Fair Representation Learning

Apr 10, 2020

Dr. Wei Wei, Prof. James Landay

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

Stanford University

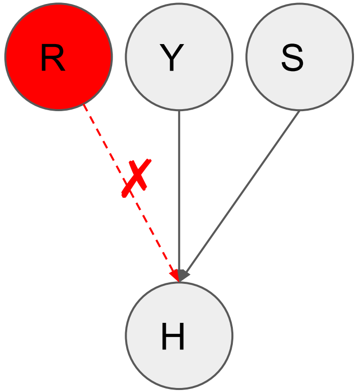
Updated Project Policies

- Maximum Number of Students For Course Projects
  - We now allow up to 3 students in a project
- Project Sharing
  - Project sharing between classes can be done under the permissions from the Instructors
- Reminder: Project Proposal Deadline
  - Apr 22, before class
  - Less than two weeks from now

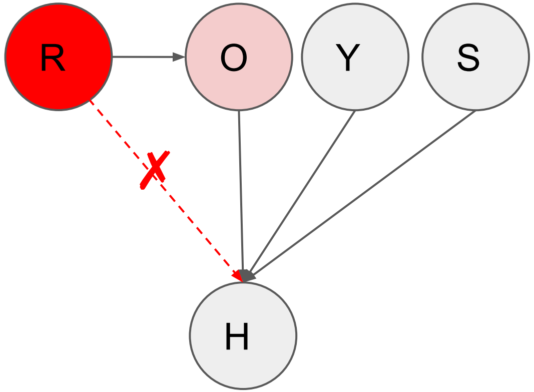
Recaps From the Previous Lecture

- Fairness Through Unawareness

Outcomes: Fair ML Model



Indirect Discrimination



R - Race
Y - Years of Exp

S = Skills
O = Often Goes to Mexico Market

Limitations

- Processing Sensitive Features
  - Fairness through unawareness requires sensitive features to be masked out
  - Not easy to do in real life
  - Referred to as individual fairness criteria


❖ Stereotypical dataset

The physician hired the secretary because he was overwhelmed with clients.

The physician hired the secretary because she was highly recommended.

❖ Anti-stereotypical dataset

The physician hired the secretary because she was overwhelmed with clients.

The physician hired the secretary because he was highly recommended.

Outline

- Major Fairness Criteria
  - Demographic Parity
  - Equality of Odds/Opportunity
  - FICO Case Study
- Fair Representation Learning
  - Prejudice Removing Regularizer

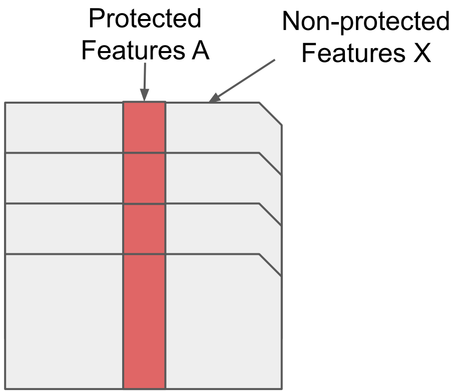
Demographic Parity

- 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

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

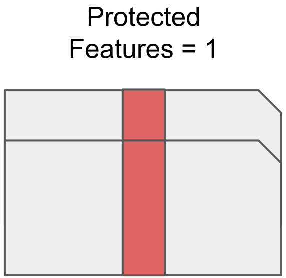
Comparisons

Individual Treatment

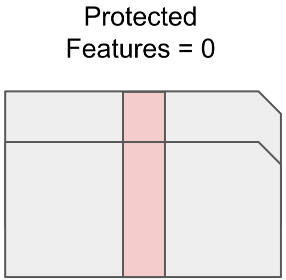


Fairness Through Unawareness
$P(\hat{Y} \mid X)$

Group Treatment



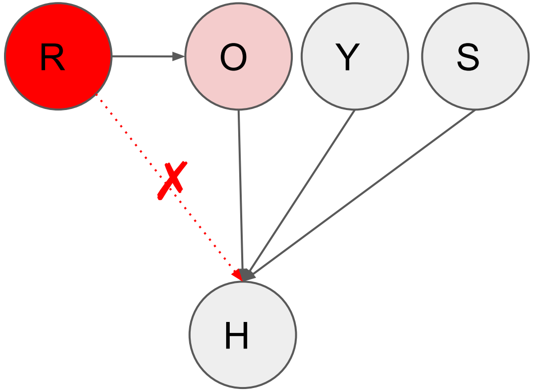
Demographic Parity
$P(\hat{Y}=1 \mid A=1)$



