Introduction to Machine Learning

Fairness in Machine Learning

Varun Chandola

Computer Science & Engineering
State University of New York at Buffalo
Buffalo, NY, USA
chandola@buffalo.edu

University at Buffalo
Department of Computer Science and Engineering
School of Engeering and Applied Sciences

Outline

Introduction to Fairness

Toy Example

Why fairness?

- Defining Fairness

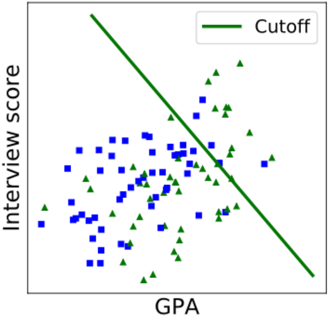
References

Introduction

- Main text - `https://fairmlbook.org` [1]
  - Solon Barocas, Moritz Hardt, Arvind Narayanan
- Other recommended resources:
  - Fairness in machine learning (NeurIPS 2017)
  - 21 fairness definitions and their politics (FAT* 2018)
  - Machine Bias - COMPAS Study
- Must read - The **Machine Learning Fairness Primer** by Dakota Handzlik
- Programming Assignment 3 and Gradiance Quiz #7
- Also see - The Mozilla Responsible Computer Science Challenge

Toy Example

- *Task*: Learn a ML based job hiring algorithm
- *Inputs*: GPA, Interview Score
- *Target*: Average performance review
- *Sensitive attribute*: Binary (denoted by □ and Δ), represents some demographic group
    - We note that GPA is correlated with the sensitive attribute



Process

1. Regression model to predict target
2. Apply a threshold (denoted by green line) to select candidates

Toy Example

- ML models does not use sensitive attribute
- Does it mean it is fair?

Toy Example

- ML models does not use sensitive attribute
- Does it mean it is fair?
- It depends on the definition of fairness

Toy Example

- ML models does not use sensitive attribute
- Does it mean it is fair?
- It depends on the definition of fairness

Fairness-as-blindness notion

- Two individuals with similar features get similar treatment
- This model is fair

What about a different definition of fairness?

- Are candidates from the two groups equally likely to be hired?

What about a different definition of fairness?

- Are candidates from the two groups equally likely to be hired?
- No - triangles are more likely to be hired than squares
- Why did the model become unfair because of this definition?
  - In the training data, average performance review is lower for squares than triangles

Why this disparity in the data?

- Many factors could have led to this:
  - Managers who score employee's performance might have a bias
  - Workplace might be biased against one group
  - Socio-economic background of one group might have resulted in poor educational outcomes
  - Some intrinsic reason
  - Combination of these factors
- Let us assume that this disparity that was learnt by the ML model is unjustified
- How do we get rid of this?

Making ML model bias-free

- Option 1: ignore GPA as a feature
  - Might result in poor accuracy of the model

Making ML model bias-free

- Option 1: ignore GPA as a feature
  - Might result in poor accuracy of the model
- Option 2: pick different thresholds for each sub-group
  - Model is no longer “blind”

Making ML model bias-free

- Option 1: ignore GPA as a feature
  - Might result in poor accuracy of the model
- Option 2: pick different thresholds for each sub-group
  - Model is no longer "blind"
- Option 3: add a diversity reward to the objective function
  - Could still result in poor accuracy

Why fairness?

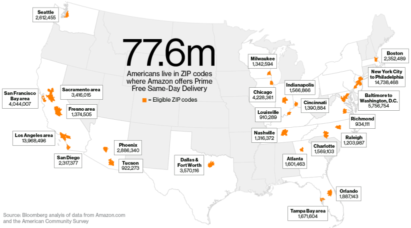
- We want/expect everything to be fair and bias-free
- Machine learning driven systems are everywhere
- Obviously we want them to be fair as well
  - Closely related are issues of ethics, trust, and accountability

What does fairness mean?

