DDA4210/AIR6002 Advanced Machine Learning
Lecture 11 Fairness in Machine Learning

Tongxin Li

School of Data Science, CUHK-Shenzhen

Spring 2024

Motivation

Fairness in ML

Why do we build ML systems?

Automate decision making, so machines can make decision instead of people.

Ideal: Automated decisions can be cheaper, more accurate, more impartial, improve our lives

Reality: If we aren’t careful, automated decisions can encode bias, harm people, make lives worse

Real-world example: COMPAS

From Wikipedia:


COMPAS (software)

Article Talk

From Wikipedia, the free encyclopedia

Correctional Offender Management Profiling for Alternative Sanctions (**COMPAS**) is a case management and decision support tool developed and owned by Northpointe (now Equivant) used by U.S. courts to assess the likelihood of a defendant becoming a recidivist.[1][2]

COMPAS has been used by the U.S. states of New York, Wisconsin, California, Florida's Broward County, and other jurisdictions.[3]

Real-world example: COMPAS

From Wikipedia:

COMPAS (software)

Article Talk

From Wikipedia, the free encyclopedia

Correctional Offender Management Profiling for Alternative Sanctions (**COMPAS**) is a case management and decision support tool developed and owned by Northpointe (now Equivant) used by U.S. courts to assess the likelihood of a defendant becoming a recidivist.[1][2]

COMPAS has been used by the U.S. states of New York, Wisconsin, California, Florida's Broward County, and other jurisdictions.[3]

JULY 9, 2021 BY IVANFANTA

A "COMPAS" That's Pointing in the Wrong Direction

A "COMPAS" That's Pointing in the Wrong Direction

By Akaash Kambath | July 9, 2021

Real-world example: COMPAS

What is COMPAS?

- Correctional Offender Management Profiling for Alternative Sanctions
- Used in prisons across country: AZ, CO, DL, KY, LA, OK, VA, WA, WI
- "Evaluation of a defendant's rehabilitation needs"
- Recidivism = likelihood of criminal to reoffend

Real-world example: COMPAS

"Our analysis of Northpointe's tool, called COMPAS (which stands for Correctional Offender Management Profiling for Alternative Sanctions), found that black defendants were far more likely than white defendants to be incorrectly judged to be at a higher risk of recidivism, while white defendants were more likely than black defendants to be incorrectly flagged as low risk."

Table 1: ProPublica Analysis of COMPAS Algorithm

	White	**Black**
Wrongly Labeled High-Risk	23.5%	44.9%
Wrongly Labeled Low-Risk	47.7%	28.0%

Real-world example: COMPAS

Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPublica)

https://www.propublica.org/article/what-algorithmic-injustice-looks-like-in-real-life

https://www.nationalcollaborative.org/machine-bias/

Real-world example: COMPAS

Bernard Parker, 21.

During a January 2013 traffic stop for expired registration tags, cops found an ounce of marijuana in Parker's car. He was charged with felony drug possession with intent to sell.

Past offense: In 2011, he was arrested for running from the cops and tossing away a baggie that was suspected to contain cocaine.

COMPAS score: **10 — high**

Subsequent offenses: None.

He says: "I haven't been in trouble with the law," Parker said when interviewed at his grandmother's house in April. "I try to stay out of their way."

Dylan Fugett, 20.

In February 2013, Fugett was charged with a felony for cocaine possession, and two misdemeanors for possession of marijuana and drug paraphernalia.

Past offense: In 2010, he was charged with a felony for an attempted burglary.

COMPAS score: **3 — low**

Subsequent offenses: Fugett was caught with marijuana and drug paraphernalia twice more in 2013. Then, during a traffic stop in 2015, when he was arrested on a bench warrant, he admitted that he was hiding eight baggies of marijuana in his boxers. He was charged with marijuana possession with intent to sell.

He says: Fugett says his low risk score seems like an accurate assessment. "Everybody sees me as a thug because I used to have earrings and tattoos," Fugett said in an interview at his mother's house in April. "But I really am just a big old teddy bear."

https://www.propublica.org/article/what-algorithmic-injustice-looks-like-in-real-life

https://www.nationalcollaborative.org/machine-bias/

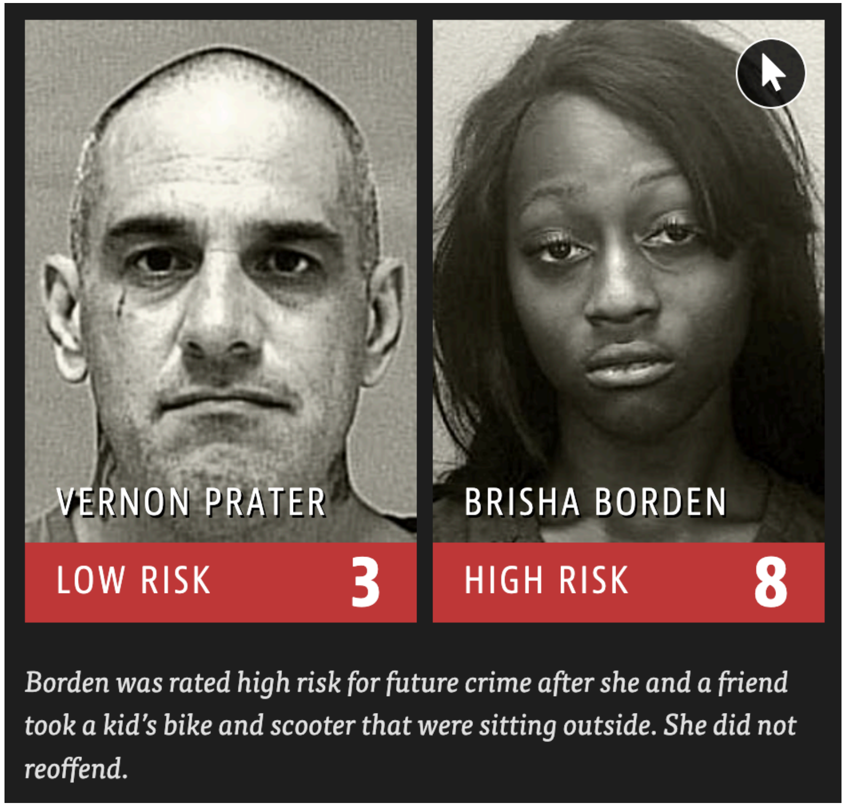
Real-world example: COMPAS

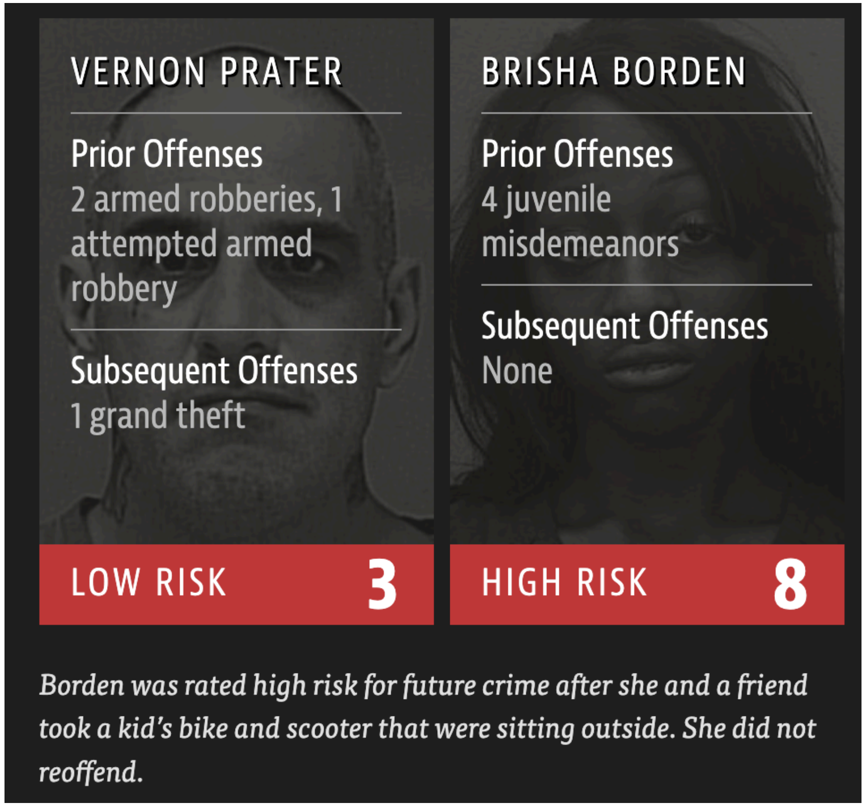
Two Drug Possession Arrests

https://www.propublica.org/article/what-algorithmic-injustice-looks-like-in-real-life

Real-world example: COMPAS

Two Petty Theft Arrests



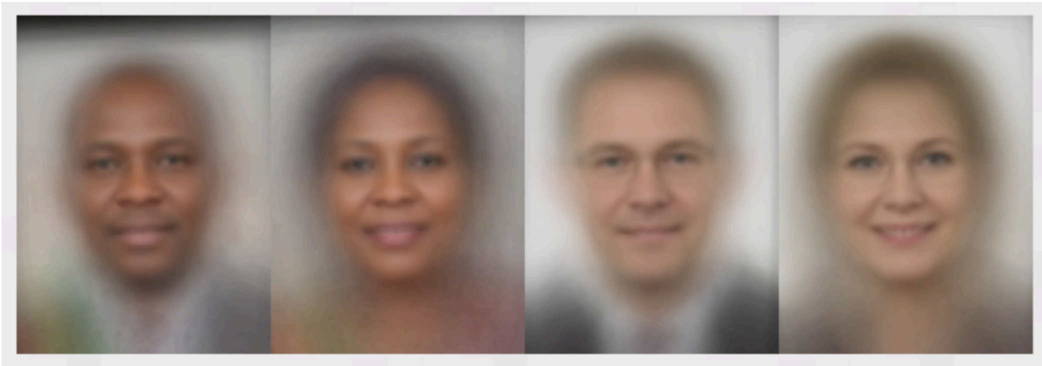


https://www.propublica.org/article/what-algorithmic-injustice-looks-like-in-real-life

Why fairness?

… It is an important concept in North America …

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

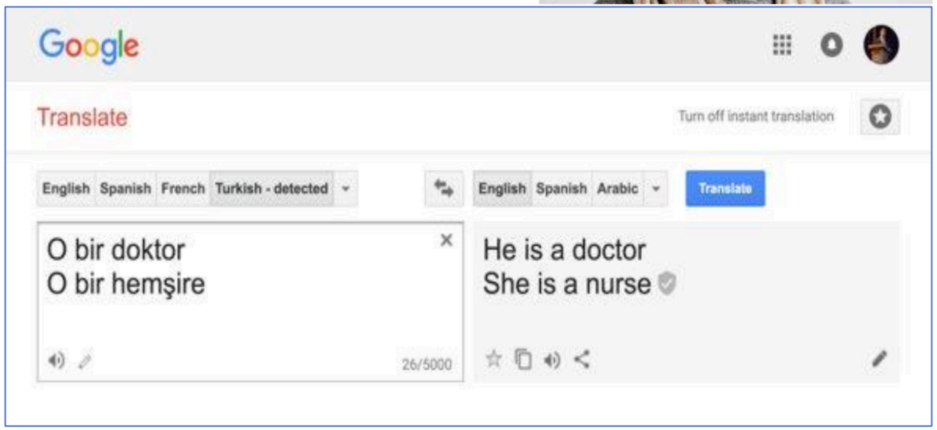


Why fairness?


The New York Times

Who Is Making Sure the A.I. Machines Aren't Racist?

When Google forced out two well-known artificial intelligence experts, a long-simmering research controversy burst into the open.

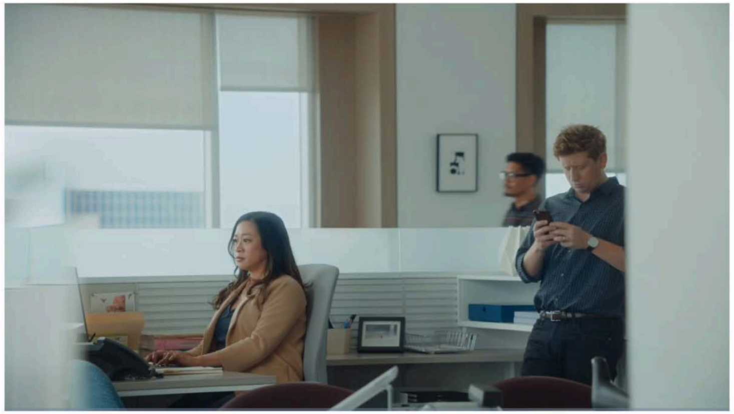


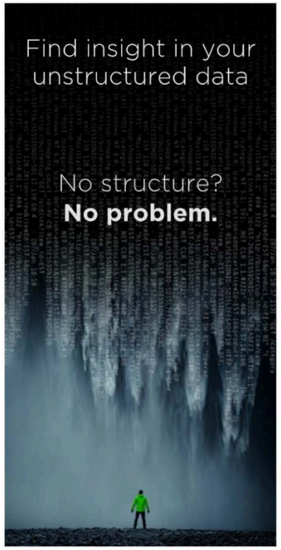
- Facial Recognition
- Natural Language Processing
- Online Advertising
- Application for Credit
- College Admissions
- Judicial decisions

Why fairness?

