Data Ethics

Jinghua

Mentor: Bryan

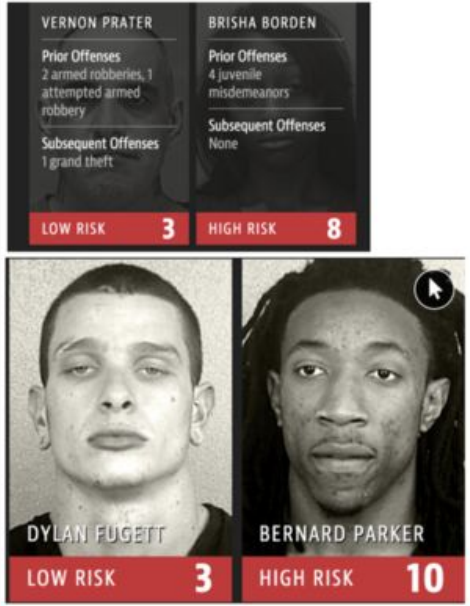
Outline

- Causality and fairness
- Algorithmic discrimination
- Fairness Penalization
  - General framework
  - Introduction to machine learning
  - Case study: *predictive policing*

Why does data ethics matter to us?

Some Examples...

Amazon scraps secret AI recruiting tool that showed bias against women



Causality and Fairness

The common use of counter-factual fairness:

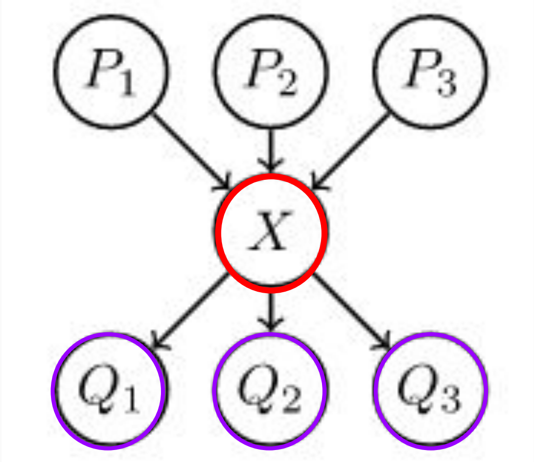
- Employment decisions
- College admissions

Race *(leaving all other attributes constant)*	Probability to be granted an offer
Asian	25%
White	36%
Hispanic	77%
African	95%

Causality and Fairness

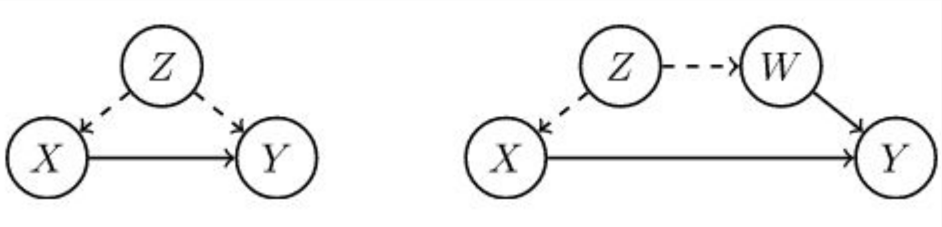
→ Attribute Flipping

Not always!!!

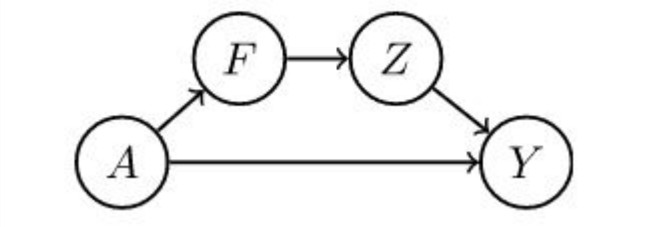


Causality and Fairness

Where problems could arise from within a causal model itself...



Unobserved Confounding



Indirect Paths

Algorithmic Discrimination

Algorithmic bias with previous examples

<u>Sources of bias:</u> In automated decision-making, such as the use of “COMPAS” in the U.S. court and Amazon’s AI recruiting tool, such algorithms run the risk of replicating or even amplifying human bias.

