Debunking the myth of the fairness-accuracy tradeoff

Joshua Loftus (loftus at nyu.edu)

New York University

10/20/2019

NYU | STERN

A common misconception

Theorem

Unconstrained optimization yields more extreme (better) optimal values of the objective

Theorem

Unconstrained optimization yields more extreme (better) optimal values of the objective

This doesn't apply to fairness

If we are using a constraint to achieve fairness, then applying this theorem is (probably) misleading at best

Fairness is a real-world, inherently complex, noisy, and contested objective. Its inclusion means we are likely in a setting where taking accuracy at face value is naive (see e.g. Goodhart's/Campbell's law)

Outline: a high level argument and three types of examples

1. The two cultures (redux (redux . . .))
2. From accuracy to utility
3. Utility under fairness constraints

This talk: my own ramblings, but inspired by

- *Counterfactual Fairness* (Kusner et al, NeurIPS 2017)
- *Causal Reasoning for Algorithmic Fairness* (Loftus et al, preprint 2018)
- *Making Decisions that Reduce Discriminatory Impacts* (Kusner et al, ICML 2019)
- Forthcoming work with student Margarita Boyarskaya

Matt Kusner[2,3] Chris Russell[1,3] Ricardo Silva[2,3]



(me) [1]Surrey [2]UCL [3]Alan Turing Institute

- Statistical modeling: The two cultures (Breiman, 2001)
- Computer Age Statistical Inference (Efron & Haste, 2016, book)
- Single objective: minimize loss function, compute a solution
- Multiple objectives: interpret parameters, quantify uncertainty (intervals / hypothesis tests), diagnose model inadequacy – tends to move slower as a result

- Statistical modeling: The two cultures (Breiman, 2001)
- Computer Age Statistical Inference (Efron & Haste, 2016, book)
- Single objective: minimize loss function, compute a solution
- Multiple objectives: interpret parameters, quantify uncertainty (intervals / hypothesis tests), diagnose model inadequacy – tends to move slower as a result

I'm a statistician, mostly interested in inference, particularly in settings where bias is a main concern (high-dimensions, post-model selection, causal inference, fairness)

- Inputs: data $\mathcal{D}$, a function class $\mathcal{H}$, and a loss function $\ell : \mathcal{H} \times \mathcal{D} \to \mathbb{R}$
- Algorithm uses training data $\mathcal{D}_{\text{train}}$ to choose $f \in \mathcal{H}$ with minimal loss on the test data $\ell(f, \mathcal{D}_{\text{test}})$
- Mathematically well-posed, can leverage academically mature literature on optimization
- Straightforward: you know what to do, even if it might be challenging to do it
- Essentially only one objective or goal–the lowest test error achievable

- Inputs: algorithms (previous slide or some variant) plus questions (about the "real world," e.g. is the algorithm fair with respect to some sensitive attribute?)
- Potentially many objectives (e.g. intervals, interpretability)
- May require untestable assumptions or not be well-posed, statistics is younger than optimization and connected to unresolved philosophical questions

"Since all models are wrong the scientist must be alert to what is importantly wrong" - George Box

"Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise" - John Tukey

"There is nothing so useless as doing efficiently that which should not be done at all" - Peter Drucker

- Combine multiple objectives and complicated sets of preferences into one utility function
- Unobservable, a conceptual framework rather than a well-posed loss function like ℓ
- Inputs: algorithms, actions/decisions
- Whose utility? (next slide)

Supervised learning intuition: outcome variables $\mathbf{Y}$ are almost always proxies for things we really care about, "utility" refers to those things

- **Z** decision variable, with $Z_i = 1$ being a desired outcome for individual i
- u_i for $i \geq 1$ utilities of individuals receiving a prediction/decision, so $u_i(1) > u_i(0)$
- Aggregate utilities for the **sample** above u_s ("customer satisfaction"), for the **population** u_p ("society"), and for the **organization**/decision-maker deploying the algorithm u_o (assume preference for algorithm accuracy)

Thesis

In algorithmic fairness settings, some/all of u_o, u_s, or u_p may take larger values when the algorithm is constrained to be fair

The rest of this talk provides examples

Tradeoff utilities, not accuracy

If people care about others, being fair increases utility!

DIOS Theorem (Diversity Is Our Strength)

Constraining the algorithm to be fair may yield higher values of

1. u_{p} if society cares about fairness
2. u_{s} if customers care about fairness
3. u_o if the organization cares about fairness (in addition to accuracy)

DIOS Theorem (Diversity Is Our Strength)

Constraining the algorithm to be fair may yield higher values of

1. u_{p} if society cares about fairness
2. u_{s} if customers care about fairness
3. u_o if the organization cares about fairness (in addition to accuracy)

Positive reinforcement

If for $x, y \in \{p, s, o\}$, we have

$$\textbf{Cooperation} : u_x \text{ increases if } u_y \text{ increases}$$

and if u_y cares about fairness, then u_x may be larger when constraining the algorithm

e.g. Inclusive organizations

Disadvantaged people may benefit more from a positive decision

Let **A** represent a sensitive attribute with $A_i = 1$ corresponding to a more privileged class, then it may be reasonable to assume a positive prediction/decision yields a greater benefit to individuals who are less privileged (similar to **diminishing marginal utility**). Recall $D = 1$ is the desirable decision: $u_i(1) > u_i(0)$.

$$\textbf{SMTE} : u_i(1) > u_j(1) \quad \text{when} \quad A_i = 0 \text{ and } A_j = 1$$

Let **A** represent a sensitive attribute with $A_i = 1$ corresponding to a more privileged class, then it may be reasonable to assume a positive prediction/decision yields a greater benefit to individuals who are less privileged (similar to **diminishing marginal utility**). Recall $D = 1$ is the desirable decision: $u_i(1) > u_i(0)$.

$$\textbf{SMTE} : u_i(1) > u_j(1) \quad \text{when} \quad A_i = 0 \text{ and } A_j = 1$$

SMTE Theorem (Strictly Monotonic Treatment Effects)

Under **SMTE**, constraining the algorithm may result in larger values of

1. u_s 2. u_p and u_o if **Cooperation** also holds

e.g. Selective school admissions

In fairness settings, it is often the case that the sampling mechanism is biased in a way that correlates with **A**.

CBS Theorem (Correcting Biased Sampling)

If the sampling mechanism is unfair, imposing fairness constraints on the algorithm can correct sampling bias and result in higher values of u_p

e.g. Feedback loops in policing. Lum & Isaac (2016), Ensign et al (2017)

Interference occurs if the decision for one individual (causally) influences the outcomes or utility of another individual:

Interference : u_i depends on D_j for some $j \neq i$

CMI Theorem (Critical Mass Interference)

For some types of **Interference**, fairness constraints may result in a **critical mass** of positive decisions for a disadvanted group yielding large increase in u_s

PEI Theorem (Positive Externality Interference)

For some types of **Interference**, fairness constraints may result in **positive externalities**

e.g. Sufficient investment in a disadvantaged community

Parting thoughts

- Utility is what matters and it could reasonably be higher with fairness constraints than without
- Revealed preference + the status quo seem to imply decision-makers don't care about fairness... (Personally, I don't buy revealed preference)
- Cooperation makes it easier to be fair, **competition** is an obstacle
- Don't take accuracy too seriously. Goodhart's law (1975) [also: Lucas critique (1976), Campbell's law (1979)]

  Any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes