FAIR, PRIVACY-PRESERVING AND SECURE AI

STRATA 2019, NEW YORK CITY

MIKIO L. BRAUN
@mikiobraun

25-09-2019
AI IS STARTING TO IMPACT OUR LIVES
Training $X \rightarrow T$

\[ \downarrow \]

Model $f$

\[ \downarrow \]

Future Data $X \rightarrow f(x)$ s.t. $l(f(x), y)$ small

Accuracy, TPR, FPR,

BIAS AND FAIRNESS
RACIAL BIAS IN CRIMINAL RISK SCORES

MACHINE BIAS

Bias in Criminal Risk Scores Is Mathematically Inevitable, Researchers Say

ProPublica’s analysis of bias against black defendants in criminal risk scores has prompted research showing that the disparity can be addressed — if the algorithms focus on the fairness of outcomes.

by Julia Angwin and Jeff Larson, Dec. 30, 2016, 4:44 p.m. EST

https://www.propublica.org/article/bias-in-criminal-risk-scores-is-mathematically-inevitable-researchers-say
GENDER BIAS IN CV SCORING

Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin

SAN FRANCISCO (Reuters) - Amazon.com Inc’s (AMZN.O) machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

RACIAL BIAS IN FACE RECOGNITION

The tech industry doesn’t have a plan for dealing with bias in facial recognition

We surveyed a dozen firms to see what they were doing about this problem. Experts say: not enough

By James Vincent | Jul 26, 2018, 11:19am EDT

We’re used to looking at performance metrics, but bias adds another dimension as it matters how models perform for different parts of the data.
Sometimes we have accurate models, but there is bias in the data.

Resulting models are most likely still as biased.

Examples of systematic bias in society:

- Gender gap in salary.
- Education reached on origin.

**What you can do:** Check data for systematic bias before training.
ML algorithms usually are designed to optimize for average performance.

Under-represented areas of data can lead to bias:

- Worse accuracy
- Under / overestimation

What you can do: make sure data is representative of application areas.
As soon as predictions are influencing how you acquire data, more opportunity for systematic bias occurs.

(Harmless) example: you will get click data only for items your recommendation algorithm shows.

**What you can do:** it's complicated.
Calibrating for all groups is highly task specific and might also not be the right thing to do.

The notion of *Equal Opportunity* states that predictions must be equally accurate irrespective of protected attribute $A$. 

Pilot Parliaments Benchmark

Test suites, e.g.
http://gendershades.org/overview.html

TEST FOR BIAS
Conference on Fairness, Accountability, Transparency in ML (since 2014) [1]

Other ideas to quantify fairness:

- Statistical Parity & Task specific similarity [1]

In ML methods:

- Measure & re-balance data sets [2]
- Working directly with the data [3]

 PRIVACY-PRESERVING AI

\[ e(x \circ y) = e(x) \circ e(y) \]

homomorphic encrypt
The general problem:

Data is *key ingredient* to ML powered AI systems, but data is also *sensitive*.

Challenges:

- Unwanted reconstruction of private data
- Using private data in training without disclosing too much data
- Using ML systems without disclosing private data
It is possible to reconstruct (private) training examples by inverting models (with a bit of preprocessing).

Recommendation: quantize confidence measures to reduce ability to invert, use decision trees.

M. Frederiksen et al, Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures
Even in aggregates and other models learned from data, you might be able to recover data if you have partial information.

Differential privacy introduces noise to reduce information, to make models more independent of individual data points.

Introduce noise: in training data, cost function, weights, etc.

We don’t want to collect private data.

Federated training combines local stochastic gradient descent with secure sharing of secrets to train “on the edge” in a secure way.

Updates are further masked by random numbers that are set up and shared as part of the protocol.


HOMOMORPHIC ENCRYPTION

When using MLaaS, you might not want to send private data.

There are ways to send encrypted data such that the model ends up computing an encrypted version of the output.

Challenges:

- Encryption uses noise (salt) for security, bounds depth of computation.
- Bootstrapping can be used to clean cipher text with certain encryption schemes.

Minelli, *Fully Homomorphic Encryption for Machine Learning*,
SECURE AI
ADVERSARIAL EXAMPLES


Simple optimization problem:

Maximize prediction for some other label while restricting the perturbation.

Okay, but why?

\[
\max_{\tilde{x}} h(\tilde{y} | \tilde{x}) \quad \text{s.t.} \quad \| x - \tilde{x} \|_\infty \leq 3
\]

for some \( \tilde{y} \neq h(x) \)
HIGH DIMENSIONAL SPACES ARE WEIRD

No data in the empty spaces within!
Make systems more resilient by ensuring that they are stable against local perturbations.

Keep in mind: high dimensional spaces are weird, and don’t expect ML to work as you would think they would.

OUTLOOK
WHAT IF AI IS BECOMING TOO POWERFUL?

Predict next word based on everything written so far

"On a gray Monday morning, he was having a coffee and getting ready for the work ahead"

Radford et al, Better Language Models and Their Implications, https://openai.com/blog/better-language-models/
EU GUIDELINES ON “TRUSTWORTHY AI”

7 recommendations

1. Human agency and oversight
2. Technical robustness and safety
3. Privacy and data governance
4. Transparency
5. Diversity, non-discrimination and fairness
6. Societal and environmental well-being
7. Accountability

SUMMARY

- AI technology ultimately affects people and impacts their life.
- There is no built-in protection against bias, if it is in the data, the data is underrepresented, or closed loop uses of predictive systems. But we can test our system for bias and uniformly good accuracy.
- AI technology like ML uses data, and there are ways to restrict the individual's data point in the training, models, and application of AI.
- ML systems work differently than us, which may lead to surprising failure modes.
MIKIO BRAUN

PRINCIPAL RESEARCH ENGINEER

mikio.braun@zalando.de

25-09-2019
This presentation and its contents are strictly confidential. It may not, in whole or in part, be reproduced, redistributed, published or passed on to any other person by the recipient.

The information in this presentation has not been independently verified. No representation or warranty, express or implied, is made as to the accuracy or completeness of the presentation and the information contained herein and no reliance should be placed on such information. No responsibility is accepted for any liability for any loss howsoever arising, directly or indirectly, from this presentation or its contents.