Building a Robot Judge: Data Science for Decision-Making

11. Fairness in AI-Supported Decision-Making

Weekly Q&A

Recap: Brazil Corruption Study

`https://padlet.com/eash44/80ofv2plt41gv7v7`

Recap: Brazil Corruption Study

Mechanism Design Issues

- With repeated audits, there could be behavioral responses by local officials.
  - could produce significant errors favoring savvy mayors.
  - Would still deter corrupt fiscal actions that are not easily substitutable.

Recap: Brazil Corruption Study

How much information to publicize about audit targeting?

Recap: Brazil Corruption Study

How much information to publicize about audit targeting?

Option 1: Give **full information** about the policy and the associated model weights.

Recap: Brazil Corruption Study

How much information to publicize about audit targeting?

Option 1: Give **full information** about the policy and the associated model weights.

- Would increase deterrence against corruption actions captured by the model, that are not substitutable.
- But would make gaming the system easier.

Recap: Brazil Corruption Study

How much information to publicize about audit targeting?

Option 1: Give **full information** about the policy and the associated model weights.

- Would increase deterrence against corruption actions captured by the model, that are not substitutable.
- But would make gaming the system easier.

Option 2: Give **no information** about how targeting is done.

- This is "the industry approach", e.g., for how google/facebook detect violations.

Recap: Brazil Corruption Study

How much information to publicize about audit targeting?

Option 1: Give **full information** about the policy and the associated model weights.

- Would increase deterrence against corruption actions captured by the model, that are not substitutable.
- But would make gaming the system easier.

Option 2: Give **no information** about how targeting is done.

- This is “the industry approach”, e.g., for how google/facebook detect violations.
- mayors might learn how algorithm works over time.
- weights could be updated in response to behavioral responses

Recap: Brazil Corruption Study

Mixing random and targeted audits

- Random audits could be maintained (along with targeted audits).
    - Preserves some deterrence incentive for all municipalities.
    - Results of random audits could be used to update algorithm parameters.

Claudio Ferraz
@claudferraz

1/3 I just came across this very interesting work by @ellliottt @sergallet and @T_Giommoni using Machine Learning to predict corrupt practices in Brazil's municipalities. They show that a ML prediction algorithm can be more effective than a random auditing....

> **Sergio Galletta** @sergallet · May 1
> In a newly released WP, together with @ellliottt and @T_Giommoni, we show how ML techniques can be used to overcome data limitations when performing policy evaluation
>
> papers.ssrn.com/sol3/papers.cf...
> Show this thread

1:03 AM · Nov 29, 2020 · Twitter Web App

10 Likes

Claudio Ferraz @claudferraz · 9h
Replying to @claudferraz
2/3 But I think they miss an important point for the practical use of ML. The random audit was politically neutral and this is why it was credible to begin with. With a ML the estimated risk based on an algorithm can, in principle, be manipulated to target places or parties

1 5

Claudio Ferraz @claudferraz · 9h
3/3 So an important discussion is how to make these ML algorithms politically unbiased and how to gain credibility and convince government officials that using these types of algorithms for policy can generate important gains in the fight against corruption

What if the AI is biased toward one of the political parties?

Outline

Overview of Fairness Problem

Formalizing Fair Classification

Pre-Processing Data to Improve Fairness

Constrained Machine Learning

“Fair ML” / “AI Fairness”

“Fair ML” / “AI Fairness”

- “ML” or “AI” refer to statistical algorithms
  - can learning algorithms be fair or not?

“Fair ML” / “AI Fairness”

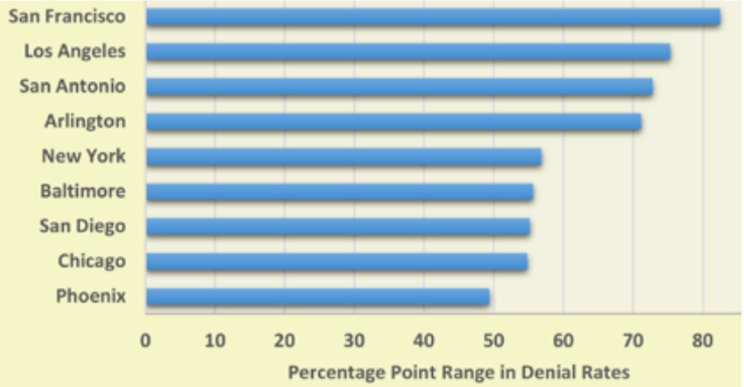
- “ML” or “AI” refer to statistical algorithms
  - can learning algorithms be fair or not?
- Rather: *fairness* is a property of *decisions*.
  - so *“AI Fairness”* should be understood as “*fairness of AI-supported decision-making*”.

