Beyond Distributive Fairness
in Algorithmic Decision Making:

Feature Selection for Procedurally Fair Learning

Nina Grgić-Hlača, Muhammad Bilal Zafar,
Krishna P. Gummadi and Adrian Weller

MAX PLANCK INSTITUTE
FOR SOFTWARE SYSTEMS

UNIVERSITY OF
CAMBRIDGE

Algorithmic Decision Making

Human decision making

Algorithmic decision making

- Algorithms help people make decisions about
  - *Hiring*
  - *Assigning social benefits*
  - *Granting bail*

Are these algorithms **fair**?

Types of Algorithmic Fairness

Distributive Fairness

Fairness of **decision making outcomes**

Example

- **Equal misclassification rates**
  - *Grant bail to high risk white defendants*
  - *Deny bail to low risk black defendants*

  → **unfair outcomes**

Procedural Fairness

Fairness of the **decision making process**

Example

- **Fairness of using features**

Is it fair to use a feature in decision making?

This Talk

- A **Notion** of Procedural Fairness: **Feature Usage Fairness**
- **Quantifying** Feature Usage Fairness
- **Mechanisms** for Achieving Feature Usage Fairness

Is it fair to use a feature?

- **Discrimination**
  - **Sensitive** (race, gender) vs **non-sensitive** features
- **Fairness beyond discrimination**
  - **Volitionality** (e.g., criminal history of defendant's father)
    - Does the feature represent the result of volitional (i.e., voluntarily chosen) decisions made by the individual (e.g., number of prior offenses); or rather is it the result of circumstances beyond their control?

Is it fair to use a feature?

- **Discrimination**
  - **Sensitive** (race, gender) vs **non-sensitive** features
- **Fairness beyond discrimination**
  - **Volitionality** (e.g., criminal history of defendant's father)
  - **Relevance** (e.g., defendant's education
    - Is the feature causally related or not to the decision outcomes?

Is it fair to use a feature?

- **Discrimination**
  - **Sensitive** (**race**, **gender**) vs **non-sensitive** features
- **Fairness beyond discrimination**
  - **Volitionality** (e.g., criminal history of defendant's father)
  - **Relevance** (e.g., defendant's education)
  - **Reliability**
    - How reliably can a feature be assessed (e.g., in credit assessments, opinions towards bankruptcy may be harder to reliably assess than number of prior bankruptcies)

Is it fair to use a feature?

- **Discrimination**
  - **Sensitive** (**race**, **gender**) vs **non-sensitive** features
- **Fairness beyond discrimination**
  - **Volitionality** (e.g., criminal history of defendant's father)
  - **Relevance** (e.g., defendant's education)
  - **Reliability**
  - **Privacy**
    - Does use of the feature give rise to a violation of the individual's privacy?

Is it fair to use a feature?

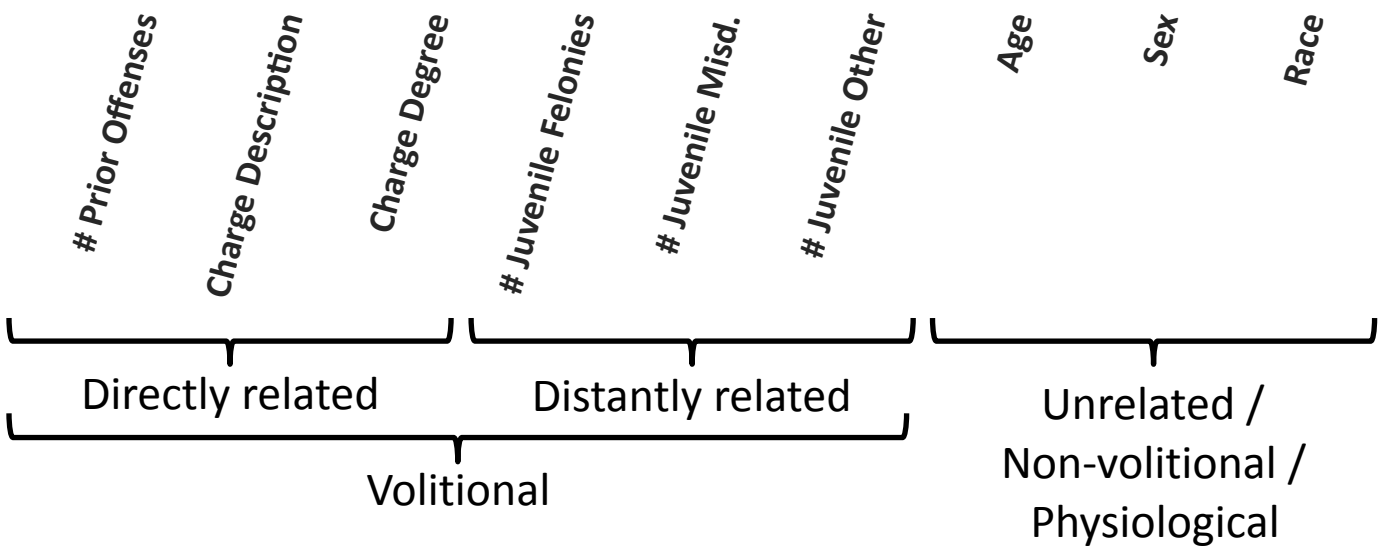
- **Discrimination**
  - **Sensitive** (**race**, **gender**) vs **non-sensitive** features
- **Fairness beyond discrimination**
  - **Volitionality** (e.g., criminal history of defendant's father)
  - **Relevance** (e.g., defendant's education)
  - **Reliability**
  - **Privacy**

