APS360 Artificial Intelligence Fundamentals

Lisa Zhang

Lecture 15; July 22, 2019

Agenda

Today:

- Ethics in Artificial Intelligence (60 min?)
- Fairness in Machine Learning (40 min?)

Goals:

- Understand the major issues in AI ethics
- Understand different approaches to measuring fairness of machine learning models
- Understand how to be a responsible AI practictioner

Ethics in Artificial Intelligence

Mindful Listening Exercise

- Find a partner and assign A and B
- There will be a question on the next slide
- A answers question, B listens (2 min)
- B answers questions, A listens (2 min)
- Discuss what you learned from each other (2 min)

What excites you about AI? Share an experience of belonging/non-belonging in the AI community.

AI Ethics Landscape

- Split into 4 groups
- Each group will have a whiteboard + a marker
- Write down as many major issues in AI ethics as you can

(5 min)

News Articles

- Split into 8 groups
- Each group will be given a news article
- Read the article (5 min)
- Each group will have 2 minutes to explain the issues in the article to the class

AI Ethics Landscape

- Go back to your whiteboards
- Revise the ethics landscape
- Look at the other whiteboards

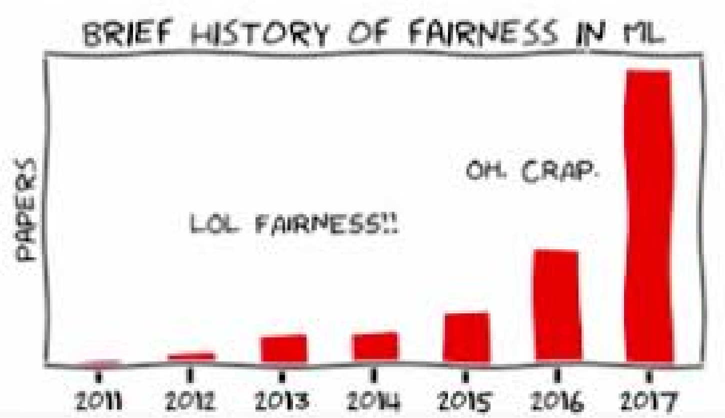
(5 min)

How can we (as AI practioners) prevent these issues?

What should we communicate about models we build?

Fairness in Machine Learning

Fairness


BRIEF HISTORY OF FAIRNESS IN ML
PAPERS
LOL FAIRNESS!!
OH. CRAP.
2011
2012
2013
2014
2015
2016
2017

Fairness

Q: How can our model from lab 3, 4 and 5 be "unfair"?

Fairness

Q: How can our model from lab 3, 4 and 5 be "unfair"?

But also, what does it mean for a machine learning model to be unfair?

Terminology

Equality:

- treating everyone the same

Equity:

- giving everyone what they need to be successful
- “equal opportunity”

Disparate Treatment

Model suffers from **disparate treatment** if decisions are correlated with the subject's sensitive attribute.

For example, in the sentencing model, does the model treat people of different ethnicities similarly?

Q: Suppose that ethnicity is not used as an input feature of the model. Does that mean that the model would treat people of different ethnicities the same way?

Disparate Impact

Model suffers from **disparate impact** if decisions disproportionally hurt people with sensitive attributes

For example, suppose we are building a model to determine whether or not to an applicant is admitted to graduate school.

Q: Does it make sense to use the same grade cutoff criteria for all applicants?

Ways of measuring fairness

There is no consensus on how to measure fairness of a model.

Different measure of fairness can contradict each other!

We'll introduce three metrics today:

- Demographic Parity
- Equalized Odds (Accuracy Parity)
- Individual Fairness

Fairness as Demographic Parity

- Acceptance **rates** of applications from both groups must be equal
- Also known as "independence" (terminology from statistics)

Problem:

- Fairness is measured at a *group* level
- Model can hire qualified people from one group, and random people from the other

Fairness as Equalized Odds (2016)

- Model should be **equally accurate** across both groups
- Also known as “accuracy parity”

Problem:

- False positives and false negatives have different impacts
- Does not help to close the gap between the two groups

Individual Fairness (2012)

- Similar individuals from different groups should be treated similarly

Problem:

- Hard to determine appropriate measure of "similarity" of inputs

Trade off

- The different definitions of fairness are inconsistent with each other
- Optimizing fairness means trading off accuracy

Ideas for more fair models

- **Pre-processing**: remove information correlated to sensitive attributes
- **Add regularization term**: add a “fairness” regularizer
- **Post-processing**: change the way we use a model to make predictions

References

[0] https://towardsdatascience.com/a-tutorial-on-fairness-in-mac
[1] http://www.cs.toronto.edu/~madras/presentations/fairness-ml-

Many thanks to Inioluwa Raji, Cindy Rottmann, Patricia Sheridan and others for helpful discussions and resources.