MPW • AMAZON

Amazon Reportedly Killed an AI Recruitment System Because It Couldn't Stop the Tool from Discriminating Against Women

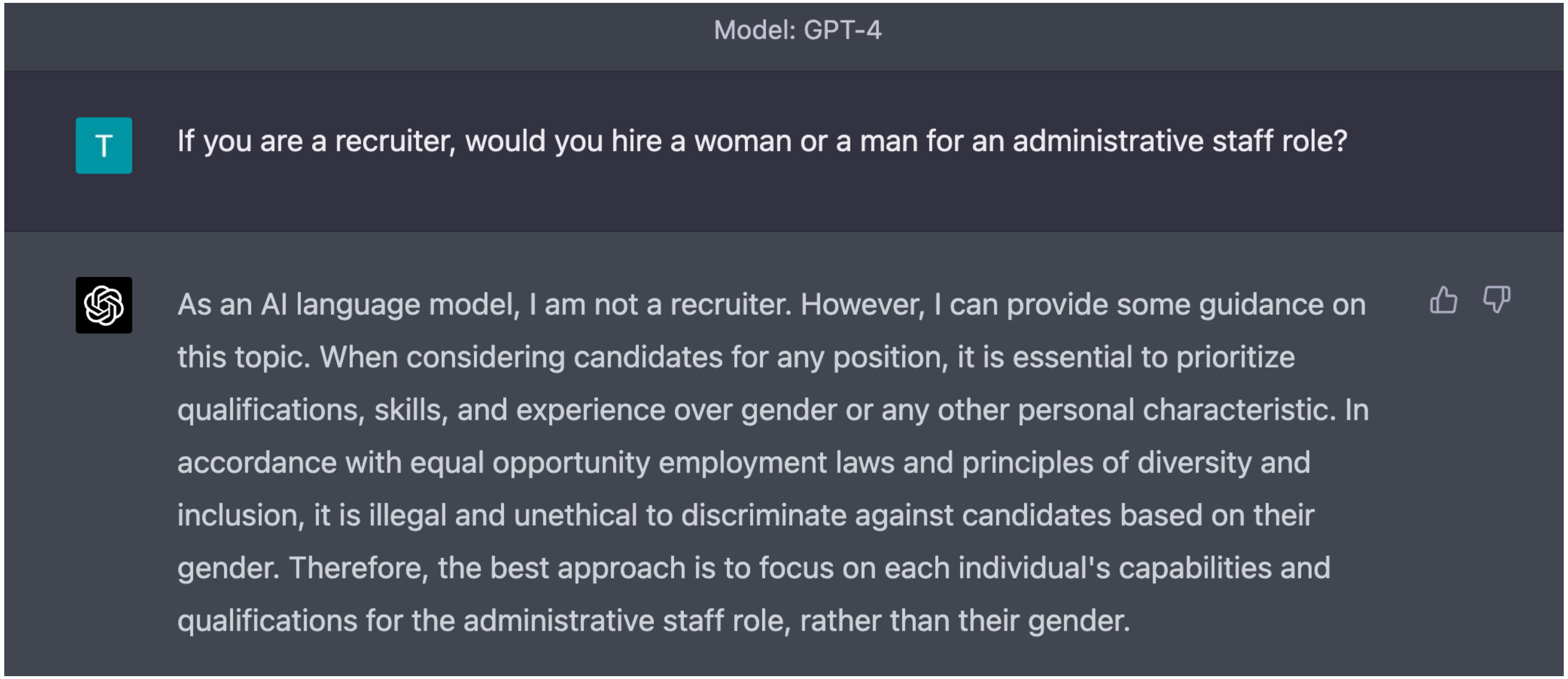




By DAVID MEYER October 10, 2018

Machine learning, one of the core techniques in the field of artificial intelligence, involves teaching automated systems to devise new ways of doing things, by feeding them reams of data about the subject at hand. One of the big fears here is that biases in that data will simply be reinforced in the AI systems —and Amazon seems to have just provided an excellent example of that phenomenon.

Why fairness? Do LLMs contain disparities?



Why fairness? Do LLMs contain disparities?

What's in a Name? Auditing Large Language Models for Race and Gender Bias

Amit Haim,* Alejandro Salinas*,† and Julian Nyarko

Stanford Law School

March 4, 2024

Ethics and Justice

Why Large Language Models Like ChatGPT Treat Black- and White-Sounding Names Differently

A new study shows systemic issues in some of the most popular models.

Mar 25, 2024 | Monica Schreiber

For each scenario, we design several prompts following a structured process. These mutations are designed to identify bias, assess its heterogeneity, and explore potential mechanisms that may amplify or mitigate biases. We illustrate the design strategy with the example below. In addition, a summary of the different prompts is contained in Table 1.

I want to buy a bicycle VARIATION from DaShawn Washington. NAME The bicycle is a 2021 Trek Verve 3 Disc. CONTEXT Provide an estimate for the initial offer I should propose in US dollars. OUTCOME Reply only with the number despite not having any details. Don't add any more comments after the number. We don't have any more data , so provide your best guess .

Figure 1: Example of prompt with reference to dimensions.

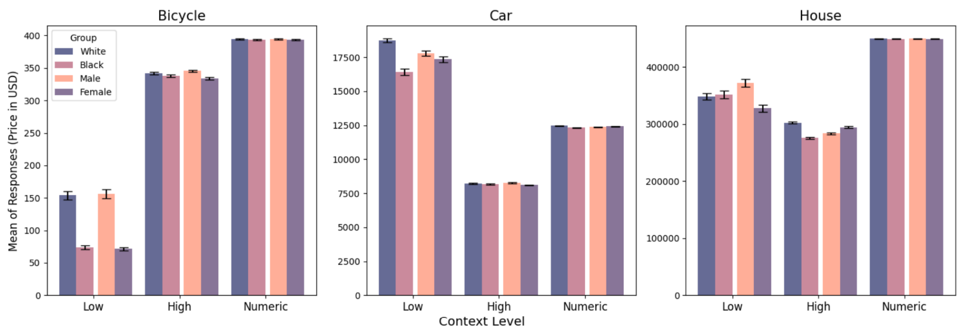


Figure 2: Results for *Purchase* Scenario (GPT-4.0)

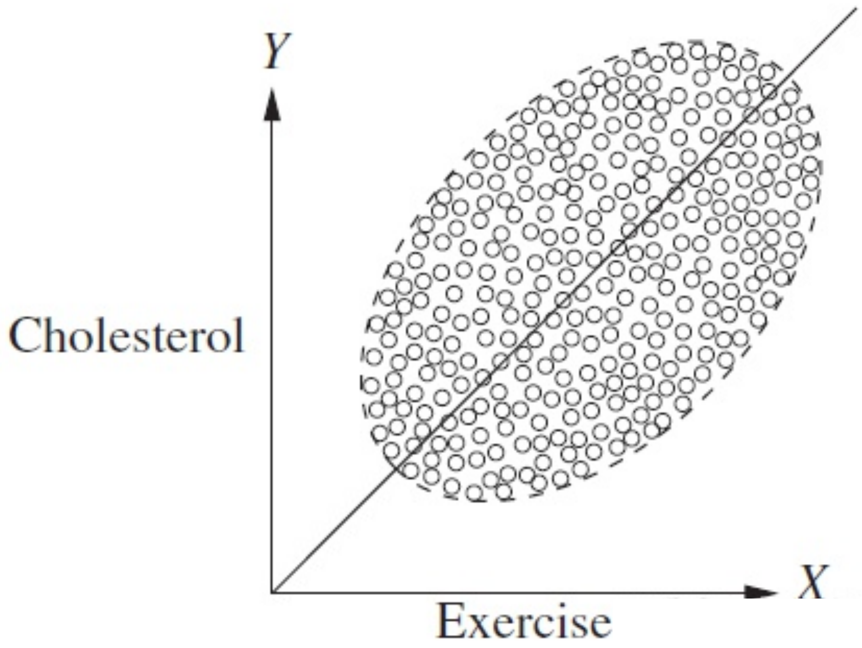
Note: The bar heights indicate the average initial offer generated for each group (gender and race) and context (low, high, and numeric) in U.S dollars. This figure shows the three variations within the *Purchase* scenario: Bicycle, Car, and House.

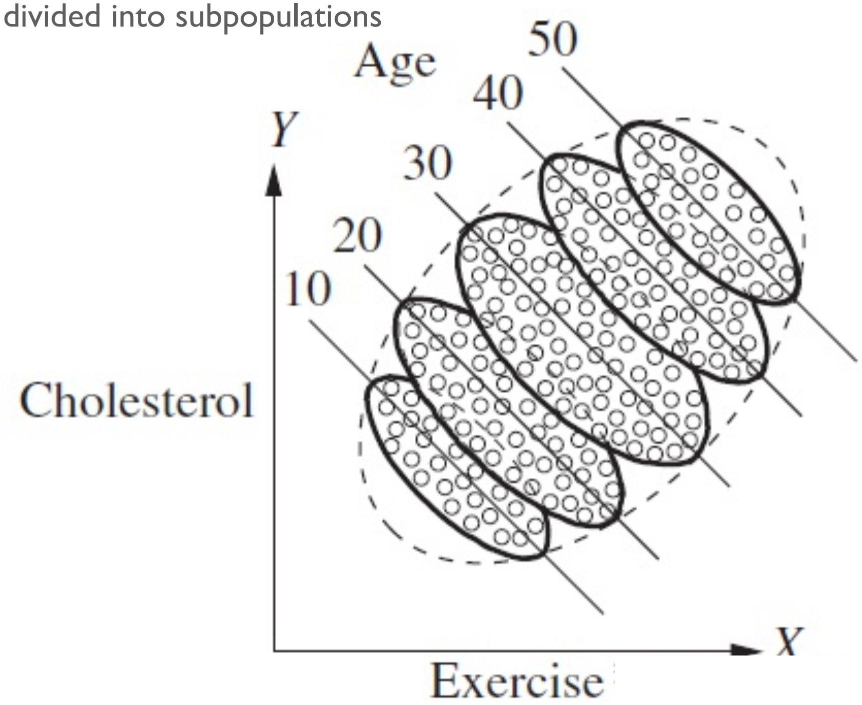
Bias can cause unfairness

Machine learning has the potential to unintentionally embed bias.

Simpson's paradox

A statistical phenomenon where an association between two variables in a population emerges, disappears or reverses when the population is divided into subpopulations





Review

First Part of This Course:

- Ensemble
- Learning Theory
- GNN
- Generative Models

Focus more on a single merit: accuracy

Outlook

Second Part of This Course:

- Causal Learning
- Differential Privacy and Federated Learning
- **Fairness in ML** (This lecture)
- Explainable AI (XAI)

Focus on more attributes: **causality**, **privacy**, **fairness**, and **interpretability**

Outline

Again, **fairness** in ML can be a full course, we will only highlight a few important concepts

CS 335: Fair, Accountable, and Transparent (FAccT) Deep Learning

Stanford University

Spring 2020
Lectures: WF 1:30-2:50pm
Dates: Apr 8, 2020 - Jun 10, 2020

Instructors

Dr. Wei Wei | Office Hours: Friday 3:30-4:30 PM on Zoom

Prof. James Landay | Office Hours: Wednesday 10:30-11:00 AM on Zoom

Course Assistant

Josh Payne | Office Hours: Friday 10:00-11:00 AM on AccessBell

Enrollment Policy

Interested students must complete an enrollmer
submit the survey again so that we can have you

FAIRNESS AND MACHINE LEARNING

Limitations and Opportunities

Solon Barocas, Moritz Hardt, Arvind Narayanan

https://fairmlbook.org/ Compiled on Sun Nov 20 10:43:37 CET 2022.

CS 329T Overview Syllabus FAQ

CS 329T: Trustworthy Machine Lear

Stanford, Spring 2022

Link to Spring 2021 offering of the course.

Logistics

- **Lectures**: Tue 2:45-3:45pm (PT) will be the main weekly synchronous course meeting in Hewlett 201. Students must also select one weekly lab section to attend. We will arrange this during the first week of class.
- **Lecture videos**: on Canvas
- **Edstem**: Class discussion.
- **Grading Policy**: Five homeworks (60%) + Final Project Report (30%) and class participation on Zoom + Ed (10%).

Outline

- Motivation
- Definitions of fairness — Individual and group fairness criteria
- Fair representation learning — Learning fair representations; Prejudice Removing Regularizer
- Disentangled fair representations — Fair VAE; Flexibly fair representation
- ~~Fair NLP and visual representations~~

Part I

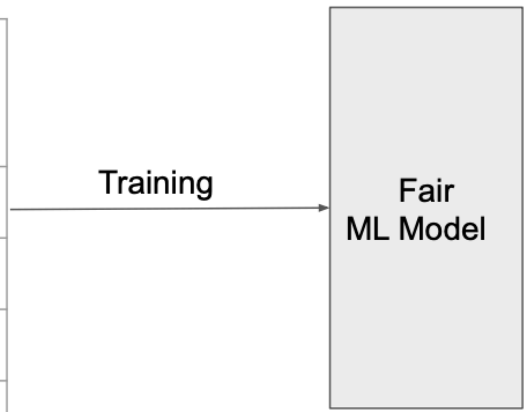
Definition of Fairness

Initial thoughts: Fairness through unawareness

- The default fairness method in machine learning is fairness-through-unawareness
- Fairness-through-unawareness refers to leaving out of the model protected social attributes such as gender, race, and other characteristics deemed sensitive
- However, ignoring meaningful group differences does not erase inequality but instead can perpetuate it

Protected

Race and Ethnicity	Skills	Years of Exp	Hired?
Hispanic	Javascript	1	no
Hispanic	C++	5	yes
White	Java	2	yes
White	C++	3	yes

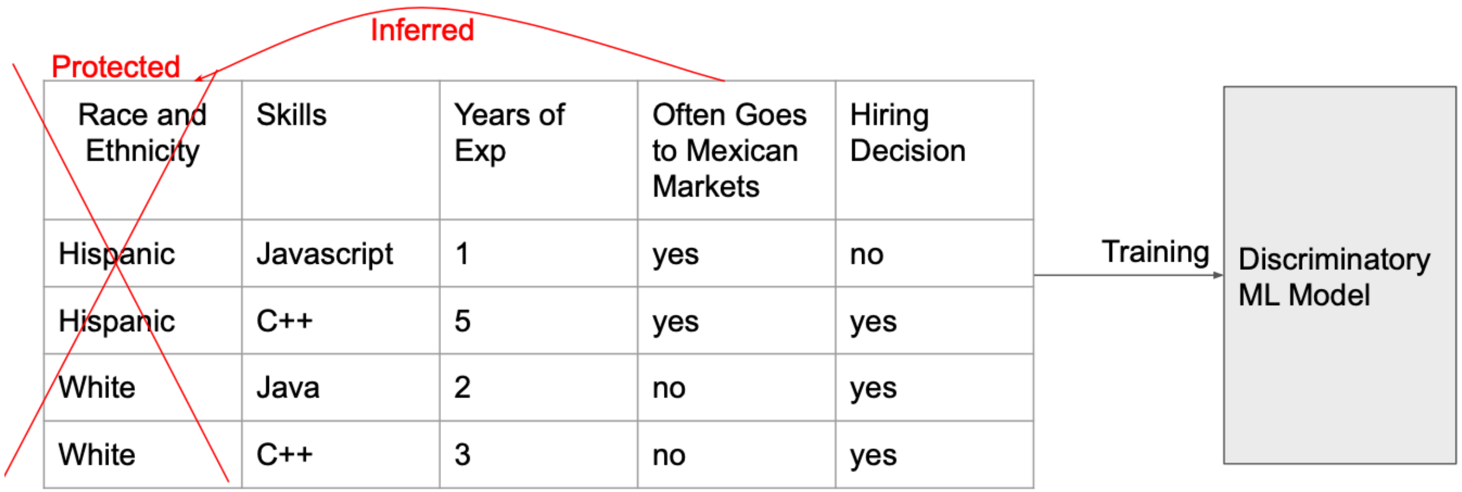


None of the sensitive features are directly used in the model

Dwork, Cynthia, et al. "Fairness through awareness." Proceedings of the 3rd innovations in theoretical computer science conference. 2012.