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

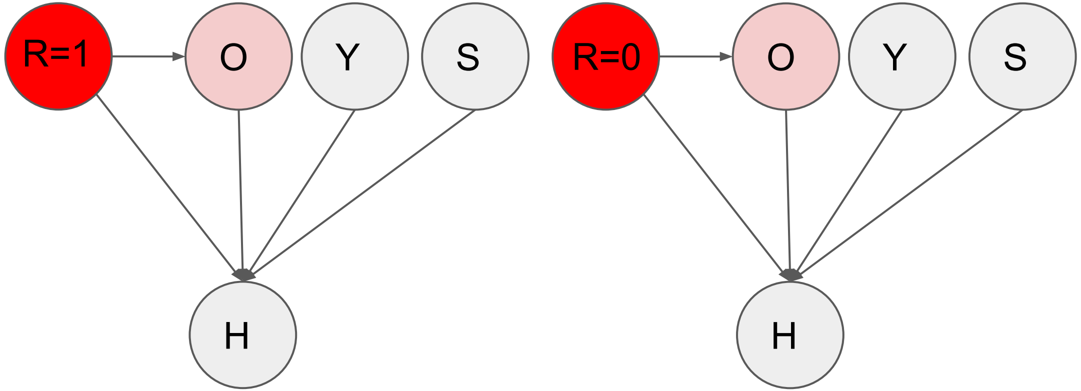
Graphical Model Explanations

Individual Treatment



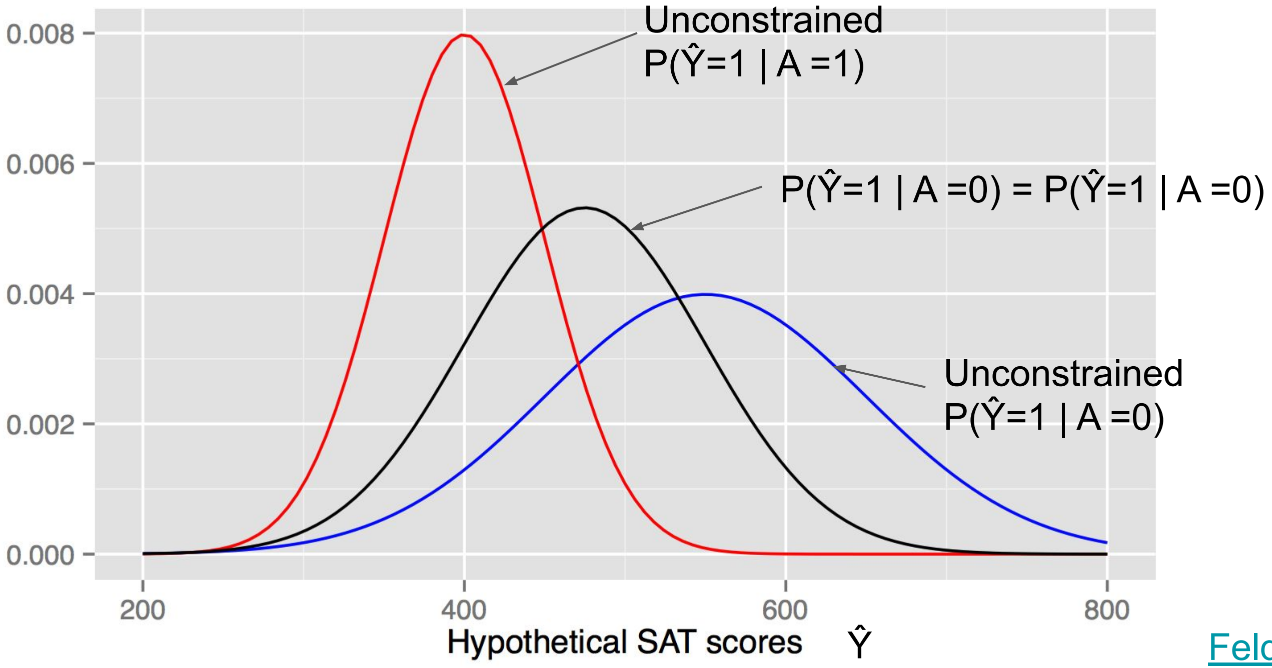
$P(H \mid O, Y, S)$

Group Treatment



$P(H=1 \mid R=1) = P(H=1 \mid R=0)$

SAT Score Prediction



Feldman et al, 2015

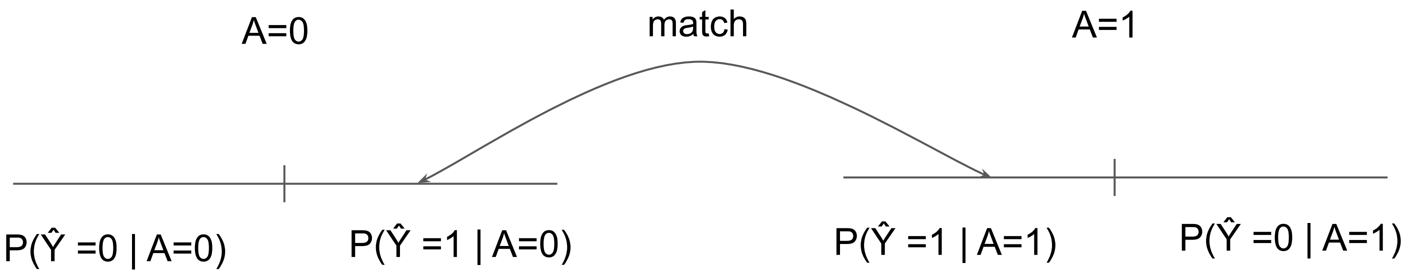
Issues With Demographic Parity

- Correlates Too Much With the Performance of the Predictor

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

predictor

predictor



Issues With Demographic Parity

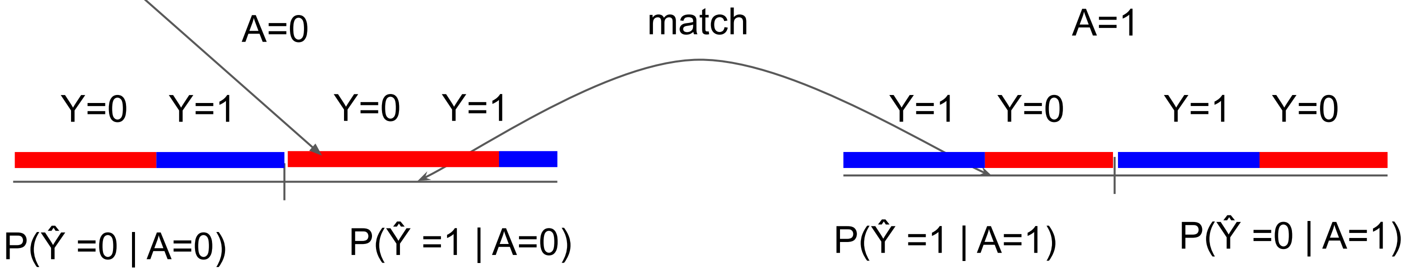
- Correlates Too Much With the Performance of the Predictor

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

predictor

predictor

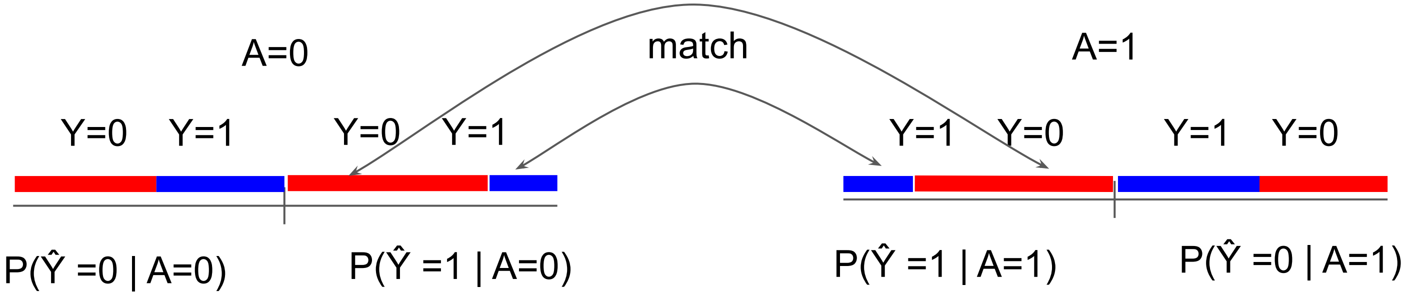
Accepted too many who are not qualified



Equality of Odds (Hardt et al, 2016)

- Equal Probabilities for Both Qualified/Unqualified People Across Protected Groups

