The right way to do the wrong thing (with ML)?

Zachary C. Lipton
Carnegie Mellon University & Amazon AI
https://arxiv.org/abs/1711.07076

Email: zlipton@cmu.edu
Shenanigans: @zacharylipton
Outline

• Philosophical considerations:
  • What sort of problem is fairness in ML?
  • What sort of inquiry is required to address it?

• Disparate learning processes
  • Treatment and impact parity
  • Reconciliation with utility

• Does mitigating ML’s impact disparity require treatment disparity?
  • Summarizing some simple results
  • Empirical

• Takeaways & discussion points
Decisions

• We all make decisions frequently
• What to eat, drive, house to buy...?
• Who to hire, lend to, release from jail....?
Fairness

• Some (esp housing, hiring, lending) loaded with ethical significance
• Decisions impact lives of those subject to them
• The very purpose of the law is to guide decisions
• Factors: intent of decision-maker, impact of actions
Anti-discrimination law

“President Lyndon B. Johnson shakes hands with Martin Luther King after signing the Civil Rights Act of 1964”
Disparate treatment

• Addresses intentional discrimination
• Includes decisions explicitly based on a *protected characteristic*
• Also intentional discrimination via proxy variables
Disparate impact

- Facially neutral practices that might nevertheless have an “unjustified adverse impact on members of a protected class”

- Complicated doctrine w 3 tests
  1. Plaintiff must demonstrate statistical disparity (e.g. 4/5 rule)
  2. Defendant must show that decisions are justified by ‘business necessity’
  3. Plaintiff must show defendant can achieve goal w ‘alternative practice’
Classic case: **Griggs vs. Duke Power**

- Decided in 1971, considered 1st case of its type
- Duke required high school diploma for higher-paid jobs
- Supreme Court: Duke must show tests "*reasonably related*" to job
Algorithmic decisions
Supervised learning

ML Algo
Choose model parameters that minimize loss function

Outputs (y)

Learned Model $f^*$

Features (x)

Features

Targets
Neglects aspects of ethical decision-making

- Short and long-term impacts
- Participation in a system of incentives
- How data came to be
- How it might change in the future
- Consequences
- Justification
Evolution of scientific disciplines
Pre-paradigmatic science

- Immature stage of inquiry
- Theory consists of philosophy
- Try random things
Normal science

• Takes place upon *acquisition of a paradigm*
• Works within a rigid set of assumptions
• Exhibits a *peculiar efficiency*
• Tends towards *puzzle solving*
Paradigm Shift

• Normal science encounters problems it cannot solve / anomalies it cannot explain

• Paradigm shifts are the “tradition-shattering complement to the tradition-bound activities of normal science”
Different problems, different tools

• Ethics remains pre-paradigmatic
  • Generally the province philosophers, legal scholars, etc.

• MLs has aspects of *normal science* (to extent it’s science at all)
  • Solves prediction problems and theory thereof
    w established set of concepts and technical tools

• Are ethical problems remediable with the tools of ML?
Basic Setup / Context
Treatment parity
Impact Parity

Model

Outputs (y)

Features (x)

Parity condition

Sensitive Feature (z)

Labels (t)
Make groups equal but how?

• Impact parity
  • \( p(y|z=0) = p(y|z=1) \)

• Treatment parity
  • The output \( y \) depends only on \( x \), not on \( z \)

• Representational parity
  • Map \( x \) to \( r(x) \) such that \( r(x) \) ind. of \( z \) (indistinguishable representations)
  • Entails impact parity

• Opportunity parity
  • False positive and false negative rates match
Problems

• If all groups are the same in every way, easy
• Otherwise mutually irreconcilable
• None capture common legal or philosophical notions
• None requires or enforces decisions be justified
• The common notions of fairness lie entirely outside the formalism of supervised learning / statistical predictive modeling
Technical vs legal (vs ethics) terminology
Does Mitigating ML’s Impact Disparity Require Treatment Disparity?

Julian McAuley

Alexandra Chouldechova

Disparate Learning Processes

Model

Features (x)

Outputs (y)

Loss

Sensitive Feature (z)

Labels (t)
Examples of DLPs

• Optimization-based
  • Numerous papers explore adding constraints ML optimization
  • Others use same idea but with regularizers

• Representation-based
  • Probabilistic mappings to fair representations
  • GAN-based learning setups
Findings

1. For reconciling impact disparity and treatment disparity, **treatment disparity is optimal** (theoretical)

2. When $x$ fully encodes $z$, for sufficiently powerful model, **DLP indistinguishable from treatment disparity** (theoretical)

3. When $x$ partially encodes $z$, DLP results in side effects (empirical)
   A. Re-orders within-group based on otherwise irrelevant characteristics
   B. Produces potentially bizarre incentive to conform to stereotype
Toy example
Case study: Gender bias in CS admissions

- **Dataset:** sample of ~9,000 students considered for admission to the MS program of a large US university over an 11-year period
- **Labels:** admissions decisions provided by a faculty admissions committee
- **Attributes:** Gender the protected attribute. Country of origin, interest area, and GRE, etc. are used as features
- **Synthetic discrimination:** applied to mimic biased training data: of all women who were admitted, we flip 25% of their labels to 0
Effects of DLP in CS admissions
Takeaways
What to do, not how to do it?

• A massive body of papers addresses hacks for how to reconcile utility (accuracy) and impact parity
• But how to do this is simple and solves (thresholding)
• Algorithmic complication contrived to make technical papers (working with incorrect definition of disparate treatment)
• What comes of this puzzle solving?
• The real question is what outcome to effect in a given context
Dynamics & equilibrium effects

• These problems consider going beyond static view of classification
• Must consider impact of policies on real-world dynamics
• Subsequent data gathered, incentives created etc.
A role for causality?

• Maybe! (in some sense)
• Requires strong assumptions
• Need to model the right entities
• Observational data cannot (always) reveal the process
When can we use wrong tools (e.g. ML)

• In many contexts ML systems, however screwed up, may be more egalitarian than human decisions

• Do we cause harm by denying application of ML (even when incorrect)?