

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?

Amazon scraps secret AI recruiting tool that showed bias against women

Some Examples...

VERNON PRATER Prior Offenses 2 armed robberies, 1 attempted armed robbery Subsequent Offenses 1 grand theft LOW RISK 3	BRISHA BORDEN Prior Offenses 4 juvenile misdemeanors Subsequent Offenses None HIGH RISK 8
---	--

DYLAN FUGETT LOW RISK 3	BERNARD PARKER HIGH RISK 10
--	--

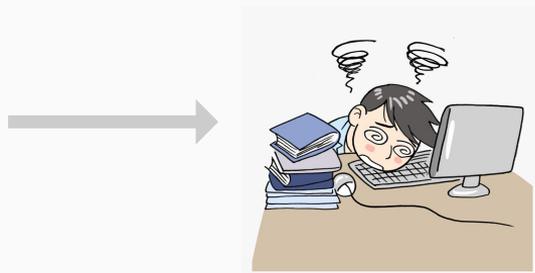
JAMES RIVELLI LOW RISK 3	ROBERT CANNON MEDIUM RISK 6
---	--

JAMES RIVELLI Prior Offenses 1 domestic violence aggravated assault, 1 grand theft, 1 petty theft, 1 drug trafficking Subsequent Offenses 1 grand theft LOW RISK 3	ROBERT CANNON Prior Offense 1 petty theft Subsequent Offenses None MEDIUM RISK 6
---	---

Causality and Fairness

The common use of **counter-factual fairness**:

- Employment decisions
- College admissions



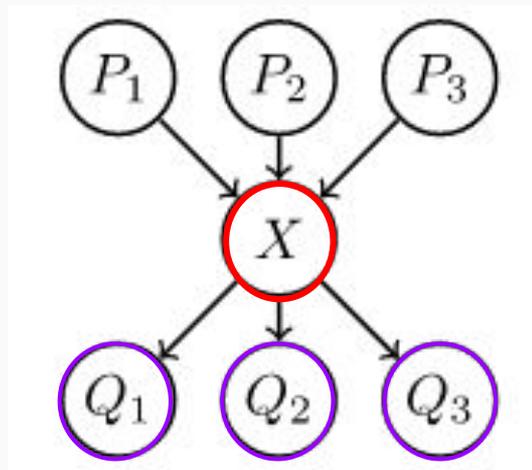
Race (<i>leaving all other attributes constant</i>)	Probability to be granted an offer
Asian	25%
White	36%
Hispanic	77%
African	95%

Causality and Fairness

→ Attribute Flipping

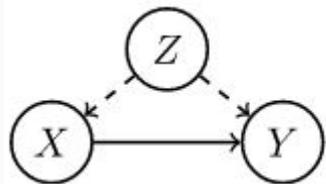


Not always!!!

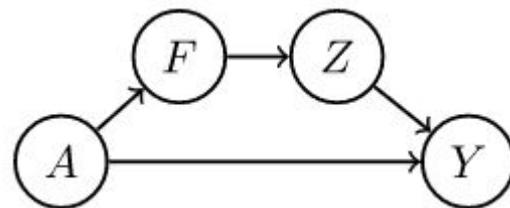
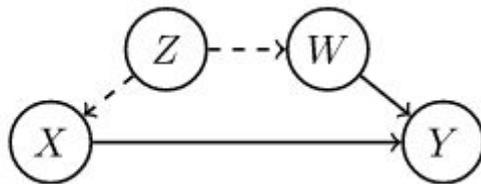


Causality and Fairness

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



Unobserved Confounding



Indirect Paths

Algorithmic Discrimination

```
int divisor = 1, temp = num;

while (num > 0) {
    num /= 10;
    divisor *= 10;
}

return (temp == 0) ? 1 : divisor;
}

// Function to Perform bubbleSort
void bubbleSort(int arr[], int n, int mod) {
    struct list arr[n];
    int divisor = get_divisor(mod);
    int i;

    // Initialize the list
    for (i = 0; i < n; ++i) {
        list arr[i].first_node_address = NULL;
        list arr[i].last_node_address = NULL;
        list arr[i].total_elements = 0;
    }

    // Inserting Elements into the list
    for (i = 0; i < n; ++i) {
        int index = (arr[i] * n) / divisor;

        if (list arr[index].last_node_address == NULL) {
            list arr[index].last_node_address = NULL;
        }

        struct node* temp = (struct node*)malloc(sizeof(struct node));
        temp->modulus = arr[i];
        temp->next = NULL;

        list arr[index].first_node_address = temp;
    }
}
```