Background knowledge on fairness of features not in the data!

- Gather **human moral judgments**

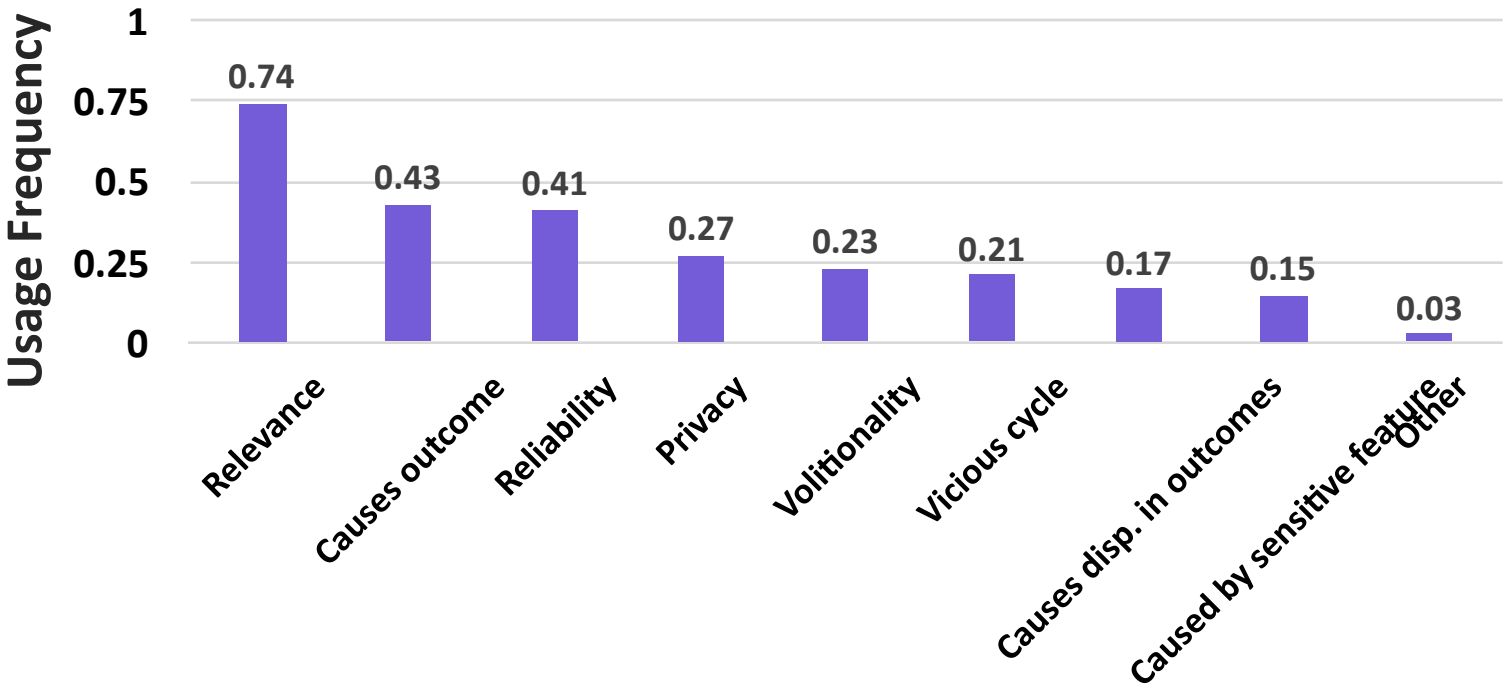
Human Judgments of Fairness

- **Case study**: COMPAS tool for **predicting criminal risk**



Reasoning About Fairness

- What determines people's moral judgments about fairness?