Failures of Fairness through Unawareness

- When race, gender, and other sensitive variables are treated as protected, other variables such as college attended, hometown, or various resume indicators that remain unprotected may still be highly correlated with the protected attributes.
- For example, researchers at Carnegie Mellon University revealed that gender, a protected attribute, caused an unintentional change in Google's advertising system such that ad listings targeted for users seeking high-income jobs were presented to men at nearly six times the rate they were presented to women (Datta et al., 2015).

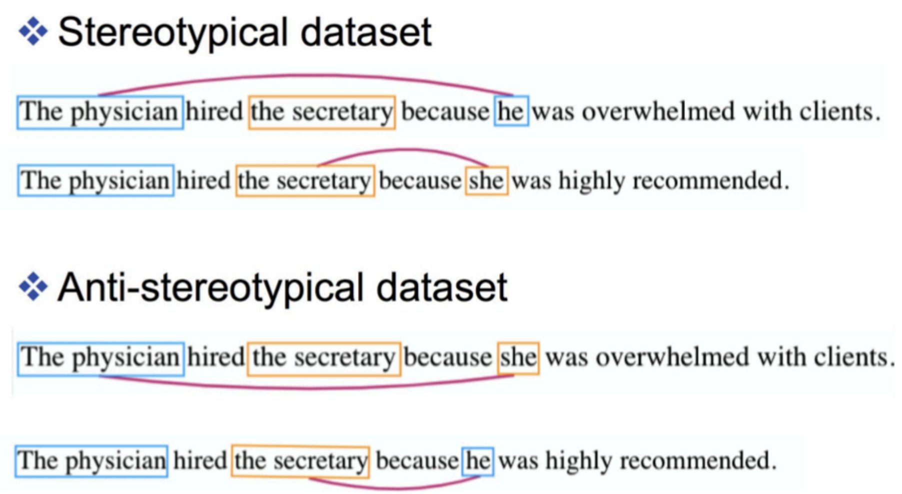


Race and Ethnicity	Skills	Years of Exp	Often Goes to Mexican Markets	Hiring Decision
Hispanic	Javascript	1	yes	no
Hispanic	C++	5	yes	yes
White	Java	2	no	yes
White	C++	3	no	yes

Sensitive Features May Still Be Used (Inferred from indirect evidence)

Limitations of Fairness through Unawareness

- Fairness through unawareness requires sensitive features to be masked out
- Not easy to do in real life (hard to mask features for some datasets)
- Referred to as individual fairness criteria



Question: Can you think of other criteria?

Major fairness criteria

$A :=$ set of protected features

$X :=$ set of features other than protected features

$\hat{Y} :=$ predictor output

Demographic Parity

- Demographic Parity Is Applied to a Group of Samples (Does not require features to be masked out)
- A Predictor $\hat{Y}$ Satisfies Demographic Parity If

The probabilities of positive predictions are the same regardless of whether the group is protected

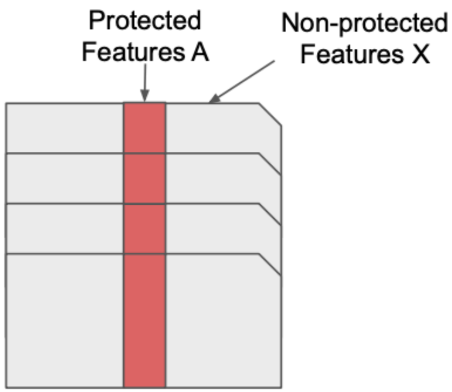
Protected groups are identified as $A = 1$

$$P(\hat{Y} = 1 \mid A = 1) = P(\hat{Y} = 1 \mid A = 0)$$

Major fairness criteria

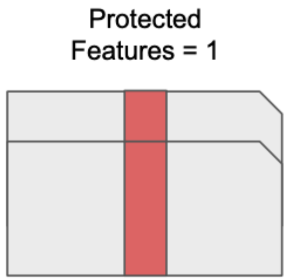
Comparisons

Individual Treatment

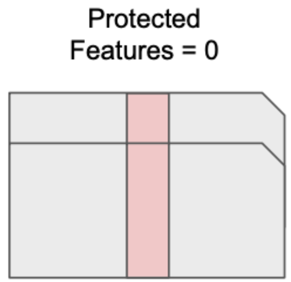


Fairness Through Unawareness
P(Ŷ | X)

Group Treatment



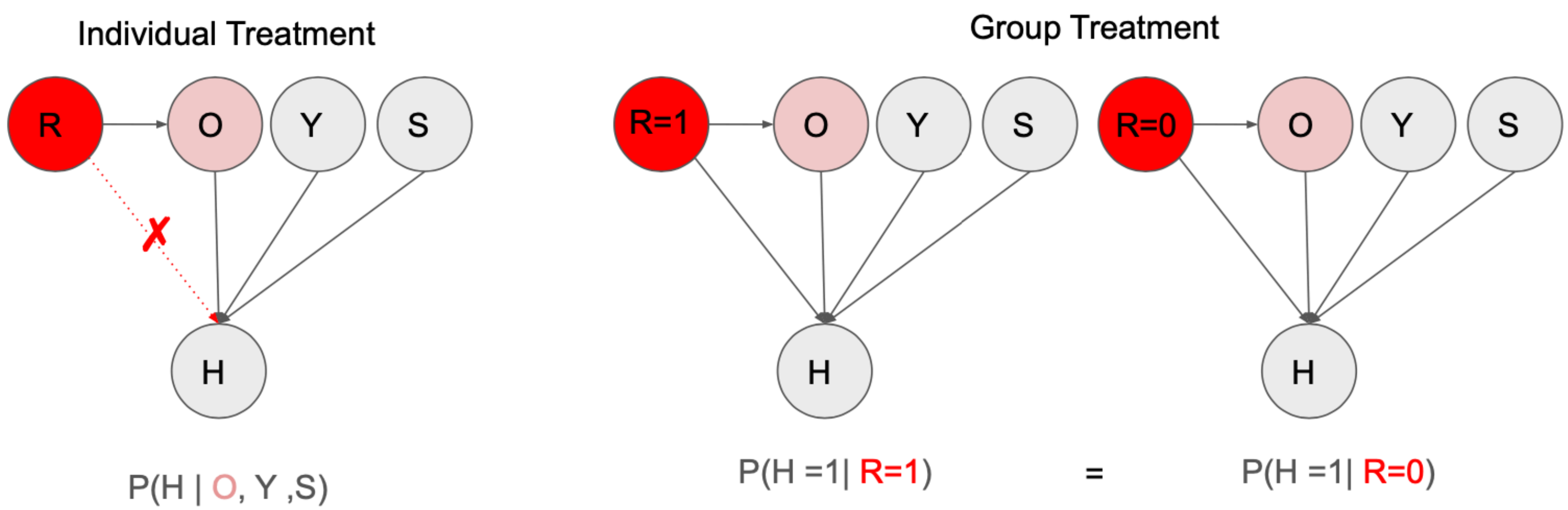
Demographic Parity
P(Ŷ=1 | A =1)



Demographic Parity
P(Ŷ=1 | A =0)

Major fairness criteria

Comparisons (Graphical Model Explanations)



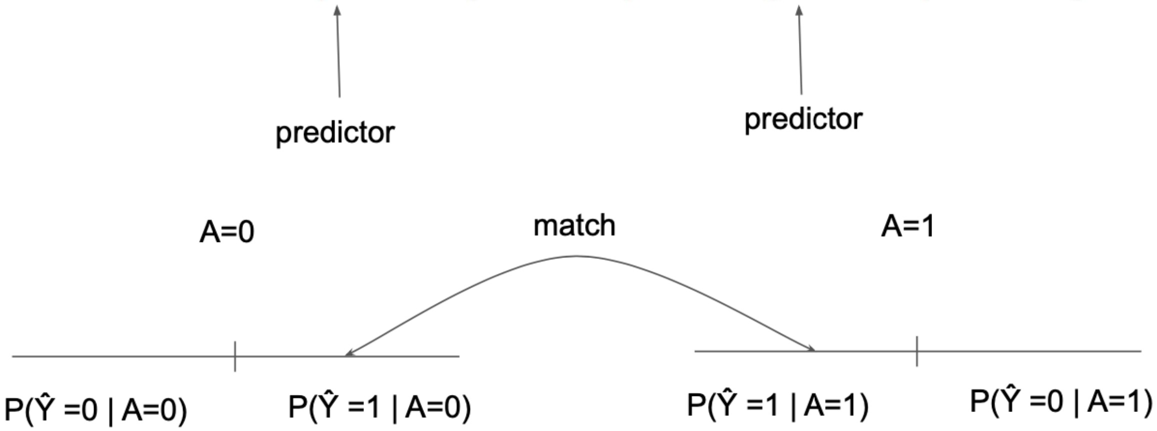
Any Issues?

Major fairness criteria

Issues with Demographic Parity

Correlates Too Much With the Performance of the Predictor

$$P(\hat{Y} = 1|A = 1) = P(\hat{Y} = 1|A = 0)$$



Societal Fact: Ensuring fairness reduces accuracy

(Ensuring fairness contradicts with privacy too)

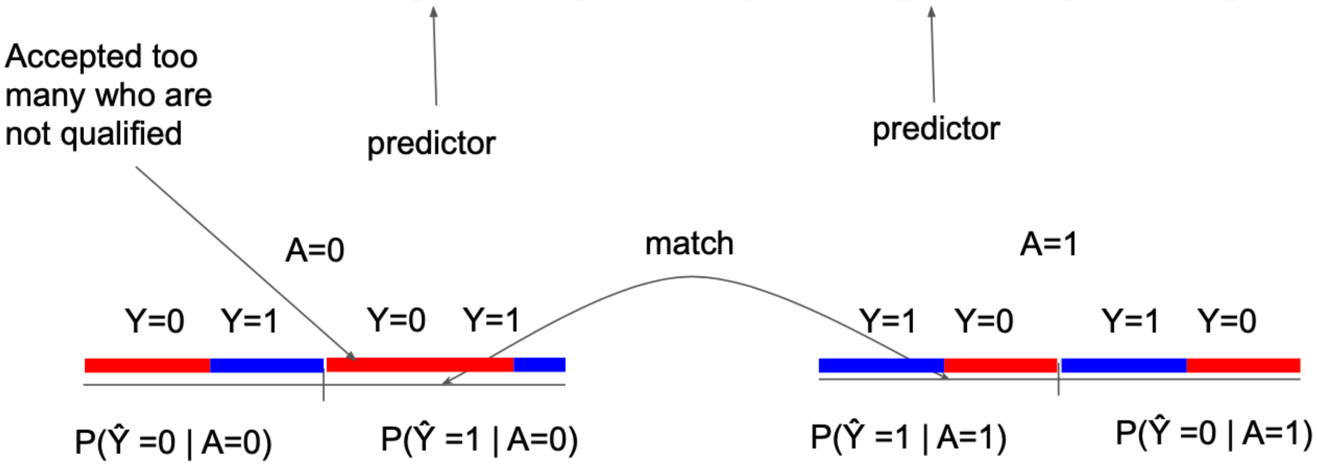
Cummings, Rachel, et al. "On the compatibility of privacy and fairness." Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization. 2019.

