

# Fairness in machine learning

Yuying Duan

# | Topics

- Bias in Automate decision making
- What is fairness in machine learning and why should we care
- Approaches to enforce fairness in machine learning

# Bias in Automate Decision Making

Machine learning systems are being implemented in decision making



## Case Study 1: Amazon Recruiting system

In 2010, Amazon built an **AI recruiting tool** that can automate the process of reviewing resumes and recommending top candidates.



## Case Study 1: Amazon Recruiting system

**In 2010**, Amazon built an **AI recruiting tool** that can automate the process of reviewing resumes and recommending top candidates.

- The system was trained on past resumes submitted to Amazon over a 10-year period.
- Predict whether a candidate fits this job.



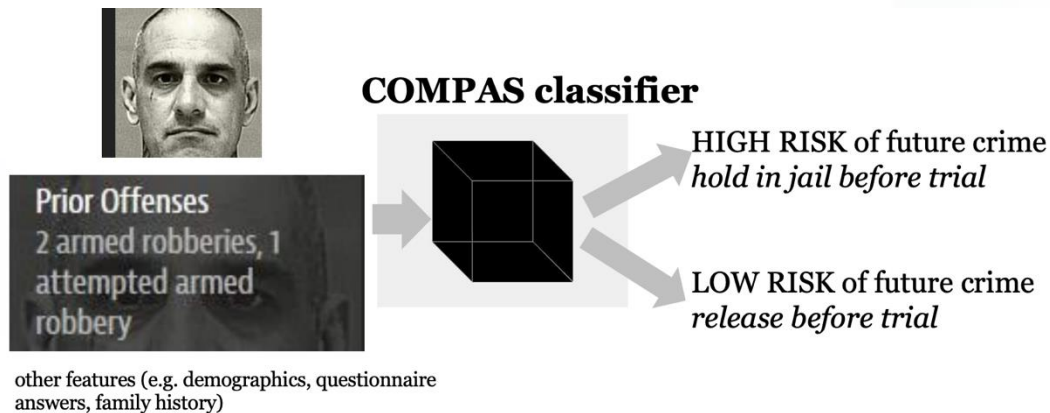
## Case Study 2: COMPAS

In the **early 2000s**, a private company called Northpointe, Inc developed Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) that is used for making decision of bail.

## Case Study 2: COMPAS

In the **early 2000s**, a private company called Northpointe, Inc developed Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) that is used for making decision of bail.

- Used in prisons across country: AZ, CO, DL, KY, LA, OK, VA, WA, WI
- The system uses answers to a 137-item questionnaire, plus data from criminal records
- Predict the likelihood that a defendant will reoffend (i.e., recidivism risk) or fail to appear in court





# Automate Decision making is a machine learning problem

Amazon trains a model to decide if the candidate can get the job based on the candidate's profile. The decision  $\hat{y}$  is a binary attribute:  $\hat{y} = 1$  indicates the candidate can get the job, if  $\hat{y} = 0$ , the candidate can not get the job.

# Automate Decision making is a machine learning problem

Amazon trains a model to decide if the candidate can get the job based on the candidate's profile. The decision  $\hat{y}$  is a binary attribute:  $\hat{y} = 1$  indicates the candidate can get the job, if  $\hat{y} = 0$ , the candidate can not get the job.

- A **generator** that generates individual's profile  $\mathbf{x} \in \mathcal{X}$ : in an i.i.d manner from  $F_{\mathbf{x}}(x)$ . An **observer** (recruit manager) that draws the target  $\mathbf{y} \in \{0, 1\}$  that indicates if one successfully get this job in an i.i.d manner from  $P_{\mathbf{y}|x}(y | x)$
- A **model set**:  $\mathcal{H} : \mathcal{X} \rightarrow \{0, 1\}$
- A **loss function**:  $L[h] = \mathbb{E}_{\mathbf{x}, \mathbf{y}}[1\{\mathbf{y} \neq h(\mathbf{x})\}]$   $1\{\cdot\}$  is the indicator function.

