Technische Universiteit
Eindhoven
University of Technology

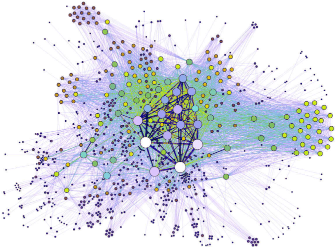
2 September 2021

Fairness in Machine Learning

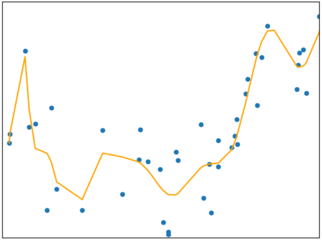
Maryam Tavakol

Machine Learning (ML)

- Turning abundance of data into powerful models to be used in various prediction and decision-making tasks

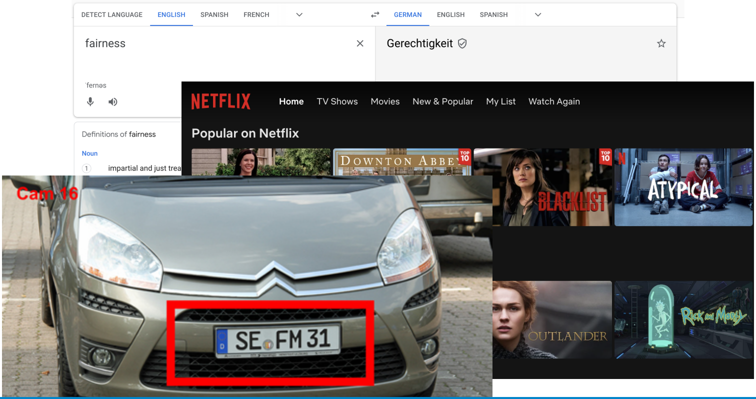


$\Rightarrow$



Real-World Examples

from improving user experience and everyday life...



Real-World Examples

...to high-stake decision-support systems

- Applications in healthcare

Real-World Examples

...to high-stake decision-support systems

- Vaccination/lockdown policies during pandemics

Real-World Examples

...to high-stake decision-support systems

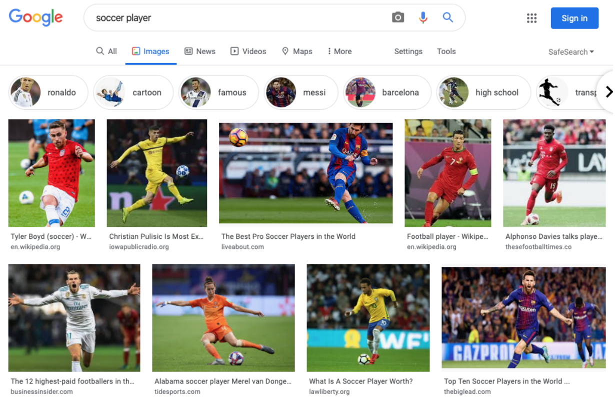
- Bail/sentencing in criminal justice

Real-World Examples

...to high-stake decision-support systems

- Loan decisions in the credit industry

Social Consequences



Social Consequences

Discrimination because of

- **Ethnicity:** loan application, criminal justice
- **Gender:** hiring system, income level
- **Age:** education, hiring system
- **Immigration/citizenship status:** healthcare, loan application
- and so on

$\rightarrow$ Sensitive attributes

Socially-Aware AI

To evaluate, model, and mitigate such biases in the AI systems toward promoting **fairness**

Socially-Aware AI

To evaluate, model, and mitigate such biases in the AI systems toward promoting **fairness**

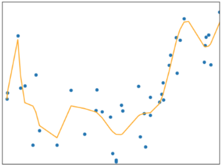
What is Fairness?

Fairness is defined as the absence of any discrimination against individuals and/or groups

Learning Pipeline

- Collecting the data (pre-processing, cleaning, etc.)
- Learning a model that fits the data (optimizing an objective)
- Output a prediction/decision/recommendation

Wife	White	Female	0	0	30	United-States	<=50K.
Unmarried	White	Female	0	0	20	United-States	<=50K.
Husband	Asian-Pac-Islander	Male	0	0	45	?	>50K.
Husband	White	Male	0	0	47	United-States	>50K.
Own-child	Black	Female	0	0	35	United-States	<=50K.
Not-in-family	White	Female	0	0	6	United-States	<=50K.
Not-in-family	White	Male	0	0	43	Peru	<=50K.
Husband	White	Male	0	0	40	United-States	<=50K.
Husband	White	Male	7298	0	90	United-States	>50K.
Own-child	White	Male	0	0	20	United-States	<=50K.
Unmarried	Black	Male	0	0	54	United-States	<=50K.



