ML Fairness Bootcamp

Day 2: Ameliorating bias

Nick Merrill

Bootcamp Timeline

Lecture 1
Introducing Bias

Lab 1
Bias in Healthcare

Lecture 2
Mitigating Bias

Lab 2
Gender Bias in Hiring

Table of contents

01

What’s the lingo?

Let’s demystify some of the jargon that experts use to talk about ML bias.

Grouping Terminology

Privileged Group

We expect this group to get the favorable outcome **more often** than they should.

Unprivileged Group

We expect this group to get the favorable outcome **less often** than they should.

		Privileged Group	Unprivileged Group
Adult Census Income	Race	White	Non-White
	Sex	Male	Non-Male
Recidivism (Compas)	Race	White	Non-White
	Sex	Female	Male

Find real-life datasets with a privileged group

5 minute break-out groups.

02

How can we find bias?

There are **concrete strategies to quantify bias** and assess its severity.

Review: Why can't we be blind to bias?

Always include the sensitive features (like race and gender)

- Sensitive features cannot be ignored.
- **If algorithms "hear no evil", we can "see no evil"**: "non-sensitive" features in our dataset may be correlated with sensitive features; if we remove those features, we will not know our algorithms exhibit bias.

Strategies to Identify Bias

- **DISPARATE IMPACT**
- STATISTICAL PARITY
- EQUAL OPPORTUNITY
- AVERAGE ODDS

Each group should have an equal opportunity of achieving the favorable outcome.
We calculate the ratio of rate of favorable outcome for unprivileged group compared to that of privileged group.

The ideal value is 1.
A value < 1 implies there is benefit toward the privileged group.

	Trait 1	Privileged	Ratio	Trait 2	Privileged	Ratio
Adult Income	Race	White	0.55	Sex	Male	0.29
Recidivism (Compas)	Race	White	0.75	Sex	Female	0.59

How does disparate impact compare to the 4/5 rule?

Strategies to Identify Bias

DISPARATE IMPACT
STATISTICAL PARITY
EQUAL OPPORTUNITY
AVERAGE ODDS

Statistical Parity:
Demographics of those receiving any classification should be the same as demographics of the underlying population.

Equal Opportunity / Average Odds
Each group should be classified (in)correctly at the same rate.

Look to the appendix to learn about these other strategies for identifying bias.

03

How do we make it better?

While there is no way to “fix” bias, **there are methods for making bias less harmful**.

Misconceptions

"Bias Starts in Data"

Bias can start **anywhere**: pre-processing, post-processing, domain specification, models...

"Algorithms Don't Create Bias – They Transmit it"

If an algorithm can be de-biased with a dataset, they can themselves be biased.

"Race & gender are the most obvious biases."

These are the least obvious. they're protected: least likely included in training data.

"Bias is Fixable"

"All data embeds a worldview & **all models have some bias**. Most interventions just try to make the model biased towards more inclusive (& non-illegal) outcomes."

Strategies to Mitigate Bias

- **FAIRNESS CONSTRAINTS**
- REWEIGHTING
- OPTIMIZED PRE-PROCESSING
- ADVERSARIAL DEBIASING
- REJECT-OPTION-BASED CLASSIFICATION

Fairness constraints allow us to specify a tradeoff between a classifier's "fairness" and its accuracy.
Sometimes, our dataset is badly biased. For example, a dataset of past hiring decisions may embed a bias against women. In this case, an "accurate" classifier would be unfair - perhaps illegally so.

To correct for this, we can set a fairness constraint (e.g., a minimum disparate impact score).
With this constraint, the classifier will be as accurate as possible while exceeding the minimum disparate impact score.

In Lab 2, you will use fairness constraints and disparate impact to ameliorate gender bias in an automated hiring system.

Strategies to Mitigate Bias

- CONSTRAINTS
- REWEIGHTING
- OPTIMIZED PRE-PROCESSING
- ADVERSARIAL DEBIASING
- REJECT-OPTION-BASED CLASSIFICATION

Look to the appendix to learn about these other strategies.

Summary

- ML bias is **sociotechnical**
  - It is not **caused** by technical problems alone, and cannot be **"solved"** by technical solutions alone

- **Technical approaches are one** way of understanding and beginning to address ML bias - a tool in the toolbox
  - Detecting bias
  - Ameliorating bias

04

What’s next?

There’s **a lot more work to do** in this space – here is a non-comprehensive list of what can come after today.

Topics to Explore Further

Accountability

Assigning responsibility for harm. Read a primer here.

Transparency & Explainability

How do algorithms make the decisions they make?

Environmental Consequences

Training large models means large energy requirements.

Reinforcement Learning

Reward function specification, existential risk/x-risk

What you can do now

Read more

ML bias: Dispelling common misconceptions by Deb Raji.

Solon Barocas, Moritz Hardt & Arvind Narayanan's Fair ML Book

Anna Lauren Hoffman. Where Fairness Fails (2019).

Bender, Gebru, McMillan- Major & Mitchell. On the Dangers of Stochastic Parrots (2021).

Get involved!

