Lecture 24

Bias & fairness in learning systems

Murat Unal

Johnson Graduate School of Management
Cornell University

11/23/2021

Agenda

Reminder 1: Assignment 4 due 11/24 11:59 PM

Reminder 2: Project deliverables due 11/30 1:00 PM

- Analysis
- Executive Summary
- Presentation Slides
- Individual Evaluation

Bias & fairness in ML

Data-driven decisions

- Our success, happiness, and wellbeing are never fully of our own making
- Others' decisions can profoundly affect the course of our lives
- Getting into a school, job, loan etc.

Data-driven decisions

- Our success, happiness, and wellbeing are never fully of our own making
- Others' decisions can profoundly affect the course of our lives
- Getting into a school, job, loan etc.
- How do we ensure that these decisions are made the right way and for the right reasons?

Data-driven decisions

- Our success, happiness, and wellbeing are never fully of our own making
- Others' decisions can profoundly affect the course of our lives
- Getting into a school, job, loan etc.
- How do we ensure that these decisions are made the right way and for the right reasons?
- Good decisions take available evidence into account.

Data-driven decisions

- In many situations data-driven decisions trounce those based on intuition or expertise.

Data-driven decisions

- In many situations data-driven decisions trounce those based on intuition or expertise.
- Machine learning promises to bring greater discipline to decision making because it offers to **uncover factors** that are relevant to decision-making that humans might overlook.

Data-driven decisions

- By exposing the computer to many examples, we hope the computer will learn the patterns that reliably distinguish different objects from one another and from the environments in which they appear.
- Learning involves **generalizing** from examples

Data-driven decisions

- Evidence-based decision making is only as reliable as the evidence on which it is based, and high quality examples are critically important to ML.

Data-driven decisions

- Evidence-based decision making is only as reliable as the evidence on which it is based, and high quality examples are critically important to ML.
- The fact that ML is evidence-based by no means ensures that it will lead to accurate, reliable, or fair decisions.

Data-driven decisions

- Our historical examples of the relevant outcomes will almost always reflect historical prejudices against certain social groups, prevailing cultural stereotypes, and existing demographic inequalities.
- And finding patterns in these data will often mean replicating these very same dynamics

Demographic disparities

- Amazon uses a data-driven system to determine the neighborhoods in which to offer free same-day delivery.
- White residents were more than twice as likely as black residents to live in one of the qualifying neighborhoods

Demographic disparities

- Amazon uses a data-driven system to determine the neighborhoods in which to offer free same-day delivery.
- White residents were more than twice as likely as black residents to live in one of the qualifying neighborhoods
- Amazon argued that its system was justified because it was designed based on efficiency and cost considerations and that race wasn't an explicit factor

Demographic disparities

- Amazon uses a data-driven system to determine the neighborhoods in which to offer free same-day delivery.
- White residents were more than twice as likely as black residents to live in one of the qualifying neighborhoods
- Amazon argued that its system was justified because it was designed based on efficiency and cost considerations and that race wasn't an explicit factor
- Nonetheless, it has the effect of providing different opportunities to consumers at racially disparate rates
- The concern is that this might contribute to the perpetuation of long-lasting cycles of inequality.

Demographic disparities

- How do we conceptualize ML systems and the responsibilities of those building them?

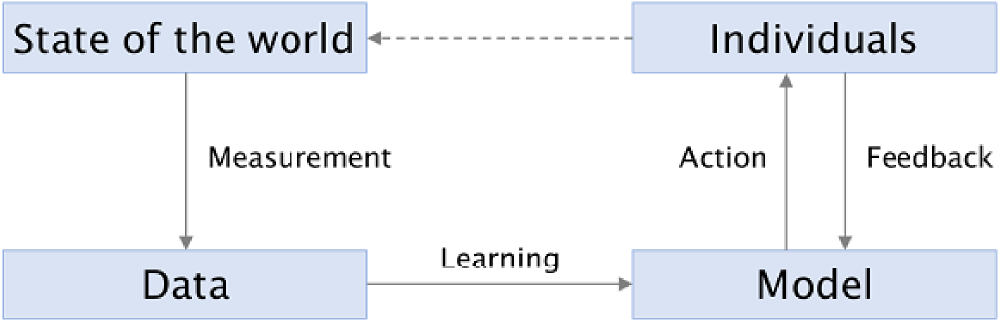
Demographic disparities

- How do we conceptualize ML systems and the responsibilities of those building them?
- Is our goal to faithfully reflect the data?

Demographic disparities

- How do we conceptualize ML systems and the responsibilities of those building them?
- Is our goal to faithfully reflect the data?
- Or do we have an obligation to **question** the data, and to design our systems to conform to some notion of **equitable behavior**, regardless of whether or not that's supported by the data currently available to us?

