded that they actually made sure to make the algorithm fair, by ensuring calibration.

Recall This Incident

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

- ProPublica showed that the algorithmic scoring rule had a false positive rate that was twice as high for one group as for another, and thus wasn't fair.
- The company that made the algorithmic scoring responded that they actually made sure to make the algorithm fair, by ensuring calibration.

Researchers subsequently showed that these two notions of fairness are fundamentally incompatible with one another:

- Outside of trivial cases, an algorithm cannot both satisfy equality of false positive rates and also be calibrated!

Impossibility Result

For each group g , define $p_g = P(Y = 1 \mid G = g)$ to be the base rate of $Y = 1$ in each group.

Proposition (Chouldechova, 2016; Kleinberg et al, 2016)

Suppose $p_r \neq p_b$. Then no scoring rule S can simultaneously satisfy calibration, equal false positive rates, and equal false negative rates.

Proof

Choose either group g , and define $p_g = P(Y = 1 \mid G = g)$,

$$FP_g = P(S = 1 \mid Y = 0, G = g)$$

$$FN_g = P(S = 0 \mid Y = 1, G = g)$$

$$PPV_g = P(Y = 1 \mid S = 1, G = g)$$

Proof

Choose either group g , and define $p_g = P(Y = 1 | G = g)$,

$$FP_g = P(S = 1 | Y = 0, G = g)$$

$$FN_g = P(S = 0 | Y = 1, G = g)$$

$$PPV_g = P(Y = 1 | S = 1, G = g)$$

Lemma

$$FP_g = \frac{p_g}{1-p_g} \times \frac{1-PPV_g}{PPV_g} \times (1 - FN_g)$$

Proof

Choose either group g , and define $p_g = P(Y = 1 | G = g)$,

$$FP_g = P(S = 1 | Y = 0, G = g)$$

$$FN_g = P(S = 0 | Y = 1, G = g)$$

$$PPV_g = P(Y = 1 | S = 1, G = g)$$

Lemma

$$FP_g = \frac{p_g}{1-p_g} \times \frac{1-PPV_g}{PPV_g} \times (1 - FN_g)$$

Proof. Expanding the statement, we have

$$\begin{aligned} P(S = 1 | Y = 0, G = g) &= \frac{P(Y = 1 | G = g)}{P(Y = 0 | G = g)} \\ &\times \frac{P(Y = 0 | S = 1, G = g)}{P(Y = 1 | S = 1, G = g)} \times P(S = 1 | Y = 1, G = g) \end{aligned}$$

Proof

$$\begin{aligned} P(S = 1 \mid Y = 0, G = g) &= \frac{P(Y = 1 \mid G = g)}{P(Y = 0 \mid G = g)} \\ &\times \frac{P(Y = 0 \mid S = 1, G = g)}{P(Y = 1 \mid S = 1, G = g)} \times P(S = 1 \mid Y = 1, G = g) \end{aligned}$$

Proof

$$P(S = 1 \mid Y = 0, G = g) = \frac{P(Y = 1 \mid G = g)}{P(Y = 0 \mid G = g)} \\ \times \frac{P(Y = 0 \mid S = 1, G = g)}{P(Y = 1 \mid S = 1, G = g)} \times P(S = 1 \mid Y = 1, G = g)$$

Multiplying both sides by $P(Y = 0 \mid G = g)$ and invoking Bayes' rule, this expression simplifies to

$$P(S = 1, Y = 0 \mid G = g) \\ = \frac{P(Y = 0 \mid S = 1, G = g)}{P(Y = 1 \mid S = 1, G = g)} \times P(S = 1, Y = 1 \mid G = g)$$

Proof

$$P(S = 1 \mid Y = 0, G = g) = \frac{P(Y = 1 \mid G = g)}{P(Y = 0 \mid G = g)} \\ \times \frac{P(Y = 0 \mid S = 1, G = g)}{P(Y = 1 \mid S = 1, G = g)} \times P(S = 1 \mid Y = 1, G = g)$$

Multiplying both sides by $P(Y = 0 \mid G = g)$ and invoking Bayes' rule, this expression simplifies to

$$P(S = 1, Y = 0 \mid G = g) \\ = \frac{P(Y = 0 \mid S = 1, G = g)}{P(Y = 1 \mid S = 1, G = g)} \times P(S = 1, Y = 1 \mid G = g)$$