“Fair ML” / “AI Fairness”

- “ML” or “AI” refer to statistical algorithms
  - can learning algorithms be fair or not?
- Rather: *fairness* is a property of *decisions*.
  - so *“AI Fairness”* should be understood as “*fairness of AI-supported decision-making*”.
- There is growing concern about social harms and disparities produced by AI decisions.
  - today: disparities in treatment
  - week 12: interpretability/explainability
  - week 13: broader social harms / policy

Humans are Inconsistent

- Before getting into bias towards particular groups, it should be emphasized that humans are "biased" in the sense that some are more/less lenient:



- A robot judge would generate consistent decisions for same evidence, correcting individual-level leniencies across judges.

Examples

- Lending laws (e.g. in the United States) prohibit practices that discriminate on the basis of race.
  - including algorithmic decision-making

Examples

- Lending laws (e.g. in the United States) prohibit practices that discriminate on the basis of race.
  - including algorithmic decision-making
- Firms using ML to screen job applicants might wish to incorporate diversity objectives.

Examples

- Lending laws (e.g. in the United States) prohibit practices that discriminate on the basis of race.
  - including algorithmic decision-making
- Firms using ML to screen job applicants might wish to incorporate diversity objectives.
- Judges might want to reduce biases in legal decisions.

List of Protected Attributes Specified in US Fair Lending Laws

- Fair Housing Acts (FHA)
- Equal Credit Opportunity ACts (ECOA)

Attribute	FHA	ECOA
Race	✓	✓
Color	✓	✓
National origin	✓	✓
Religion	✓	✓
Sex	✓	✓
Familial status	✓	
Disability	✓	
Exercised rights under CCPA		✓
Marital status		✓
Recipient of public assistance		✓
Age		✓

- Machine learning researchers take these as given.

1. Task definition
2. Data collection

$\}$ Data Unfairness

3. Model specification
4. Model fitting/training

$\}$ Algorithmic Unfairness

5. Model testing
6. Deployment in real-world

$\}$ Impact Unfairness

Data can be biased

- Education:
  - SAT scores might be used to guide college admissions, but some students get SAT prep courses

Data can be biased

- Education:
  - SAT scores might be used to guide college admissions, but some students get SAT prep courses
  - Teachers (grading essays) might be biased against some students $\rightarrow$ so will automated essay graders based on those grades.

Data can be biased

- Education:
  - SAT scores might be used to guide college admissions, but some students get SAT prep courses
  - Teachers (grading essays) might be biased against some students $\rightarrow$ so will automated essay graders based on those grades.
- Criminal risk scoring (Skeem and Lovenkamp 2016):
  - Blacks and whites who are otherwise identical are treated the same;
  - But blacks tend to be rated as more risky due to longer criminal histories (**which were produced by biased system**).

Data can be biased

- Education:
  - SAT scores might be used to guide college admissions, but some students get SAT prep courses
  - Teachers (grading essays) might be biased against some students $\rightarrow$ so will automated essay graders based on those grades.
- Criminal risk scoring (Skeem and Lovenkamp 2016):
  - Blacks and whites who are otherwise identical are treated the same;
  - But blacks tend to be rated as more risky due to longer criminal histories (**which were produced by biased system**).
  - similarly: we measure recidivism as "is re-arrested" rather than "commits more crimes". some people more likely to be re-arrested due to policing bias.