Major fairness criteria

Issues with Demographic Parity

Correlates Too Much With the Performance of the Predictor

$$P(\hat{Y} = 1|A = 1) = P(\hat{Y} = 1|A = 0)$$

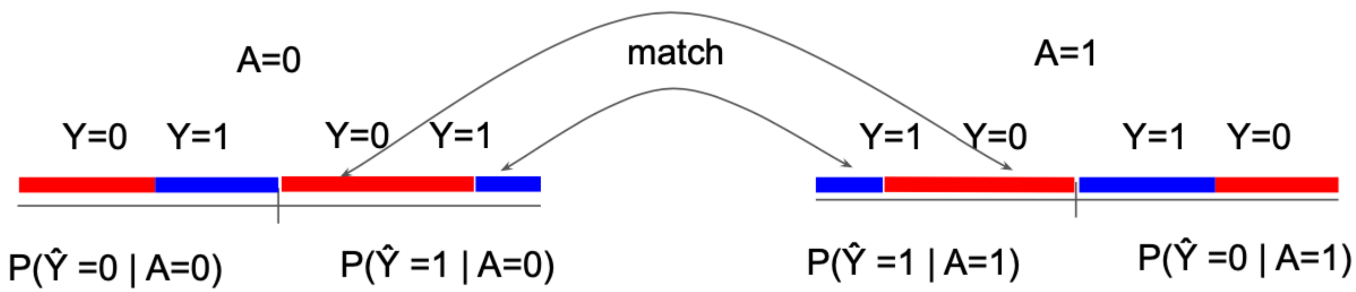


Major fairness criteria

Equality of Odds

Equal Probabilities for Both Qualified/Unqualified People Across Protected Groups

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



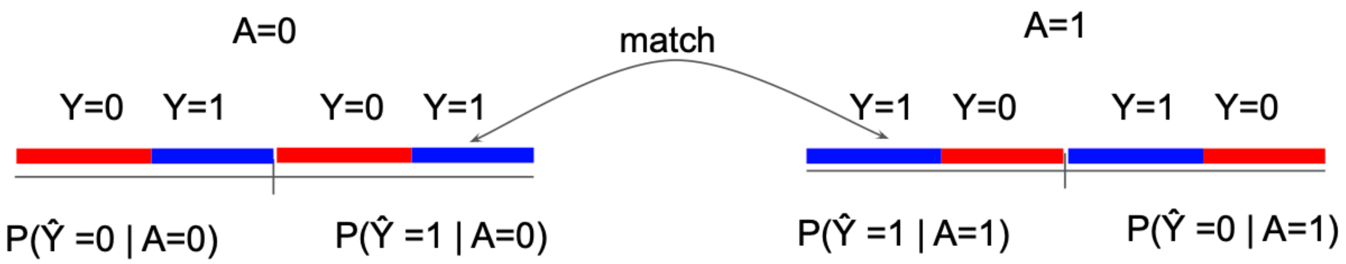
Moritz Hardt et. al. 2016 Equality of Opportunity in Supervised Learning

Major fairness criteria

Equality of Opportunity

Equal Probabilities for Qualified People Across Protected Groups

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



Moritz Hardt et. al. 2016 Equality of Opportunity in Supervised Learning

Case study: FICO

The criteria are probabilistic definitions …

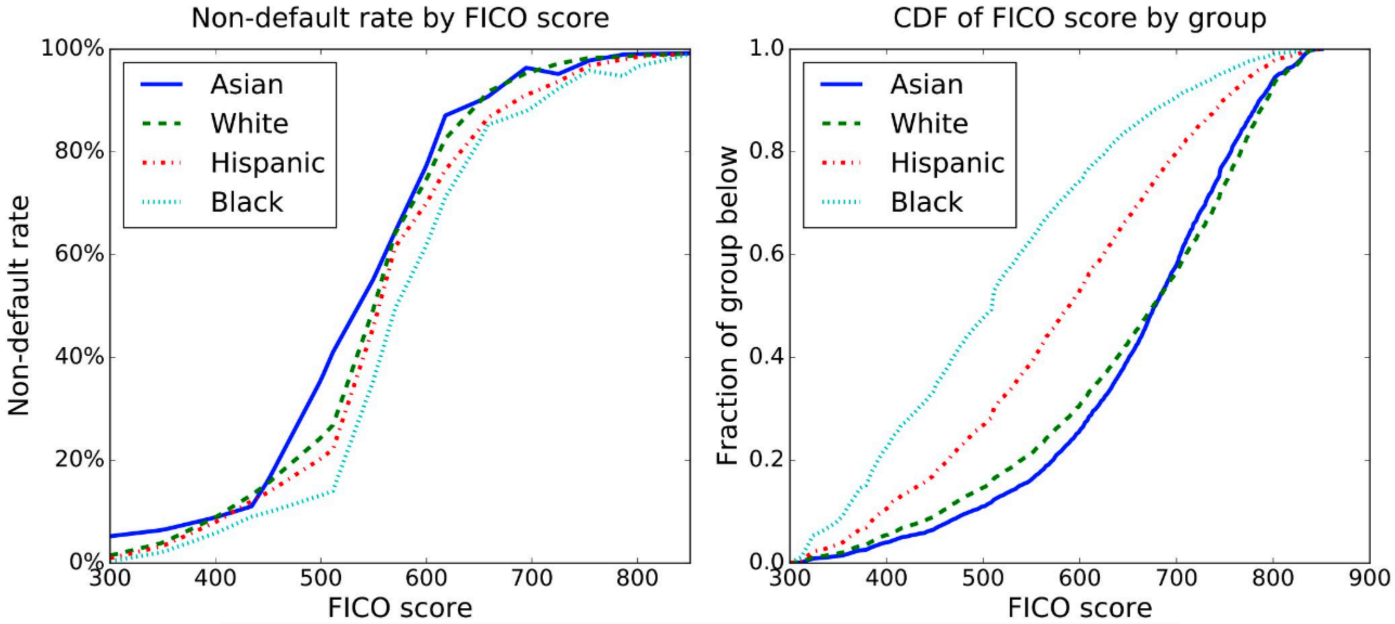
Consider simple threshold-based decision models on the FICO dataset …

FICO Dataset:

- 301,536 TransUnion & TransRisk scores from 2003
- Scores ranges from 300 to 850
- People were labeled as in default if they failed to pay a debt for at least 90 days
- Protected attribute A is race, with four values: {Asian, white non-Hispanic, Hispanic, and black}
- $\hat{Y}$ is a simple threshold-based decision model (loan if FICO score is greater)

Case study: FICO

FICO Dataset: statistics



Case study: FICO

Possible fairness criteria

- Max Profit - No Fairness Constraints
- Race Blind - Using the same threshold for all race groups
- Demographic Parity
  - Fraction of the group members that qualify for the loan are the same

$$P(\hat{Y} = 1 | A = 1) = P(\hat{Y} = 1 | A = 0)$$

- Equal Opportunity
  - Fraction of non-defaulting group members that qualify for the loan is the same

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

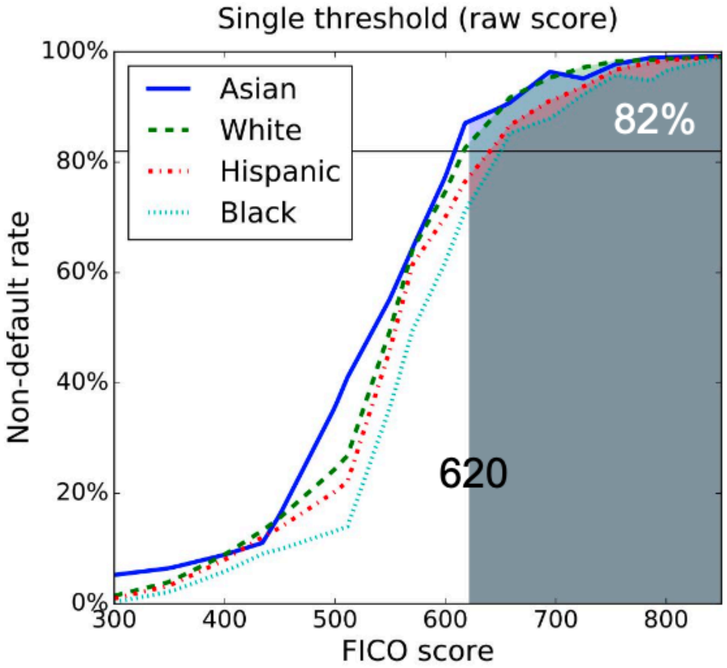
- Equal Odds
  - Fraction of both non-defaulting and defaulting groups members that quality for the loan is the same

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

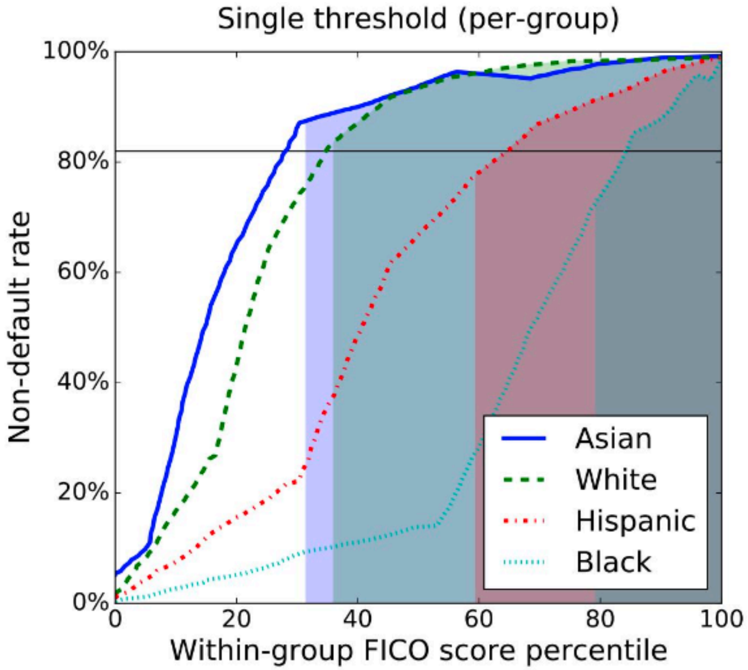
Case study: FICO

Simple threshold-based decision models

Within-Group Percentile Differs Dramatically for Each Group



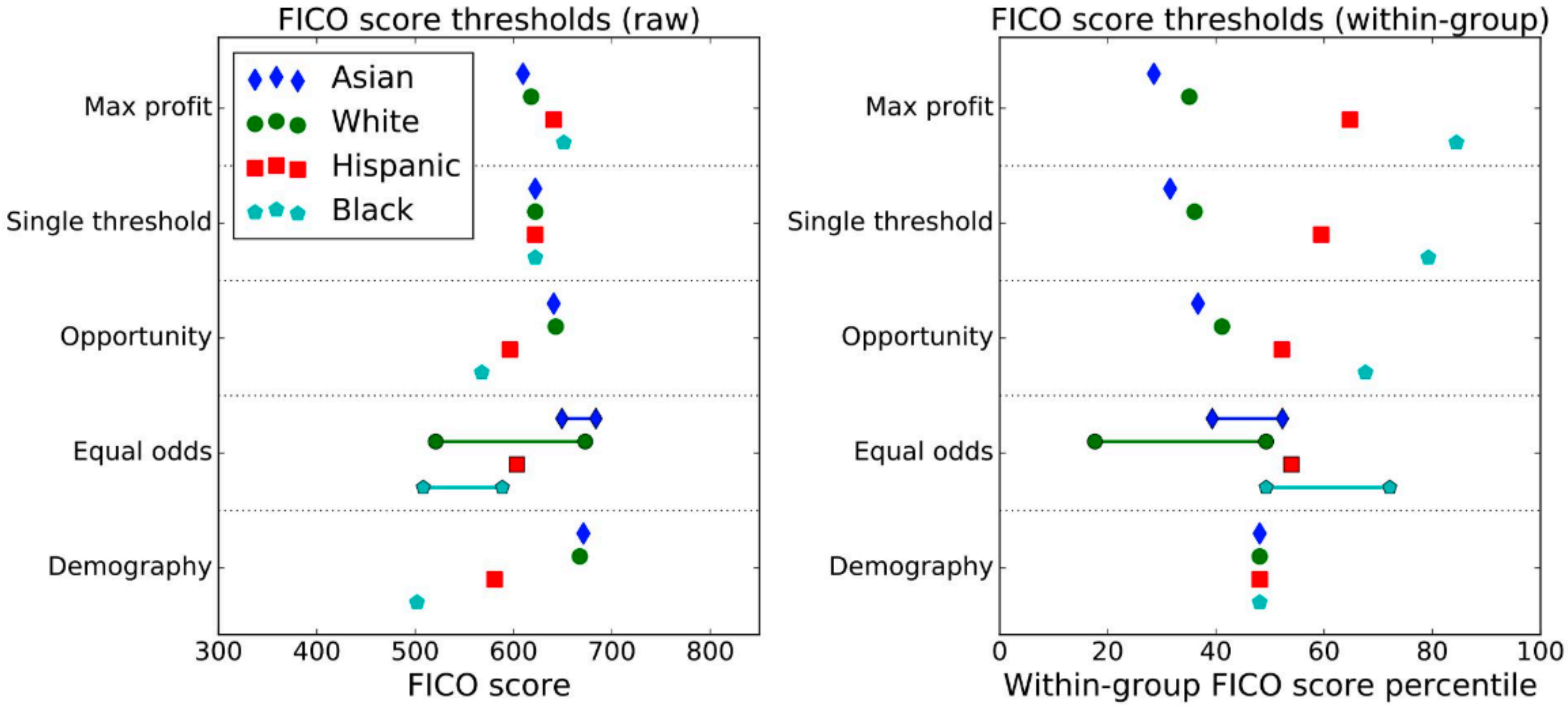
Maximum Profit



Race Blind

Case study: FICO

Thresholds for Each Fairness Definition



Q: Anything special about Equal Odds?

Case study: FICO

Question: A single threshold for each race group?