# Automate Decision making is a machine learning problem

Amazon trains a model to decide if the candidate can get the job based on the candidate's profile. The decision  $\hat{y}$  is a binary attribute:  $\hat{y} = 1$  indicates the candidate can get the job, if  $\hat{y} = 0$ , the candidate can not get the job.

- A **generator** that generates individual's profile  $\mathbf{x} \in \mathcal{X}$ : in an i.i.d manner from  $F_{\mathbf{x}}(x)$ . An **observer** (recruit manager) that draws the target  $\mathbf{y} \in \{0, 1\}$  that indicates if one successfully get this job in an i.i.d manner from  $P_{\mathbf{y}|x}(y | x)$
- A **model set**:  $\mathcal{H} : \mathcal{X} \rightarrow \{0, 1\}$
- A **loss function**:  $L[h] = \mathbb{E}_{\mathbf{x}, \mathbf{y}}[1\{\mathbf{y} \neq h(\mathbf{x})\}]$   $1\{\cdot\}$  is the indicator function.

For a **machine learning** problem, we will pick a model (recruit policy) that minimizes the loss function:

$$\min_{h \in \mathcal{H}} : L[h] = \mathbb{E}_{\mathbf{x}, \mathbf{y}}[1\{\mathbf{y} \neq h(\mathbf{x})\}]$$

# Concerns in Automate decision making



Hiring could become faster and less expensive by using ML models.

# Concerns in Automate decision making



Hiring could become faster and less expensive by using ML models.

But there are problems

# Bias in ML models

In 2018 [Reuters](#) reported: Amazon developed an AI-powered recruiting tool that exhibited bias against women.

Support The Guardian

Search jobs Dating Sign in Search UK edition

Contribute → Subscribe →

The Guardian

News Opinion Sport Culture Lifestyle More

UK World Business Football UK politics Environment Education Society Science Tech Global development Cities Obituaries

Amazon


## Amazon ditched AI recruiting tool that favored men for technical jobs

Specialists had been building computer programs since 2014 to review résumés in an effort to automate the search process

Reuters

Thu 11 Oct 2018 00:42 BST

309 This article is over 1 month old



Amazon's automated hiring tool was found to be inadequate after penalizing the résumés of female candidates. Photograph: Brian Snyder/Reuters

Amazon's machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

Advertisement

Santander Corporate & Commercial

Download the report

# Bias in ML models

In 2018 [Reuters](#) reported: Amazon developed an AI-powered recruiting tool that exhibited bias against women.

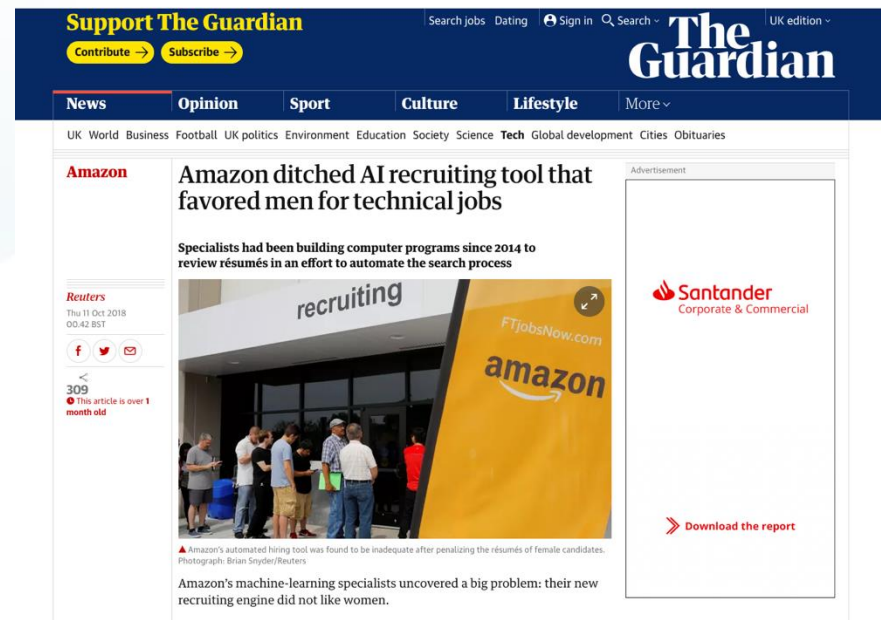


Table 1: The employment rates of male and female applicants in the field of information technology at Amazon.