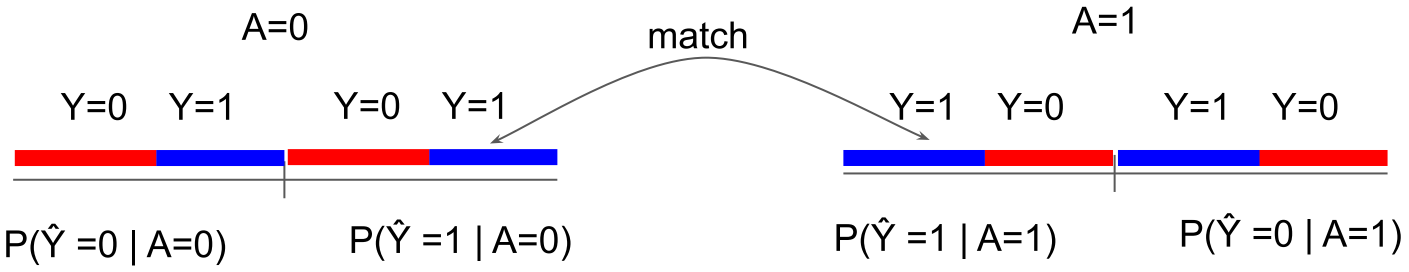
$$P(\hat{Y} = 1 | A = 0, Y) = P(\hat{Y} = 1 | A = 1, Y)$$



Equality of Opportunity (Hardt et al, 2016)

- Equal Probabilities for Qualified People Across Protected Groups

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

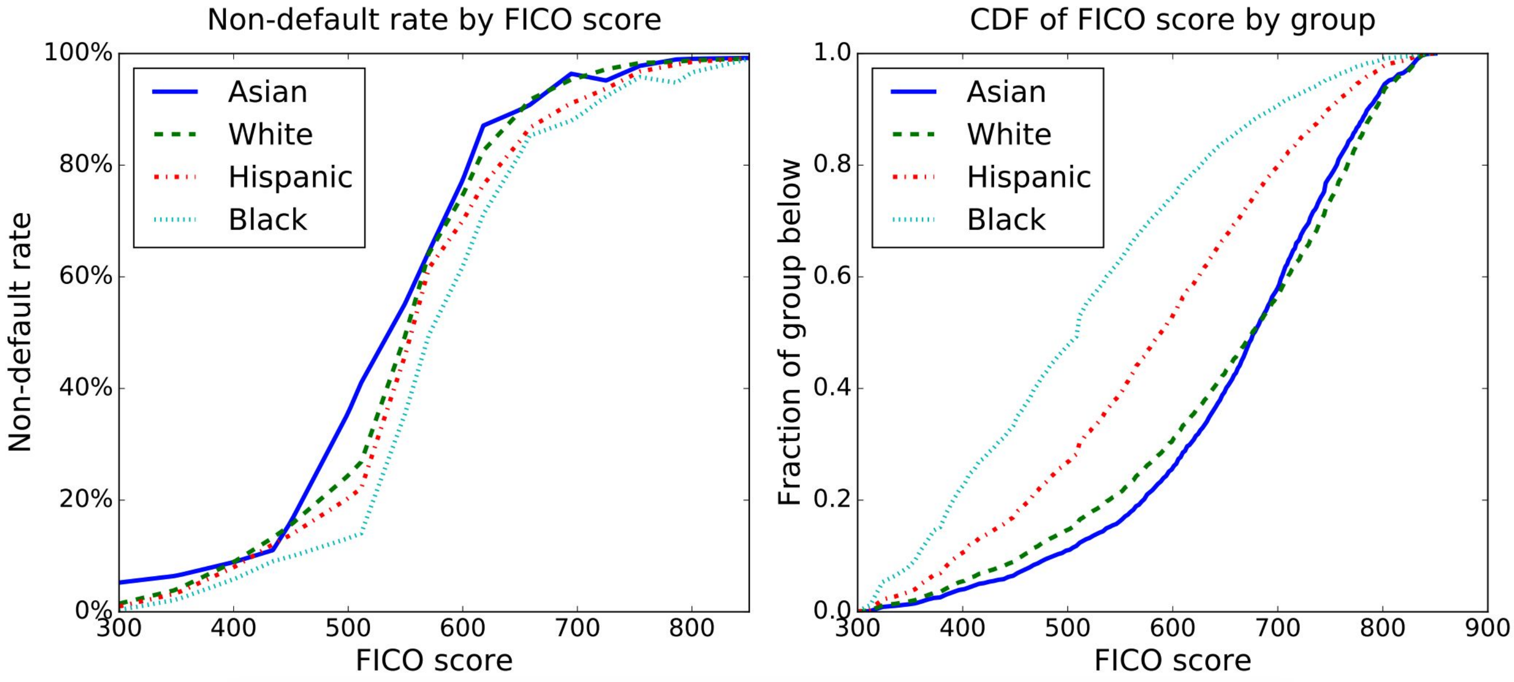


Case Study on FICO

- 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}

FICO Scores

- 18% Default Rate on Any Accounts Corresponds to a 2% Default Rate for New Loans



Making Lending Decisions Without Discriminating

- Requirement: **Default Rate < 18%**, Simple Threshold Model
  - Max Profit - No Fairness Constraints
  - Race Blind - Using the same threshold for all race groups

Making Lending Decisions Without Discriminating

- Requirement: **Default Rate < 18%**, Simple Threshold Model
  - 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

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

Making Lending Decisions Without Discriminating

- Requirement: **Default Rate < 18%, Simple Threshold Model**
  - 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

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

Making Lending Decisions Without Discriminating

- Requirement: **Default Rate < 18%**, Simple Threshold Model
  - 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

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

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

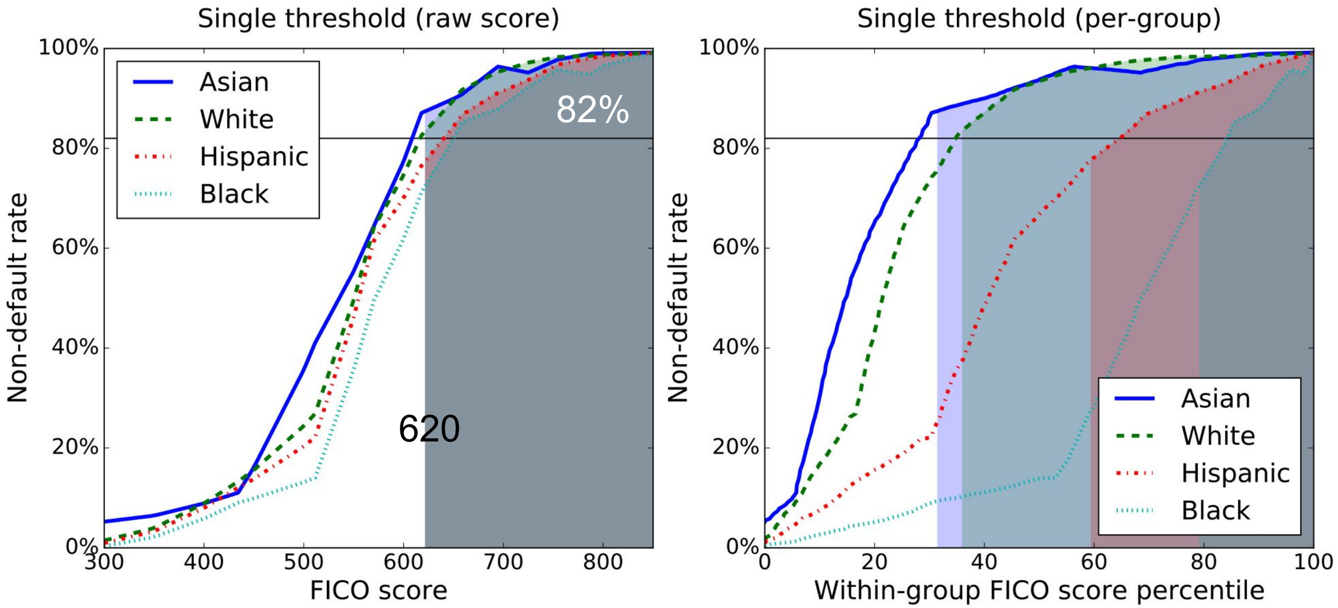
$$P(\underline{\hat{Y} = 1}|\underline{A = 0}, \underline{Y = 1}) = P(\underline{\hat{Y} = 1}|\underline{A = 1}, \underline{Y = 1})$$

