Trustworthy Machine Learning

Fairness in Learning 1

Sangdon Park

POSTECH

Contents from

FAIRNESS AND MACHINE LEARNING

Limitations and Opportunities

Solon Barocas, Moritz Hardt, Arvind Narayanan

Equality of Opportunity in Supervised Learning

Moritz Hardt Eric Price Nathan Srebro

October 11, 2016

Abstract

We propose a criterion for discrimination against a specified sensitive attribute in supervised learning, where the goal is to predict some target based on available features. Assuming data about the predictor, target, and membership in the protected group are available, we show how to optimally *adjust* any learned predictor so as to remove discrimination according to our definition. Our framework also improves incentives by shifting the cost of poor classification from disadvantaged groups to the decision maker, who can respond by improving the classification accuracy.

In line with other studies, our notion is *oblivious*: it depends only on the joint statistics of the predictor, the target and the protected attribute, but not on interpretation of individual features. We study the inherent limits of defining and identifying biases based on such oblivious measures, outlining what can and cannot be inferred from different oblivious tests.

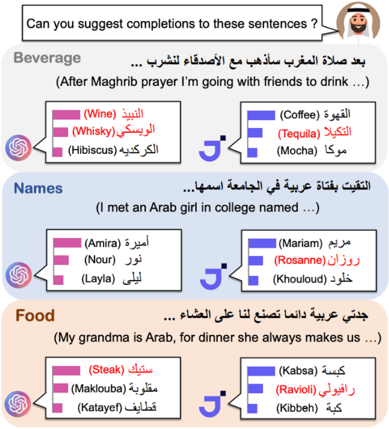
We illustrate our notion using a case study of FICO credit scores.

- and contents partially from slides by Roger Grosse at University of Toronto.

Why Fairness in Learning?

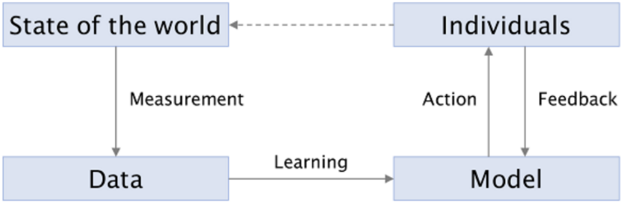
- Translation from English to Turkish, then back to English injects gender bias.

Why Fairness in Learning?



- Training sets introduce cultural bais [Naous et al., 2023]

Why Fairness in Learning?



- The machine learning loop
- Biased models enforce the bias of the world.

Fairness in Learning: Overview

Goal

Identify and mitigate "bias" in ML-based decision making.

Source of bias:

- Data
  - imbalanced data (*e.g.*, rare data, gender-biased data)
  - incorrect data (*e.g.*, noisy data, data with historical bias)
- Model
  - modeling error
  - bias in loss

Credit: Richard Zemel

Fairness in Learning: Definitions

- Known definitions
  - Demographic parity
  - Equalized odds
  - Equal opportunity
  - Equal (weak) calibration
  - Equal (strong) calibration
  - Fair subgroup accuracy
  - ...
- Definitions are controversial and should be used depending on applications.

Setup

- Supervised learning for binary classification
- f: a classifier
- $Y \in \{0, 1\}$: an outcome
- X: features
- $A \in \{0, 1\}$: a protected attribute (*e.g.*, "woman" or not)
- $\widehat{Y} := f(X, A) \in \{0, 1\}$: a prediction

Demographic Parity

Definition (demographic parity)

A predictor $\widehat{Y}$ satisfies demographic parity with respect to the protected attribute A if

$$\mathbb{P}\left\{\widehat{Y} = 1 \;\middle|\; A = 0\right\} = \mathbb{P}\left\{\widehat{Y} = 1 \;\middle|\; A = 1\right\}$$

- Its variants appears in many papers.