Thus, the statement is equivalent to

$$\frac{P(S = 1, Y = 0 \mid G = g)}{P(S = 1, Y = 1 \mid G = g)} = \frac{P(Y = 0 \mid S = 1, G = g)}{P(Y = 1 \mid S = 1, G = g)}$$

Proof

$$\frac{P(S = 1, Y = 0 \mid G = g)}{P(S = 1, Y = 1 \mid G = g)} = \frac{P(Y = 0 \mid S = 1, G = g)}{P(Y = 1 \mid S = 1, G = g)}$$

Proof

$$\frac{P(S = 1, Y = 0 \mid G = g)}{P(S = 1, Y = 1 \mid G = g)} = \frac{P(Y = 0 \mid S = 1, G = g)}{P(Y = 1 \mid S = 1, G = g)}$$

But the RHS can be rewritten

$$\frac{P(Y = 0 \mid S = 1, G = g)}{P(Y = 1 \mid S = 1, G = g)}$$

Proof

$$\frac{P(S = 1, Y = 0 \mid G = g)}{P(S = 1, Y = 1 \mid G = g)} = \frac{P(Y = 0 \mid S = 1, G = g)}{P(Y = 1 \mid S = 1, G = g)}$$

But the RHS can be rewritten

$$\begin{aligned} & \frac{P(Y = 0 \mid S = 1, G = g)}{P(Y = 1 \mid S = 1, G = g)} \\ &= \frac{P(S = 1, Y = 0 \mid G = g)/P(S = 1 \mid G = g)}{P(S = 1, Y = 1 \mid G = g)/P(S = 1 \mid G = g)} \end{aligned}$$

Proof

$$\frac{P(S = 1, Y = 0 \mid G = g)}{P(S = 1, Y = 1 \mid G = g)} = \frac{P(Y = 0 \mid S = 1, G = g)}{P(Y = 1 \mid S = 1, G = g)}$$

But the RHS can be rewritten

$$\begin{aligned} & \frac{P(Y = 0 \mid S = 1, G = g)}{P(Y = 1 \mid S = 1, G = g)} \\ &= \frac{P(S = 1, Y = 0 \mid G = g) / P(S = 1 \mid G = g)}{P(S = 1, Y = 1 \mid G = g) / P(S = 1 \mid G = g)} \\ &= \frac{P(S = 1, Y = 0 \mid G = g)}{P(S = 1, Y = 1 \mid G = g)} \end{aligned}$$

Proof

$$\frac{P(S = 1, Y = 0 \mid G = g)}{P(S = 1, Y = 1 \mid G = g)} = \frac{P(Y = 0 \mid S = 1, G = g)}{P(Y = 1 \mid S = 1, G = g)}$$

But the RHS can be rewritten

$$\begin{aligned} & \frac{P(Y = 0 \mid S = 1, G = g)}{P(Y = 1 \mid S = 1, G = g)} \\ &= \frac{P(S = 1, Y = 0 \mid G = g) / P(S = 1 \mid G = g)}{P(S = 1, Y = 1 \mid G = g) / P(S = 1 \mid G = g)} \\ &= \frac{P(S = 1, Y = 0 \mid G = g)}{P(S = 1, Y = 1 \mid G = g)} \end{aligned}$$

So the statement is equivalent to

$$\frac{P(S = 1, Y = 0 \mid G = g)}{P(S = 1, Y = 1 \mid G = g)} = \frac{P(S = 1, Y = 0 \mid G = g)}{P(S = 1, Y = 1 \mid G = g)}$$

and must therefore always be true.

Proof

We've shown

$$FP_r = \frac{p_r}{1 - p_r} \times \frac{1 - PPV_r}{PPV_r} \times (1 - FN_r)$$
$$FP_b = \frac{p_b}{1 - p_b} \times \frac{1 - PPV_b}{PPV_b} \times (1 - FN_b)$$

Proof

We've shown

$$FP_r = \frac{p_r}{1 - p_r} \times \frac{1 - PPV_r}{PPV_r} \times (1 - FN_r)$$
$$FP_b = \frac{p_b}{1 - p_b} \times \frac{1 - PPV_b}{PPV_b} \times (1 - FN_b)$$

Thus if

$$FP_r = FP_b$$

$$PPV_r = PPV_b$$

$$FN_r = FN_b$$

Proof

We've shown

$$FP_r = \frac{p_r}{1 - p_r} \times \frac{1 - PPV_r}{PPV_r} \times (1 - FN_r)$$
$$FP_b = \frac{p_b}{1 - p_b} \times \frac{1 - PPV_b}{PPV_b} \times (1 - FN_b)$$

Thus if

$$FP_r = FP_b$$
$$PPV_r = PPV_b$$
$$FN_r = FN_b$$

it must follow that $p_r = p_b$!