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

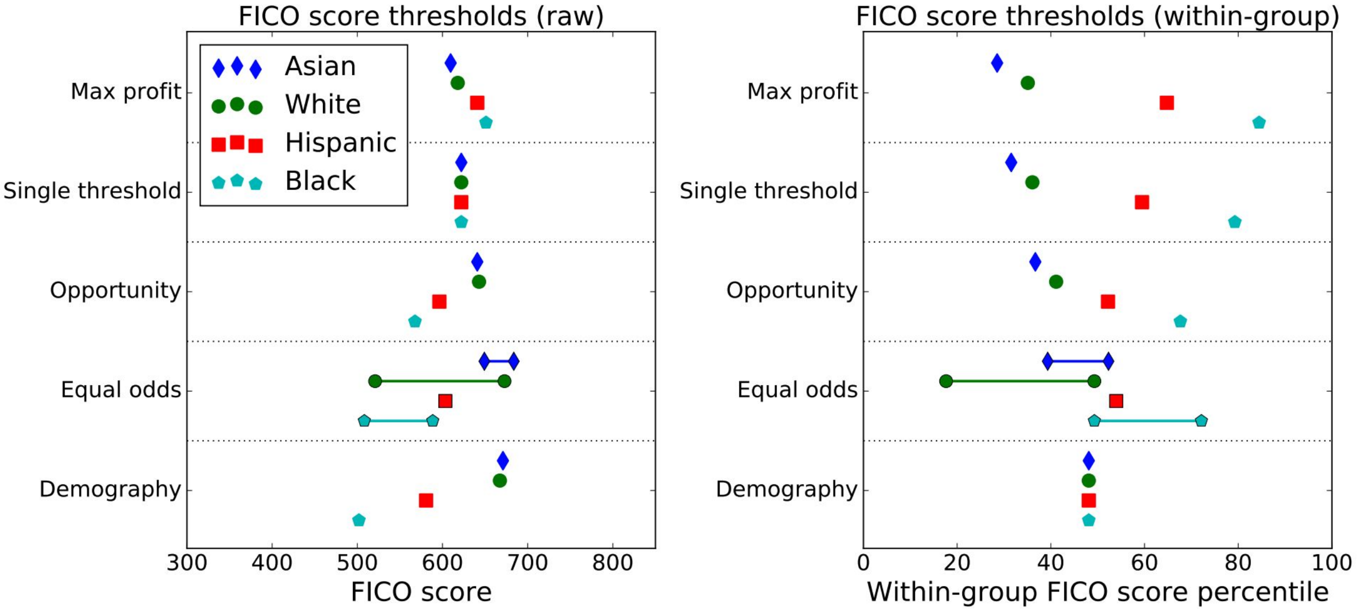
$$P(\underline{\hat{Y} = 1}|\underline{A = 0}, \underline{Y}) = P(\underline{\hat{Y} = 1}|\underline{A = 1}, \underline{Y})$$

Credit Modeling Using A Single Threshold

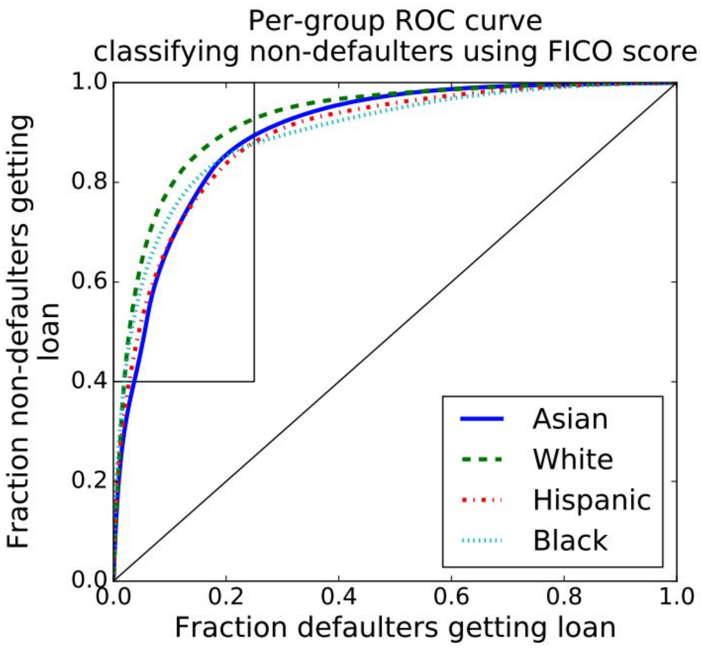
- Within-Group Percentile Differs Dramatically for Each Group

