ORIE 4741: Learning with Big Messy Data

Fairness

Professor Udell
Operations Research and Information Engineering
Cornell

September 13, 2021
Q: Why should algorithm designers think about fairness?

Fairness in big data

- Algorithms bias available information (search, recommendations, social media)
- Algorithms can have big impacts (parole, credit)
- Avoid unintended (and often unobservable) negative consequences

Important questions for algorithm designers:
- What is the harm of false positives? False negatives?
- How can errors change the data distribution in the future?
Q: Why should algorithm designers think about fairness?

- algorithms bias available information (search, recommendations, social media)
- algorithms can have big impacts (parole, credit)
- avoid unintended (and often unobservable) negative consequences
- legal requirements
**Fairness in big data**

**Q:** Why should algorithm designers think about fairness?

- algorithms bias available information (search, recommendations, social media)
- algorithms can have big impacts (parole, credit)
- avoid unintended (and often unobservable) negative consequences
- legal requirements

Important questions for algorithm designers:

- What is the harm of false positives? False negatives?
- How can errors change the data distribution in the future?
Examples

- Credit decisioning: HDMA https://www.consumerfinance.gov/data-research/hmda/explore, apple credit card
- Criminal justice: COMPAS https://github.com/propublica/compas-analysis/
- Advertising (eg, job openings)
- Hiring https://www.businessinsider.com/amazon-built-ai-to-hire-people-discriminated-against-
- Information feeds and recommender systems
- College admissions
- Medical diagnosis or treatment recommendation

Important questions for algorithm designers:

- What is the harm of false positives? False negatives?
- How can errors change the data distribution in the future?
Fairness: definitions

Fairness vs discrimination

**Q:** What does it mean for a classifier to be unfair?

- What groups or individuals should be protected from discrimination?
- How can we tell if the algorithm is unfair or discriminatory?
- What information is it permissible for the algorithm to use?
**Fairness: legal definitions**

**Housing:** The Equal Housing Opportunity Act states that someone seeking to rent a home has the right to expect that housing will be available to them without discrimination or other limitations based on race, sex, color, religion, sex, disability, familial status, or nationality.
Fairness: legal definitions

**Housing:** The Equal Housing Opportunity Act states that someone seeking to rent a home has the right to expect that housing will be available to them without discrimination or other limitations based on race, sex, color, religion, sex, disability, familial status, or nationality.

**Credit and banking:** The Equal Credit Opportunity Act makes discrimination unlawful in credit applications based on race, color, religion, national origin, sex, marital status, age, or because all or part of the applicant’s income comes from a public assistance program.
Fairness: legal definitions

**Housing:** The Equal Housing Opportunity Act states that someone seeking to rent a home has the right to expect that housing will be available to them without discrimination or other limitations based on race, sex, color, religion, sex, disability, familial status, or nationality.

**Credit and banking:** The Equal Credit Opportunity Act makes discrimination unlawful in credit applications based on race, color, religion, national origin, sex, marital status, age, or because all or part of the applicant’s income comes from a public assistance program.

**Labor market:** The Equal Employment Opportunity Act prohibits employment discrimination in its programs based on race, color, national origin, sex, religion, age, disability, political beliefs, and marital or familial status.
Definitions and notation

**Definition**

A **protected attribute** is a feature on which discrimination is prohibited (e.g., race, color, national origin, sex, religion, age, disability, political beliefs, and marital or familial status).

**Notation:** (assume classification problem)

- **binary outcomes** \( y \in \mathcal{Y} = \{0, 1\} \)
- **covariates** \( x \in \mathbb{R}^d \)
- **binary protected attribute** \( a \in \{0, 1\} \)
- **prediction** \( \hat{y} \)
A classifier is unaware of the protected attribute if the prediction is independent of the protected attribute given other covariates:

\[ \hat{y} \perp a | x \quad \iff \quad \mathbb{P}(\hat{y} | x, a = 0) = \mathbb{P}(\hat{y} | x, a = 1). \]