Proof

We've shown

$$FP_r = \frac{p_r}{1 - p_r} \times \frac{1 - PPV_r}{PPV_r} \times (1 - FN_r)$$
$$FP_b = \frac{p_b}{1 - p_b} \times \frac{1 - PPV_b}{PPV_b} \times (1 - FN_b)$$

Thus if

$$FP_r = FP_b$$
$$PPV_r = PPV_b$$
$$FN_r = FN_b$$

it must follow that $p_r = p_b$!

i.e., if calibration, equal false positive rates, and equal false negative rates are all satisfied, the base rate of $Y = 1$ must be identical across groups.

Proof

Converse:

If the base rate of $Y = 1$ is **not** identical, then at least one of calibration, equal false positive rates, and equal false negative rates must fail.

Fairness Gerrymandering

- Previous definition was defined given a primitive set of groups.
But what is the relevant set of groups?

Fairness Gerrymandering

- Previous definition was defined given a primitive set of groups. But what is the relevant set of groups?
- Kearns et al. (2008) on “fairness gerrymandering”: Suppose each individual is equally likely to be red or blue, and square or triangle.
- There are four protected groups: “red,” “blue,” “square,” and “triangle.”

Fairness Gerrymandering

- Previous definition was defined given a primitive set of groups. But what is the relevant set of groups?
- Kearns et al. (2008) on “fairness gerrymandering”: Suppose each individual is equally likely to be red or blue, and square or triangle.
- There are four protected groups: “red,” “blue,” “square,” and “triangle.”
- An ML classifier has false positive rates given as follows:

	square	triangle
red	1	0
blue	0	1

Fairness Gerrymandering

- Previous definition was defined given a primitive set of groups. But what is the relevant set of groups?
- Kearns et al. (2008) on “fairness gerrymandering”: Suppose each individual is equally likely to be red or blue, and square or triangle.
- There are four protected groups: “red,” “blue,” “square,” and “triangle.”
- An ML classifier has false positive rates given as follows:

	square	triangle
red	1	0
blue	0	1
- Then false positive rates across the protected groups are the same (50%), but clearly this is violated if we look at group conjunctions.

Individual Fairness

- Another approach is to define fairness over individuals instead of groups.

Individual Fairness

- Another approach is to define fairness over individuals instead of groups.
- Dwork et al. (2012): Fix a metric d on the space of covariates \mathcal{X} . Let A be a space of actions/outcomes. Fix a metric D on the space of measures $\Delta(A)$.

Individual Fairness

- Another approach is to define fairness over individuals instead of groups.
- Dwork et al. (2012): Fix a metric d on the space of covariates \mathcal{X} . Let A be a space of actions/outcomes. Fix a metric D on the space of measures $\Delta(A)$.

Definition

A mapping $M : \mathcal{X} \rightarrow \Delta(A)$ satisfies the (D, d) -Lipschitz property if

$$D(M(x), M(x')) \leq d(x, x') \quad \forall x, x' \in \mathcal{X}$$

i.e., similar individuals should receive similar distributions over outcomes.

Individual Fairness

- Another approach is to define fairness over individuals instead of groups.
- Dwork et al. (2012): Fix a metric d on the space of covariates \mathcal{X} . Let A be a space of actions/outcomes. Fix a metric D on the space of measures $\Delta(A)$.

Definition

A mapping $M : \mathcal{X} \rightarrow \Delta(A)$ satisfies the (D, d) -Lipschitz property if

$$D(M(x), M(x')) \leq d(x, x') \quad \forall x, x' \in \mathcal{X}$$

i.e., similar individuals should receive similar distributions over outcomes.

- But this essentially kicks the can down to the question of what is d . (Is similarity in group membership part of the definition of d ?)