| Gender | Number of Applicants | Employment Rate |
|--------|----------------------|-----------------|
| Male   | 100                  | 40%             |
| Female | 50                   | 20%             |

# Bias in ML models

In 2018 [Reuters](#) reported: Amazon developed an AI-powered recruiting tool that exhibited bias against women.

- The tool, trained on resumes submitted over a decade, predominantly from male applicants, learned to favor male candidates for technical roles.
- It penalized resumes containing the word "women's".

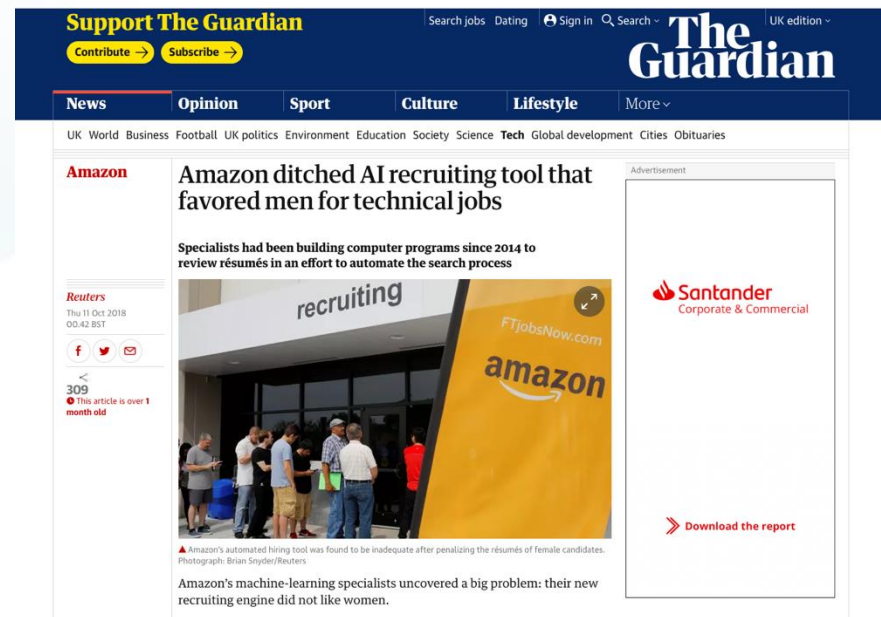


Table 1: The employment rates of male and female applicants in the field of information technology at Amazon.

| Gender | Number of Applicants | Employment Rate |
|--------|----------------------|-----------------|
| Male   | 100                  | 40%             |
| Female | 50                   | 20%             |



# Bias in ML models

In 2016, [ProPublica](#) reported that COMPAS was:

- **Biased against Black defendants:** Black defendants were **more likely** to be incorrectly predicted as high risk for reoffending.
- **Favoring white defendants:** White defendants were **more likely** to be incorrectly labeled as low risk.

# Bias in ML models

In 2016, [ProPublica](#) reported that COMPAS was:

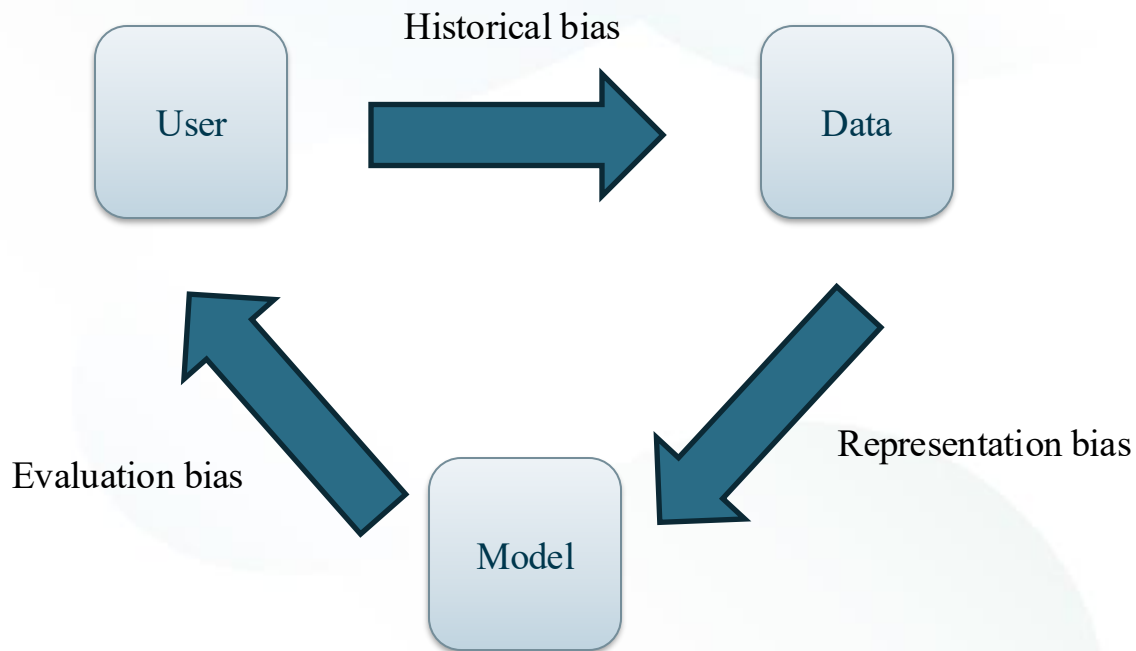
- **Biased against Black defendants:** Black defendants were **more likely** to be incorrectly predicted as high risk for reoffending.
- **Favoring white defendants:** White defendants were **more likely** to be incorrectly labeled as low risk.



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

# Where Does the Bias Come From?

The process of the automate decision making:



# Where Does the Bias Come From?

Historical bias:

- Historical bias is the already existing bias and socio-technical issues in the world and can seep into from the data generation process.

# Where Does the Bias Come From?

## Historical bias:

- An example of this type of bias can be found in a 2018 image search result, where searching for women CEOs ultimately resulted in fewer female CEO images due to the fact that only 5% of Fortune 500 CEOs were woman, which would cause the search results to be biased towards male CEOs.



Philip N Cohen ✓  
@familyunequal

I get 5 women in my first 154 images for "CEO" in Google images. (That includes CEO Barbie.)



# Where Does the Bias Come From?

Historical bias:

- The search algorithm reflected the reality caused by historical inequalities in access to leadership opportunities and resources between men and women.

# Where Does the Bias Come From?

## Representation bias:

The model trained on the given dataset demonstrates better generalization for the majority class but underperforms on the minority class.

## Evaluation bias:

The data distribution changes during inference.

- What is fairness in machine learning and why should we care



**| Why we care if a ML model have bias or not?**

# Why we care if a ML model have bias or not?

Preventing bias against specific groups is a legal requirement:

The **Equal Credit Opportunity Act (ECOA)** is codified in the **United States Code** at:

[15 U.S. Code § 1691 - Purpose of the Act](#)

“ It shall be unlawful for any creditor to discriminate against any applicant, with respect to any aspect of a credit transaction-

(1) on the basis of race, color, religion, national origin, sex or marital status, or age (provided the applicant has the capacity to contract);

(2) because all or part of the applicant's income derives from any public assistance program; or

(3) because the applicant has in good faith exercised any right under this chapter.”