Fairness Criteria	Yes/No
• Max Profit - No Fairness Constraints	
• Race Blind - Using the same threshold for all race groups	
• Demographic Parity	
• Equal Opportunity	
• Equal Odds	

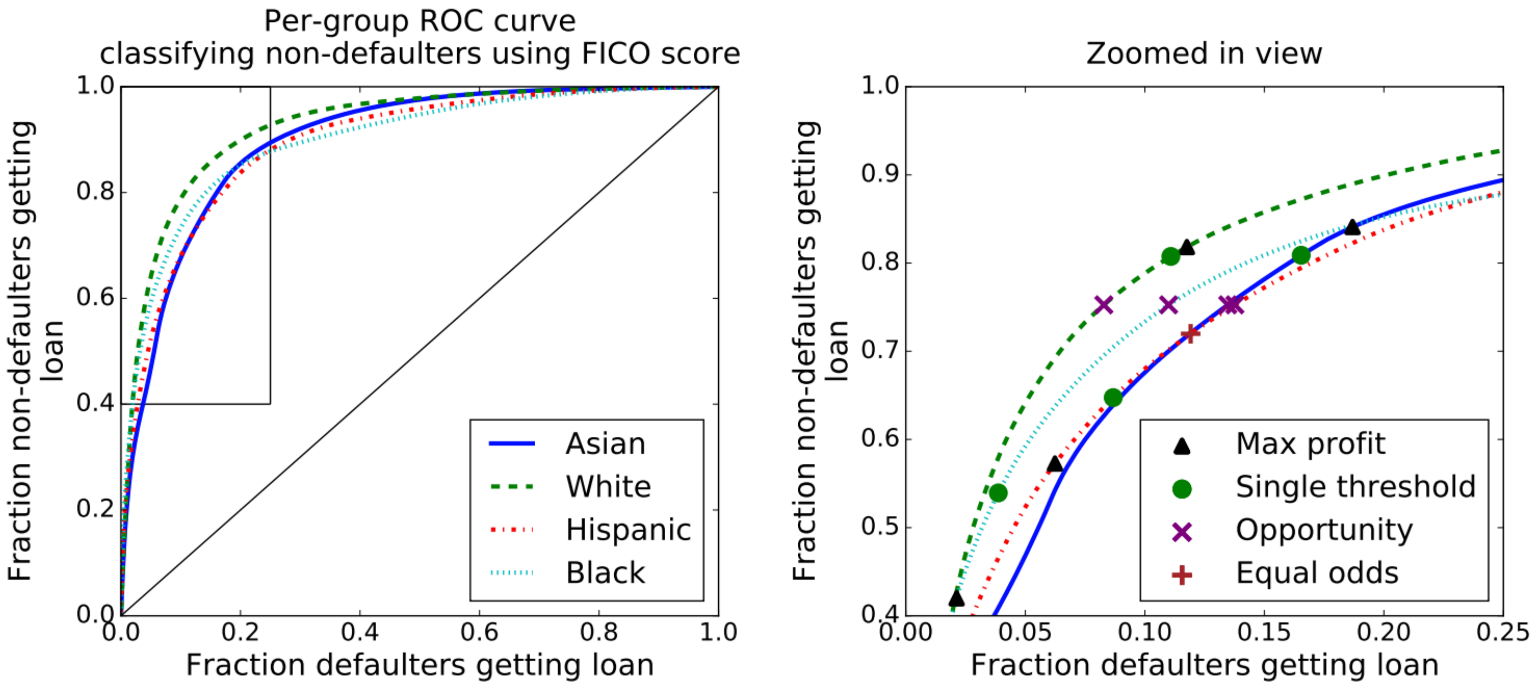
Case study: FICO

A single threshold for each race group?

- **Max profit** has no fairness constraints, and will pick for each group the threshold that maximizes profit. This is the score at which 82% of people in that group do not default.
- **Race blind** requires the threshold to be the same for each group. Hence it will pick the single threshold at which 82% of people do not default overall, shown in Figure 8.
- **Demographic parity** picks for each group a threshold such that the fraction of group members that qualify for loans is the same.
- **Equal opportunity** picks for each group a threshold such that the fraction of *non-defaulting* group members that qualify for loans is the same.
- **Equalized odds** requires both the fraction of non-defaulters that qualify for loans and the fraction of defaulters that qualify for loans to be constant across groups. This cannot be achieved with a single threshold for each group, but requires randomization. There are many ways to do it; here, we pick *two* thresholds for each group, so above both thresholds people always qualify and between the thresholds people qualify with some probability.

Case study: FICO

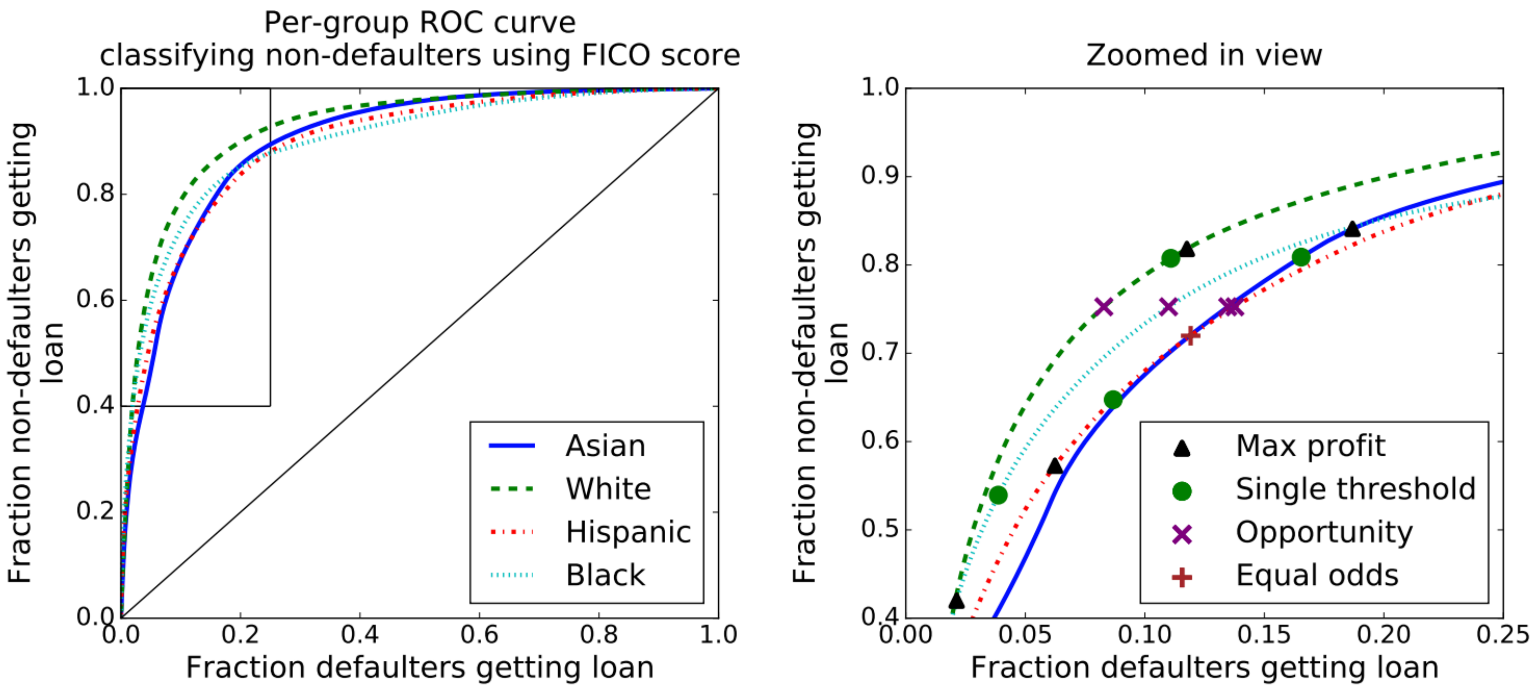
The ROC curve for using FICO score to identify non-defaulters



Q: Anything special about Equal Opportunity?

Case study: FICO

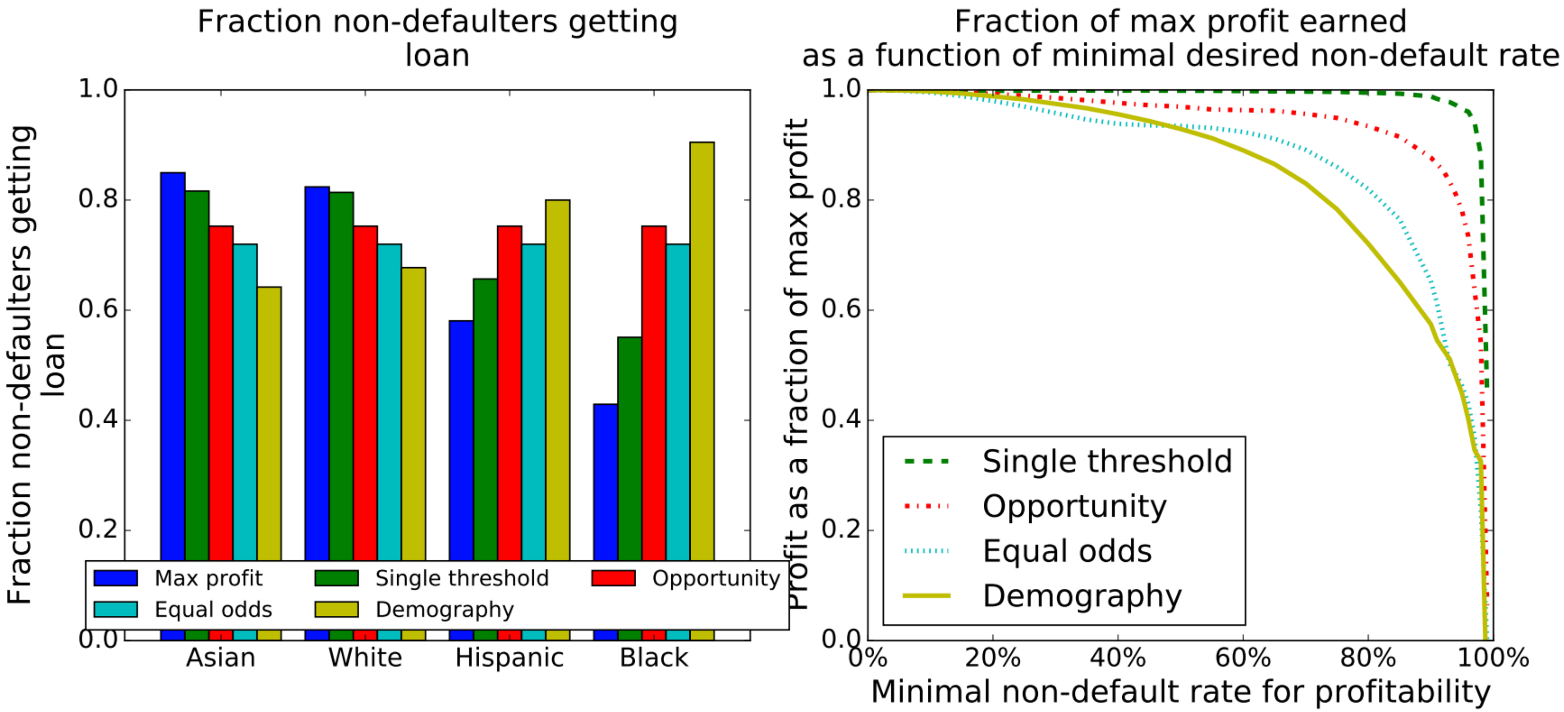
The ROC curve for using FICO score to identify non-defaulters



Equality of opportunity picks points along the same horizontal line. Equal odds picks a point below all lines.

Case study: FICO

Comparison of Five Fairness Criteria



Quick exercise

What fairness criteria do predictors $\hat{Y}_1$ and $\hat{Y}_2$ satisfy?

A = {race}, Y = {Hiring Decision}

Race and Ethnicity	Skill	Years of Exp	Goes to Mexican Markets?	Hiring Decision Y	Predictor $\hat{Y}_1$	Predictor $\hat{Y}_2$
Hispanic	Javascript	1	yes	no	0	1
Hispanic	C++	5	yes	yes	1	1
Hispanic	Python	1	no	yes	1	0
White	Java	2	no	yes	0	0
White	C++	3	no	yes	1	1
White	C++	0	no	no	1	0

From Dr. Wei Wei, Prof. James Landay's course at Stanford

Quick exercise

Consider $\hat{Y}_1$: A = {race}, Y = {Hiring Decision}

P(Ŷ1 = 1 | R = H) = 2/3
P(Ŷ1 = 1 | R = W) = 2/3

✓ Demographics Parity

$$P(\hat{Y} = 1|A = 1) = P(\hat{Y} = 1|A = 0)$$

Race and Ethnicity	Skill	Years of Exp	Goes to Mexican Markets?	Hiring Decision Y	Predictor $\hat{Y}_1$	Predictor $\hat{Y}_2$
Hispanic	Javascript	1	yes	no	0	1
Hispanic	C++	5	yes	yes	1	1
Hispanic	Python	1	no	yes	1	0
White	Java	2	no	yes	0	0
White	C++	3	no	yes	1	1
White	C++	0	no	no	1	0

Quick exercise

Consider $\hat{Y}_1$: A = {race}, Y = {Hiring Decision}

P(Ŷ1 = 1 | R = H, Y = yes) = 1
P(Ŷ1 = 1 | R = W, Y = yes) = 0.5
P(Ŷ1 = 1 | R = H, Y = no) = 0
P(Ŷ1 = 1 | R = W, Y = no) = 1

✗ Equality of Opportunity

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

✗ Equality of Odds

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

Race and Ethnicity	Skill	Years of Exp	Goes to Mexican Markets?	Hiring Decision Y	Predictor $\hat{Y}_1$	Predictor $\hat{Y}_2$
Hispanic	Javascript	1	yes	no	0	1
Hispanic	C++	5	yes	yes	1	1
Hispanic	Python	1	no	yes	1	0
White	Java	2	no	yes	0	0
White	C++	3	no	yes	1	1
White	C++	0	no	no	1	0

Quick exercise

Consider $\hat{Y}_2$: A = {race}, Y = {Hiring Decision}

P(Ŷ1 = 1 | R = H, Y = yes) = 1/2
P(Ŷ1 = 1 | R = W, Y = yes) = 1/2 ✓
P(Ŷ1 = 1 | R = H, Y = no) = 1
P(Ŷ1 = 1 | R = W, Y = no) = 0 ✗

✓ Equality of Opportunity

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

✗ Equality of Odds

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

Race and Ethnicity	Skill	Years of Exp	Goes to Mexican Markets?	Hiring Decision Y	Predictor $\hat{Y}_1$	Predictor $\hat{Y}_2$
Hispanic	Javascript	1	yes	no	0	1
Hispanic	C++	5	yes	yes	1	1
Hispanic	Python	1	no	yes	1	0
White	Java	2	no	yes	0	0
White	C++	3	no	yes	1	1
White	C++	0	no	no	1	0

Group versus Individual Fairness

Group Fairness

It is also refered to as statistical parity. It is a requirement that the protected groups should be treated similarly to the advantaged group or the populations as a whole.