- **Consequential decision making**: ML system makes a decision that impacts individuals
  - admissions, job offers, bail granting, loan approvals
- Should use factors that are *relevant* to the outcome of interest

Amazon same-day delivery

- A data-driven system to determine neighborhoods to offer *same-day delivery* service



Source: Bloomberg analyis of data from Amazon.com and the American Community Survey

  - In many U.S. cities, white residents were more than twice as likely as black residents to live in one of the qualifying neighborhoods.
  - *Src:* - `https://www.bloomberg.com/graphics/2016-amazon-same-day/`

ML - Antithesis to fairness

- Machine learning algorithms are based on *generalization*
- Trained on historical data which can be unfair
    - Our society has always been unfair
- Can perpetuate historical prejudices

Continuing with the Amazon example

- Amazon claims that *race* was not a factor in their model (not a feature)
- Was designed based on efficiency and cost considerations
- Race was *implicitly* coded

When is there a fairness issue?

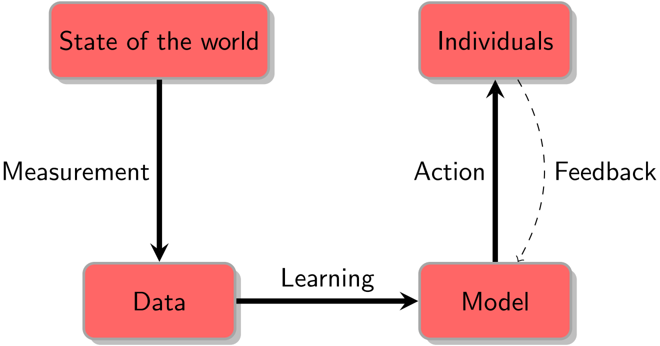
- What if the Amazon system was such that zip codes ending in an odd digit are selected for same-day delivery?
- It is biased and maybe unfair to individuals living in the even numbered zipcodes
- But will that trigger a similar reaction?
- Is the system unfair?

What do we want to do?

- Make machine learning algorithms fair
- Need a quantifiable fairness metric
  - Similar to other performance metrics such as precision, recall, accuracy, etc.
- Incorporate the fairness metric in the learning process
- Often leads to a tension with other metrics

How does an ML algorithm becomes unfair?

- The "ML for People" Pipeline



State of the world

Individuals

Measurement

Action

Feedback

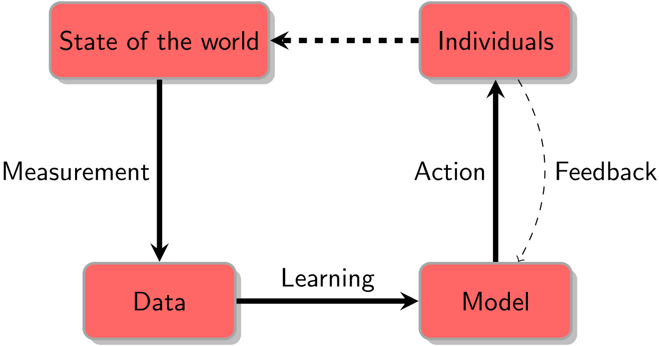
Data

Learning

Model

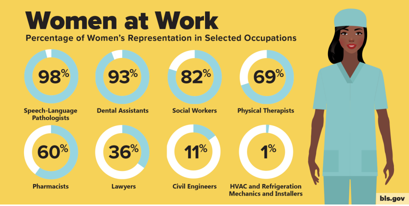
How does an ML algorithm becomes unfair?

- The “ML for People” Pipeline



Issues with the state of the society

- Most ML applications are about people
  - Even a pothole identification algorithm
- Demographic disparities exist in society
- These get embedded into the training data
- As ML practitioners we are not focused on removing these disparities
- We do not want ML to reinforce these disparities
- The dreaded **feedback loops** [2]



Measurement Issues

- Measurement of data is fraught with subjectivity and technical issues
- Measuring race, or any categorical variable, depends on how the categories are defined
- Most critical - defining the target variable
  - Often this is "made up" rather than measured objectively
  - credit-worthiness of a loan applicant
  - attractiveness of a face (beauty.ai, FaceApp)

Criminal Risk Assessment

1. Target variable - bail or not?
2. Target variable - will commit a crime later or not (recidivism)?

Measurement Issues

- Technical issues can often lead to bias
  - Default settings of cameras are usually optimized for lighter skin tones [3]

- Most images data sets used to train object recognition systems are biased relative to each other
  - http://people.csail.mit.edu/torralba/research/bias/

How to fix the measurement bias?