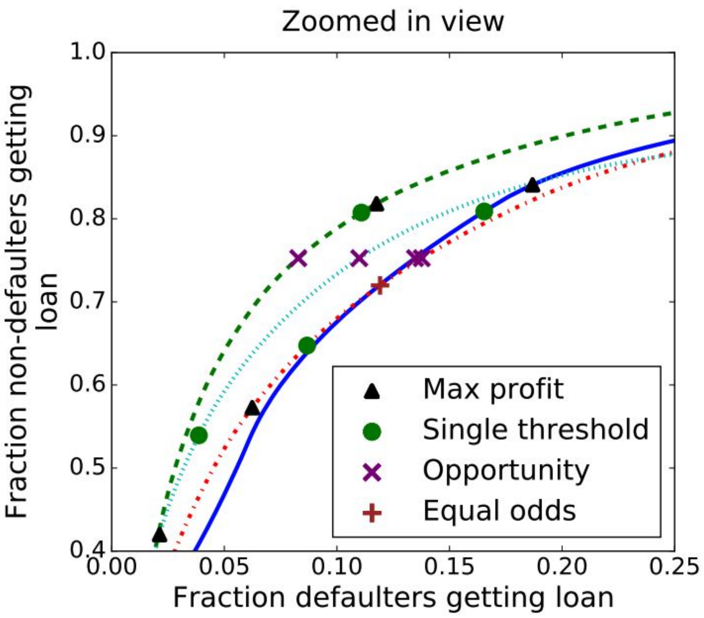


Found Thresholds for Each Fairness Definitions

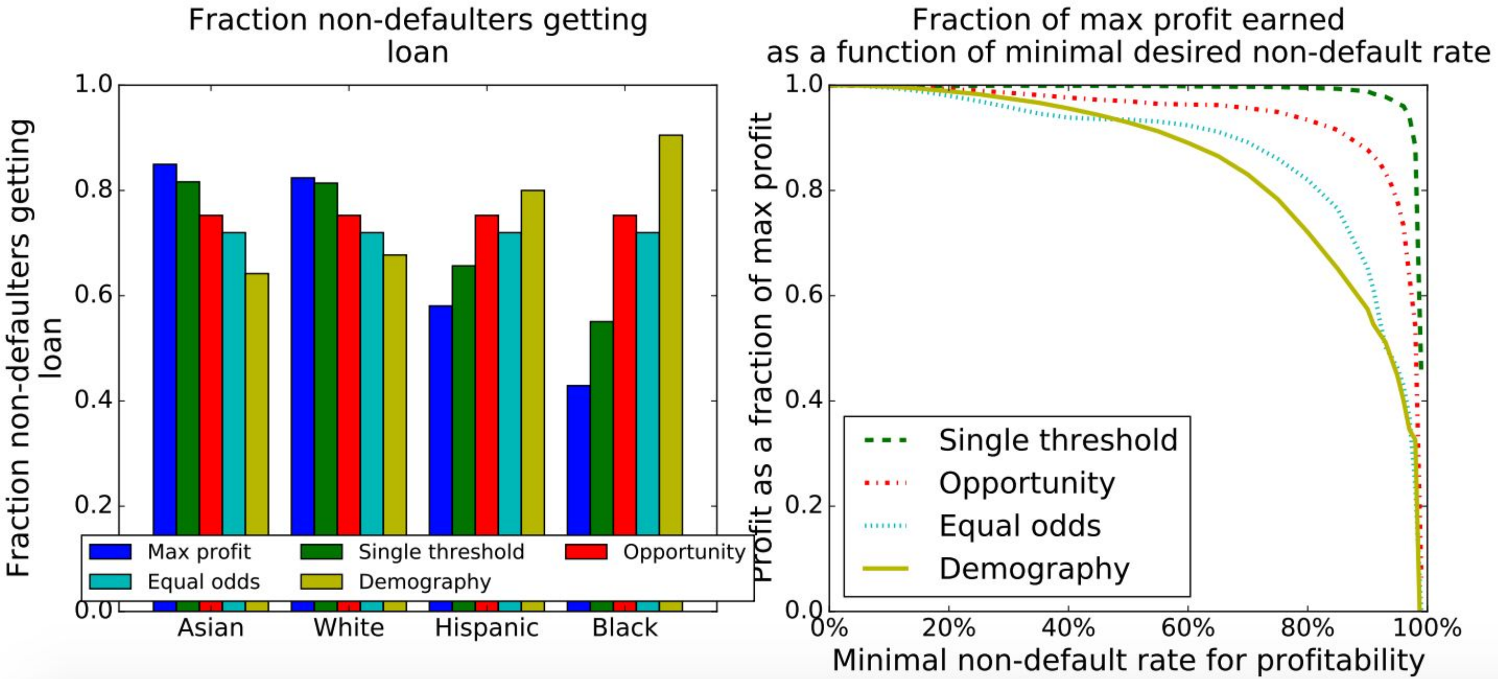


Identifying Non-Defaulters





Non-Defaulters and Max Profits



Practice Question

- Find out the Fairness Criteria that $\hat{Y}1$, 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

Demographic Parity for Predictor $\hat{Y}1$

- $P(\hat{Y}1 = 1 \mid R = H) = 2/3$
- $P(\hat{Y}1 = 1 \mid R = W) =$

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

Demographic Parity for Predictor Ŷ1

- $P(\hat{Y}1 = 1 \mid R = H) = 2/3$
- $P(\hat{Y}1 = 1 \mid R = W) = 2/3$

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

Demographic Parity for Predictor Ŷ1

- $P(\hat{Y}1 = 1 \mid R = H) = 2/3$
- $P(\hat{Y}1 = 1 \mid R = W) = 2/3$

✅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

Equality of Opportunity/Odds for Predictor $\hat{Y}1$

- $P(\hat{Y}1 = 1 \mid R = H, Y = yes) = 1$
- $P(\hat{Y}1 = 1 \mid R = W, Y = yes) =$
- $P(\hat{Y}1 = 1 \mid R = H, Y = no) =$
- $P(\hat{Y}1 = 1 \mid R = W, Y = no) =$

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

Equality of Opportunity/Odds for Predictor $\hat{Y}1$

- $P(\hat{Y}1 = 1 \mid R = H, Y = yes) = 1$
- $P(\hat{Y}1 = 1 \mid R = W, Y = yes) = 0.5$
- $P(\hat{Y}1 = 1 \mid R = H, Y = no) =$
- $P(\hat{Y}1 = 1 \mid R = W, Y = no) =$

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

Equality of Opportunity/Odds for Predictor Ŷ1

- 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) =

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

Equality of Opportunity/Odds for Predictor $\hat{Y}1$

- $P(\hat{Y}1 = 1 \mid R = H, Y = yes) = 1$
- $P(\hat{Y}1 = 1 \mid R = W, Y = yes) = 0.5$
- $P(\hat{Y}1 = 1 \mid R = H, Y = no) = 0$
- $P(\hat{Y}1 = 1 \mid R = W, Y = no) = 1$

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

Equality of Opportunity/Odds for Predictor Ŷ1

- 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

Demographic Parity for Predictor $\hat{Y}2$

- $P(\hat{Y}1 = 1 \mid R = H) = 2/3$
- $P(\hat{Y}1 = 1 \mid R = W) =$

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

Demographic Parity for Predictor $\hat{Y}2$

- $P(\hat{Y}1 = 1 \mid R = H) = 2/3$
- $P(\hat{Y}1 = 1 \mid R = W) = 1/3$

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

Demographic Parity for Predictor Ŷ2

- $P(\hat{Y}1 = 1 \mid R = H) = 2/3$
- $P(\hat{Y}1 = 1 \mid R = W) = 1/3$

✗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

Equality of Opportunity/Odds for Predictor Ŷ2

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

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

Equality of Opportunity/Odds for Predictor $\hat{Y}2$

- $P(\hat{Y}1 = 1 \mid R = H, Y = \text{yes}) = 1/2$
- $P(\hat{Y}1 = 1 \mid R = W, Y = \text{yes}) = 1/2$
- $P(\hat{Y}1 = 1 \mid R = H, Y = \text{no}) =$
- $P(\hat{Y}1 = 1 \mid R = W, Y = \text{no}) =$