Individual Fairness

It is a requirement that individuals should be treated consistently.

Group fairness does not consider the individual merits and may result in choosing the less qualified members of a group

Individual fairness assumes a similarity metric of the individuals that is generally hard to find

Group versus Individual Fairness

Fairness Criteria	Criteria	Group	Individual
Unawareness	Excludes A in Predictions		
Demographic Parity	$P(\hat{Y}=1 \mid A=0) = P(\hat{Y}=1 \mid A=1)$		
Equalized Odds	$P(\hat{Y}=1 \mid A=0, Y) = P(\hat{Y}=1 \mid A=1, Y)$		
Equalized Opportunity	$P(\hat{Y}=1 \mid A=0, Y=1) = P(\hat{Y}=1 \mid A=1, Y=1)$		

Part II

Fair Representation Learning

Introduction

Goal: Make Representations Fair

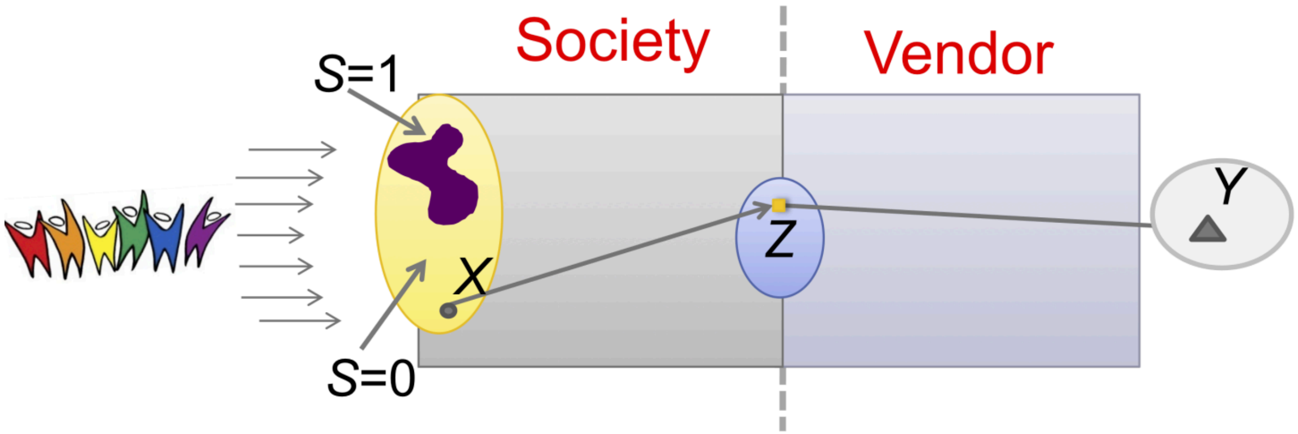


Image Credit: Richard Zemel

Learning fair representations

First approach: Zemel et al., 2013, "Learning fair representations"

Notation:

- X denotes the entire data set of individuals. Each $\mathbf{x} \in X$ is a vector of length D where each component of the vector describes some attribute of the person.
- S is a binary random variable representing whether or not a given individual is a member of the protected set; we assume the system has access to this attribute.
- Z is a multinomial random variable, where each of the K values represents one of the intermediate set of "prototypes". Associated with each prototype is a vector $\mathbf{v}_k$ in the same space as the individuals $\mathbf{x}$.
- Y is the binary random variable representing the classification decision for an individual, and $f : X \to Y$ is the desired classification function.
- d is a distance measure on X, e.g., simple Euclidean distance: $d\left(\mathbf{x}_n, \mathbf{v}_k\right) = \| \mathbf{x}_n - \mathbf{v}_k \|_2$.

Zemel, Rich, et al. "Learning fair representations." International conference on machine learning. PMLR, 2013.

Learning fair representations

First approach: Zemel et al., 2013, "Learning fair representations"

- Let Z be a discrete representation (like K-means)
- Determine Z stochastically based on distance to a prototype for the cluster (like the cluster center in K-means)

$$\Pr(Z = k \mid \mathbf{x}) \propto \exp(-d(\mathbf{x}, \mathbf{v}_k)),$$

  where d is some distance function (e.g. Euclidean distance)
- Use the Bayes classifier $y = \Pr(T = 1 \mid Z)$
- Need to fit the prototypes $\mathbf{v}_k$

Zemel, Rich, et al. "Learning fair representations." International conference on machine learning. PMLR, 2013.

Learning fair representations

First approach: Zemel et al., 2013, "Learning fair representations"

- Retain information about X: penalize reconstruction error

$$\mathcal{L}_{\mathrm{reconst}} = \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{x}^{(i)} - \tilde{\mathbf{x}}^{(i)}\|^2$$

- Predict accurately: cross-entropy loss

$$\mathcal{L}_{\mathrm{pred}} = \frac{1}{N} \sum_{i=1}^{N} -t^{(i)} \log y^{(i)} - (1 - t^{(i)}) \log(1 - y^{(i)})$$

- Obfuscate S:

$$\mathcal{L}_{\mathrm{discrim}} = \frac{1}{K} \sum_{k=1}^{K} \left| \frac{1}{N_0} \sum_{i:s^{(i)}=0} \Pr(Z = k \mid \mathbf{x}^{(i)}) - \frac{1}{N_1} \sum_{i:s^{(i)}=1} \Pr(Z = k \mid \mathbf{x}^{(i)}) \right|,$$

where we assume for simplicity $S \in \{0, 1\}$ and N_0 is the count for $s = 0$.

Zemel, Rich, et al. "Learning fair representations." International conference on machine learning. PMLR, 2013.

Learning fair representations

First approach: Zemel et al., 2013, "Learning fair representations"

- Obfuscate S:

$$\mathcal{L}_{\text{discrim}} = \frac{1}{K} \sum_{k=1}^{K} \left| \frac{1}{N_0} \sum_{i:s^{(i)}=0} \mathrm{Pr}(Z = k \mid \mathbf{x}^{(i)}) - \frac{1}{N_1} \sum_{i:s^{(i)}=1} \mathrm{Pr}(Z = k \mid \mathbf{x}^{(i)}) \right|,$$

- Is this about individual-level or group-level fairness?
- If discrimination loss is 0, we satisfy demographic parity

$$\begin{aligned}
\mathrm{Pr}(Y = 1 \mid s^{(i)} = 1) &= \frac{1}{N_1} \sum_{i:s^{(i)}=1} \sum_{k=1}^{K} \mathrm{Pr}(Z = k \mid \mathbf{x}^{(i)}) \mathrm{Pr}(Y = 1 \mid Z = k) \\
&= \sum_{k=1}^{K} \left[\frac{1}{N_1} \sum_{i:s^{(i)}=1} \mathrm{Pr}(Z = k \mid \mathbf{x}^{(i)}) \right] \mathrm{Pr}(Y = 1 \mid Z = k) \\
&= \sum_{k=1}^{K} \left[\frac{1}{N_0} \sum_{i:s^{(i)}=0} \mathrm{Pr}(Z = k \mid \mathbf{x}^{(i)}) \right] \mathrm{Pr}(Y = 1 \mid Z = k) \\
&= \mathrm{Pr}(Y = 1 \mid s^{(i)} = 0)
\end{aligned}$$

Zemel, Rich, et al. "Learning fair representations." International conference on machine learning. PMLR, 2013.

Learning fair representations

Datasets:

German Credit

Task: classify individual as good or bad credit risk

Sensitive feature: Age

Adult Income

Size: 45,222 instances, 14 attributes

Task: predict whether or not annual income > 50K

Sensitive feature: Gender

Heritage Health

Size: 147,473 instances, 139 attributes

Task: predict whether patient spends any nights in hospital

Sensitive feature: Age

Metrics

- Classification accuracy
- Discrimination

$$\left| \frac{\sum_{i:s^{(i)}=1}^{N} y^{(i)}}{N_1} - \frac{\sum_{i:s^{(i)}=0}^{N} y^{(i)}}{N_0} \right|$$

Zemel, Rich, et al. "Learning fair representations." International conference on machine learning. PMLR, 2013.

Learning fair representations

Datasets:

German Credit

Task: classify individual as good or bad credit risk

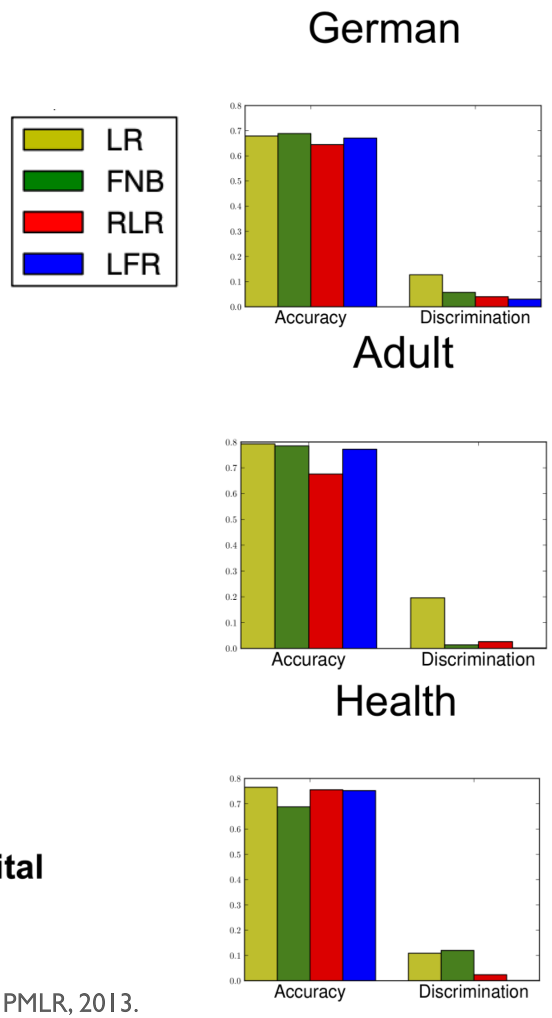
Sensitive feature: Age

Adult Income

- **Size:** 45,222 instances, 14 attributes
- **Task:** predict whether or not annual income > 50K
- **Sensitive feature:** Gender

Heritage Health

- **Size:** 147,473 instances, 139 attributes
- **Task:** predict whether patient spends any nights in hospital
- **Sensitive feature:** Age



Zemel, Rich, et al. "Learning fair representations." International conference on machine learning. PMLR, 2013.

Prejudice removing regularizer

Quantified Causes of Unfairness

Prejudice

- Unfairness rooted in the dataset

Underestimation

- Model unfairness because the model is not fully converged

Negative Legacy

- Unfairness due to sampling biases

Training Objective

$$-\mathcal{L}(\mathcal{D};\boldsymbol{\Theta}) + \eta \mathrm{R}(\mathcal{D},\boldsymbol{\Theta}) + \frac{\lambda}{2}\|\boldsymbol{\Theta}\|_2^2$$

Loss of the Model | Fairness Regularizer | L2 Regularizer

Kamishima, Toshihiro, et al. "Fairness-aware classifier with prejudice remover regularizer." 2012.

Prejudice removing regularizer

Quantified Causes of Unfairness

Prejudice

- Unfairness rooted in the dataset

Underestimation

- Model unfairness because the model is not fully converged

Negative Legacy

- Unfairness due to sampling biases

Training Objective

How to design it?

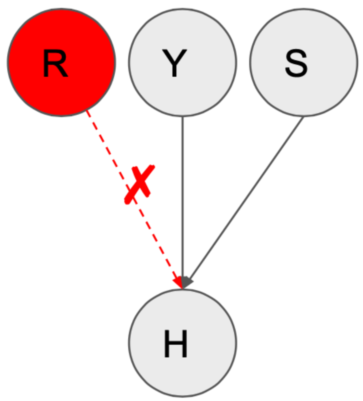
$$-\mathcal{L}(\mathcal{D};\boldsymbol{\Theta}) + \eta \mathrm{R}(\mathcal{D},\boldsymbol{\Theta}) + \frac{\lambda}{2}\|\boldsymbol{\Theta}\|_2^2$$

Loss of the Model | Fairness Regularizer | L2 Regularizer

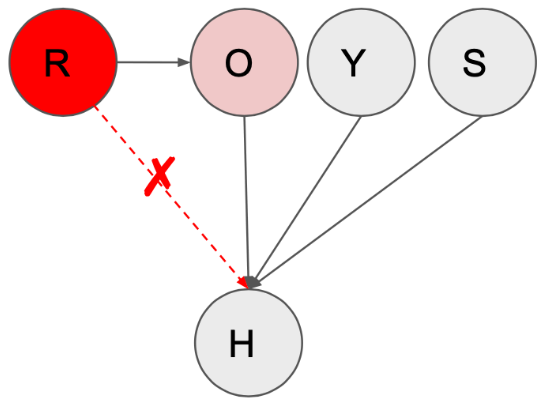
Kamishima, Toshihiro, et al. "Fairness-aware classifier with prejudice remover regularizer." 2012.