# Why we care if a ML model have bias or not?

Preventing bias against specific groups is a legal requirement:

- ▶ Credit (Equal Credit Opportunity Act)
- ▶ Education (Civil Rights Act of 1964; Education Amendments of 1972)
- ▶ Employment (Civil Rights Act of 1964)
- ▶ Housing (Fair Housing Act)

## ■ Why we care if a ML model have bias or not?

Preventing bias against **specific groups** is a legal requirement:

# Why we care if a ML model have bias or not?

Preventing bias against **specific groups** is a legal requirement:

- ▶ Race
- ▶ Sex
- ▶ Religion
- ▶ National origin
- ▶ Citizenship
- ▶ Pregnancy
- ▶ Disability status
- ▶ Genetic information
- ▶ Others depend on the application

# What that mean fairness in ML?

# What that mean fairness in ML?

- really hard question
- start with the common concrete statistical definition

# What that mean fairness in ML?

## Def. 1 Statistical Parity

- Decide whether someone should be hired ( $\hat{Y} = 1$ ) or not ( $\hat{Y} = 0$ )
- Sensitive attribute: male  $S = 0$ , female:  $S = 1$

Table 1: The employment rates of male and female applicants in the field of information technology at Amazon.

| Gender | Number of Applicants | Employment Rate |
|--------|----------------------|-----------------|
| Male   | 100                  | 40%             |
| Female | 50                   | 20%             |



Historical bias



# What that mean fairness in ML?

## Def. 1 Statistical Parity

- Decide whether someone should be hired ( $\hat{Y} = 1$ ) or not ( $\hat{Y} = 0$ )
- Sensitive attribute: male  $S = 0$ , female:  $S = 1$

Statistical parity:

$$Pr(\hat{Y} = 1 \mid S = 0) = Pr(\hat{Y} = 1 \mid S = 1)$$

Male and female have the same probability getting hired.

Table 1: The employment rates of male and female applicants in the field of information technology at Amazon.

| Gender | Number of Applicants | Employment Rate |
|--------|----------------------|-----------------|
| Male   | 100                  | 40%             |
| Female | 50                   | 20%             |



Historical bias

# What that mean fairness in ML?

## Def. 2 Equal Opportunity

- Ground-truth label that indicates one's qualification  $Y = 0$  (not qualified)  $Y = 1$  (qualified)
- Decide whether someone should get resource ( $\hat{Y} = 1$ ) or not ( $\hat{Y} = 0$ )
- Sensitive attribute: African-American  $S = 0$ , white:  $S = 1$

# What that mean fairness in ML?

## Def. 2 Equal Opportunity

- Ground-truth label that indicates one's qualification  $Y = 0$  (not qualified)  $Y = 1$  (qualified)
- Decide whether someone should get resource ( $\hat{Y} = 1$ ) or not ( $\hat{Y} = 0$ )
- Sensitive attribute: African-American  $S = 0$ , white:  $S = 1$

Equal Opportunity:

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

African-American and white who are qualified/ who are deserved have the same probability getting hired.

# What that mean fairness in ML?

## Def. 3 Equalized Odds

- Ground-truth label that indicate whether the defendant actually reoffends ( $Y = 1$ ) or not ( $Y = 0$ )
- Decide whether in the jail e some before tail ( $\hat{Y} = 1$ ) or not ( $\hat{Y} = 0$ )
- Sensitive attribute: African- American  $S = 0$ , white:  $S = 1$

# What that mean fairness in ML?

## Def. 3 Equalized Odds

- Ground-truth label that indicate whether the defendant actually reoffends ( $Y = 1$ ) or not ( $Y = 0$ )
- Decide whether in the jail e some before tail ( $\hat{Y} = 1$ ) or not ( $\hat{Y} = 0$ )
- Sensitive attribute: African- American  $S = 0$ , white:  $S = 1$

## Equalized Odds:

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

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

African-American and white individuals **who are reoffenders** have the same probability of **not being released**.  
African-American and white individuals **who are not reoffenders** have the same probability of **being released**.

# Debate on using which fairness notion?

**Statistical Parity** corrects historical bias, but may significantly reduce model accuracy and lead to overcompensation.

**Equal Opportunity / Equalized Odds** Corrects model-induced bias, results in less accuracy loss, has the problem of "bias in, bias out."