Looking Inside the Data

Whether individuals earn an income higher or lower than $50k^*$

Wife	White	Female	0	0	30	United-States	<=50K.
Unmarried	White	Female	0	0	20	United-States	<=50K.
Husband	Asian-Pac-Islander	Male	0	0	45	?	>50K.
Husband	White	Male	0	0	47	United-States	>50K.
Own-child	Black	Female	0	0	35	United-States	<=50K.
Not-in-family	White	Female	0	0	6	United-States	<=50K.
Not-in-family	White	Male	0	0	43	Peru	<=50K.
Husband	White	Male	0	0	40	United-States	<=50K.
Husband	White	Male	7298	0	90	United-States	>50K.
Own-child	White	Male	0	0	20	United-States	<=50K.
Unmarried	Black	Male	0	0	54	United-States	<=50K.

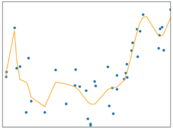
*Adult income data

What Happens?

All the **missteps** in the historical data will be **retained** in the model and **reflected** in the final decision

Wife	White	Female	0	0	30	United-States	<=50K.
Unmarried	White	Female	0	0	20	United-States	<=50K.
Husband	Asian-Pac-Islander	Male	0	0	45	?	>50K.
Husband	White	Male	0	0	47	United-States	>50K.
Own-child	Black	Female	0	0	35	United-States	<=50K.
Not-in-family	White	Female	0	0	6	United-States	<=50K.
Not-in-family	White	Male	0	0	43	Peru	<=50K.
Husband	White	Male	0	0	40	United-States	<=50K.
Husband	White	Male	7298	0	90	United-States	>50K.
Own-child	White	Male	0	0	20	United-States	<=50K.
Unmarried	Black	Male	0	0	54	United-States	<=50K.

$\rightarrow$



$\Rightarrow$

> 50k

< 50k

Various Types of Biases

- **Historical:** due to socio-technical issues in the world
- **Representation:** depend on how to define a population
- **Measurement:** how to choose, utilize, and measure a feature
- **Sampling:** due to non-random sampling of subgroups
- **Algorithmic:** only added by the algorithm
- and many more...

Fairness in Machine Learning

Why:

to have more responsible AI and trustworthy decision-support systems that can be used in *real life*

Goal:

to develop models without any discrimination against individuals or groups, while preserving the utility/performance

Fairness in Machine Learning

How:

- Define fairness measures/constraints
- Alter the data/learning/model to satisfy fairness
- Evaluate the model for balancing performance vs. fairness

How to Define Fairness?

Everybody has an opinion!

Fairness Definitions

Fairness measures per individual:

- **Fairness through unawareness:** sensitive attribute not used
- **Individual fairness:** same outcome for similar individuals
- **Counterfactual fairness:** same outcome for factual and counterfactual situations
- etc.

Fairness Definitions

Fairness measures per group:

- **Demographic parity:** acceptance rate independent of sensitive attribute
- **Equal opportunity:** equal true positive rates
- **Equalized odds:** equal true positive and false positive rates
- etc.

Which Measure to Choose?

A model can not satisfy all measures at the same time

- Individual fairness is a more strong measure, **but...**

Which Measure to Choose?

A model can not satisfy all measures at the same time

- Individual fairness is a more strong measure, **but...**
- who will define them?
- how to specify the appropriate similarity metric?
- and so on

Which Measure to Choose?

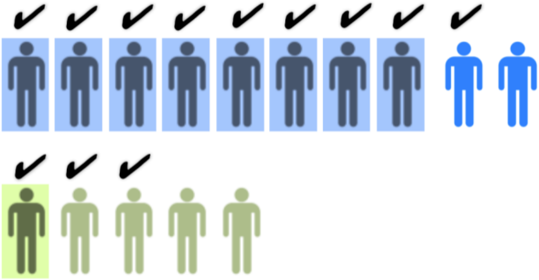
A model can not satisfy all measures at the same time

- Individual fairness is a more strong measure, **but...**
- who will define them?
- how to specify the appropriate similarity metric?
- and so on

$\rightarrow$ group fairness is more plausible due to its amenability to statistical analysis

