APS360 Fundamentals of AI

Lisa Zhang

Lecture 15; March 25, 2019

Office Hours

Office Hours This Week:

- Monday: 4pm-5pm BA2197 (Kingsley)
- Tuesday: 1pm-2pm BA5287A (Hojjat)
- Wednesday: 3pm-4pm BA4161 (Andrew)
- Thursday: 3pm-4pm BA2197 (Lisa)
- Friday: 1pm-2pm BA5287A (Bibin)

You can also get help from your TA mentor by email.

Agenda

Today:

- Fairness in Machine Learning
- Presentation

Thursday:

- Machine Learning Ethic

Both topics will be on the exam

Lecture Structure

- “Unfair” machine learning models in the news
- Terminology, definitions, ideas
- Look at previous models & code

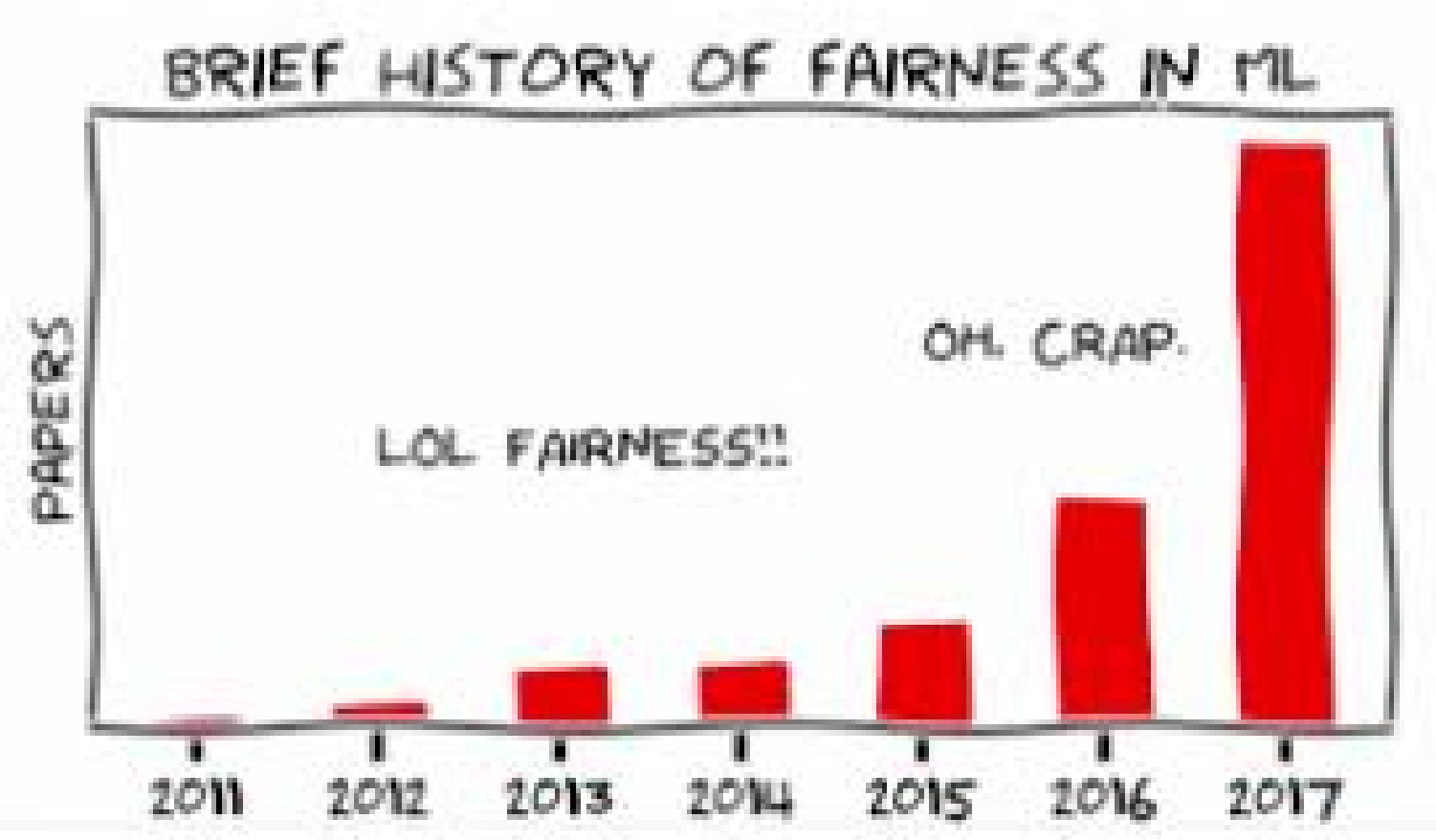
Required Reading for Thursday

Model Cards for Model Reporting

https://arxiv.org/pdf/1810.03993.pdf

Fairness in Machine Learning

Fairness


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

In the news...

- AI-Driven Dermatology Could Leave Dark-Skinned Patients Behind (Aug 2018)
  - https://www.theatlantic.com/health/archive/2018/08/machine-learning-dermatology-skin-color/567619/
- How Amazon Accidentally Invented a Sexist Hiring Algorithm (Oct 2018)
  - https://www.inc.com/guadalupe-gonzalez/amazon-artificial-intelligence-ai-hiring-tool-hr.html
- Google "fixed" its racist algorithm by removing gorillas from its image-labeling tech (Jan 2018)
  - https://www.theverge.com/2018/1/12/16882408/google-racist-gorillas-photo-recognition-algorithm-ai

Predicting Recidivism

COMPAS, the algorithm used for recidivism prediction produces much higher false positive rate for black people than white people (2016).

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. *(Source: ProPublica analysis of data from Broward County, Fla.)*

What could have caused this issue?

Ranking

XING, a job platform, rank less qualified male candidates higher than more qualified female candidates.

Search query	Work experience	Education experience	Profile views	Candidate	Xing ranking
Brand Strategist	146	57	12992	male	1
Brand Strategist	327	0	4715	female	2
Brand Strategist	502	74	6978	male	3
Brand Strategist	444	56	1504	female	4
Brand Strategist	139	25	63	male	5
Brand Strategist	110	65	3479	female	6