- Approaches to enforce fairness in machine learning

# Three different approaches to enforce Fairness in ML

- Pre-processing: adjusts the features space to be uncorrelated with the sensitive attribute.
- In-processing: works with non-discrimination criterion as a regularization term in the model training process.
- Post-processing: adjusts learned classifiers so the resulting classifier is uncorrelated with the sensitive attribute.



# Pre-processing

Data:  $x \in \mathcal{X}, y \in \mathcal{Y}, s \in \mathcal{S}$

Model: one generator and two discriminators

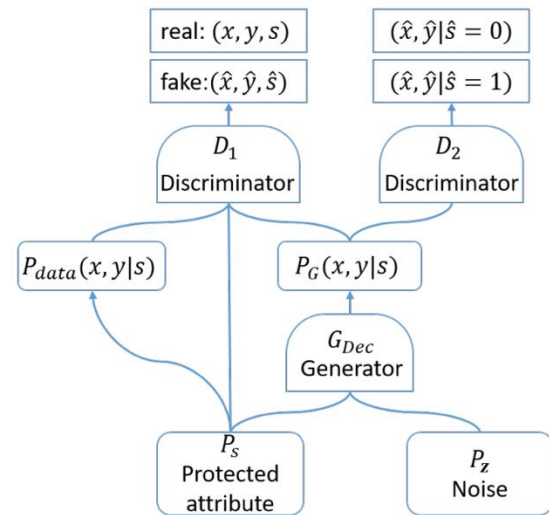
Loss function:

$$\min_{G_{Dec}} \max_{D_1, D_2} V(G_{Dec}, D_1, D_2) = V_1(G_{Dec}, D_1) + \lambda V_2(G_{Dec}, D_2),$$

where

$$\begin{aligned} V_1(G_{Dec}, D_1) &= \mathbb{E}_{s \sim P_{data}(s), (\mathbf{x}, y) \sim P_{data}(\mathbf{x}, y | s)} [\log D_1(\mathbf{x}, y, s)] \\ &+ \mathbb{E}_{\hat{s} \sim P_G(s), (\hat{\mathbf{x}}, \hat{y}) \sim P_G(\mathbf{x}, y | s)} [\log(1 - D_1(\hat{\mathbf{x}}, \hat{y}, \hat{s}))], \end{aligned}$$

$$\begin{aligned} V_2(G_{Dec}, D_2) &= \mathbb{E}_{(\hat{\mathbf{x}}, \hat{y}) \sim P_G(\mathbf{x}, y | s=1)} [\log D_2(\hat{\mathbf{x}}, \hat{y})] \\ &+ \mathbb{E}_{(\hat{\mathbf{x}}, \hat{y}) \sim P_G(\mathbf{x}, y | s=0)} [\log(1 - D_2(\hat{\mathbf{x}}, \hat{y}))], \end{aligned}$$



# | Pre-processing

- Pre-processing transforms the feature space so it is independent of the sensitive attribute.
- This approach is agnostic to what we do with these features later on and so it can ensure independence under any training process on the new space.
- Pre-processing typically uses adversarial training, so it has high computational cost.

# In-processing

Data:  $x \in \mathcal{X}, y \in \mathcal{Y}, s \in \mathcal{S}$

Model: a classifier  $h : \mathcal{X} \rightarrow [0, 1]$

Loss function:  $\mathcal{L}_\lambda(h(x), y) = -[y \log(h(x)) + (1 - y) \log(1 - h(x))] + \lambda \cdot \text{Cov}(h(x), s)$

where:  $\text{Cov}(s, h(x)) = \mathbb{E}[(s - \bar{s})h(x)] - \mathbb{E}[(s - \bar{s})]h(\bar{x})$