- There is **more to fairness than discrimination!**

Quantifying Fairness of a Classifier

- **Feature usage fairness**
  - The fraction of people that consider using that feature fair

- **Feature usage fairness of a classifier**
  - Fraction of people that consider **all** of its **features fair**

Fairness – Accuracy Tradeoff

- **Intuitively**
  - Adding features: higher accuracy, lower fairness
  - Removing features: lower accuracy, higher fairness

- There is a **tradeoff** between **feature usage fairness** & **accuracy**

Fair Feature Selection

- We want to **select a subset of features** that leads to
  - **High accuracy**
  - **High feature usage fairness**

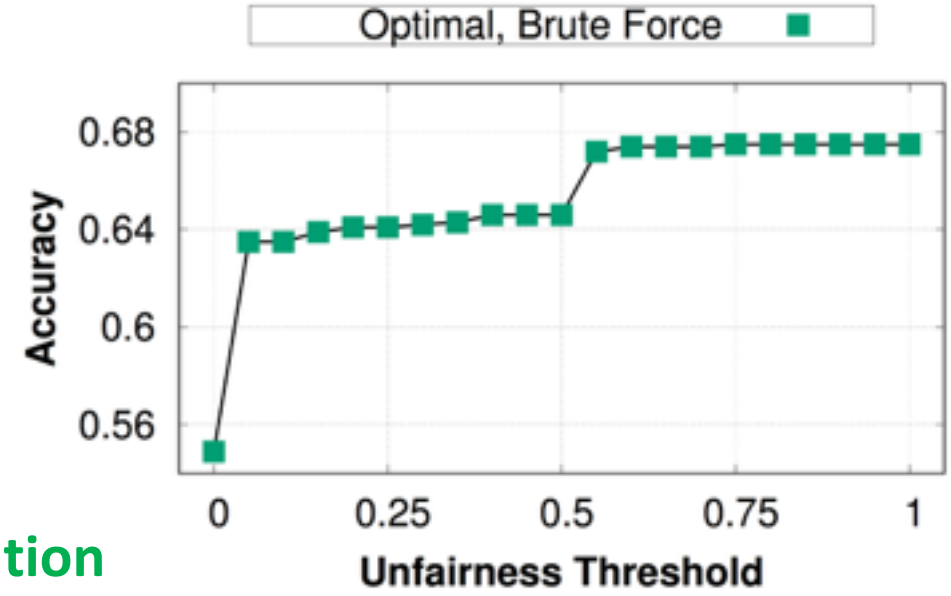
- **Formulation**

$$\begin{aligned} \underset{S \subseteq \mathcal{F}}{\text{maximize}} \quad & accuracy(\mathcal{S}) \\ \text{subject to} \quad & unfairness(\mathcal{S}) \leq t \end{aligned}$$

- **How do we do this?**

Naïve Approach

- **Brute force**
  - **Train 2^n classifiers**, n = number of features



- **Optimal Solution**
- **Not scalable!** 30 features = more than 1 billion classifiers
- **Is there an efficient alternative?**

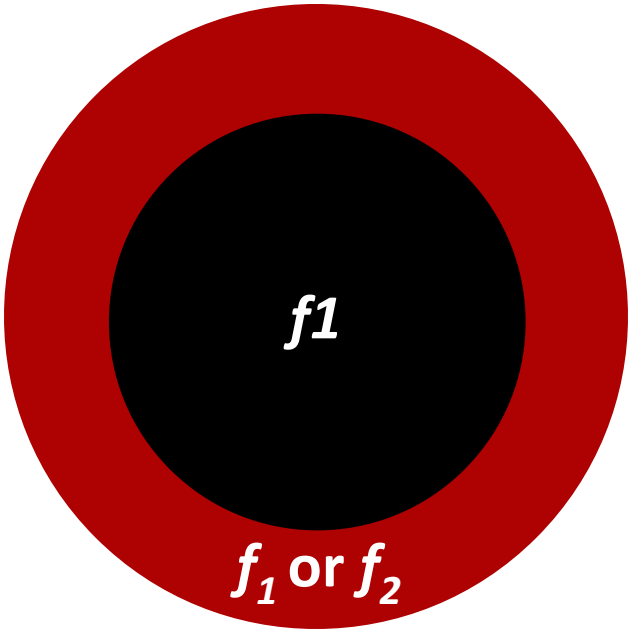
Submodular Optimization

- **Feature usage unfairness** is **submodular** & **monotone**

Fairness Properties - Monotonicity

- Feature unfairness is **monotone non-decreasing**

- *Intuition*
  - A set function is monotone non-decreasing if **adding elements to a set cannot decrease its value**