Data can be biased

- Education:
  - SAT scores might be used to guide college admissions, but some students get SAT prep courses
  - Teachers (grading essays) might be biased against some students $\rightarrow$ so will automated essay graders based on those grades.
- Criminal risk scoring (Skeem and Lovenkamp 2016):
  - Blacks and whites who are otherwise identical are treated the same;
  - But blacks tend to be rated as more risky due to longer criminal histories (**which were produced by biased system**).
  - similarly: we measure recidivism as "is re-arrested" rather than "commits more crimes". some people more likely to be re-arrested due to policing bias.
  - selective labeling:
    - predictive policing – produces evidence of more crimes in the neighborhoods where police want to go.
    - only observe recidivism if released.

Data can be biased

- Education:
  - SAT scores might be used to guide college admissions, but some students get SAT prep courses
  - Teachers (grading essays) might be biased against some students $\rightarrow$ so will automated essay graders based on those grades.
- Criminal risk scoring (Skeem and Lovenkamp 2016):
  - Blacks and whites who are otherwise identical are treated the same;
  - But blacks tend to be rated as more risky due to longer criminal histories (**which were produced by biased system**).
  - similarly: we measure recidivism as "is re-arrested" rather than "commits more crimes". some people more likely to be re-arrested due to policing bias.
  - selective labeling:
    - predictive policing – produces evidence of more crimes in the neighborhoods where police want to go.
    - only observe recidivism if released.
- a subjective label, such as "harmful to self or others", when made by a human, could be biased (and so would teaching an ML model to reproduce that label)

Data can be biased

- Education:
  - SAT scores might be used to guide college admissions, but some students get SAT prep courses
  - Teachers (grading essays) might be biased against some students $\rightarrow$ so will automated essay graders based on those grades.
- Criminal risk scoring (Skeem and Lovenkamp 2016):
  - Blacks and whites who are otherwise identical are treated the same;
  - But blacks tend to be rated as more risky due to longer criminal histories (**which were produced by biased system**).
  - similarly: we measure recidivism as "is re-arrested" rather than "commits more crimes". some people more likely to be re-arrested due to policing bias.
  - selective labeling:
    - predictive policing – produces evidence of more crimes in the neighborhoods where police want to go.
    - only observe recidivism if released.
- a subjective label, such as "harmful to self or others", when made by a human, could be biased (and so would teaching an ML model to reproduce that label)

These types of problems cannot be fixed by ML.
But ML can help diagnose them.

Systematic Human Biases

- Say a standard ML model $\hat{Y}(X)$ is trained to replicate the decision the modal human judge would make given case features X.

Systematic Human Biases

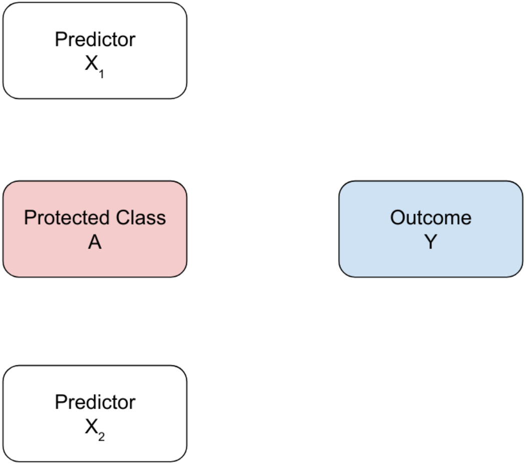
- Say a standard ML model $\hat{Y}(X)$ is trained to replicate the decision the modal human judge would make given case features X.
- *Systematic* biases across all judges will *not* be corrected:
  - These could be reproduced or even *amplified* in the automated decisions.

Systematic Human Biases

- Say a standard ML model $\hat{Y}(X)$ is trained to replicate the decision the modal human judge would make given case features X.
- *Systematic* biases across all judges will *not* be corrected:
  - These could be reproduced or even *amplified* in the automated decisions.

