Fairness in AI

Paul-Louis Pröve – ARIC Brown Bag Session - 03.02.2022

tensora

Responsible AI

Fairness

Mitigate bias in the algorithm and data

Explainability

Understanding how decisions are made

Reliability

Uncertainty of the model is realistic

Security

Protected from third parties

Privacy

Confidential user data is always protected

Accountability

Who is responsible if something goes wrong

tensora

Sources of Unfairness

Unfairness can emerge at every step of the ML pipeline

tensora

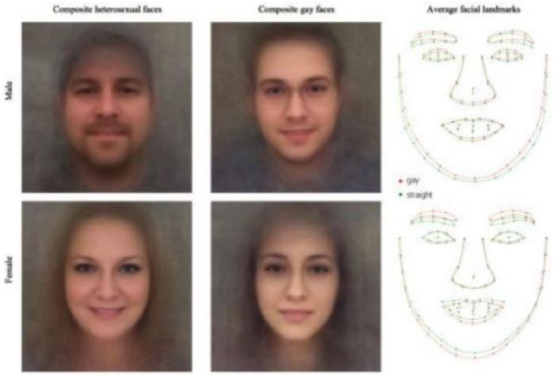
Unfairness from task definitions

Play stupid games, win stupid prizes

AI can tell from photo whether you're gay or straight

Stanford University study acertained sexuality of people on a dating site with up to 91 per cent accuracy

Fri, Sep 8, 2017, 10:14 | Updated: Fri, Sep 8, 2017, 11:11



Facial recognition to 'predict criminals' sparks row over AI bias

24 June 2020

tensora

Unfairness from data collection

Be aware of the sampling bias or it will haunt your dreams

Google Translate's gender bias pairs "he" with "hardworking" and "she" with lazy, and other examples

FROM OUR OBSESSION
Language
We explore how language helps us make sense of a changing world.

By **Nikhil Sonnad**
Reporter

Published November 29, 2017 • This article is more than 2 years old.

In the Turkish language, there is one pronoun, "o," that covers every

Boston releases Street Bump app that automatically detects potholes while driving

By DAILY MAIL REPORTER
PUBLISHED: 00:37 GMT, 21 July 2012 | **UPDATED:** 01:01 GMT, 21 July 2012

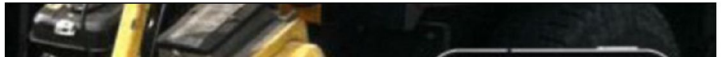
Share

3
View comments

The next time your car hits a pothole, a new technology could help you immediately tell someone who can do something about it.

Boston officials are testing an app called Street Bump that allows drivers to automatically report the road hazards to the city as soon as they hear that unfortunate 'thud,' with their smartphones doing all the work.

The app's developers say their work has already sparked interest from other cities in the U.S., Europe, Africa and elsewhere that are imagining other ways to harness the technology.



tensora

Unfairness from data labelling

When your “ground truth” is actually an opinion

More States Opting To 'Robo-Grade' Student Essays By Computer

June 30, 2018 · 8:13 AM ET
Heard on Weekend Edition Saturday

TOVIA SMITH

5-Minute Listen + PLAYLIST

arXiv.org > cs > arXiv:1901.04966

Search... All fields Search
Help | Advanced Search

Computer Science > Machine Learning

[Submitted on 15 Jan 2019]

Identifying and Correcting Label Bias in Machine Learning

Heinrich Jiang, Ofir Nachum

Datasets often contain biases which unfairly disadvantage certain groups, and classifiers trained on such datasets can inherit these biases. In this paper, we provide a mathematical formulation of how this bias can arise. We do so by assuming the existence of underlying, unknown, and unbiased labels which are overwritten by an agent who intends to provide accurate labels but may have biases against certain groups. Despite the fact that we only observe the biased labels, we are able to show that the bias may nevertheless be corrected by re-weighting the data points without changing the labels. We show, with theoretical guarantees, that training on the re-weighted dataset corresponds to training on the unobserved but unbiased labels, thus leading to an unbiased machine learning classifier. Our procedure is fast and robust and can be used with virtually any learning algorithm. We evaluate on a number of standard machine learning fairness datasets and a variety of fairness notions, finding that our method outperforms standard approaches in achieving fair classification.

Subjects: **Machine Learning (cs.LG)**; Artificial Intelligence (cs.AI); Machine Learning (stat.ML)
Cite as: arXiv:1901.04966 **[cs.LG]**
(or arXiv:1901.04966v1 **[cs.LG]** for this version)

Submission history
From: Heinrich Jiang [view email]
[v1] Tue, 15 Jan 2019 18:40:06 UTC (227 KB)

Bibliographic Tools | Code & Data | Related Papers | About arXivLabs

Download:
- PDF
- Other formats

(license)

Current browse context:
cs.LG
< prev | next >
new | recent | 1901

Change to browse by:
cs
cs.AI
stat
stat.ML

References & Citations
- NASA ADS
- Google Scholar
- Semantic Scholar

1 blog link (what is this?)