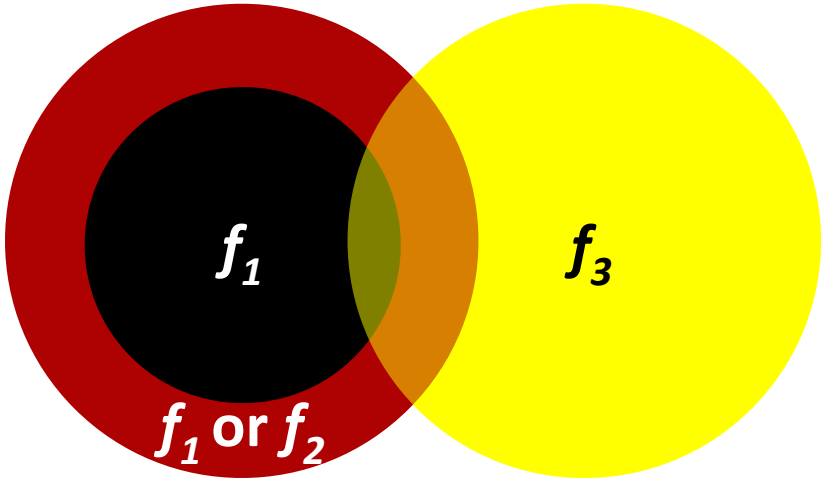
- *Definition*

$$g(\mathcal{F}_i \cup \{f\}) \geq g(\mathcal{F}_i),$$
$$\forall \mathcal{F}_i \subseteq \mathcal{F}, f \in \mathcal{F} \setminus \mathcal{F}_i$$

Fairness Properties - Submodularity

- Feature unfairness is **submodular**

- *Intuition*
  - A set function is submodular if it exhibits **diminishing marginal returns**

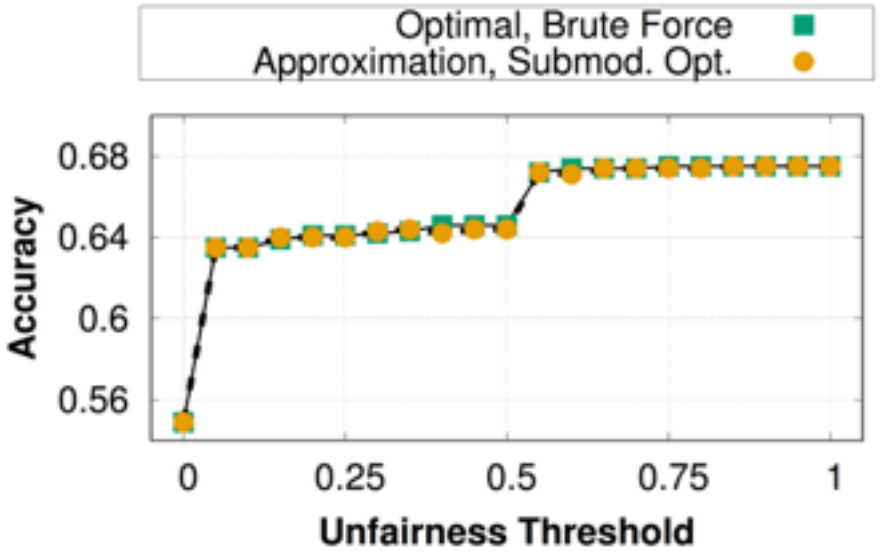


- *Definition*

$$g(\mathcal{F}_A \cup \{f\}) - g(\mathcal{F}_A) \geq g(\mathcal{F}_B \cup \{f\}) - g(\mathcal{F}_B),$$
$$\mathcal{F}_A \subseteq \mathcal{F}_B \subset \mathcal{F}, f \in \mathcal{F} \setminus \mathcal{F}_B$$

Submodular Optimization

- **Feature usage unfairness** is **submodular** & **monotone**
- **Submodular cost submodular knapsack problem**
  - Approximate using **ISK** algorithm (Iyer and Bilmes, NIPS 2013)



- **Efficient & scalable approximation**
- **Near optimal results**

ISK algorithm

$$\begin{aligned} \underset{S \subseteq \mathcal{F}}{\text{maximize}} \quad & accuracy(\mathcal{S}) \\ \text{subject to} \quad & unfairness(\mathcal{S}) \leq t \end{aligned}$$