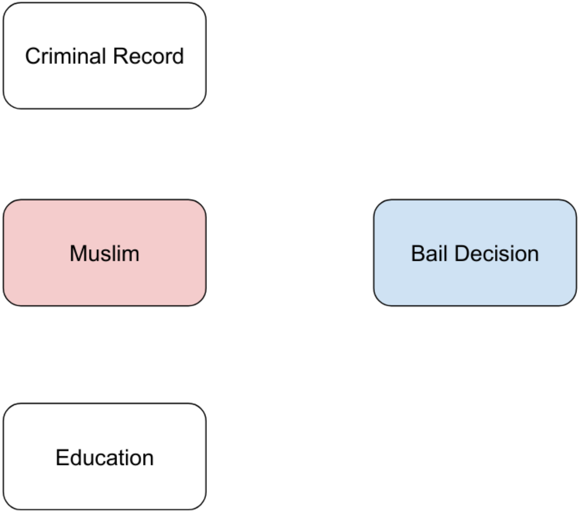
However, again, diagnosing bias is easier with ML than with humans.

Overview: Fairness in Decision-Making

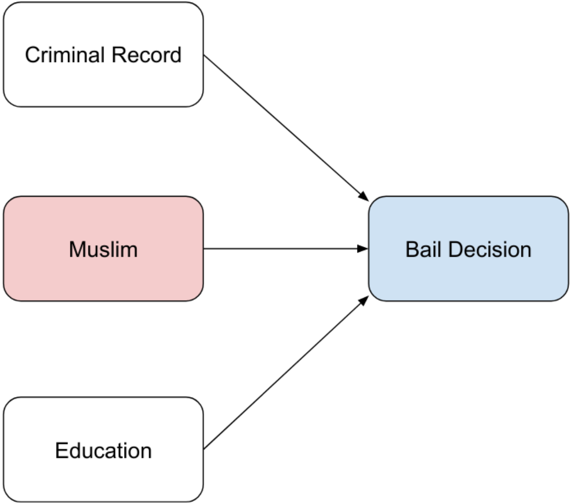


- $A \in \{0,1\}$= protected class, X= other predictors, Y= outcome.
- let $\hat{Y}(X,A)$ be our model predictions.

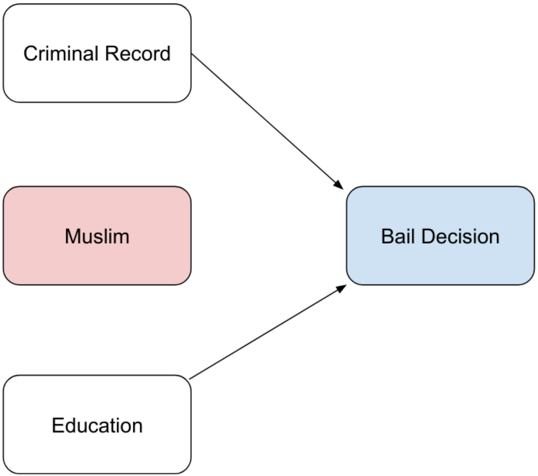
For example:



Standard Approach: Use All Data

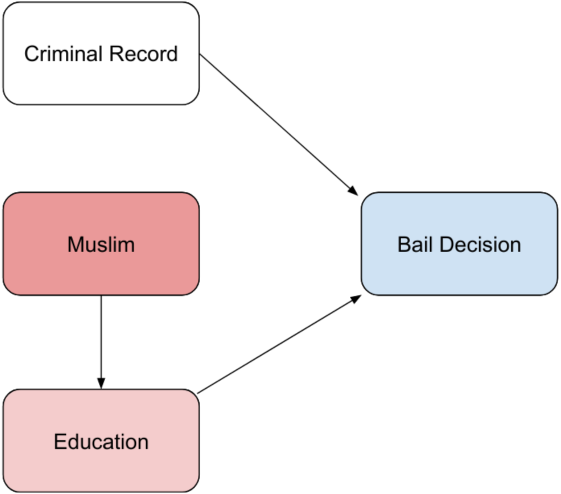


Fairness through Unawareness



- **Fairness through unawareness:** protected attributes are not explicitly used in the prediction process.
  - that is, $\hat{Y}(X,0) = \hat{Y}(X,1)$, $\forall X$.

Problem: Indirect Discrimination



- sensitive factors are implicitly being used by the model, to the extent that they are correlated with included predictors.
  - e.g., muslims have lower education than rest of population.