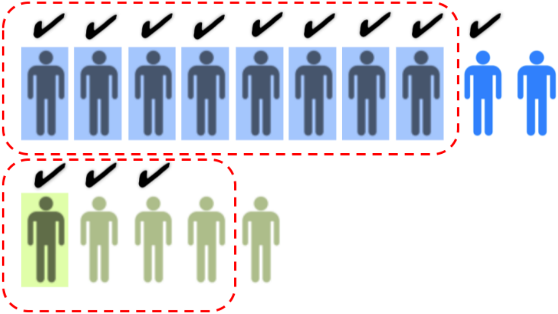
Group Fairness

In a binary classification with binary sensitive attribute:



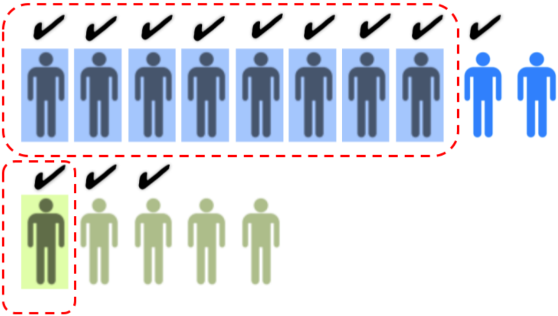
Group Fairness

- Demographic parity



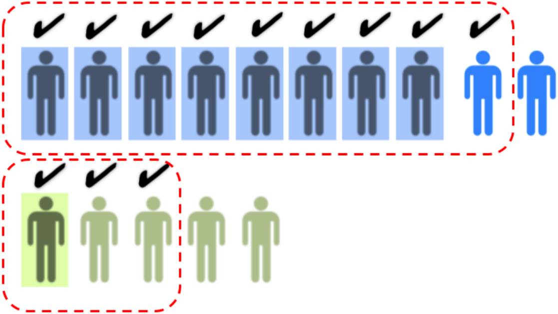
Group Fairness

- Equal opportunity



Group Fairness

- Equalized odds



Equalized Odds

Both protected and non-protected groups should have equal true positive rates and false positive rates

$$P(\hat{y} = 1 | s = 0, y) = P(\hat{y} = 1 | s = 1, y), \quad y \in \{0, 1\}$$

s is a binary sensitive attribute

Fair Learning Techniques

- Pre-processing: transform data to remove discrimination
- Post-processing: incorporate an additional re-assigning step to the obtained models
- We focus on **in-processing:** modify learning algorithms during model training process
    - in classification scenarios

Fair Classification

Zafar et al. propose to modify the objective function in logistic regression (or SVM):

$$\begin{aligned} \text{min} \quad & \textit{classification\_loss} \\ \text{s.t.} \quad & \textit{fairness\_constraint} \end{aligned}$$

Limitations of this kind of methods:

specific loss and fainess measure, convex-based

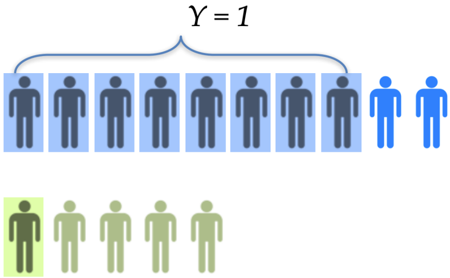
Zafar et al., Fairness Constraints: Mechanisms for Fair Classification, AISTATS 2017

A Different Perspective

- ML methods often depend of **factual** reasoning

A Different Perspective

- ML methods often depend of **factual** reasoning
- i.e., the data/observations are considered the facts



A Different Perspective

Use **counterfactual reasoning** instead...

- to take into account conditions that *could have happened*
- but they didn't, so we cannot observe them

A Different Perspective

Use **counterfactual reasoning** instead...

- to take into account conditions that *could have happened*
- but they didn't, so we cannot observe them

Counterfactual learning

to evaluate and learn what will happen if the data was created, sampled, or labeled differently

Example

- **Observed data:**

treatments

	A	B	C	outcome
patient 1		1		✓
patient 2	1			×
patient 3		1		×
patient 4			1	✓
...		...		...
patient n	1			×

Example

- **Counterfactual model:**

treatments

	A	B	C	outcome
patient 1		~~1~~	1	?
patient 2	1			×
patient 3		1		×
patient 4			1	✓
...		...		...
patient n	1			×

Aims and Conditions

- Model situations that *could have happened* if the data was created or sampled differently
- Evaluate new policies only from available partial feedback (bandit labels)
- Learn a policy that optimizes the outcome
- Make sure the learned policy has low **bias** and **variance** w.r.t. the behavior (sampling) policy

The Proposed Idea

- **Fairness**-aware learning: to learn impartial models from biased data
- **Counterfactual** learning: to evaluate and learn new/optimal policies from logged data