Demographic Parity

Definition (demographic parity)

A predictor $\widehat{Y}$ satisfies demographic parity with respect to the protected attribute A if

$$\mathbb{P}\left\{\widehat{Y} = 1 \,\middle|\, A = 0\right\} = \mathbb{P}\left\{\widehat{Y} = 1 \,\middle|\, A = 1\right\}$$

- Its variants appears in many papers.
- Is this definition okay?

Demographic Parity

Definition (demographic parity)

A predictor $\widehat{Y}$ satisfies demographic parity with respect to the protected attribute A if

$$\mathbb{P}\left\{\widehat{Y}=1 \,\middle|\, A=0\right\}=\mathbb{P}\left\{\widehat{Y}=1 \,\middle|\, A=1\right\}$$

- Its variants appears in many papers.
- Is this definition okay?
  - ✓ Intuitive
  - ✗ Actually not quite fair (in some common sense)
    - ⋆ A classifier accepts qualified applicants in $A=0$ but unqualified applicants in $A=1$.
    - ⋆ *e.g.*, when we don't have enough training samples for $A=1$, this constraint forces to have $\widehat{Y}=1$ for $A=1$.
  - ✗ This definition does not allow the perfect predictor $\widehat{Y}=Y$.

Better Fairness Definitions

Definition (equalized odd)

We say that a predictor $\widehat{Y}$ satisfies equalized odds with respect to the protected attribute A and outcome Y if $\widehat{Y}$ and A are conditionally independent given Y, *e.g.*,

$$\mathbb{P}\left\{\widehat{Y}=1 \,\middle|\, A=0, Y=y\right\}=\mathbb{P}\left\{\widehat{Y}=1 \,\middle|\, A=1, Y=y\right\} \quad \forall y \in\{0,1\}.$$

- The definition is applicable to other setups, *e.g.*, multi-class classification.

Better Fairness Definitions

Definition (equalized odd)

We say that a predictor $\widehat{Y}$ satisfies equalized odds with respect to the protected attribute A and outcome Y if $\widehat{Y}$ and A are conditionally independent given Y, *e.g.*,

$$\mathbb{P}\left\{\widehat{Y}=1 \,\middle|\, A=0, Y=y\right\} = \mathbb{P}\left\{\widehat{Y}=1 \,\middle|\, A=1, Y=y\right\} \quad \forall y \in \{0,1\}.$$

- The definition is applicable to other setups, *e.g.*, multi-class classification.
- If $y=1$, this constrains equalizes *true positive rates* (TPR) for both $A=0$ and $A=1$.

Better Fairness Definitions

Definition (equalized odd)

We say that a predictor $\widehat{Y}$ satisfies equalized odds with respect to the protected attribute A and outcome Y if $\widehat{Y}$ and A are conditionally independent given Y, *e.g.*,

$$\mathbb{P}\left\{\widehat{Y}=1 \,\middle|\, A=0, Y=y\right\}=\mathbb{P}\left\{\widehat{Y}=1 \,\middle|\, A=1, Y=y\right\} \quad \forall y \in\{0,1\}.$$

- The definition is applicable to other setups, *e.g.*, multi-class classification.
- If $y=1$, this constrains equalizes *true positive rates* (TPR) for both $A=0$ and $A=1$.
- If $y=0$, this constraint equalizes *false positive rates* (FPR) for both $A=0$ and $A=1$.

Better Fairness Definitions

Definition (equalized odd)

We say that a predictor $\widehat{Y}$ satisfies equalized odds with respect to the protected attribute A and outcome Y if $\widehat{Y}$ and A are conditionally independent given Y, *e.g.*,

$$\mathbb{P}\left\{\widehat{Y}=1 \,\middle|\, A=0, Y=y\right\}=\mathbb{P}\left\{\widehat{Y}=1 \,\middle|\, A=1, Y=y\right\} \quad \forall y \in\{0,1\}.$$

- The definition is applicable to other setups, *e.g.*, multi-class classification.
- If $y=1$, this constrains equalizes *true positive rates* (TPR) for both $A=0$ and $A=1$.
- If $y=0$, this constraint equalizes *false positive rates* (FPR) for both $A=0$ and $A=1$.
- Is this enough?