A deeper problem: Unobserved confounders

A deeper problem: Unobserved confounders

- There is a deeper problem with this approach:
  - protected attributes (e.g. race) are confounded with many unobserved factors, which are correlated with outcomes and predictors.

A deeper problem: Unobserved confounders

- There is a deeper problem with this approach:
  - protected attributes (e.g. race) are confounded with many unobserved factors, which are correlated with outcomes and predictors.
  - as we have seen in our causal inference lectures, resulting estimates for fairness formulas/criteria will be statistically biased.

A deeper problem: Unobserved confounders

- There is a deeper problem with this approach:
  - protected attributes (e.g. race) are confounded with many unobserved factors, which are correlated with outcomes and predictors.
  - as we have seen in our causal inference lectures, resulting estimates for fairness formulas/criteria will be statistically biased.
- Counterfactual fairness (e.g. Kusner et al 2018): "had the protected attributes (e.g., race) of the individual been different, other things being equal, the decision would have remained the same."

A deeper problem: Unobserved confounders

- There is a deeper problem with this approach:
  - protected attributes (e.g. race) are confounded with many unobserved factors, which are correlated with outcomes and predictors.
  - as we have seen in our causal inference lectures, resulting estimates for fairness formulas/criteria will be statistically biased.
- Counterfactual fairness (e.g. Kusner et al 2018): "had the protected attributes (e.g., race) of the individual been different, other things being equal, the decision would have remained the same."
  - e.g., had a defendant been from a different race, he would have had different education, different residence location, etc..

Counterfactual Fairness: Structural Approach

- Assume a structural causal model relating class to all other predictors:
  1. estimate parameters of structural model
  2. flip the protected attribute (e.g.)
  3. compute counterfactual values of predictors (e.g. education) based on changed attribute from structural model
  4. check if predicted outcome changes for counterfactual instance. fairness $\leftrightarrow$ no change.

Counterfactual Fairness: Structural Approach

- Assume a structural causal model relating class to all other predictors:
  1. estimate parameters of structural model
  2. flip the protected attribute (e.g.)
  3. compute counterfactual values of predictors (e.g. education) based on changed attribute from structural model
  4. check if predicted outcome changes for counterfactual instance. fairness $\leftrightarrow$ no change.
- there are now a lot of papers coming out applying this approach, but they rely on extremely strong assumptions.

Outline

Zoom Poll: Classification Metrics Review

Assessing Fair Machine Learning Models

	Predicted Positive	Predicted negative
Actual Positive	$TP = \#$ true positives	$FN = \#$ false negatives
Actual Negative	$FP = \#$ false positives	$TN = \#$ true negatives

- $Y \in \{0,1\}$= outcome label, e.g. reoffends or not.
- $A \in \{0,1\}$= protected class, e.g. gender, X= other predictors
- $\hat{Y}(X,A)$ = the model output
    - a class label (zero or one) and a predicted probability between zero and one.

1. Equality of accuracy

accuracy ($\frac{TP+TN}{\textbf{sample size}}$) is the same across groups

- e.g. men and women have same model accuracy

1. Equality of accuracy

accuracy ($\frac{TP+TN}{\textbf{sample size}}$) is the same across groups

- e.g. men and women have same model accuracy
- pros:
  - intuitive
- cons:
  - weights false positives equal to false negatives
  - minority class usually has lower accuracy, forces that on majority

2. Statistical parity

average predicted outcome ($\frac{\text{\# predicted positive}}{\text{sample size}}$) is the same for each group

- also called "demographic parity".

2. Statistical parity

average predicted outcome ($\frac{\#\text{ predicted positive}}{\text{sample size}}$) is the same for each group

- also called "demographic parity".
- Pros:
  - simple and intuitive
  - sometimes legally required (e.g. EEOC's four-fifths rule)
- Cons:
  - usually reduces accuracy
  - if decision to grant bail is based on $\hat{Y}$, can lead to undesirable outcomes, such as imprisoning a lot more women who are not risky.

3. Error rate balance

false positive rate ($\frac{\text{\# false positives}}{\text{\# actual negatives}}$) and false negative rate ($\frac{\text{\# false negatives}}{\text{\# actual positives}}$) are the same for each groups

- i.e.: Conditioning on the known outcome, is $\hat{Y}(\cdot)$ equally accurate across groups?