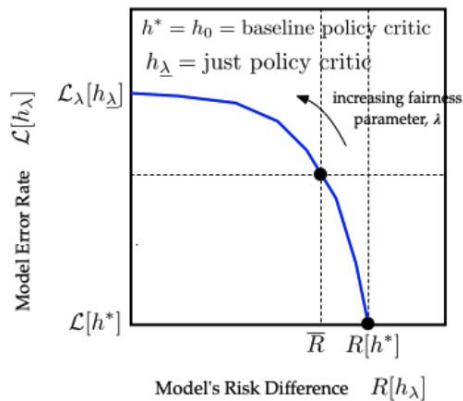
Fairness regularization

# In-processing

Data:  $x \in \mathcal{X}, y \in \mathcal{Y}, s \in \mathcal{S}$

Model: a classifier  $h : \mathcal{X} \rightarrow [0, 1]$

Loss function:  $\mathcal{L}_\lambda(h(x), y) = -[y \log(h(x)) + (1 - y) \log(1 - h(x))] + \lambda \cdot \text{Cov}(h(x), s)$



Fairness-accuracy tradeoff

# | In-processing

- In-processing introduces the non-discrimination criterion as a regularization constraint during model training.
- The issue of in-processing is the the regularization constraint may greatly slow down the convergence of the training algorithm.

# Post-processing

Data:  $x \in \mathcal{X}, y \in \mathcal{Y}, s \in \mathcal{S}$

Model: a classifier  $h : \mathcal{X} \rightarrow [0, 1]$

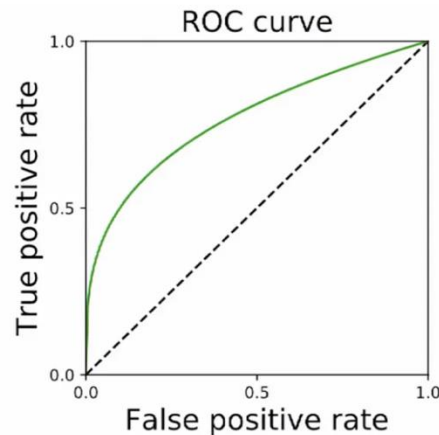
Loss function:  $\mathcal{L}(h(x), y) = -[y \log(h(x)) + (1 - y) \log(1 - h(x))]$

# Post-processing

- $h(x)$  indicates the probability that a sample  $x$  is classified as class 1.
- The optimal predictor takes value of:

$$\hat{Y} = \begin{cases} 1 & \text{if } h(x) > 1/2 \\ 0 & \text{otherwise} \end{cases}$$

If I plot the TPR against FPR for all possible thresholds :

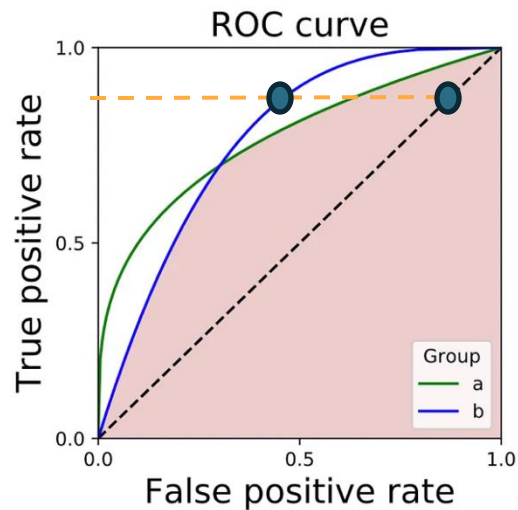


# Post-processing

- $h(x)$  indicates the probability that a sample  $x$  is classified as class 1.
- The optimal predictor takes value of:

$$\hat{Y} = \begin{cases} 1 & \text{if } h(x) > 1/2 \\ 0 & \text{otherwise} \end{cases}$$

If I plot the TPR against FPR for all possible thresholds group a and group b :





# Post-processing

- Post-processing refers to the process of taking a trained classifier and adjusting it using a randomization procedure to enforce fairness.
- Post-processing does not impact the model original training pipeline.
- Post processing's advantage is that it works with trained classifiers and therefore does not need access to the raw data.

# Message taken from this lecture

- Fairness is a legal requirement in automate decision making.
- Fairness is hard to define and evaluate, this lecture introduces three notation of fairness:  
Statistical parity, Equal Opportunity and Equalized Odds
- Three different approach to enforce fairness:  
Pre-processing, in-processing, post-processing