The ML loop

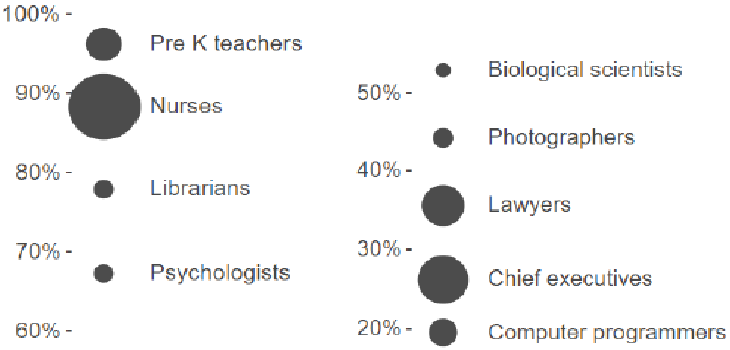

State of the world
Individuals
Measurement
Action
Feedback
Learning
Data
Model

State of the world

- Most ML applications involve data about people
- In these applications, the available training data will likely encode the **demographic disparities** that exist in our society.

State of the world

[4] The percentage of women in a sample of occupations in the United States. The area of the bubble represents the number of workers.



State of the world

- Automated Essay Scoring
- Seeks algorithms that attempt to match the scores of human graders of student essays.
- What can go wrong?

State of the world

- Targeted ads in marketing
- How can they perpetuate bias?

State of the world

- Boston:"Street Bump"
- A project by the city of Boston to crowdsource data on potholes. The smartphone app automatically detects pot holes using data from the smartphone's sensors and sends the data to the city.
- What can go wrong?

State of the world

- Human society is full of **demographic disparities**, and training data will likely reflect these.
- As we integrate ML into decision-making, we should be careful to ensure that ML doesn't become a part of this feedback loop.

Measurement

- Measurement involves defining your variables of interest and turning your observations into numbers
- Usually ML practitioners don't think about these steps, because someone else has already done those things.
- And yet it is crucial to understand the provenance of the data.

Measurement

- Measurement is fraught with subjective decisions and technical difficulties
- "Even with Affirmative Action, Blacks and Hispanics Are More Underrepresented at Top Colleges Than 35 Years Ago" (https://www.nytimes.com/interactive/2017/08/24/us/affirmative-action.html, 2017).

Measurement

- Defining the **target variable** is particularly challenging
- How do you define **creditworthiness**?
- How do you rank **physical attractiveness**?
- How do you define **good employee**?

Measurement

- We might use performance review scores to quantify it.

Measurement

- We might use performance review scores to quantify it.
- This means that our data inherits any biases present in managers' evaluations of their reports.

Measurement

- We might use performance review scores to quantify it.
- This means that our data inherits any biases present in managers' evaluations of their reports.
- Instead of relying on performance reviews for (say) a sales job, we might rely on the number of sales closed.
- But is that an objective measurement?

From data to models

- Training data reflects the disparities, distortions, and biases from the real world and the measurement process.
- When we learn a model from such data, are these disparities preserved, mitigated, or exacerbated?

From data to models

- Predictive models trained with supervised learning methods are often good at calibration: ensuring that the model's prediction subsumes all features in the data for the purpose of predicting the outcome.
- But calibration also means that by default, we should expect our models to faithfully **reflect disparities** found in the input data.

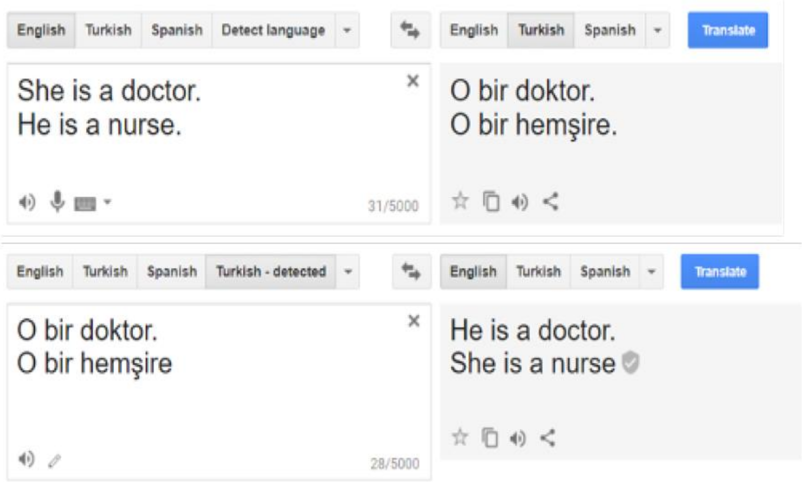
From data to models

- Patterns we wish to **learn** from data: smoking is associated with cancer
- Patterns we wish to **avoid**: girls like pink and boys like blue

From data to models

- Patterns we wish to **learn** from data: smoking is associated with cancer
- Patterns we wish to **avoid**: girls like pink and boys like blue
- But learning algorithms have no general way to distinguish between these two types of patterns, because they are the result of social norms and moral judgments.
- Absent specific intervention, machine learning will extract stereotypes, including incorrect and harmful ones, in the same way that it extracts knowledge.

From data to models



The pitfalls of action

- If a model is calibrated—it faithfully captures the patterns in the underlying data
- Predictions made using that model will inevitably have disparate error rates for different groups, if those groups have different base rates.
- Understanding the properties of a prediction requires understanding not just the model, but also the population differences between the groups on which the predictions are applied.