3. Error rate balance

false positive rate ($\frac{\text{\# false positives}}{\text{\# actual negatives}}$) and false negative rate ($\frac{\text{\# false negatives}}{\text{\# actual positives}}$) are the same for each groups

- i.e.: Conditioning on the known outcome, is $\hat{Y}(\cdot)$ equally accurate across groups?
- Berk et al call this "conditional procedure accuracy equality". Similar metrics have been called "equalized odds" or "equality of opportunity."

4. Predictive parity

positive predictive value ($\frac{\#\text{ true positives}}{\#\text{ predicted positives}}$) and negative predictive value ($\frac{\#\text{ true negatives}}{\#\text{ predicted positives}}$) are the same for each groups

- i.e., conditional on predicted to be a particular class, is accuracy equal across groups?
  - Berk et al call this "conditional use accuracy equality".

5. Treatment equality

The ratio of false positives to false negatives ($\frac{\#\text{ false positives}}{\#\text{ false negatives}}$) is equal across groups

5. Treatment equality

The ratio of false positives to false negatives ($\frac{\#\text{ false positives}}{\#\text{ false negatives}}$) is equal across groups

- important when positive/negative predictions imply different decisions (e.g. jail or release).
- advantage over (3) and (4): a single criterion rather than two.

Total Fairness

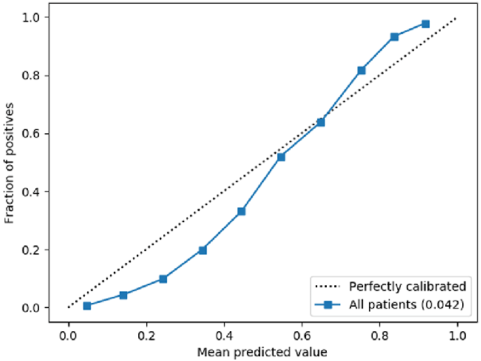
$\hat{Y}(X, A)$ **has achieved total fairness if it satisfies:**

1. equality of accuracy
2. statistical parity
3. error rate balance
4. predictive parity
5. treatment equality

- impossible except in highly artificial datasets.

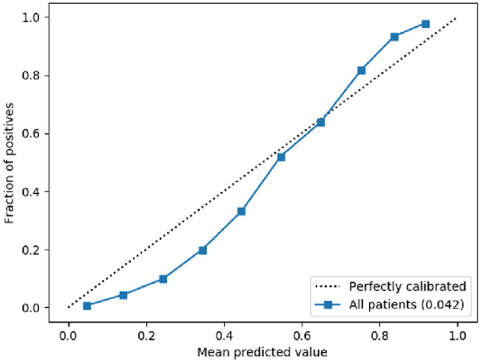
Trade-off 1: Calibration vs. Error Rate Balance

- recall that in a well-calibrated model, we can bin observations by their predicted outcome probabilities, and the outcome rates should roughly match in those bins.
- good calibration requires equalizing false positive and false negative rates in the aggregate.



Trade-off 1: Calibration vs. Error Rate Balance

- recall that in a well-calibrated model, we can bin observations by their predicted outcome probabilities, and the outcome rates should roughly match in those bins.
- good calibration requires equalizing false positive and false negative rates in the aggregate.



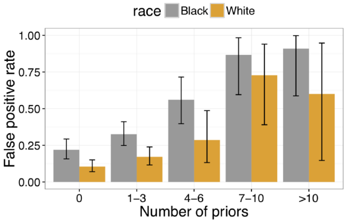
Trade-off: If base rates differ by group, error rate balance (equality of FPR/FNR across groups) precludes calibration.