The image shows three overlapping, semi-transparent blue forms titled "LOAN APPLICATION". The forms are arranged in a perspective view, with the front-most form being the most prominent. The front form displays the following information:

- Name: Gerard Washington
- Credit Score: 553
- Household Income: \$45K - \$50K
- Likelihood of Default: 67%

A large, bold, red "REJECT" stamp is overlaid on the bottom right of the front form. The background of the image is dark, and the forms have a slight reflection effect.

Algorithmic bias with previous examples

Sources of bias: 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

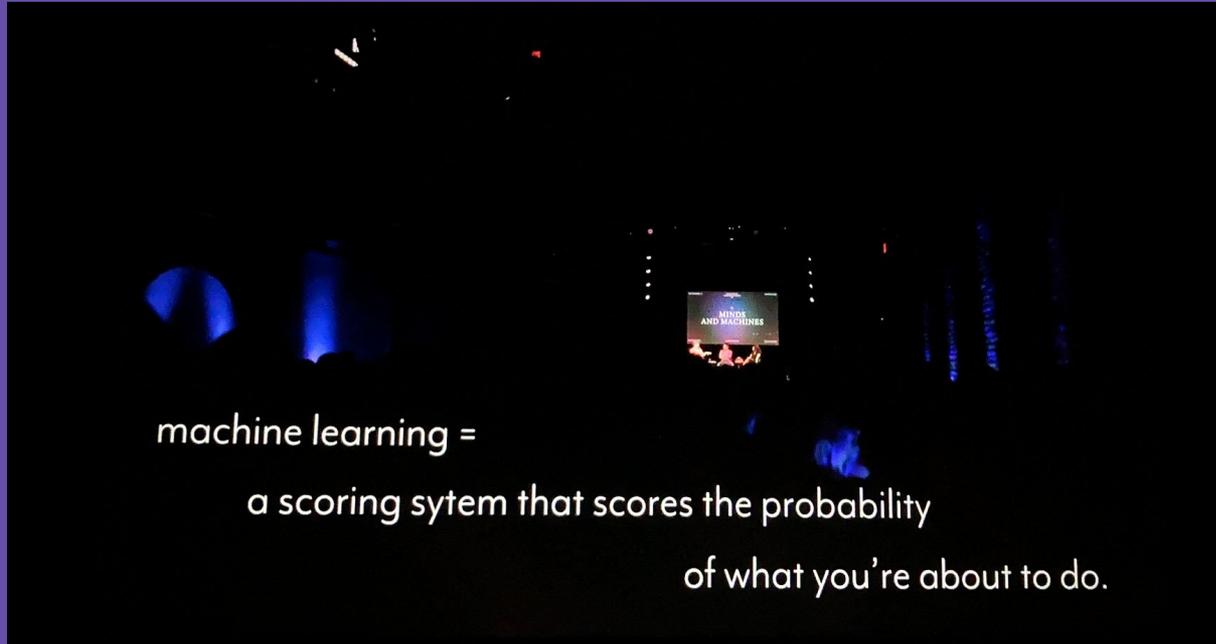
Individual Fairness

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

Group Fairness

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

A Brief Introduction to Machine Learning



machine learning =

a scoring system that scores the probability

of what you're about to do.