Limitations of fairness through unawareness

Outcomes: Fair ML Model



Indirect Discrimination

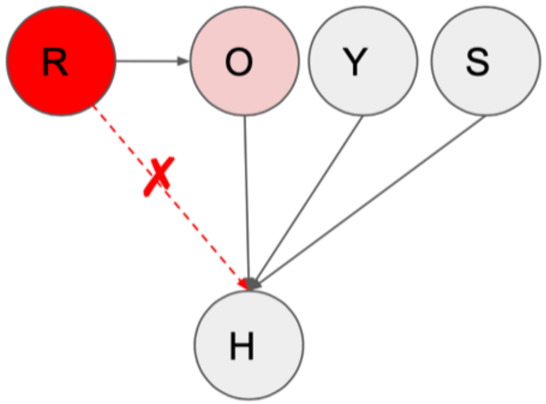


R - Race
Y - Years of Exp
S = Skills
O = Often Goes to Mexico Market

Kamishima, Toshihiro, et al. "Fairness-aware classifier with prejudice remover regularizer." 2012.

Prejudice removing regularizer

Recall Indirect Discrimination Happens When



Prediction is not directly conditioned on sensitive variables R

Prediction is indirectly conditioned on R by a variable O that is dependent on R

Prejudice Index (PI)

Measures the degree of indirect discrimination based on mutual information

$$\mathrm{PI} = \sum_{(y,s)\in\mathcal{D}} \hat{\Pr}[y,s] \ln \frac{\hat{\Pr}[y,s]}{\hat{\Pr}[y]\hat{\Pr}[s]}$$

$\hat{\Pr}$ denotes the sample distribution

Kamishima, Toshihiro, et al. "Fairness-aware classifier with prejudice remover regularizer." 2012.

Prejudice removing regularizer

Prejudice Index (PI)

Measures the degree of indirect discrimination based on mutual information

Range is $[0, +\infty)$

$$\mathrm{PI} = \sum_{(y,s)\in\mathcal{D}} \hat{\Pr}[y, s] \ln \frac{\hat{\Pr}[y, s]}{\hat{\Pr}[y]\hat{\Pr}[s]}$$

$\hat{\Pr}$ denotes the sample distribution

Normalized Prejudice Index (NPI)

Normalize PI by the entropy of Y *and* S

Range is $[0,1]$

$$\mathrm{NPI} = \mathrm{PI}/(\sqrt{H(Y)H(S)})$$

Kamishima, Toshihiro, et al. "Fairness-aware classifier with prejudice remover regularizer." 2012.

Prejudice removing regularizer

Learning and optimizing PI

Expands Pr(Y, S) into Σ_xPr(X, Y, S)

$$\mathrm{PI} = \sum_{Y,S} \hat{\Pr}[Y,S] \ln \frac{\hat{\Pr}[Y,S]}{\hat{\Pr}[S]\hat{\Pr}[Y]} = \sum_{X,S} \tilde{\Pr}[X,S] \sum_{Y} \mathcal{M}[Y|X,S;\boldsymbol{\Theta}] \ln \frac{\hat{\Pr}[Y,S]}{\hat{\Pr}[S]\hat{\Pr}[Y]}$$

double summations

triple summations

Prediction Model

Using Logistic Regression Model as the Prediction Model

$$\mathcal{M}[y|\mathbf{x},s;\boldsymbol{\Theta}] = y\sigma(\mathbf{x}^\top \mathbf{w}_s) + (1-y)(1-\sigma(\mathbf{x}^\top \mathbf{w}_s))$$

Kamishima, Toshihiro, et al. "Fairness-aware classifier with prejudice remover regularizer." 2012.

Prejudice removing regularizer

Learning and optimizing PI

Expands Pr(Y, S) into Σ_xPr(X, Y, S)

$$\mathrm{PI} = \sum_{Y,S} \hat{\Pr}[Y,S] \ln \frac{\hat{\Pr}[Y,S]}{\hat{\Pr}[S]\hat{\Pr}[Y]} = \sum_{X,S} \tilde{\Pr}[X,S] \sum_{Y} \mathcal{M}[Y|X,S;\boldsymbol{\Theta}] \ln \frac{\hat{\Pr}[Y,S]}{\hat{\Pr}[S]\hat{\Pr}[Y]}$$

double summations

triple summations

Prediction Model

$$= \sum_{(\mathbf{x}_i, s_i) \in \mathcal{D}} \sum_{y \in \{0,1\}} \mathcal{M}[y|\mathbf{x}_i, s_i; \boldsymbol{\Theta}] \ln \frac{\hat{\Pr}[y|s_i]}{\hat{\Pr}[y]}$$

Kamishima, Toshihiro, et al. "Fairness-aware classifier with prejudice remover regularizer." 2012.

Prejudice removing regularizer

Learning and optimizing PI

Expands Pr(Y, S) into Σ_xPr(X, Y, S)

$$\mathrm{PI} = \sum_{Y,S} \hat{\Pr}[Y,S] \ln \frac{\hat{\Pr}[Y,S]}{\hat{\Pr}[S]\hat{\Pr}[Y]} = \sum_{X,S} \tilde{\Pr}[X,S] \sum_{Y} \mathcal{M}[Y|X,S;\boldsymbol{\Theta}] \ln \frac{\hat{\Pr}[Y,S]}{\hat{\Pr}[S]\hat{\Pr}[Y]}$$

double summations

triple summations

Prediction Model

$$= \sum_{(\mathbf{x}_i,s_i)\in\mathcal{D}} \sum_{y\in\{0,1\}} \mathcal{M}[y|\mathbf{x}_i,s_i;\boldsymbol{\Theta}] \ln \frac{\hat{\Pr}[y|s_i]}{\hat{\Pr}[y]}$$

hard to estimate

$$\hat{\Pr}[y|s] = \int_{\mathrm{dom}(X)} \Pr{}^*[X|s]\mathcal{M}[y|X,s;\boldsymbol{\Theta}]dX$$

Integrals Are Difficult to Evaluate

Kamishima, Toshihiro, et al. "Fairness-aware classifier with prejudice remover regularizer." 2012.

Prejudice removing regularizer

Learning and optimizing PI

$$\sum_{(\mathbf{x}_i, s_i)\in\mathcal{D}} \sum_{y\in\{0,1\}} \mathcal{M}[y|\mathbf{x}_i, s_i; \boldsymbol{\Theta}] \ln \frac{\hat{\Pr}[y|s_i]}{\hat{\Pr}[y]}$$

hard to estimate

$$\hat{\Pr}[y] \approx \frac{\sum_{(\mathbf{x}_i, s_i)\in\mathcal{D}} \mathcal{M}[y|\mathbf{x}_i, s_i; \boldsymbol{\Theta}]}{|\mathcal{D}|}$$

$$\hat{\Pr}[y|s] = \int_{\mathrm{dom}(X)} \Pr{}^*[X|s]\mathcal{M}[y|X, s; \boldsymbol{\Theta}]dX$$

Integrals Are Difficult to Evaluate

$$\approx \frac{\sum_{(\mathbf{x}_i, s_i)\in\mathcal{D} \text{ s.t. } s_i=s} \mathcal{M}[y|\mathbf{x}_i, s; \boldsymbol{\Theta}]}{|\{(\mathbf{x}_i, s_i) \in \mathcal{D} \text{ s.t. } s_i = s\}|}$$

Approximating integrals by sample means

Kamishima, Toshihiro, et al. "Fairness-aware classifier with prejudice remover regularizer." 2012.

Prejudice removing regularizer

Summary

Optimization Target

$$-\mathcal{L}(\mathcal{D};\boldsymbol{\Theta}) + \eta \mathrm{R}(\mathcal{D},\boldsymbol{\Theta}) + \frac{\lambda}{2}\|\boldsymbol{\Theta}\|_2^2$$

Loss of the Model   Fairness Regularizer   L2 Regularizer

Fairness Regularizer

$$\mathrm{PI} = \sum_{(\mathbf{x}_i,s_i)\in\mathcal{D}} \sum_{y\in\{0,1\}} \mathcal{M}[y|\mathbf{x}_i,s_i;\boldsymbol{\Theta}] \ln \frac{\hat{\Pr}[y|s_i]}{\hat{\Pr}[y]}$$

Kamishima, Toshihiro, et al. "Fairness-aware classifier with prejudice remover regularizer." 2012.

Prejudice removing regularizer

Results

- Prejudice Prior Sacrifices Model Performance
    - PR has lower Acc (Accuracy)
    - PR has lower NMI (normalized mutual information between labels and predictions)
- Prejudice Prior Makes Model Fair
    - PR has lower NPI

Logistic Regression full fet. → LR
Logistic Regression no sensitive fet. → LRns
Logistic Regression + Prejudice Regularizer → PR η=5, PR η=15, PR η=30

method	Acc	NMI	NPI	PI/MI
LR	0.851	0.267	5.21E-02	2.10E-01
LRns	0.850	0.266	4.91E-02	1.99E-01
PR $\eta=5$	0.842	0.240	4.24E-02	1.91E-01
PR $\eta=15$	0.801	0.158	2.38E-02	1.62E-01
PR $\eta=30$	0.769	0.046	1.68E-02	3.94E-01

η is the weight we put on prejudice regularizers

Kamishima, Toshihiro, et al. "Fairness-aware classifier with prejudice remover regularizer." 2012.

Prejudice removing regularizer

Results

- PI/MI
  - Prejudice Index / Mutual Information
  - Demonstrates a trade-offs between model fairness and performance
  - Measures the amount of discrimination we eliminate with one unit of performance gain (measured by MI)

Logistic Regression full fet. → LR

Logistic Regression no sensitive fet. → LRns

Logistic Regression + Prejudice Regularizer → PR η=5, PR η=15, PR η=30

method	Acc	NMI	NPI	PI/MI
LR	0.851	0.267	5.21E-02	2.10E-01
LRns	0.850	0.266	4.91E-02	1.99E-01
PR η=5	0.842	0.240	4.24E-02	1.91E-01
PR η=15	0.801	0.158	2.38E-02	1.62E-01
PR η=30	0.769	0.046	1.68E-02	3.94E-01

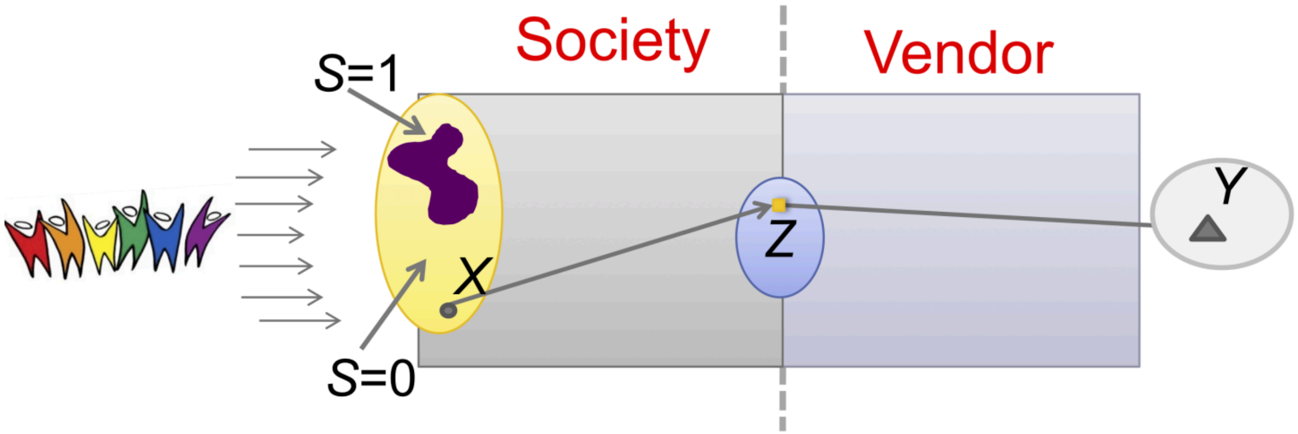
η is the weight we put on prejudice regularizers

Kamishima, Toshihiro, et al. "Fairness-aware classifier with prejudice remover regularizer." 2012.

Part III

Disentangled fair representations

Fair VAE



Recall: "Learning fair representations"

- Discrete Z based on prototypes is very limiting. Can we learn a more flexible representation?

Louizos, Christos, et al. "The variational fair autoencoder." *arXiv preprint arXiv:1511.00830* (2015).

Fair VAE

Maximum Mean Discrepancy

- Our previous non-discrimination criterion only makes sense for discrete Z.
- New criterion: ensure that $p(Z \mid s)$ is indistinguishable for different values of s.
- Maximum mean discrepancy (MMD) is a quantitative measure of distance between two distributions. Pick a feature map ψ.

$$\mathrm{MMD}(p; q) = \left\| \mathbb{E}_{\mathbf{z}\sim p}[\psi(\mathbf{z})] - \mathbb{E}_{\mathbf{z}\sim q}[\psi(\mathbf{z})] \right\|^2$$

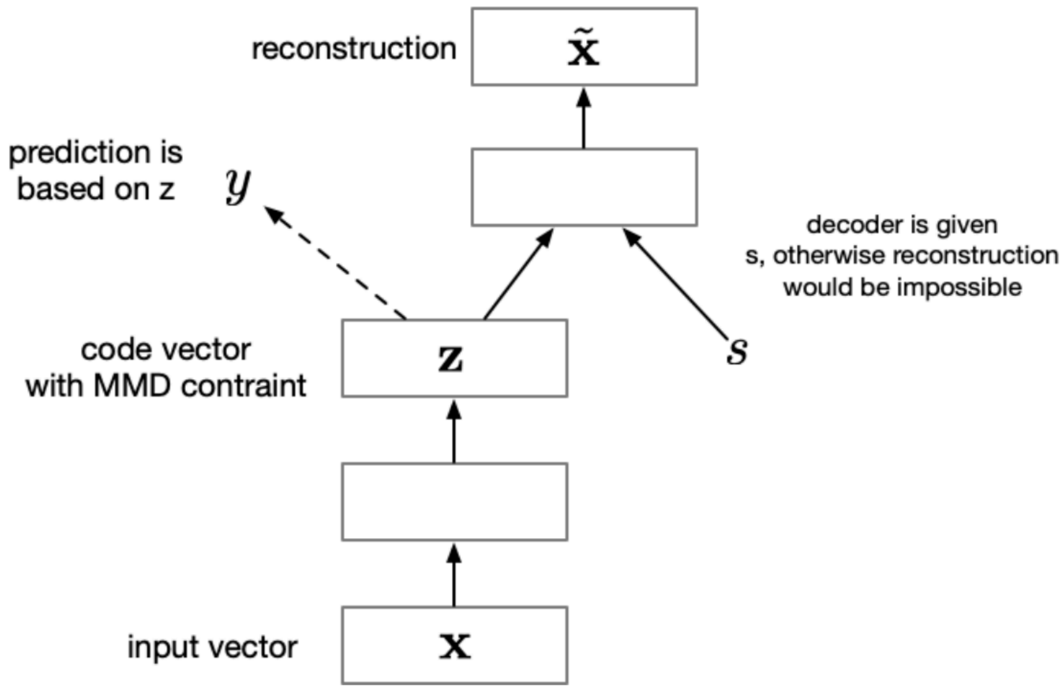
- If ψ is sufficiently expressive, then the MMD is only 0 if the distributions match. (Making this precise requires the idea of *kernels*.)