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

Equality of Opportunity/Odds for Predictor Ŷ2

- 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) =

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

Equality of Opportunity/Odds for Predictor $\hat{Y}2$

- $P(\hat{Y}1 = 1 \mid R = H, Y = yes) = 1/2$
- $P(\hat{Y}1 = 1 \mid R = W, Y = yes) = 1/2$
- $P(\hat{Y}1 = 1 \mid R = H, Y = no) = 1$
- $P(\hat{Y}1 = 1 \mid R = W, Y = no) = 0$

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

Equality of Opportunity/Odds for Predictor Ŷ2

- 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

Summary of Fairness Criteria

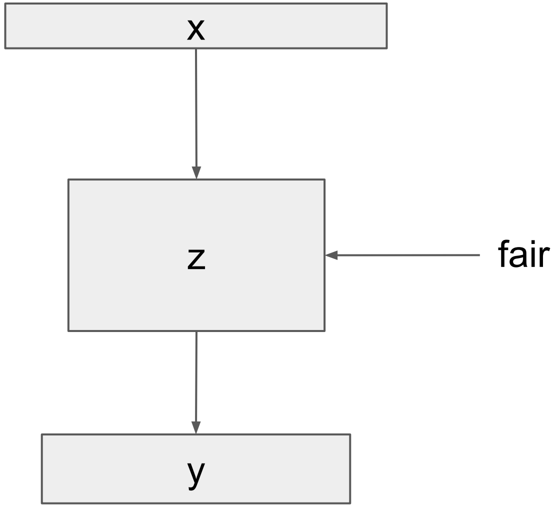
Fairness Criteria	Criteria	Group	Individual
Unawareness	Excludes A in Predictions		✓
Demographic Parity	$P(\hat{Y}=1\|A=0)=P(\hat{Y}=1\|A=1)$		
Equalized Odds	$P(\hat{Y}=1\|A=0,Y)=P(\hat{Y}=1\|A=1,Y)$	✓	
Equalized Opportunity	$P(\hat{Y}=1\|A=0,Y=1)=P(\hat{Y}=1\|A=1,Y=1)$	✓	

Outline

- Major Fairness Criteria
  - Demographic Parity
  - Equality of Odds/Opportunity
  - FICO Case Study
- Fair Representation Learning
  - Prejudice Removing Regularizer

Fair Representation Learning

- Make Representations Fair
  - Ensure fairness up to a certain level



Prejudice Remover Regularizer (Kamishima et al, 2012)

- 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

$$-\underbrace{\mathcal{L}(\mathcal{D};\boldsymbol{\Theta})}_{\text{Loss of the Model}} + \underbrace{\eta \mathrm{R}(\mathcal{D},\boldsymbol{\Theta})}_{\text{Fairness Regularizer}} + \underbrace{\frac{\lambda}{2}\|\boldsymbol{\Theta}\|_2^2}_{\text{L2 Regularizer}}$$

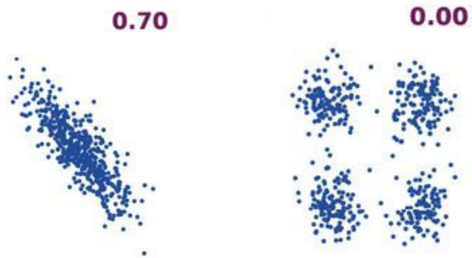
Prejudice Index (PI)

- Recall that Indirect Discrimination Happens When
  - Prediction is not directly conditioned on sensitive variables S
  - Prediction is indirectly conditioned on S by a variable O that is dependent on S
  - P(Ŷ | O), and O ~ P(O | S)

- Prejudice Index (PI)
  - Measures the degree of indirect discrimination based on mutual information

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

prediction model



some prejudice little prejudice

Kamishima et al, 2012

Normalized Prejudice Index (NPI)

- Prejudice Index (PI)
  - Measures the degree of indirect discrimination based on mutual information
  - Ranges in [0, +∞)

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

- Normalized Prejudice Index (NPI)
  - Normalize PI by the entropy of Y and S
  - Ranges in [0, 1]

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

Kamishima et al, 2012

Optimizing PI

- Learning PI

$$\mathrm{PI} = \sum_{Y,S} \hat{\Pr}[Y,S] \ln \frac{\hat{\Pr}[Y,S]}{\hat{\Pr}[S]\hat{\Pr}[Y]}$$

Kamishima et al, 2012

Optimizing PI

- Learning 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 et al, 2012]

Optimizing PI

- Learning PI

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

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

- Using Logistic Regression Model as the Prediction Model

difficult to evaluate

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

Kamishima et al, 2012

Optimizing PI

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

[Kamishima et al, 2012](#)

Optimizing PI

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

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

Integrals Are Difficult to Evaluate

Kamishima et al, 2012

Optimizing PI

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

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

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

Approximating integrals by sample means

Kamishima et al, 2012

Optimizing PI

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

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

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

Approximating integrals by sample means

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

Kamishima et al, 2012

Putting Things Together

- Optimization Target

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

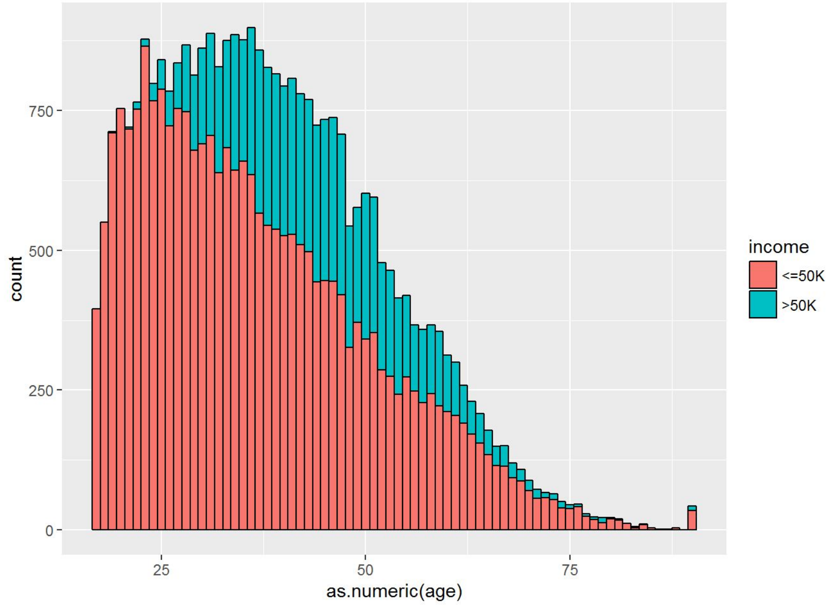
Loss of the Model Fairness Regularizer L2 Regularizer