General Models

Fairness Penalization

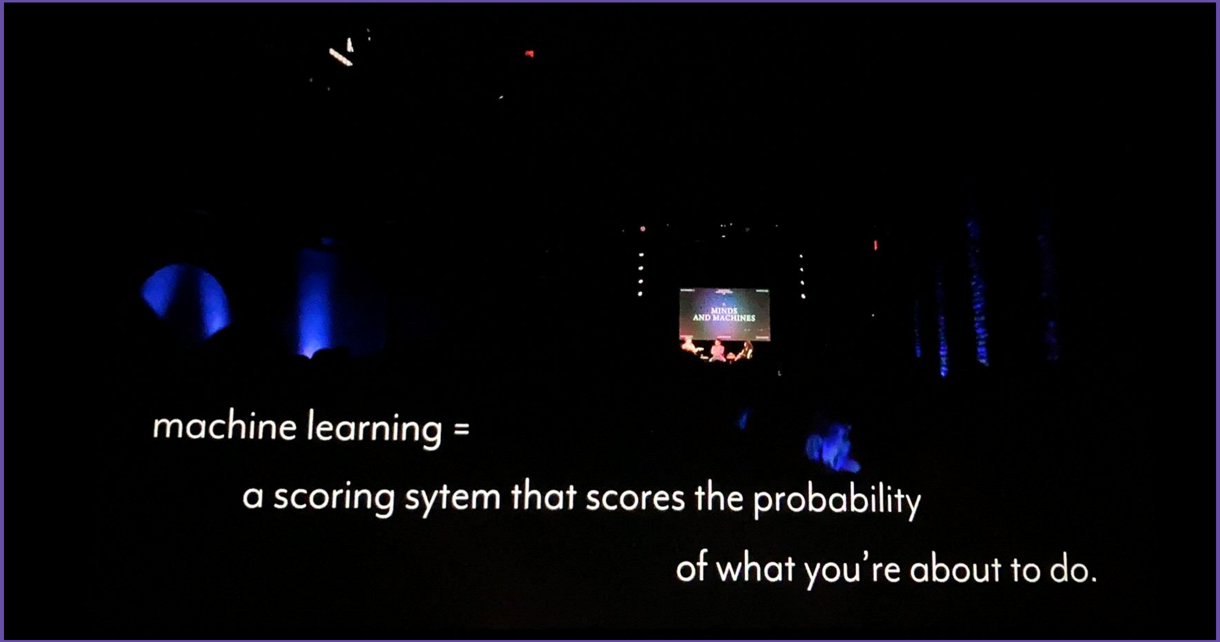
Individual Fairness

$$f_1(\mathbf{w}, S) = \frac{1}{n_1 n_2} \sum_{\substack{(\mathbf{x}_i, y_i) \in S_1 \\ (\mathbf{x}_j, y_j) \in S_2}} d(y_i, y_j)\left(\mathbf{w} \cdot \mathbf{x}_i - \mathbf{w} \cdot \mathbf{x}_j\right)^2$$

Group Fairness

$$f_2(\mathbf{w}, S) = \left(\frac{1}{n_1 n_2} \sum_{\substack{(\mathbf{x}_i, y_i) \in S_1 \\ (\mathbf{x}_j, y_j) \in S_2}} d(y_i, y_j)\left(\mathbf{w} \cdot \mathbf{x}_i - \mathbf{w} \cdot \mathbf{x}_j\right) \right)^2$$

A Brief Introduction to Machine Learning



The Lasso Regression

Lasso

$$(1) \qquad \boxed{\min_{\beta\in\mathbb{R}^p} \frac{1}{2N}||y-\mathbf{X}\beta||_2^2} \qquad \boxed{\text{s.t. } ||\beta||_1 \leq t}$$

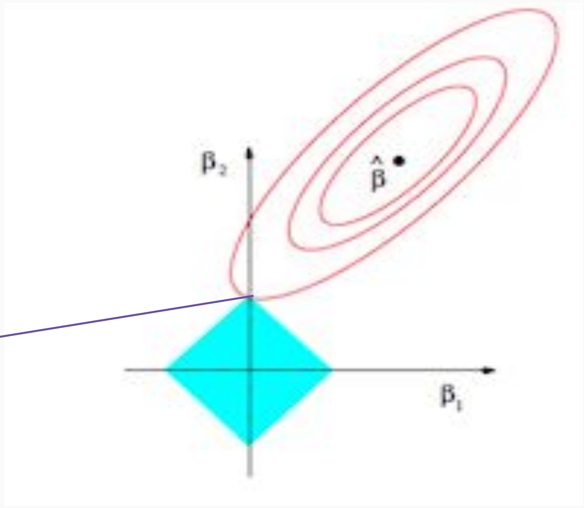
$$(2) \qquad \boxed{\min_{\beta\in\mathbb{R}^p} \frac{1}{2N}||y-\mathbf{X}\beta||_2^2} + \boxed{\lambda \cdot ||\beta||_1}$$