Better Fairness Definitions

Definition (equalized odd)

We say that a predictor $\widehat{Y}$ satisfies equalized odds with respect to the protected attribute A and outcome Y if $\widehat{Y}$ and A are conditionally independent given Y, *e.g.*,

$$\mathbb{P}\left\{\widehat{Y}=1 \,\middle|\, A=0, Y=y\right\} = \mathbb{P}\left\{\widehat{Y}=1 \,\middle|\, A=1, Y=y\right\} \quad \forall y \in \{0,1\}.$$

- The definition is applicable to other setups, *e.g.*, multi-class classification.
- If $y=1$, this constrains equalizes *true positive rates* (TPR) for both $A=0$ and $A=1$.
- If $y=0$, this constraint equalizes *false positive rates* (FPR) for both $A=0$ and $A=1$.
- Is this enough?
  - ✓ Intuitive – controlling TPR and FPR is common.
  - ✗ The accuracy is equally high for all demographics $\rightarrow$ a model good at the majority will be penalized.

Better Fairness Definitions

Definition (equal opportunity)

We say that a binary predictor $\widehat{Y}$ satisfies *equal opportunity* with respect to A and Y if

$$\mathbb{P}\left\{\widehat{Y}=1 \,\middle|\, A=0, Y=1\right\} = \mathbb{P}\left\{\widehat{Y}=1 \,\middle|\, A=1, Y=1\right\}.$$

- Suppose $Y=1$ is the "advantaged" outcome.
- At least provides equal oppertunities for the advantaged option!
- Equal opportunity is weaker than equalized odd but typically allows stronger utility.

Better Fairness Definitions

Definition (equal opportunity)

We say that a binary predictor $\widehat{Y}$ satisfies *equal opportunity* with respect to A and Y if

$$\mathbb{P}\left\{\widehat{Y} = 1 \;\middle|\; A = 0, Y = 1\right\} = \mathbb{P}\left\{\widehat{Y} = 1 \;\middle|\; A = 1, Y = 1\right\}.$$

- Suppose $Y = 1$ is the "advantaged" outcome.
- At least provides equal oppertunities for the advantaged option!
- Equal opportunity is weaker than equalized odd but typically allows stronger utility.
  - Why weaker?

How to Build a Fair Classifier?

A Score-based Predictor

A score-based predictor

$$\widehat{Y} = \mathbb{1}\left(\widehat{R} > t\right)$$

- We consider a real valued score $\widehat{R} \in [0, 1]$, from which a classifier decides a label.
- *e.g.*, a neural network with a single output neuron: $R = f_{\mathsf{NN}}(X)$
- Here, we suppose a pre-trained model is given and fixed; only change the threshold.

How to Build a Fair Classifier?

A Score-based Predictor

A score-based predictor

$$\widehat{Y} = \mathbb{1}\left(\widehat{R} > t\right)$$

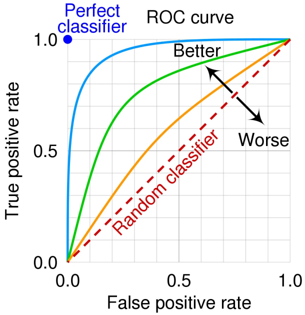
- We consider a real valued score $\widehat{R} \in [0, 1]$, from which a classifier decides a label.
- *e.g.*, a neural network with a single output neuron: $R = f_{\mathsf{NN}}(X)$
- Here, we suppose a pre-trained model is given and fixed; only change the threshold.
- The equalized odds and equal opportunity definitions are characterized by true positive and false positive rates, which is controlled by the threshold, *i.e.*,

$$\text{(FPR)} = \mathbb{P}\left\{\widehat{R} > t \;\middle|\; A = a, Y = 0\right\}$$
$$\text{(TPR)} = \mathbb{P}\left\{\widehat{R} > t \;\middle|\; A = a, Y = 1\right\}.$$