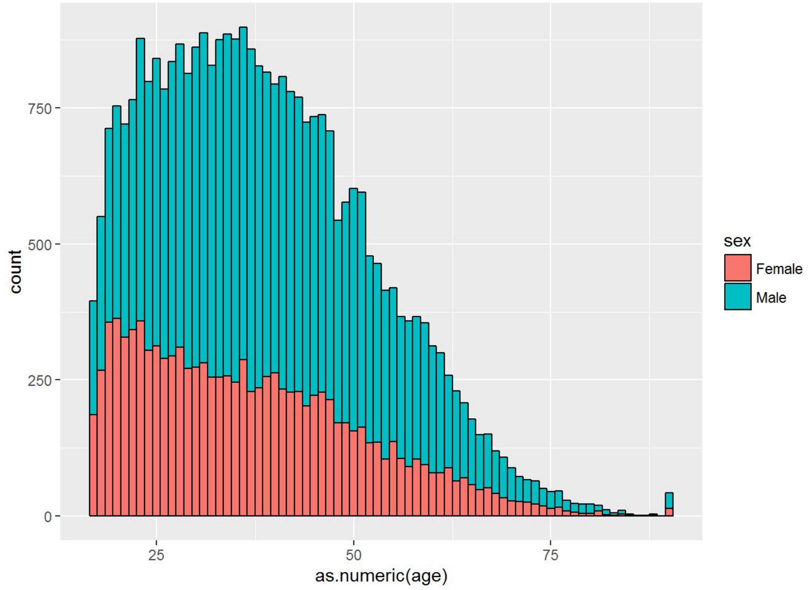
- Fairness Regularizer

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

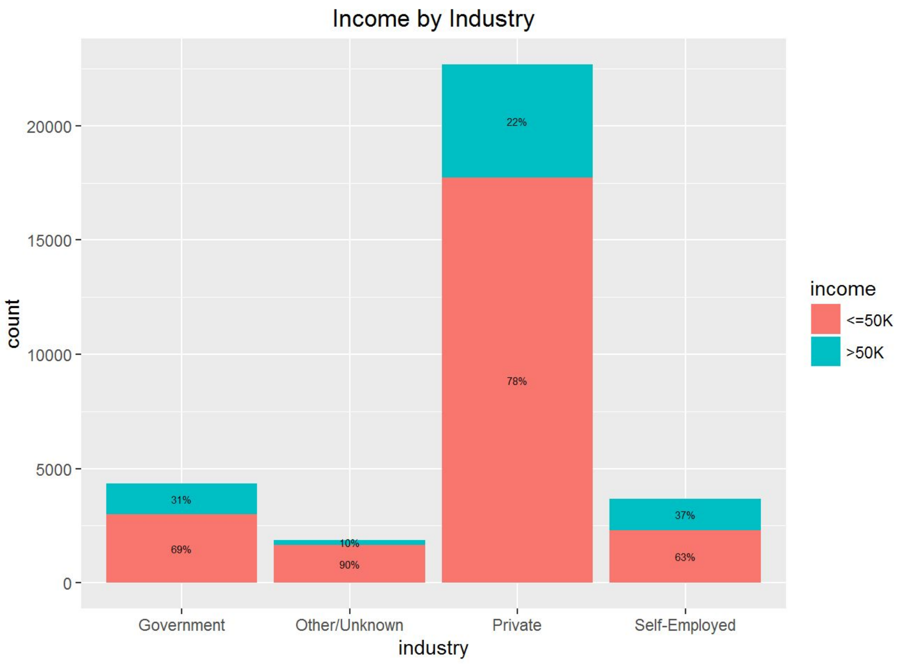
Kamishima et al, 2012

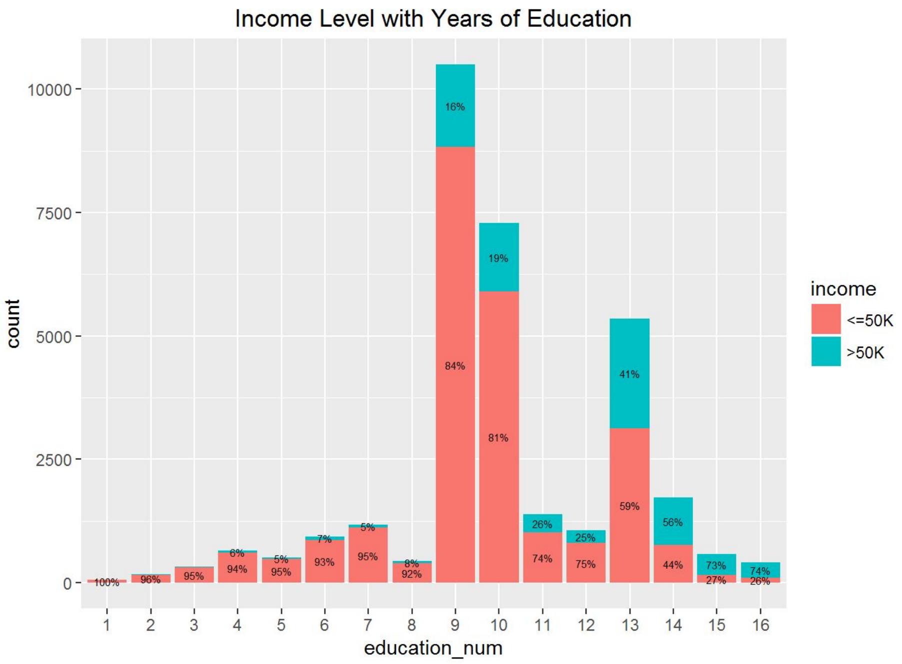
Adult Income Dataset (Kohavi 1996)





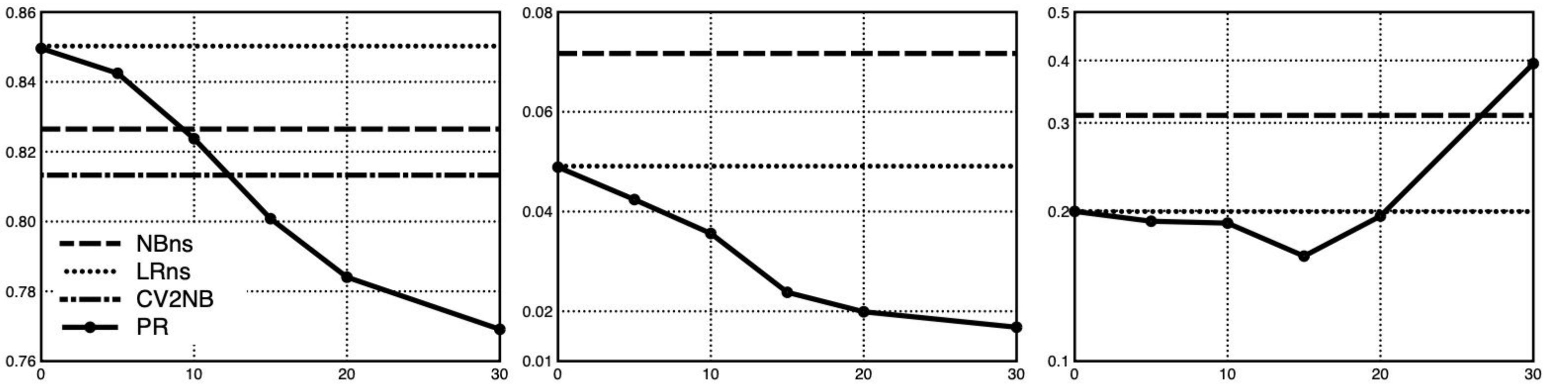
Adult Income Dataset (Kohavi 1996)





