Lecture 18: Fairness in Machine Learning

10 December 2019

Learning Outcomes

- Understand how a dataset can be biased
- Understand the way in which machine learning algorithms can amplify bias
- Understand the implications of bias and why developers must be conscious of this

Warning

- We will discuss issues of fairness in machine learning
- Some examples of unfairness will be presented
- These will involved real examples of discrimination based on protected characteristics, including gender and race.

Two very simple examples

- Two classes – one class has 20 data points, one class has 5.
- Is *k*nn a fair algorithm?
- How about LDA?

The Problem with Priors

- Bayes rule: $P(B|A)=P(A|B)P(B)/P(A)$
- Explicitly relies on prior belief
- Intrinsically biases inference to reflect and amplify our existing beliefs
- Especially problematic if those beliefs are wrong

Argument

- Algorithms only reflect the truth in the underlying data
- If decisions reflect the relative class priors, is that a problem?

- Where has the data come from?
- Does the data itself reflect historic bias?

The Matthew Effect

- “For to every one who has will more be given, and he will have abundance; but from him who has not, even what he has will be taken away.”
  - Matthew 25:29, Revised Standard Version of the Bible

- Noted in science by Merton (Science **159**(3810), 56-63 (1968))

- A danger of relying on data to make decisions

The Matthew Effect – Consequences

- Datasets can reflect historic bias
- Certain characteristics can be weighted to reflect this

Gender Bias in Word Embeddings

- A mapping from words/phrases to vectors of real numbers.
  - word2vec, GloVe
- Unsupervised – learns associations of words from large text corpora
- Lazy example: King – Man + Woman = Queen
- But also: Doctor – Man + Woman = ?
- Societal bias may be inherent in the data