The Proposed Idea

- **Fairness**-aware learning: to learn impartial models from biased data
- **Counterfactual** learning: to evaluate and learn new/optimal policies from logged data

Connect two concepts:

design non-discriminatory models by learning unbiased policies in counterfactual settings

Counterfactual Learning

Swaminathan & Joachims propose to model counterfactual learning as a risk minimization problem, given

- context $\boldsymbol{x}$ drawn i.i.d. from $P(\mathcal{X})$
- decision y chosen from sampling policy $\pi_0 : \mathcal{X} \rightarrow Y$
- and partial feedback $r : \mathcal{X} \times Y \rightarrow \mathbb{R}$

Swaminathan & Joachims, Batch learning from logged bandit feedback through counterfactual risk minimization, JMLR 2015

Counterfactual Learning (cont.)

Goal:

to find an optimal policy π^* which minimizes the loss of prediction on offline data

Counterfactual Learning (cont.)

Goal:

to find an optimal policy π^* which minimizes the loss of prediction on offline data

1. **Evaluation:** to estimate the loss of any policy π

$$R(\pi) = \mathbb{E}_{\boldsymbol{x}} \mathbb{E}_{y \sim \pi(y|\boldsymbol{x})} \mathbb{E}_r [r]$$

2. **Learning:** to optimize the objective over all possible policies

$$\pi^* = \arg\min_{\pi \in \Pi} [R(\pi)]$$

Off-policy Evaluation

Idea:

Fix the mismatch between π_0 that generated the data and π that we aim to evaluate

- Inverse propensity scoring (IPS)
- Self-normalize IPS estimator
- Doubly robust estimator
- and so on.

Learning Algorithm

POEM (Policy Optimization for Exponential Models) is introduced by Swaminathan & Joachims

- an efficient algorithm for structured output prediction
- possible choice of **off-policy estimator** to compute an unbiased estimate of a new policy
- possible choice of **regularizer** to avoid a high-variance policy

Fairness in Counterfactual Setting

Idea:

turn the biased (unfair) **classification** into the task of learning from logged **bandit** data

class label

	$y = 0$	$y = 1$	is fair
$\boldsymbol{x}_1$		1	✓
$\boldsymbol{x}_2$		1	×
$\boldsymbol{x}_3$	1		✓
...	...	...	...
$\boldsymbol{x}_n$	1		×

Properties

- Decisions/observations from the data are **not final**
- **Any** fairness measure can be used to evaluate decisions
- Can be extended to **multi-class** classification problems
- Aims to trade-off between the performance of classification vs. fairness

Counterfactual Framework

Main components:

- context $\boldsymbol{x}$ drawn i.i.d. from $P(\mathcal{X})$
- decision y chosen from sampling policy $\pi_0 : \mathcal{X} \to Y$
- and partial feedback $r : \mathcal{X} \times Y \to \mathbb{R}$

Counterfactual Framework

Main components:

- context $\boldsymbol{x}$ drawn i.i.d. from $P(\mathcal{X})$
- decision y chosen from sampling policy $\pi_0 : \mathcal{X} \rightarrow Y$
- and partial feedback $r : \mathcal{X} \times Y \rightarrow \mathbb{R}$

$\rightarrow$ $\boldsymbol{x}$ and y are available from the data

Counterfactual Framework

Main components:

- context $\boldsymbol{x}$ drawn i.i.d. from $P(\mathcal{X})$
- decision y chosen from sampling policy $\pi_0 : \mathcal{X} \rightarrow Y$
- and partial feedback $r : \mathcal{X} \times Y \rightarrow \mathbb{R}$

$\rightarrow$ $\boldsymbol{x}$ and y are available from the data

ToDo:

We need to estimate the **behavior** (sampling) policy π_0 and formulate the **reward** function r

Behavior Policy

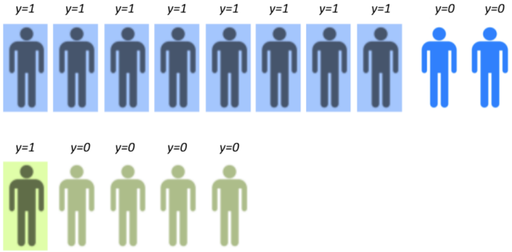
- The true class labels are the sampling (unfair) policy π_0 –**known & deterministic**
- We aim at re-labelling the samples in order to additionally satisfy fairness –**learn** π^*

Behavior Policy

- The true class labels are the sampling (unfair) policy π_0 –**known & deterministic**
- We aim at re-labelling the samples in order to additionally satisfy fairness –**learn** π^*
- Therefore, π_0 is (re-)estimated as a **stochastic** policy to identify the decisions with low probability
  - later used in characterizing the feedback

