STAT 302: Quantitative Fairness

Sheridan Grant

University of Washington Statistics Department

slgstats@uw.edu

June 1, 2020

Types of Discrimination

Direct Discrimination

Algorithm 1 Hiring algorithm that discriminates directly

```
Require: person
  if person$race == 'white' then
    return True
  else
    return False
  end if
```

Direct Discrimination

Algorithm 2 Hiring algorithm that discriminates directly

```
Require: person
  if person$race == 'white' then
    return True
  else
    return False
  end if
```

Such algorithms are more often found in humans than programmed into computers, and their problems are pretty easy to diagnose.

Proxy Discrimination

Algorithm 3 Interview algorithm that discriminates by proxy

```
Require: resume
  if soundsWhite(resume$name) then
    return True
  else
    return False
  end if
```

Proxy Discrimination

Algorithm 4 Interview algorithm that discriminates by proxy

```
Require: resume
  if soundsWhite(resume$name) then
    return True
  else
    return False
  end if
```

Oversimplified, but happens in the real world (Bertrand and Mullainathan 2003)

More Proxy Discrimination

Redlining

- A *proxy* is a variable strongly associated with the *sensitive attribute* (race, gender, nationality, etc.) but substantively different from it.
- In the 1930s, federal mortgage loans were denied to people in "high-risk" zip codes, which were disproportionately black.
- Sensitive attribute: race
- Proxy: zip code
- WaPo article on Redlining's impacts today

More Proxy Discrimination

Redlining

- A *proxy* is a variable strongly associated with the *sensitive attribute* (race, gender, nationality, etc.) but substantively different from it.
- In the 1930s, federal mortgage loans were denied to people in "high-risk" zip codes, which were disproportionately black.
- Sensitive attribute: race
- Proxy: zip code
- WaPo article on Redlining's impacts today

College Admissions

- Students for Fair Admissions sued Harvard, arguing admissions process discriminated against Asians
- One complaint: measures of "well-roundedness" and "uniqueness" were racially biased, downweighting academic credentials.
- "Unique extracurriculars" were allegedly a proxy for non-Asian
- Supreme Court ruled in favor of Harvard

Preventing Discrimination

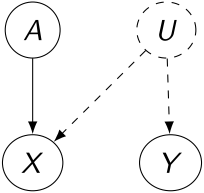
Fairness through Awareness

Fairness through *Unawareness* is a naive but initially appealing approach: simply don't consider race when making decisions (human OR algorithmic).

Previous slides makes clear this doesn't work. Often, the only fair thing to do is to explicitly consider race: Fairness through *Awareness*.

Fairness through Awareness

Fairness through *Unawareness* is a naive but initially appealing approach: simply don't consider race when making decisions (human OR algorithmic).
Previous slides makes clear this doesn't work. Often, the only fair thing to do is to explicitly consider race: Fairness through *Awareness*. Example: Y = accident rate; X = color of car (1 if red); A = race (1 if black).



Fairness through Awareness

Unfair approaches:

- Model relationship between car color and accidents, charge black people more even though race doesn't affect accident risk

Fair approaches:

Fairness through Awareness

Unfair approaches:

- Model relationship between car color and accidents, charge black people more even though race doesn't affect accident risk

Fair approaches:

- Randomly price insurance (lose money)

Fairness through Awareness

Unfair approaches:

- Model relationship between car color and accidents, charge black people more even though race doesn't affect accident risk

Fair approaches:

- Randomly price insurance (lose money)
- Model relationship between accidents and race, find none, randomly price, lose money

Fairness through Awareness

Unfair approaches:

- Model relationship between car color and accidents, charge black people more even though race doesn't affect accident risk

Fair approaches:

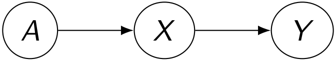
- Randomly price insurance (lose money)
- Model relationship between accidents and race, find none, randomly price, lose money
- Model relationship between accidents vs. race AND car color, charge red cars more, give black people fair "discount" that accounts for association with accident-prone trait

Muddying the Waters

Fairness through Awareness doesn't always work. Consider a job where qualifications are directly related to race. For example, Hispanic people may tend to be better-qualified to be Spanish translators than non-Hispanics. Then accounting for race in hiring, so that Hispanics and non-Hispanics are hired at the same rates, is clearly discriminatory towards Hispanics.

Muddying the Waters

Fairness through Awareness doesn't always work. Consider a job where qualifications are directly related to race. For example, Hispanic people may tend to be better-qualified to be Spanish translators than non-Hispanics. Then accounting for race in hiring, so that Hispanics and non-Hispanics are hired at the same rates, is clearly discriminatory towards Hispanics.



Muddying the Waters: Thresholds

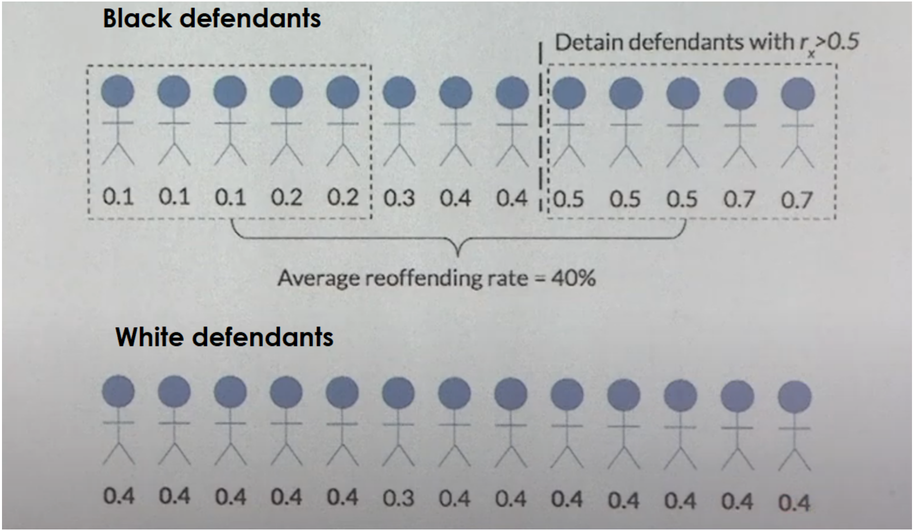
- *Recidivism*: criminal re-offense
- *Posting bail*: when a judge allows a defendant to leave jail until trial if they pay a large sum of money
- *Parole*: when you're let out of jail early because they think you'll behave well

Muddying the Waters: Thresholds

- *Recidivism*: criminal re-offense
- *Posting bail*: when a judge allows a defendant to leave jail until trial if they pay a large sum of money
- *Parole*: when you're let out of jail early because they think you'll behave well

Ignoring the underlying flaws in the system for the moment, we want to let as many people who are unlikely to recidivate out on bail and parole as possible. How should we do this? Statistical model! Estimate recidivism probability from covariates (age, prior offenses, type of offense, mitigating factors, **race???**, **gender???**)

Muddying the Waters: Prediction Rules



Credit: Dr. Sharad Goel, Simons Institute Workshop on Fairness

Human vs. Machine

If the idea of an algorithm determining liberty makes you uncomfortable, you're not alone. What happens in reality is the algs make *recommendations* to judges, who make the final parole/bail decision.

Human vs. Machine

If the idea of an algorithm determining liberty makes you uncomfortable, you're not alone. What happens in reality is the algs make *recommendations* to judges, who make the final parole/bail decision.

...But are the human judges any better?

Human vs. Machine

If the idea of an algorithm determining liberty makes you uncomfortable, you're not alone. What happens in reality is the algs make *recommendations* to judges, who make the final parole/bail decision.

...But are the human judges any better?

Why not both? In the Quora example from last Wednesday, I said our Troll prediction algorithm would flag questions *for human review*.

Human in the Loop

Disadvantages

- Expensive
- Slow
- Human bias (judges may discriminate consciously or unconsciously)
- Human judgment errors (humans make calculation mistakes, computers don't)

Advantages

- **Transparent.** Best prediction algorithms are incomprehensible to humans, and we'd rather know why the decisions are being made the way they are
- Machines can temper human biases and vice-versa (judges extend leniency to young defendants, incorporate other circumstantial info that algs don't; **stevenson˙algorithmic˙2019**)

Other Concerns

Biased Labels:

- In the previous example's training data, what is the label we have? What is the label we *want*?

Other Concerns

Biased Labels:

- In the previous example's training data, what is the label we have? What is the label we *want*?
- We have whether or not someone was *convicted* of a re-offense, not whether or not they *actually re-offended*.
- What would this imply if certain groups were more likely to be caught committing crimes than others, or more likely to be falsely convicted?

Other Concerns

Biased Labels:

- In the previous example's training data, what is the label we have? What is the label we *want*?
- We have whether or not someone was *convicted* of a re-offense, not whether or not they *actually re-offended*.
- What would this imply if certain groups were more likely to be caught committing crimes than others, or more likely to be falsely convicted?

Feedback Loops:

- Suppose my algorithm dispatches more police where more crimes are found. What will happen over time?

Other Concerns

Biased Labels:

- In the previous example's training data, what is the label we have? What is the label we *want*?
- We have whether or not someone was *convicted* of a re-offense, not whether or not they *actually re-offended*.
- What would this imply if certain groups were more likely to be caught committing crimes than others, or more likely to be falsely convicted?

Feedback Loops:

- Suppose my algorithm dispatches more police where more crimes are found. What will happen over time?
- Amazon trained a recruiting algorithm on historical data. But historically, hiring has been sexist, so the algorithm simply learned to down-rate women from the historical data.

Garbage in—garbage out.

Quantitative Fairness in Affirmative Action

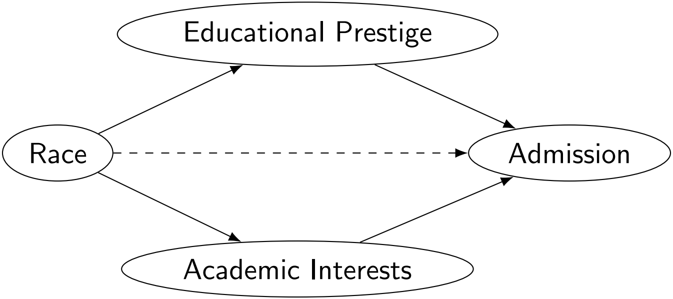
QF gives us a framework for analyzing fairness and discrimination. This framework organizes passionate perspectives dispassionately. Consider two opposing perspectives on affirmative action:

- Fairness through Unawareness (do not use race in admissions decision): strongest anti-affirmative action stance.
- "Statistical Parity": universities should admit all groups at equal rates (strong pro-affirmative action stance).

These two stances are irreconcilable, but using the quantitative fairness framework yields a way forward.

Quantitative Fairness in Affirmative Action

Consider the following diagram:


Educational Prestige
Race
Admission
Academic Interests