Brand Strategist	12	73	846	male	7
Brand Strategist	99	41	3019	male	8
Brand Strategist	42	51	1359	female	9
Brand Strategist	220	102	17186	female	10

TABLE II: Top k results on www.xing.com (Jan 2017) for the job serach query "Brand Strategist".

What could have caused this issue?

Causes

- Skewed data set: training on an unrepresentative data set
  - e.g. Assignment 2, 3, 4
- Bias in human generated labels
  - e.g. Assignment 5?

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 partly based on subject's sensitive attribute

- e.g. what if XING uses *gender* as an attribute in its decision making?
- e.g. what if COMPAS uses *race* as an attribute in its decision making?

Question:

If we removed sensitive features from those models, would those models treat sensitive groups the same way?

Disparate Treatment

Model suffers from **disparate treatment** if decisions are partly based on subject's sensitive attribute

- e.g. what if XING uses *gender* as an attribute in its decision making?
- e.g. what if COMPAS uses *race* as an attribute in its decision making?

Question:

If we removed sensitive features from those models, would those models treat sensitive groups the same way?

Answer: No, other features can correlate with *gender* and *race*.

Disparate Impact

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

Question: How do we measure disparate treatment, disparate impact, and *fairness*?

Disparate Impact

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

Question: How do we measure disparate treatment, disparate impact, and *fairness*?

Answer: No real consensus

Fairness as Demographic Parity

- Acceptance rates of applications from both groups must be equal
- Also known as "independence" (why?)

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

Tradeoff

- 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

Coding

- Bias in word embeddings
- Let's jump to PyTorch!

References

[0] https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb [1] http://www.cs.toronto.edu/~madras/presentations/fairness-ml-uaig.pdf

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