- Maps to **Submodular Cost Submodular Knapsack** problem

- *Guarantees & hardness*

Approx. factor*
Bi-Criterion factor#
$[1 - e^{-1}, \frac{K_f}{1+(K_f-1)(1-\kappa_f)}]^{\#}$

- *Algorithm*
  - Iteratively **finding modular approximations of submodular functions** & solving the resulting knapsack problems

Accuracy Properties

- Accuracy is **weakly submodular**

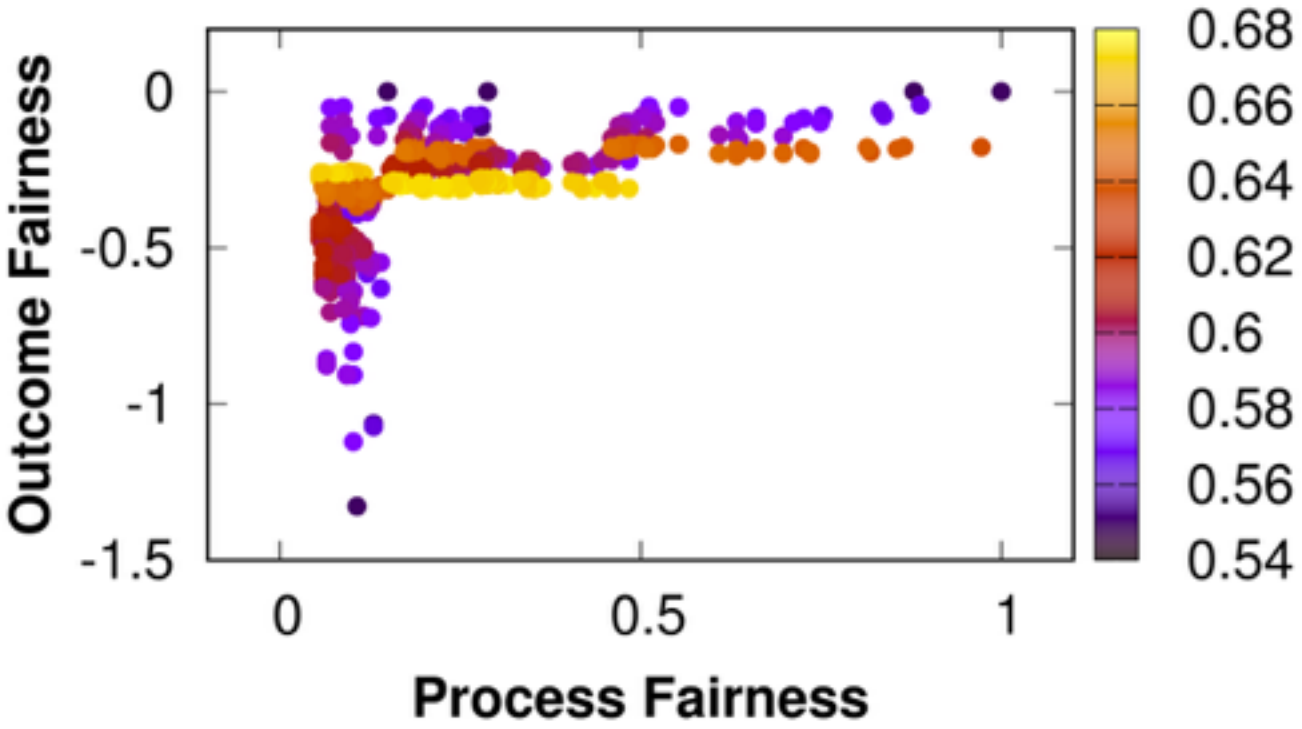
- *More precisely*
  - Logistic loss with l2 regularization exhibits restricted strong convexity, which implies it's weakly submodular

- *Intuition on why this approach performs well*
  - **Greedy algorithms preform well in practice** for logistic loss with l2 regularization

Procedural vs Distributive Fairness

- In the **ProPublica COMPAS dataset**:

high process fairness → high outcome fairness



Key Points

- A **Notion** of Procedural Fairness: **Feature Usage Fairness**
  - Relies on people's moral judgments
  - Beyond discrimination: volitionality, relevance, reliability…
- **Quantifying** Feature Usage Fairness of a Decision Making System
  - Fraction of people that consider all features fair
- **Mechanisms** for Achieving Feature Usage Fairness
  - Control tradeoffs between fairness and accuracy
  - Submodular measure → scalable fair feature selection