Results

- Changes of Performance With η
  - Model performance decreases (Acc)
  - Discrimination Decreases (NPI)
  - "Fairness Efficiency" (PI/MI) Increases

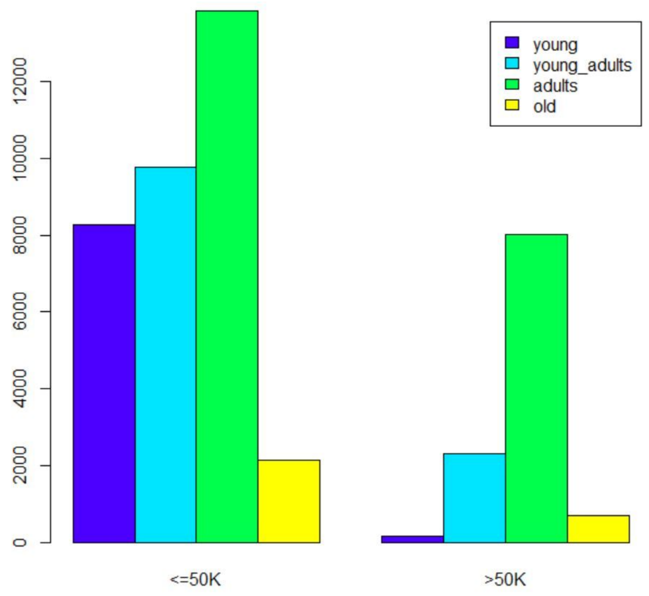


(a) Acc (b) NPI (c) PI/MI

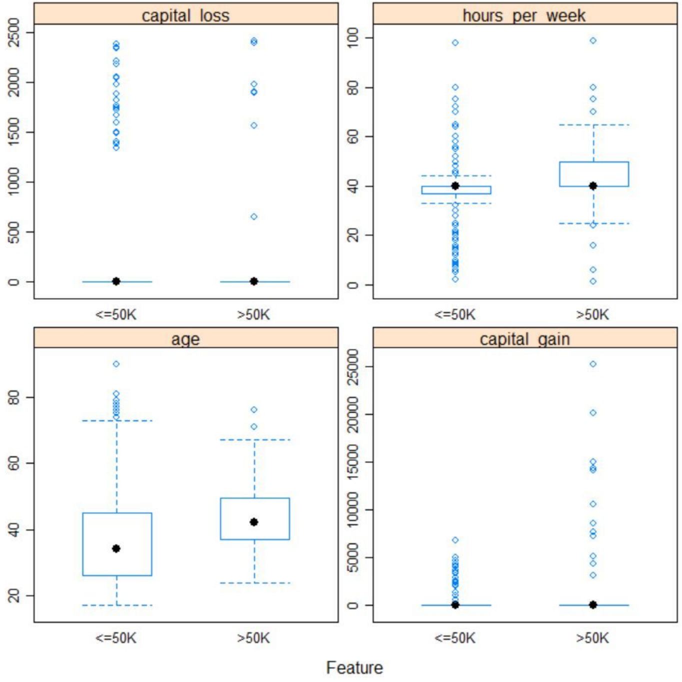
[Kamishima et al, 2012]

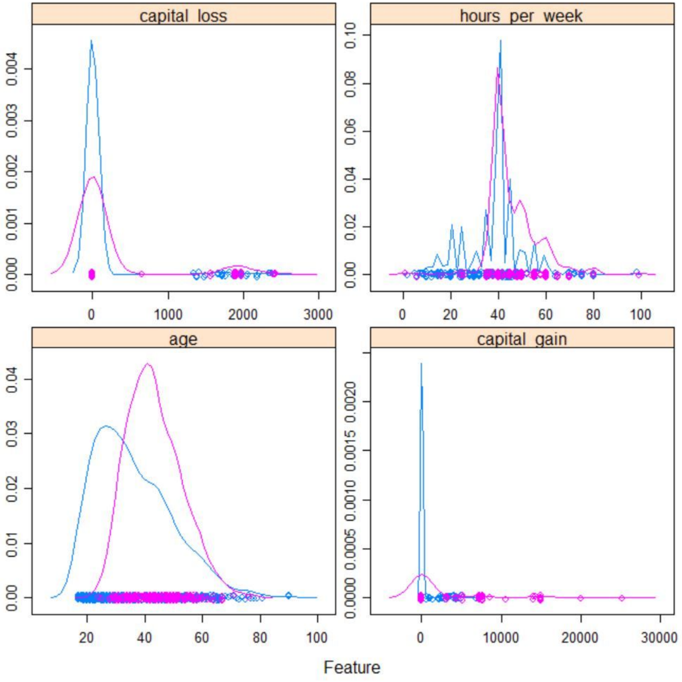
Adult Income Dataset (Kohavi 1996)

- Predict Whether Income Exceeds $50K/yr Based on Census Data



Adult Income Dataset (Kohavi 1996)





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

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

Logistic Regression full fet.

Logistic Regression no sensitive fet.

Logistic Regression + Prejudice Regularizer

η is the weight we put on prejudice regularizers

Kamishima et al, 2012

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)

	method	Acc	NMI	NPI	PI/MI
Logistic Regression full fet.	LR	0.851	0.267	5.21E-02	2.10E-01
Logistic Regression no sensitive fet.	LRns	0.850	0.266	4.91E-02	1.99E-01
Logistic Regression + Prejudice Regularizer	PR η=5	0.842	0.240	4.24E-02	1.91E-01
Logistic Regression + Prejudice Regularizer	PR η=15	0.801	0.158	2.38E-02	1.62E-01
Logistic Regression + Prejudice Regularizer	PR η=30	0.769	0.046	1.68E-02	3.94E-01

η - weight put on the prejudice regularizer

Kamishima et al, 2012

Reading Assignments (Pick One)

- A. Beutel, J. Chen, Z. Zhao, and E. H. Chi, Data decisions and theoretical implications when adversarially learning fair representations, FAT 2017
- Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores, ArXiv, 2016
- Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. Fairgan: Fairness-aware generative adversarial networks. IEEE International Conference on Big Data (Big Data), 2018
- Creager, E., Madras, D., Jacobsen, J. H., Weis, M. A., Swersky, K., Pitassi, T., & Zemel, R. Flexibly fair representation learning by disentanglement, ICML 2019
- Jiang, R., Pacchiano, A., Stepleton, T., Jiang, H., & Chiappa, S. Wasserstein Fair Classification. UAI, 2019

Next Lecture

Interpretability and Transparency

Questions?