Trade-off 2: Error Rate Balance vs Predictive Parity

- If base rates differ by group, these requirements cannot hold simultaneously:
  - error rate balance (equality of FPR/FNR across groups)
  - predictive parity (equality of PPV/NPV across groups)

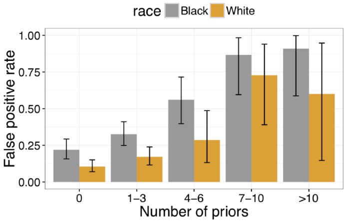
Example: COMPAS

FPR is higher for black defendants! (Chouldechova'17):

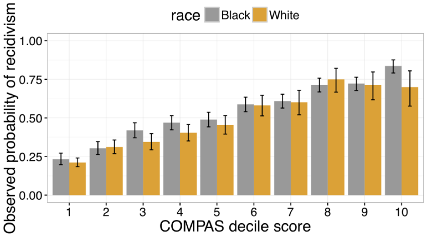


Example: COMPAS

FPR is higher for black defendants! (Chouldechova'17):



But the scores are well-calibrated (or PPV similar across all groups)! (Chouldechova'17):



COMPAS: Dressel and Farid (2018)

COMPAS has higher false positive rate and lower false negative rate for black defendants.

- errors disfavor black defendants.

Dressel and Farid (2018):

- also asked human annotators to produce recidivism predictions, and race info was not provided.

COMPAS: Dressel and Farid (2018)

COMPAS has higher false positive rate and lower false negative rate for black defendants.

- errors disfavor black defendants.

Dressel and Farid (2018):

- also asked human annotators to produce recidivism predictions, and race info was not provided.
- humans were almost identically biased.

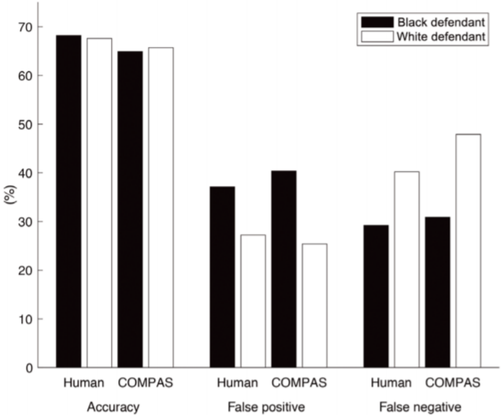


Fig. 1. Human (no-race condition) versus COMPAS algorithmic predictions (see also Table 1).

- giving the human annotators information on the race of the defendant made no difference.

Summary: Group Fairness Theory

- No universally accepted definitions of group fairness.
  - they all make implicit moral assumptions
  - mutually inconsistent

Summary: Group Fairness Theory

- No universally accepted definitions of group fairness.
  - they all make implicit moral assumptions
  - mutually inconsistent
- Still, useful as diagnostic tools.
- Practically, the constrained ML approaches seem to focus on statistical parity.

Outline

Overview of Fairness Problem

Formalizing Fair Classification

Pre-Processing Data to Improve Fairness

Constrained Machine Learning

Pre-Processing: What if we adjust a predictor for the protected class?

- E.g., take education, residualize it on muslim:
  - for each predictor j, regress X_j on A, produce $\tilde{X}_j = X_j - \hat{X}_j$, then use $\tilde{X}_j$ in the ML model.

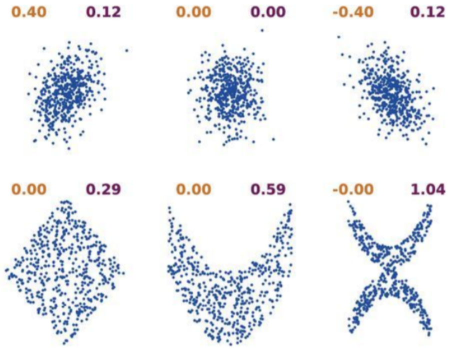
Pre-Processing: What if we adjust a predictor for the protected class?

- E.g., take education, residualize it on muslim:
  - for each predictor j, regress X_j on A, produce $\tilde{X}_j = X_j - \hat{X}_j$, then use $\tilde{X}_j$ in the ML model.
  - Then $\text{corr}(\tilde{X}_j, A) = 0$ by construction.

Pre-Processing: What if we adjust a predictor for the protected class?

- E.g., take education, residualize it on muslim:
  - for each predictor j, regress X_j on A, produce $\tilde{X}_j = X_j - \hat{X}_j$, then use $\tilde{X}_j$ in the ML model.
  - Then $\text{corr}(\tilde{X}_j, A) = 0$ by construction.