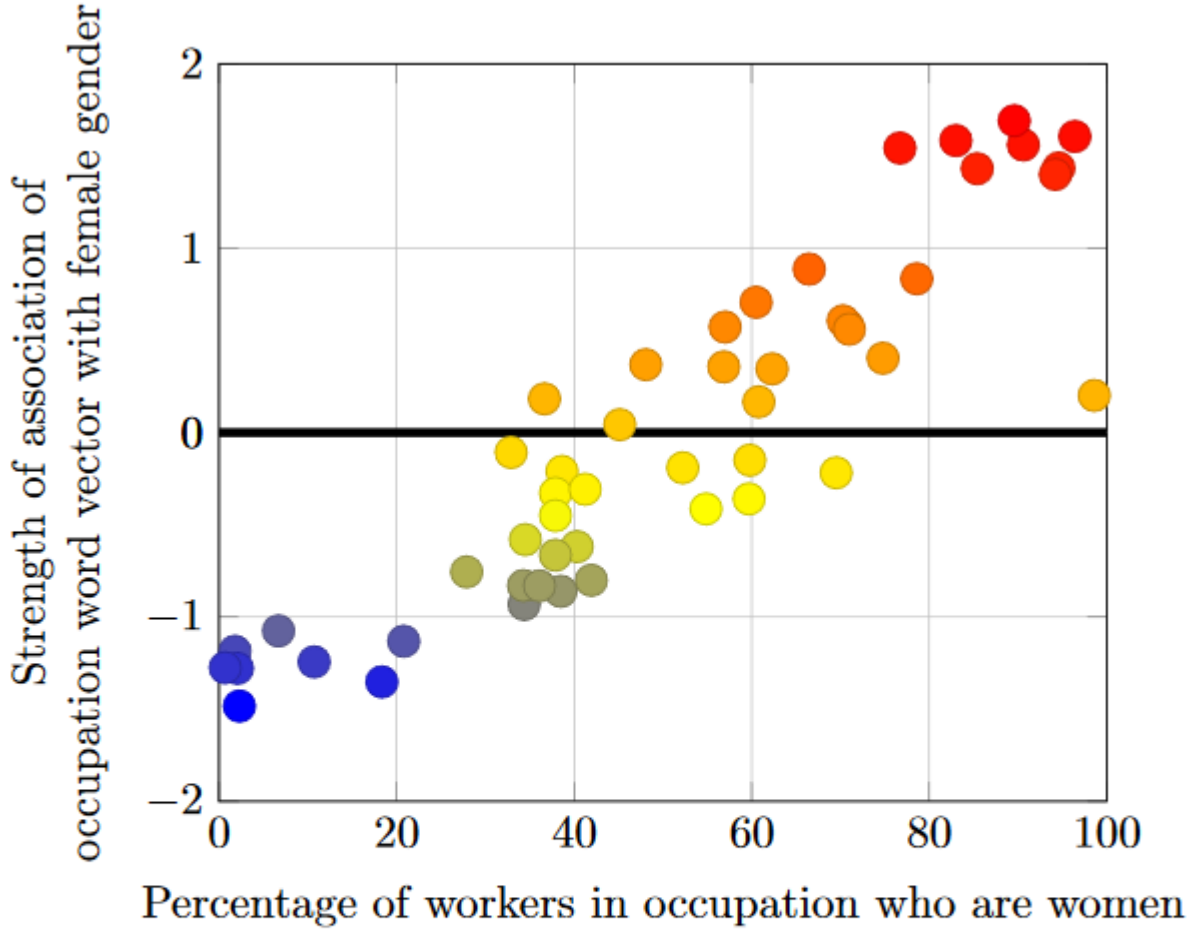

Figure 1: Occupation-gender association. Pearson's correlation coefficient $\rho = 0.90$ with p-value $< 10^{-18}$.

Caliskan, Bryson, and Narayanan https://science.sciencemag.org/content/356/6334/183

Millions of black people affected by racial bias in health-care algorithms

Study reveals rampant racism in decision-making software used by US hospitals — and highlights ways to correct it.

https://www.nature.com/articles/d41586-019-03228-6

Dissecting racial bias in an algorithm used to manage the health of populations

Ziad Obermeyer[1,2,*], Brian Powers[3], Christine Vogeli[4], Sendhil Mullainathan[5,*,†]

+ See all authors and affiliations

Science 25 Oct 2019:
Vol. 366, Issue 6464, pp. 447-453
DOI: 10.1126/science.aax2342

Abstract

Health systems rely on commercial prediction algorithms to identify and help patients with complex health needs. We show that a widely used algorithm, typical of this industry-wide approach and affecting millions of patients, exhibits significant racial bias: At a given risk score, Black patients are considerably sicker than White patients, as evidenced by signs of uncontrolled illnesses. Remedying this disparity would increase the percentage of Black patients receiving additional help from 17.7 to 46.5%. The bias arises because the algorithm predicts health care costs rather than illness, but unequal access to care means that we spend less money caring for Black patients than for White patients. Thus, despite health care cost appearing to be an effective proxy for health by some measures of predictive accuracy, large racial biases arise. We suggest that the choice of convenient, seemingly effective proxies for ground truth can be an important source of algorithmic bias in many contexts.

https://science.sciencemag.org/content/366/6464/447.abstract

Bias in Facial Recognition

“The study found that Amazon had an error rate of 31% when identifying the gender of images of women with dark skin.

This compared with a 22.5% rate from Kairos, which offers a rival commercial product, and a 17% rate from IBM.

By contrast Amazon, Microsoft and Kairos all successfully identified images of light-skinned men 100% of the time.”

https://www.bbc.co.uk/news/technology-47117299

Raji, Inioluwa Deborah, and Joy Buolamwini. "Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products." AAAI/ACM Conf. on AI Ethics and Society. Vol. 1. 2019.

Right to an explanation

- European Union General Data Protection Regulation

The data subject should have the right not to be subject to a decision, which may include a measure, evaluating personal aspects relating to him or her which is based solely on automated processing and which produces legal effects concerning him or her or similarly significantly affects him or her, such as automatic refusal of an online credit application or e-recruiting practices without any human intervention.

...

In any case, such processing should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision.

Much current debate about what form an explanation should take and what the legislation really means

The Key Messages

- Datasets, especially constructed from human decisions or annotations, can include our implicit and explicit biases
- Some decision-making algorithms can reinforce and even amplify those biases
- These decisions can have serious implications for both individuals and for whole groups of people
- Understanding the algorithm is not enough – you must understand the limits of your data

Further Reading

Much recent literature, I have found the following especially useful

- The Trouble with Bias: Kate Crawford's keynote talk at NIPS 2017
  - https://www.youtube.com/watch?v=fMym_BKWQzk
- Fairness and machine learning. Limitations and Opportunities
  - Solon Barocas, Moritz Hardt, Arvind Narayanan
  - https://fairmlbook.org/
- Datasheets for Datasets
  - Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumeé III, Kate Crawford
  - https://arxiv.org/abs/1803.09010
  - All of the authors of this paper have produced many excellent works on bias in ML.