Receiver Operator Characteristic (ROC) Curves

A-conditional ROC Curves

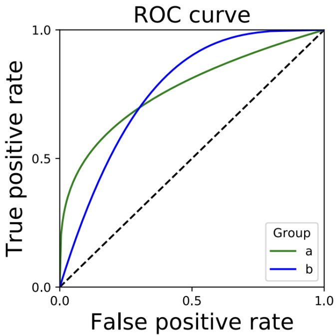
$$C_a(t) := \Big(\underbrace{\mathbb{P}\left\{\widehat{R} > t \;\middle|\; A = a, Y = 0\right\}}_{\text{false positive rate (FPR)}}, \underbrace{\mathbb{P}\left\{\widehat{R} > t \;\middle|\; A = a, Y = 1\right\}}_{\text{true positive rate (TPR)}}\Big)$$



Picture Credit: Ilyurek Kilic

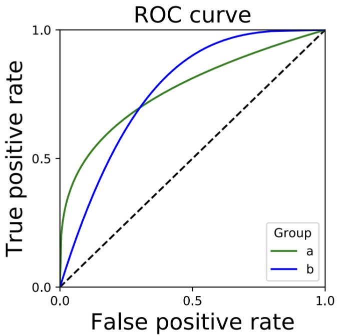
- $t \uparrow \rightarrow$ FPR $\downarrow$ and TPR $\downarrow$.

Algorithm for Equalized Odds



- As two ROC curves are intersected, let an intersecting point be $(\mathrm{FPR}^*, \mathrm{TPR}^*)$
- Find (t_0, t_1) such that $C_0(t_0) = (\mathrm{FPR}^*, \mathrm{TPR}^*)$ and $C_1(t_1) = (\mathrm{FPR}^*, \mathrm{TPR}^*)$.
- Our classifier is $\widehat{Y} := \mathbb{1}\left(\widehat{R} > t_a\right)$, *i.e.*, an attribute specific t_a.

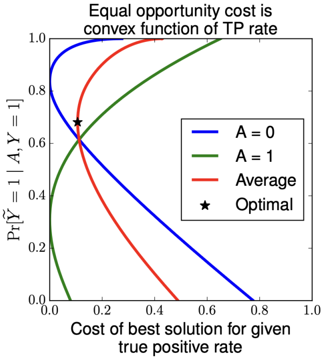
Algorithm for Equalized Odds



- As two ROC curves are intersected, let an intersecting point be $(\mathsf{FPR}^*, \mathsf{TPR}^*)$
- Find (t_0, t_1) such that $C_0(t_0) = (\mathsf{FPR}^*, \mathsf{TPR}^*)$ and $C_1(t_1) = (\mathsf{FPR}^*, \mathsf{TPR}^*)$.
- Our classifier is $\widehat{Y} := \mathbb{1}\left(\widehat{R} > t_a\right)$, *i.e.*, an attribute specific t_a.

✗ The accuracy is determined; when the accuracy is poor, no room to tune.

Algorithm for Equal Opportunity



- Recall that our classifier is $\widehat{Y} := \mathbb{1}\left(\widehat{R} > t_a\right)$, where t_a is a threshold for $A = a$.
- The algorithm solves the following constraint minimization with some loss ℓ.

$$\min_{t_0, t_1} \quad \mathbb{E}\, \ell(\widehat{Y}, Y) \quad \text{s.t.} \quad \mathsf{TPR}_0(\widehat{Y}) = \mathsf{TPR}_1(\widehat{Y})$$

Conclusion

- Fairness definitions – *no winner*
  1. Demographic parity
  2. Equalized Odds
  3. Equal Opportunity
- Fairness algorithms
  1. Algorithm for Equalized Odds
  2. Algorithm for Equal Opertunity
- There are neither "(ε, δ)-fairness" nor the proof of fairness; why?
  - Proving the fairness may be impossible without clearly understanding on domain-specific knowledge.
  - Fairness through Awareness!

Reference I

T. Naous, M. J. Ryan, A. Ritter, and W. Xu. Having beer after prayer? measuring cultural bias in large language models. *arXiv preprint arXiv:2305.14456*, 2023.