The pitfalls of action

- If a model is calibrated—it faithfully captures the patterns in the underlying data
- Predictions made using that model will inevitably have disparate error rates for different groups, if those groups have different base rates.
- Understanding the properties of a prediction requires understanding not just the model, but also the population differences between the groups on which the predictions are applied.
- Subpopulations change differently over time, that can introduce disparities as well.

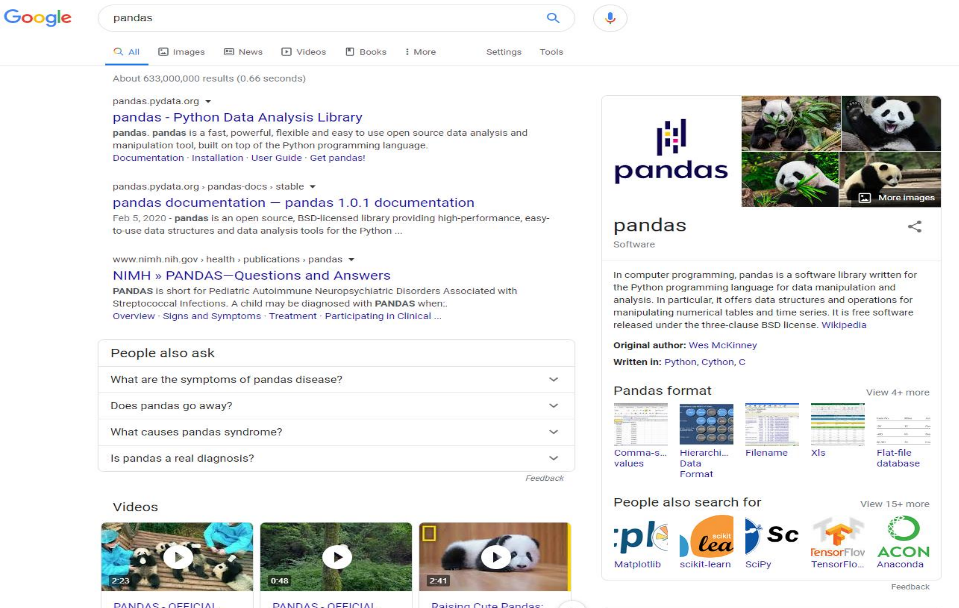
The pitfalls of action

- A major limitation of machine learning is that it only reveals correlations, but we often use its predictions as if they reveal causation
- Can patients with asthma have lower risk of dying from pneumonia?
- "Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital 30- Day Readmission," (Proc. 21st ACM SIGKDD, 2015, 1721–30).

Feedback and feedback loops

- Many systems receive feedback when they make predictions
- But feedback is tricky to interpret correctly.
- If a user clicked on the first link on a page of search results, is that simply because it was first, or because it was in fact the most relevant?
- Even feedback that's designed into systems can lead to unexpected or undesirable biases

Feedback and feedback loops



Feedback and feedback loops

pandas

ALL | IMAGES | VIDEOS | MAPS | NEWS | SHOPPING | MY SAVES

What do you want to do with pandas?

install pandas | save pandas

PANDAS is an acronym for "pediatric autoimmune neuropsychiatric disorders associated with streptococcal infections.". It is a fairly recently described disorder (1990s). An autoimmune response to a streptococcal infection is the leading theory as to the cause of PANDAS.

What Is PANDAS Syndrome? Symptoms and Treatment
www.medicinenet.com/pandas/article.htm

Was this helpful?

PEOPLE ALSO ASK

What is so special about giant pandas and pandas?

What is the difference between Pandas and Panda Bears?

What is the primary food for PANDAS?

What are dangers for PANDAS?

Feedback

pandas - Python Data Analysis Library
https://pandas.pydata.org

pandas pandas is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool, built on top of the Python programming language. **pandas** is ...

Installation · Get Pandas · Pandas 0.24.2 Documentation · What's New in 0.25.3

PANDAS

Pediatric autoimmune neuropsychiatric disorders associated with streptococcal infections is a hypothesis that there exists a subset of children with rapid onset of obsessive-compulsive disorder or tic disorders and these symptoms are caused by group A beta-hemolytic streptococcal infections. The proposed link between infection and these disorders is that an initial autoimmune reaction to a GABHS infection produces antibodies that interfere with basal ganglia function, causing symptom exacerbations. It has been proposed that this autoimmune response can result in a broad range of neuropsychiatric symptoms. PANDAS is a subset of the pediatric autoimmune neuropsychiatric syndrome hypothesis.

Wikipedia

People also search for

Sydenham's chorea

Rheumatic Fever

Tourette Syndrome

Group A streptococcal infection

Strep Throat

See more

Data from: Wikipedia

Suggest an edit

Feedback and feedback loops

- Self-fulfillling predictions
- Predictions that affect the training set
- Initial bias could be amplified by a feedback loop

Is this a fair classifier?