Algorithms & the mathematics of fairness

1. Background & Motivation

COMPAS (Northpointe / Equivant) - predict recicidivism (probability of reoffending)
correctional offender management profiling for alternative sanctions

- gives an offender three scores 1-10
  - ↳ general recidivism risk
  - ↳ violent recidivism risk
  - ↳ failure to appear
- designed for pretrial screening / probation

2016 ProPublica article : Machine Bias ProPublica - non-profit investigative journalism
There's a software used across the country to predict future criminals. And it's bias against blacks.

studied COMPAS results & subsequent recidivism rates of 10000 pretrial criminal defendants from Broward Country, Florida.

Results : • Black offenders more likely to be labelled high risk.

- accuracy for both races about the same 0.68 - 2009 study
- $\frac{FP}{FP+TN}$ — false positive rate for Black defendants higher B 0.45 W 0.23
- $\frac{FN}{FN+TP}$ — false negative rate for White defendants lower B 0.28 W 0.48

Influences prison sentencing Paul Zilly 2 Years ⟶ 18 months
see some cases on ProPublica article.

2017 Flores, Lowenkamp, Bechtel - Rejoinder

- Flores ↓ professor at California State uni
- Lowenkamp ↓ admin office united states courts
- Bechtel ↘ non profit in justice sector

5 contentions. Including :

AUC-ROC scores

Claim: differences in mean score between races is not bias.
In particular (simplifying) PPV similar across races

2. Impossibility of Fairness

Suppose gps A_1, A_2 with $P(x \in R \mid x \in A_i) = p_i$ $i = 1, 2$
Want to predict if $x \in R$. Label H or L.
Suppose

$P(x \in R \mid x \in H) = q_1$ PPV positive predictive value
$P(x \notin R \mid x \in L) = q_2$ NPV negative predictive value

Want: same across groups.

Calculate

(TPR =) $P(x \in H \mid x \in R, x \in A_i) = q_1 \frac{1}{p_i} \left(\frac{p_i + q_2 - 1}{q_1 + q_2 - 1} \right)$

FNR = 1 - TPR

lack of statistical parity

Note if $q_2 \neq 1$ & $p_1 > p_2$ then $TPR_2 > TPR_1 \Rightarrow FNR_1 > FNR_2$

not perfect predictor

Similarly for FPR, & discrepency is cts in $q_1, q_2, p_1 - p_2$

In the US, recividism rates for Black offenders is higher than that of White people.

probublica 51%
39%

'18 study U.S. department of Justice
Update on prisoner recidivism
A 9-year follow up period
9 years B 87%
W + Hispanic 81%

Q: Are you happy with this?

- Does fairness mean statistical parity?
- What is the lack of statistical parity telling you?

Note: (i) address / income / highest levels of education are all proxies for race and predictors of certain crimes.

(ii) reoffending rates ≠ rearrest rates

Terminology demystifying algorithmic fairness

- Separation people of same real life outcomes should get treated similarly by algorithm, regardless of what group they are in

FPR / FNR equal across gps (1 - FPR called positive recall)
FNR negative

training algorithm optimally on outcome of interest will likely cause it to not satisfy separation

- Sufficiency risk scores should indicate same real life outcome, regardless of group

Precision $\left(\frac{TP}{TP+FP}, \frac{TN}{TN+FN} \right)$ same across groups

usually fine.

- Independence A given score should be equally likely across groups

$P(H \mid A_i) = P(H \mid A_j) \quad \forall i,j$

Example. (of failure) advertising depending on post / ZIP codes in a way that targets / excludes certain races. Redlining

2021 study Chang et al.
college scholarship adds in NY.

We have shown

more analysis on disparate impacts

Thm. (Impossibility of fairness, Chouldechova '16) If distribution of outcomes unequal and you do not have a perfect predictor, then it is not possible to satisfy both sufficiency & separation.

Moreover, approximate fairness can only simultaneously hold under ε-approximate equal base rates or ε-approximate perfect performance.

Algorithmic decision making & the cost of fairness, Corbett-Davis et al 2017

Question: what do you care about?
algorithmic accuracy or individuals being falsely labelled?

3. What now?

We live in a society. Can't just go 'oh no its impossible & hide in a hole'

- Fix the problem e.g. institutionalised racism
- Do you need the algorithm?
  - ↳ if we know there will be problems, should we implement?
  - ↳ no algorithm not strictly better than yes algorithm
  - ↳ things can be intractable to do w/o algorithms e.g. google
- Throw something out
  - ↳ what is important depends on context
- Compromise
  Pushing the limits of fairness impossibility: who's the fairest of them all
  Hsu et al '22
  - ↳ turn it into an optimisation problem
- Accept that different metrics measure different things that contribute to fairness
  Instead of 'is this algorithm fair?' ask 'in what way is this algorithm fair?'

  Different metrics can be used to inform decision making & catch when something is going wrong.
  - ↳ COMPAS tells you what is going wrong
  - ↳ could be used to target interventions

⚠ This is not a maths problem.
We (humanity) are far from a consensus on what the most basic definitions of fairness, equality & discrimination are. Fairness in Machine learning: Lessons from political philosophy

We studied binary classifier & two groups - important case
But many other types of algorithms exist. There has been lots of work on developing metrics for fairness & studying how they behave / implementing in practice

Examples. 1. "The outcome shouldn't depend on whether you are in a protected gp"
↑ race, gender, disability.

ordinary least squares regression
Do OLS regression to the data { w/ characteristic
w/o
are compare

independence

2. "Treat similar individuals similarly" ⤳ Lipschitz condition
classifier $x \longmapsto M(x)$
Require $d(M(x), M(y) < d(x,y)$ $\forall x, y \in V$

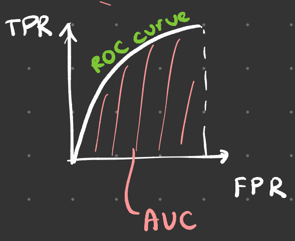
Dwork et al '11 Fairness through Awareness (how to add this constraint)

separation

What is similar?

3. "Predictor should be equally good for different gps'
Standard test for classifiers ROC/AUC score.
with sliding scale

compare for different gps. Sufficiency



Survey: Zliobaite '17

4. Closing Remarks & Open Questions

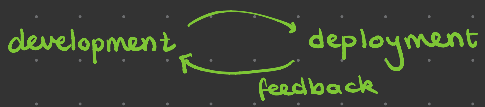
- There is no such thing as a perfectly fair algorithm

- Context is always important. Cannot discuss COMPAS & its issues fully without understanding the US criminal justice system, institutionalised racism, & its effects on marginalised communities
↳ garnered attention during Black Lives Matter

- Different metrics measure different things
What is fair depends on what you care about

- Beaware of feedback loops

Open Questions
- How do we ensure that algorithms are fair



- What responsibility do developers have in understanding & conveying the limitations of the tools they create?

- What does fairness mean to you?