$y \in \mathbb{R}^N$, design matrix $\mathbf{X} \in R^{N\times p}$

Components of the Lasso:

- RSS
- Penalizing term
  - L1 norm

Point of intersection

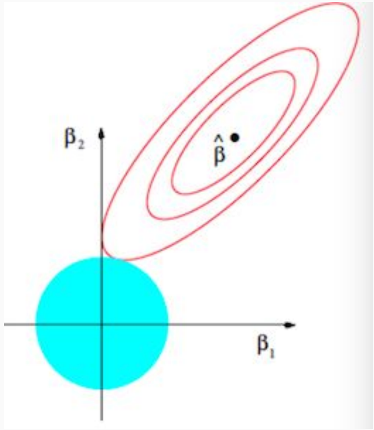


The Ridge Regression

Ridge regression

$$(1) \qquad \min_{\beta \in \mathbb{R}^p} ||y - \mathbf{X}\beta||_2^2 \qquad \boxed{\text{s.t. } ||\beta||_2 \leq t}$$

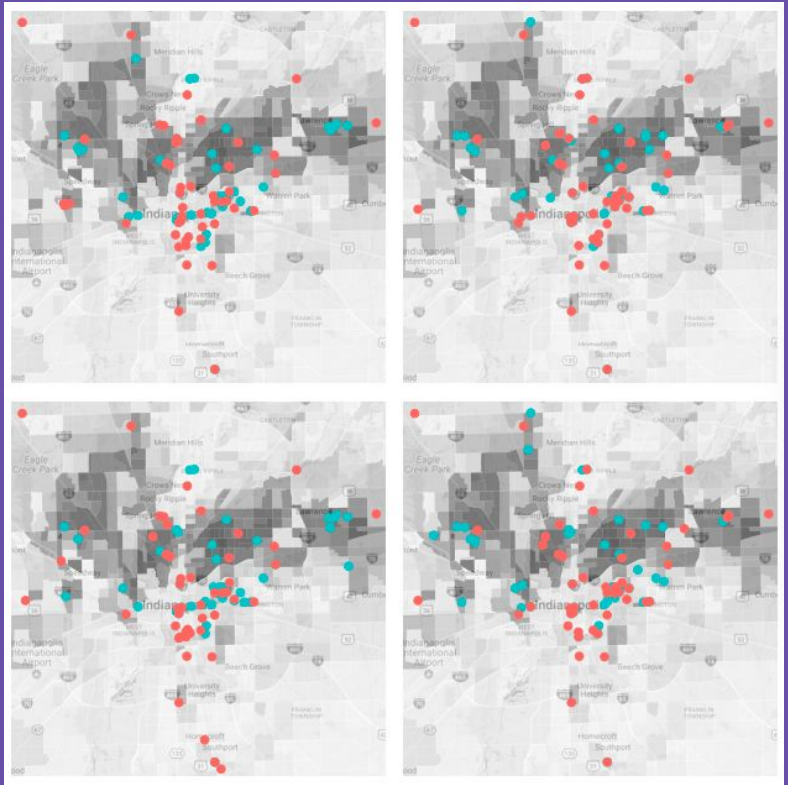
$$(2) \qquad \min_{\beta \in \mathbb{R}^p} ||y - \mathbf{X}\beta||_2^2 + \lambda||\beta||_2^3 \quad \Longrightarrow \quad (\mathbf{X}^T\mathbf{X} + \lambda\mathbb{I})^{-1}\mathbf{X}^T Y$$



Components of the Lasso:

- RSS
- Penalizing term
    - L2 norm

Predictive Policing --> a case study of fair regression



Penalty term

Notion of fairness (F): calculated by comparing the amount of patrol received between a pair of groups (grouped based on race)

Purpose: penalizing the original likelihood function to achieve a “fair” model where police patrol level in a certain racial group matches exactly the true demographic representation of the group

Predictive policing algorithms

Maximizing the likelihood, L, using a log function for prediction of crime rates

$$L(\vec{a}, \omega, \theta) = \boxed{\sum_{i=1}^{N} \log(\lambda_{g_i}(t_i)) - \sum_{g \in G} \int_0^T \lambda_g(t)dt,}$$

Neutral

Fair

$$\boxed{\sum_{i=1}^{N} \log(\lambda_{g_i}(t_i)) - \sum_{g \in G} \int_0^T \lambda_g(t)dt} - \boxed{\chi F},$$

Penalizing the original log-likelihood function by varying the coefficient X to 0 or 10^8

Wrapping up...