The Lasso Regression

Lasso

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

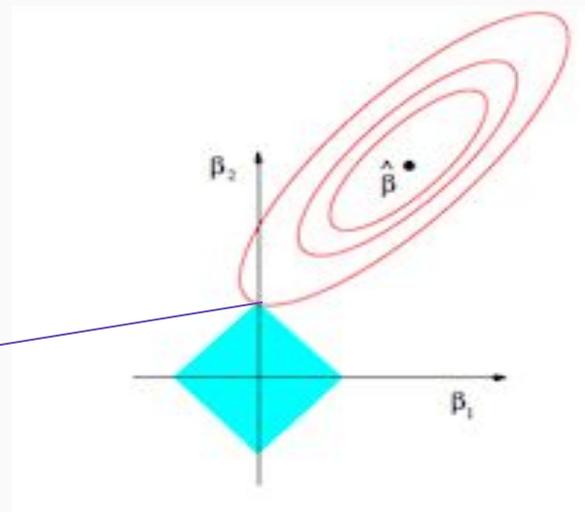
$$(2) \quad \min_{\beta \in \mathbb{R}^p} \frac{1}{2N} \|y - \mathbf{X}\beta\|_2^2 + \lambda \cdot \|\beta\|_1$$

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

Components of the Lasso:

- RSS
- Penalizing term
 - L1 norm

Point of intersection



The Ridge Regression

Ridge regression

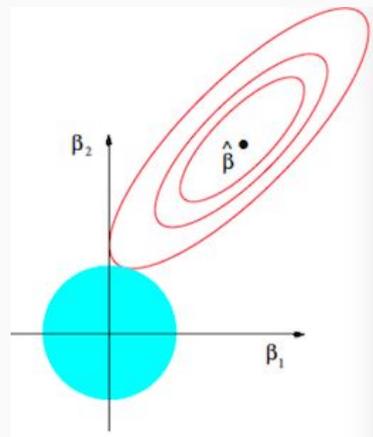
$$(1) \quad \min_{\beta \in \mathbb{R}^p} \|y - \mathbf{X}\beta\|_2^2$$

$$\text{s.t. } \|\beta\|_2 \leq t$$

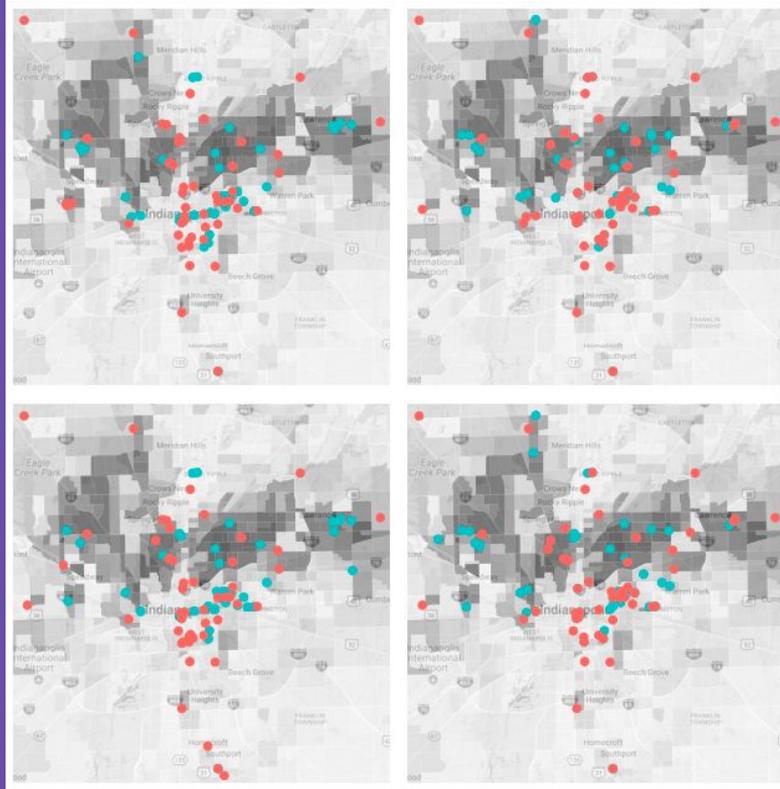
$$(2) \quad \min_{\beta \in \mathbb{R}^p} \|y - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_2^2 \implies (\mathbf{X}^T \mathbf{X} + \lambda \mathbb{I})^{-1} \mathbf{X}^T Y$$

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

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

Neutral

Fair

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

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

Wrapping up...