Rediet Abebe's Mechanism Design for Social Good

Black in AI

Algorithmic Justice League

AFOG (Algorithmic Fairness and Opacity Working Group

Bootcamp Timeline

Lecture 1

Introducing Bias

Lab 1

Bias in Healthcare

Lecture 2

Mitigating Bias

Lab 2

Gender Bias in Hiring

05

Lab 2

Improving classifier fairness in the context of **gender bias in hiring**.

Background

Employers like Amazon have attempted to **use algorithms to automate hiring**. But these algorithms have **shown a bias against hiring women** (See: Reuters, 2018).

Why?

- In the past, human managers have discriminated against women in hiring.
- So, if we train an algorithm on humans' past decisions, **it will learn the same bias**.
- In other words, an "accurate" classifier would be biased: it would "accurately" capture managers' bias against hiring women.

What can we do?

- Pick an acceptable fairness score (in this lab, we'll use **disparate impact**).
- Use **fairness constraints** to make the classifier **as accurate as possible while meeting our metric of fairness**.
- One frame: there is a tradeoff between fairness and accuracy.
- Another frame: the data is incorrectly labeled due to managers' bias. We are trying correct their labeling.

Answer: Because there's a gender pay gap, **the algorithm will simply learn gender from income**, to which gender is correlated.

Let’s get started!

https://colab.research.google.com/drive/1GhRPfQ9gcG1JiAyP0TUx9gMkGLkGG5Sy

Run in browser: File > Save a copy in Drive

Run as Jupyter notebook: File > Download > Save as .ipynb (Requires wget, python3 and pip)

06

Appendix

For further information!

Strategies to Identify Bias

- DISPARATE IMPACT
- **STATISTICAL PARITY**
- EQUAL OPPORTUNITY
- AVERAGE ODDS

Demographics of those receiving any classification should be the same as demographics of the underlying population.
We take the difference of rate of favorable outcomes by rate of favorable outcomes by unprivileged group.

The ideal value is 0.
A value < 0 implies there is benefit toward the privileged group.

	Trait 1	Privileged	Ratio	Trait 2	Privileged	Ratio
Adult Income	Race	White	-0.18	Sex	Male	-0.33
Recidivism (Compas)	Race	White	-0.18	Sex	Female	-0.36

When might statistical parity and disparate impact disagree?

Strategies to Identify Bias

- DISPARATE IMPACT
- STATISTICAL PARITY
- **EQUAL OPPORTUNITY**
- AVERAGE ODDS

Each group should be 'equally' incorrectly classified.
We take the difference of true positive rates between unprivileged and privileged groups.

The ideal value is 0.
A value < 0 implies there is benefit toward the privileged group.

	Trait 1	Privileged	Ratio	Trait 2	Privileged	Ratio
Adult Income	Race	White	-0.06	Sex	Male	-0.14
Recidivism (Compas)	Race	White	-0.12	Sex	Female	-0.30

Where is there the **most bias** in these 2 datasets?

Strategies to Identify Bias

- DISPARATE IMPACT
- STATISTICAL PARITY
- EQUAL OPPORTUNITY
- **AVERAGE ODDS**

Each group should be ‘equally’ incorrectly classified.
We take the average difference of false positive rate and true positive rate between unprivileged and privileged groups.

The ideal value is 0.
A value < 0 implies there is benefit toward the privileged group.

	Trait 1	Privileged	Ratio	Trait 2	Privileged	Ratio
Adult Income	Race	White	-0.09	Sex	Male	-0.19
Recidivism (Compas)	Race	White	-0.16	Sex	Female	-0.35

Where is there the **most bias** in these 2 datasets?

Strategies to Mitigate Bias

- **REWEIGHTING**
- OPTIMIZED PRE-PROCESSING
- ADVERSARIAL DEBIASING
- REJECT-OPTION-BASED CLASSIFICATION

Weights the examples in each (group, label) combination differently to ensure fairness before classification.

Strategies to Mitigate Bias

- REWEIGHTING
- **OPTIMIZED PRE-PROCESSING**
- ADVERSARIAL DEBIASING
- REJECT-OPTION-BASED CLASSIFICATION

Learns a probabilistic transformation that can modify the features and the labels in the training data.

Strategies to Mitigate Bias

- REWEIGHTING
- OPTIMIZED PRE-PROCESSING
- **ADVERSARIAL DEBIASING**
- REJECT-OPTION-BASED CLASSIFICATION

Learns a classifier that maximizes prediction accuracy and simultaneously reduces an adversary's ability to determine the protected attribute from the predictions.

Since the predictions cannot carry any group discrimination information that the adversary can exploit, the classifier must be fair (right?).

Strategies to Mitigate Bias

- REWEIGHTING
- OPTIMIZED PRE-PROCESSING
- ADVERSARIAL DEBIASING
- **REJECT-OPTION-BASED CLASSIFICATION**

Changes predictions from a classifier to make them fairer.

Provides favorable outcomes to unprivileged groups and unfavorable outcomes to privileged groups in a confidence band around the decision boundary with the highest uncertainty.