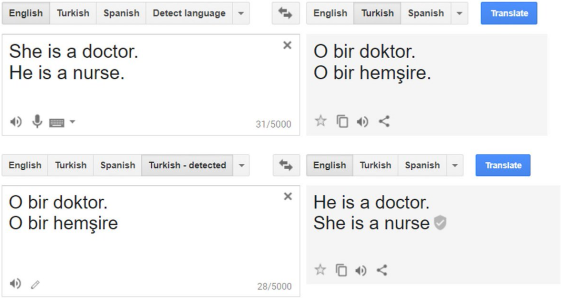
- Understand the provenance of the data
  - Even though you (ML practitioner) are working with data “given” to you
- “Clean” the data

Issues with models

- We know the training data can have biases
- Will the ML model preserve, mitigate or exacerbate these biases?
- ML model will learn a pattern in the data that assists in optimizing the objective function
- Some patterns are useful - *smoking is associated with cancer*, some are not - *girls like pink and boys like blue*
- But ML algorithm has not way of distinguishing between these two types of patterns
  - established by social norms and moral judgements
- Without a specific intervention, the ML algorithm will extract stereotypes

An Example

- Machine translation



She is a doctor.
He is a nurse.

O bir doktor.
O bir hemşire.

O bir doktor.
O bir hemşire

He is a doctor.
She is a nurse

How to make the ML model more fair

- Model reflects biases in the data
- Withold sensitive attributes (gender, race, . . .)
- Is that enough?

How to make the ML model more fair

- Model reflects biases in the data
- Withold sensitive attributes (gender, race, . . .)
- Is that enough?

Unfortunately not

- There could be *proxies* or *redundant encodings*
- Example - Using "programming experience in years" might indirectly encode gender bias
  - Age at which someone starts programming is well-known to be correlated with gender

How to make the ML model more fair

- Better objective functions that are fair to all sub-groups
  - More about this in next lecture
- Ensure equal error rate for all sub-groups

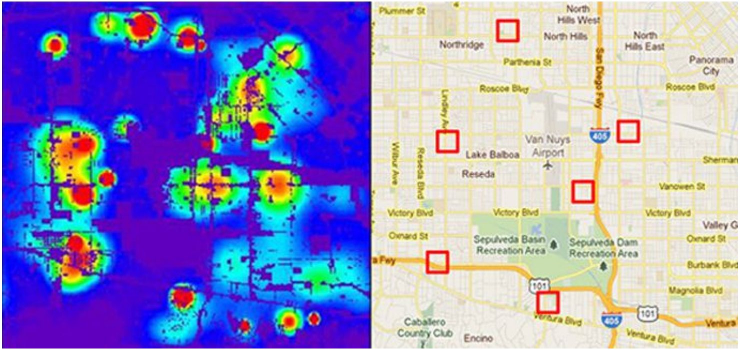
The Nymwars Controversy

- Google, Facebook and other companies blocking users with uncommon names (presumably *fake*)
- Higher error rate for cultures with a diverse set of names

The pitfalls of action

- While as ML practitioners our world ends after we have trained a *good* model
- But this model will impact people
- Need to understand that impact in the larger socio-technical system
  - Are there disparities in the error across different sub-groups?
  - How do these disparities change over time (drift)?
  - What is the perception of society about the model?
    - Ethics, trustworthiness, accountability
    - Explainability and interpretability
    - **Correlation is not causation**

The perils of feedback loops



- The "actions" made by individuals based on the predictions of the ML model could be fed back into the system, either explicitly or implicitly
  - Self-fulfilling predictions
  - Predictions impacting the training data
  - Predictions impacting the society

References

S. Barocas, M. Hardt, and A. Narayanan.
Fairness and Machine Learning.
fairmlbook.org, 2019.
http://www.fairmlbook.org.

D. Ensign, S. A. Friedler, S. Neville, C. Scheidegger, and S. Venkatasubramanian.
Runaway feedback loops in predictive policing.
In *Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA*, volume 81 of *Proceedings of Machine Learning Research*, pages 160–171. PMLR, 2018.

L. Roth.
Looking at shirley, the ultimate norm: Colour balance, image technologies, and cognitive equity.
Canadian Journal of Communication, 34:111–136, 2009.