Problem: Uncorrelated $\neq$ Independent (e.g. Ince et al 2016)



correlation $\neq$ mutual information

- relations could be non-linear
- could be interactions between predictors, $X_j X_k$, $j \neq k$, correlated with A.
- X_j and A could have an interaction effect on Y.

Purging information on the protected class

Goal: remove any dependence between X and A while preserving information in X that is predictive for Y.

- See Zemel et al (2013), "Learning fair representations" and follow-up papers for sophisticated approach to this problem.
- Double ML methods would seem to also work, but I have not seen that (potential project idea).

Wang et al (adversarial approach using gender and images)

Figure 6. Images after adversarial removal of gender in image space by using a U-Net based autoencoder as inputs to the recognition model. While people are clearly being obscured from the image, the model selectively chooses to obscure only parts that would reveal gender such as faces but tries to keep information that is useful to recognize objects or verbs. 1st row: WWWM MMWW; 2nd row: MWWW WMWW; 3rd row: MMMW MMWM; 4th row: MMMW WWMM. W: woman; M: man.

Outline

Overview of Fairness Problem

Formalizing Fair Classification

Pre-Processing Data to Improve Fairness

Constrained Machine Learning

Prejudice Remover Regularizer (Kamashima et al 2012)

- Let D be the dataset, Θ be the learnable parameters.
- Modified training objective:

$$\underbrace{L(D,\Theta)}_{\text{model loss}} + \underbrace{\eta R(D,\Theta)}_{\text{fairness regularizer}} + \underbrace{\lambda||\Theta||_2^2}_{\text{ridge penalty}}$$

Prejudice Remover Regularizer (Kamashima et al 2012)

- Let D be the dataset, Θ be the learnable parameters.
- Modified training objective:

$$\underbrace{L(D,\Theta)}_{\text{model loss}} + \underbrace{\eta R(D,\Theta)}_{\text{fairness regularizer}} + \underbrace{\lambda ||\Theta||_2^2}_{\text{ridge penalty}}$$

- The "prejudice index" for outcome Y and protected class A is

$$R(\cdot) = \sum_{Y,S} \widehat{\Pr}(Y|X,A) \log \underbrace{\frac{\widehat{\Pr}(Y,A)}{\widehat{\Pr}(Y)\widehat{\Pr}(A)}}_{\text{mutual info}}$$

where $\widehat{\Pr}(Y|X,A)$ is an auxiliary prediction model, e.g. logit, trained for each category of the protected attribute.

Prejudice Remover Regularizer (Kamashima et al 2012)

- Let D be the dataset, Θ be the learnable parameters.
- Modified training objective:

$$\underbrace{L(D,\Theta)}_{\text{model loss}} + \underbrace{\eta R(D,\Theta)}_{\text{fairness regularizer}} + \underbrace{\lambda||\Theta||_2^2}_{\text{ridge penalty}}$$

- The "prejudice index" for outcome Y and protected class A is

$$R(\cdot) = \sum_{Y,S} \widehat{\Pr}(Y|X,A) \log \underbrace{\frac{\widehat{\Pr}(Y,A)}{\widehat{\Pr}(Y)\widehat{\Pr}(A)}}_{\text{mutual info}}$$

where $\widehat{\Pr}(Y|X,A)$ is an auxiliary prediction model, e.g. logit, trained for each category of the protected attribute.

- Model penalizes model errors and indirect discrimination based on mutual information.

The "Reductions" approach (Agarwal 2018)

- Approach: solve a series of cost-sensitive classification problems using off-the-shelf methods.
  - more flexible than the regularizer approach, see paper for details.

The "Reductions" approach (Agarwal 2018)

- Approach: solve a series of cost-sensitive classification problems using off-the-shelf methods.
  - more flexible than the regularizer approach, see paper for details.
- in general, there appear to be dozens of approaches and no firm consensus.

Constrained Optimization in TensorFlow Keras

- The TFCO package in TensorFlow integrates constrained optimization into the training process.
- not that easy to use yet – check out the notebooks linked the syllabus.