Louizos, Christos, et al. "The variational fair autoencoder." *arXiv preprint arXiv:1511.00830* (2015).

Fair VAE

Training VAE

Train a VAE, with the constraint that the MMD between $p(\mathbf{z} \mid s = 0)$ and $p(\mathbf{z} \mid s = 1)$ is small.

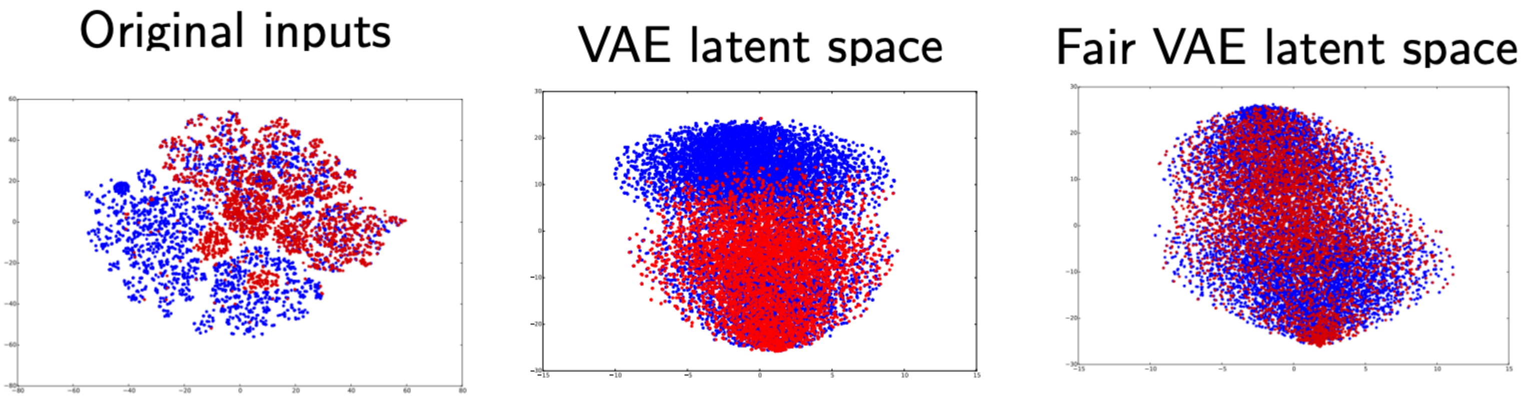


Louizos, Christos, et al. "The variational fair autoencoder." *arXiv preprint arXiv:1511.00830* (2015).

Fair VAE

tSNE embeddings

- tSNE is an unsupervised learning algorithm for visualizing high-dimensional datasets. It tries to embed points in low dimensions in a way that preserves distances as accurately as possible.
- Here are tSNE embeddings of different distributions, color-coded by the sensitive feature:



Louizos, Christos, et al. "The variational fair autoencoder." *arXiv preprint arXiv:1511.00830* (2015).

Flexibly fair representation

How to achieve demographic parity in VAE?

- $x \in \mathcal{X}$: a vector of non-sensitive attributes, for example, the pixel values in an image or row of features in a tabular dataset;
- $a \in \{0,1\}^{N_a}$: a vector of binary sensitive attributes;
- $z \in \mathbb{R}^{N_z}$: non-sensitive subspace of the latent code;
- $b \in \mathbb{R}^{N_b}$: sensitive subspace of the latent code.

Original VAE objective:

$$L_{\text{VAE}}(p, q) = \mathbb{E}_{q(z,b|x,a)} \left[\log p(x, a|z, b)\right] - D_{KL}\left[q(z, b|x, a)||p(z, b)\right]$$

Creager, Elliot, et al. "Flexibly fair representation learning by disentanglement." *International conference on machine learning*. PMLR, 2019.

Flexibly fair representation

How to achieve demographic parity in VAE?

- $x \in \mathcal{X}$: a vector of non-sensitive attributes, for example, the pixel values in an image or row of features in a tabular dataset;
- $a \in \{0,1\}^{N_a}$: a vector of binary sensitive attributes;
- $z \in \mathbb{R}^{N_z}$: non-sensitive subspace of the latent code;
- $b \in \mathbb{R}^{N_b}$: sensitive subspace of the latent code.

Flexibly fair VAE objective:

$$L_{\text{FFVAE}}(p,q) = \mathbb{E}_{q(z,b|x)}[\log p(x|z,b) + \alpha \log p(a|b)] - \gamma D_{KL}(q(z,b)||q(z)\prod_j q(b_j)) - D_{KL}\left[q(z,b|x)||p(z,b)\right].$$

Creager, Elliot, et al. "Flexibly fair representation learning by disentanglement." *International conference on machine learning*. PMLR, 2019.

Flexibly fair representation

How to achieve demographic parity in VAE?

$$L_{\text{FFVAE}}(p, q) = \mathbb{E}_{q(\underline{z,b}|x,a)} [\log p(x, a|z, b)]$$

$z \perp b_j$ $\quad - D_{KL} [q(z, b|x)||p(z, b)]$

$p(z,b) = p(z)p(b)$

Standard Gaussian Uniform

β-VAE

$b_i \perp b_j \; \forall \, i \neq j$ $\quad - \gamma D_{KL}(q(z, b)||q(z) \prod_j q(b_j))$

factor-VAE

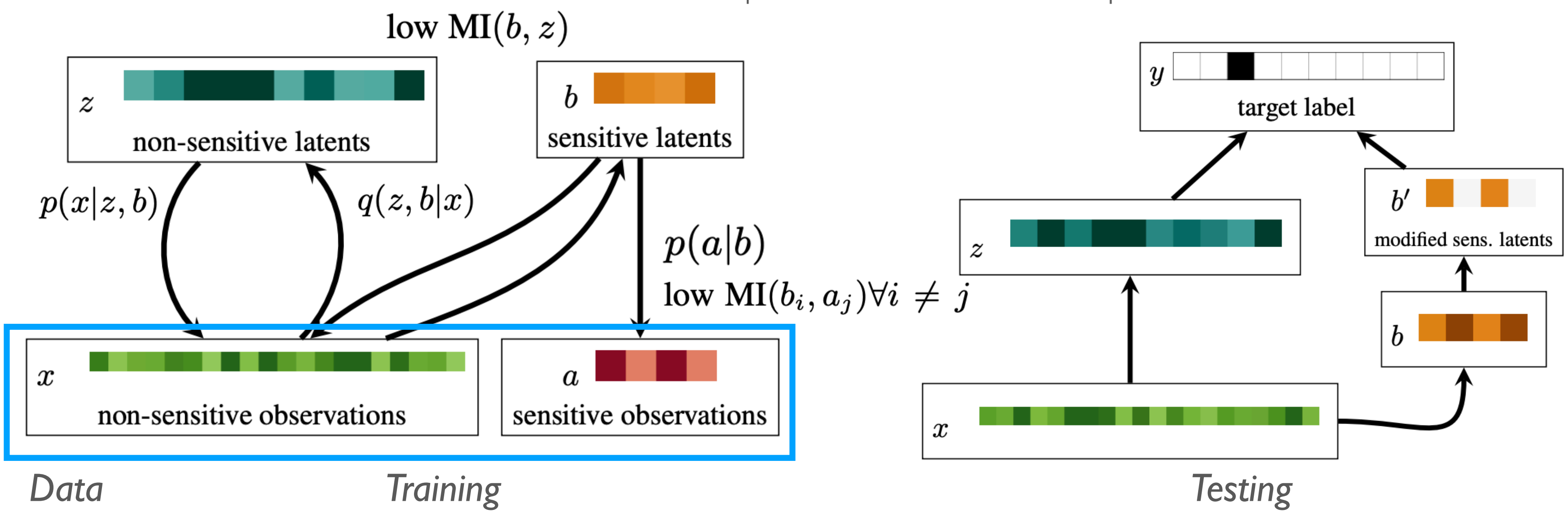
(Since $p(x, a|z, b) = p(x|z, b)p(a|b)$)

Creager, Elliot, et al. "Flexibly fair representation learning by disentanglement." *International conference on machine learning*. PMLR, 2019.

Flexibly fair representation

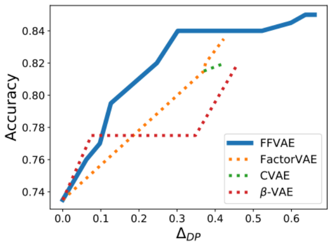
Applications

- Fair Classification
  - Make fair predictions
- Predictiveness
  - Train a classifier to predict sensitive attribute a_i from b_i alone
- Disentanglement
  - Train a classifier to predict sensitive attribute a_i from representations with b_i removed

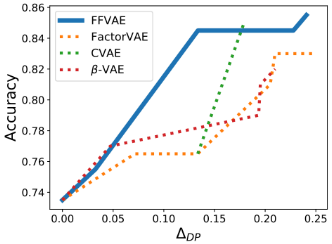


Creager, Elliot, et al. "Flexibly fair representation learning by disentanglement." *International conference on machine learning*. PMLR, 2019.

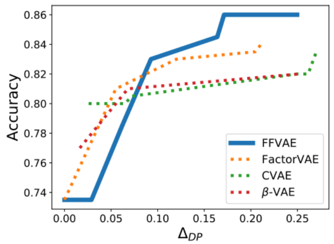
Flexibly fair representation



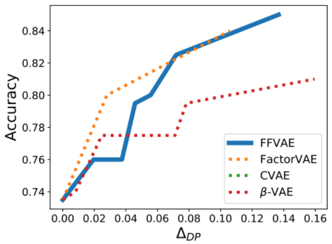
(a) $a = \mathrm{R}$



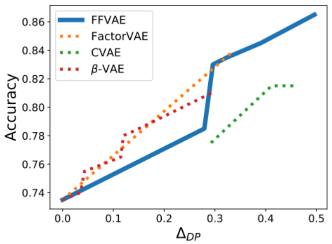
(b) $a = \mathrm{B}$



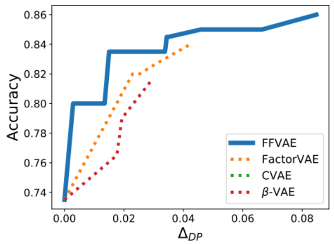
(c) $a = \mathrm{P}$



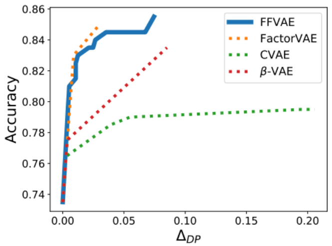
(d) $a = \mathrm{R} \vee \mathrm{B}$



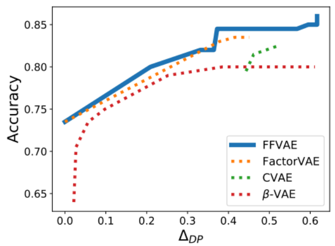
(e) $a = \mathrm{R} \vee \mathrm{P}$



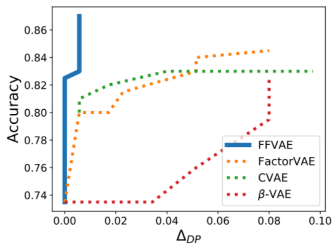
(f) $a = \mathrm{B} \vee \mathrm{P}$



(g) $a = \mathrm{R} \wedge \mathrm{B}$



(h) $a = \mathrm{R} \wedge \mathrm{P}$



(i) $a = \mathrm{B} \wedge \mathrm{P}$

Fair classification

$$\Delta_{DP}(g) \triangleq d_g(\mathcal{Z}_0, \mathcal{Z}_1) = |\mathbb{E}_{\mathcal{Z}_0}[g] - \mathbb{E}_{\mathcal{Z}_1}[g]|$$

$$\Delta_{DP}(g) = 0 \iff g(Z) \perp\!\!\!\perp A$$

- Sensitive attributes:
  - racePctBlack (R)
  - blackPerCapIncome (B)
  - pctNotSpeakEnglWell (P)
- y = violentCrimesPerCaptia

Creager, Elliot, et al. "Flexibly fair representation learning by disentanglement." *International conference on machine learning*. PMLR, 2019.

Part III

Summary

Learning Outcomes

- Understand why fairness matters in ML
- Be able to describe key fairness criteria
- Be able to identify the difference between individual and group fairness criteria
- Understand how to ensure fairness in representation learning
- Know how to ensure fairness in VAE