DBLP - CS Bibliography
listing | bibtex
Heinrich Jiang
Ofir Nachum

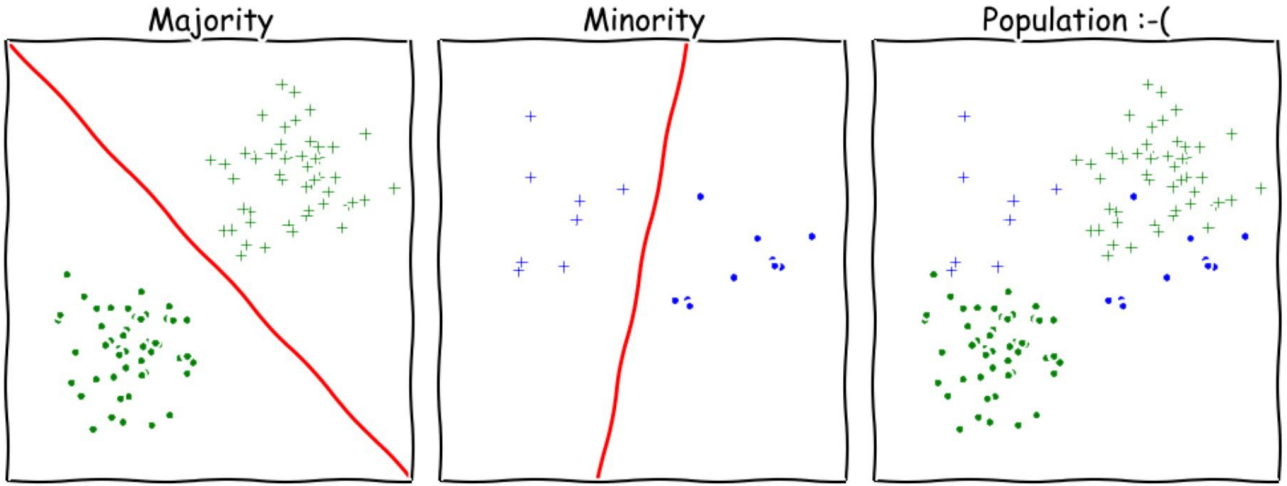
Export Bibtex Citation

Bookmark

tensora

Unfairness from the wrong model

Understand the assumptions you introduce with your model



tensora

Evaluating Fairness

Looking at fairness from the statistical perspective

Demographic Parity	Equality of Opportunity	Predictive Parity
$P(\hat{Y} = 1 \mid S = 0) = P(\hat{Y} = 1 \mid S = 1)$	$P(\hat{Y} = 1 \mid S = 0, Y = 1) = P(\hat{Y} = 1 \mid S = 1, Y = 1)$	$P(Y = 1 \mid S = 0, \hat{Y} = 1) = P(Y = 1 \mid S = 1, \hat{Y} = 1)$
Both groups have the same probability for a **positive prediction**	Both groups have the same **recall** or **sensitivity**	Both groups have the same **precision** and therefore the same FDR

tensora

S: sensitive variable, $\hat{Y}$: prediction, Y: ground truth

Example: ARIC University

2500 people apply to ARIC University

1000 people from group **A**

1500 people from group **B**

650 qualified

350 not qualified

525 qualified

975 not qualified

We only have **1100** seats ☹
Who should we accept?

tensora

Demographic Parity

We are all created equal

A n=1000	Qualified	Not Qualified
Accept	440	0
Reject	210	350

440 accepted

B n=1500	Qualified	Not Qualified
Accept	525	135
Reject	0	840

660 accepted

- The same percentage is accepted in each group (44%)
- It is completely ignored how qualified someone is
- This can cause a problem but it also may be wanted

tensora

Equality of Opportunity

If you're qualified, you'll have equal chances

A n=1000	Qualified	Not Qualified
Accept	600	20
Reject	200	180

620 accepted

B n=1500	Qualified	Not Qualified
Accept	225	255
Reject	75	945

480 accepted

- Both groups have the same probability of being accepted, given that they are qualified (75%)
- Equality of opportunity only looks at the qualified cases

tensora

Predictive Parity

Everyone we pick is equally likely to be qualified

A n=1000	Qualified	Not Qualified
Accept	600	200
Reject	50	150

800 accepted

B n=1500	Qualified	Not Qualified
Accept	225	75
Reject	300	900

300 accepted

- Out of those that were accepted, both groups have the same probability of being qualified (75%)
- The same ratio is also true for the negative cases

tensora

The slide where I tell you which metric is the best overall

tensora

Fairness in AI
Summary

- Unfairness can emerge at every step of the data science process
- Removing sensitive features is not enough
- Fairness can be expressed mathematically
- There are different types of fairness and we cannot guarantee them all at once
- Which one to use depends on the purpose

tensora

Thank You

References & Further Reading

- Novi Quadrianto – Fairness in ML Workshop
- Thomas Kehrenberg – Fairness/Accuracy Tradeoff
- ACM/IEEE 2018 – Fairness Definitions Explained
- [1609.07236] On the (im)possibility of fairness
- Cathy O'Neil – Weapons of Math Destruction

Contact

Paul-Louis Pröve
Geschäftsführer

paul-louis.proeve@tensora.co
linkedin.com/in/gopietz/

Tensora GmbH
Axel-Springer-Platz 3
20355 Hamburg
tensora.co

tensora