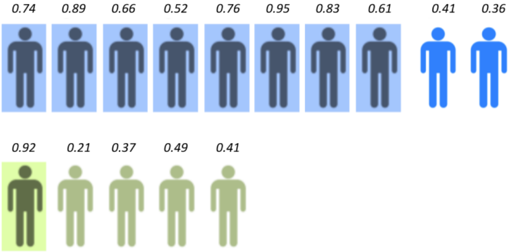
Stochastic Decisions

To better distinguish between the quality of different samples



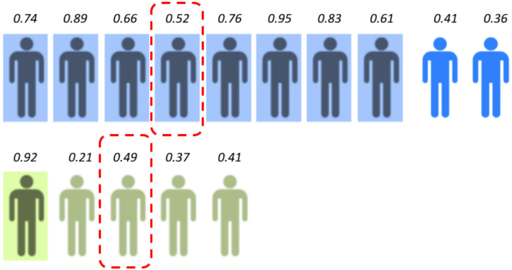
Stochastic Decisions

To better distinguish between the quality of different samples



Stochastic Decisions

To better distinguish between the quality of different samples



Reward Function

- Recall equalized odds

$$P(\hat{y} = 1 | s = 0, y) = P(\hat{y} = 1 | s = 1, y), \quad y \in \{0, 1\}$$

- In order to satisfy fairness measure, find k such that

$$\frac{\sum_{i=1}^{n} \mathbb{1}\{y_i = 1 \wedge s_i = 1\} + k}{\sum_{i=1}^{n} \mathbb{1}\{s_i = 1\}} = \frac{\sum_{i=1}^{n} \mathbb{1}\{y_i = 1 \wedge s_i = 0\} - k}{\sum_{i=1}^{n} \mathbb{1}\{s_i = 0\}}$$

Reward Function (cont.)

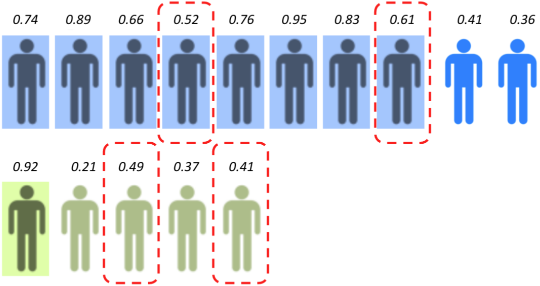
- B_k^+: set of k **positive** samples from non-protected group ($s = 0$) with lowest sampling probabilities, $\hat{\pi}_0(y = 1|\boldsymbol{x})$
- B_k^-: set of k **negative** samples from protected group ($s = 1$) with lowest sampling probabilities, $\hat{\pi}_0(y = 0|\boldsymbol{x})$

$$r_i = \begin{cases} 0 & i \in \{\mathbb{B}_k^+ \vee \mathbb{B}_k^-\} \\ -1 & \text{otherwise} \end{cases}$$

- penalize k most-likely unfair decisions from each group

Example

For $k = 2$:



Learning Overview

1. Learn a stochastic sampling policy from a fraction of data
2. Convert the classification data into bandit data
3. Compute bandit feedback from fairness measure (other definitions or their combination also possible)
4. Learn a counterfactual policy that trades-off classification performance vs. fairness

In Practice

- Adult income data of $\sim$45k subjects, with binary label of a high or low income, and **gender** as sensitive attribute
- Training via POEM algorithm with self-normalized estimator and empirical variance reqularizer (by model selection)
- π_0 estimated using logistic regression with LBFGS solver and l_2-norm regularizer
- **Baseline:** method from Zafar et al.

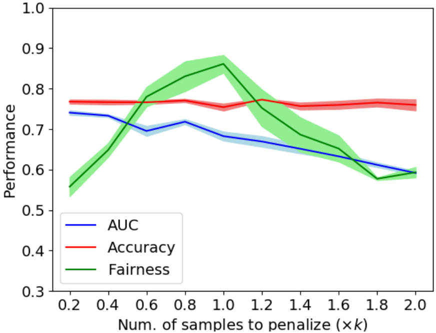
Performance Evaluation

- Classification performance: Area under the ROC curve (AUC) and accuracy
- Fairness measure

$$\min\left(\frac{P(\hat{y}=1|s=0,y)}{P(\hat{y}=1|s=1,y)},\frac{P(\hat{y}=1|s=1,y)}{P(\hat{y}=1|s=0,y)}\right)$$

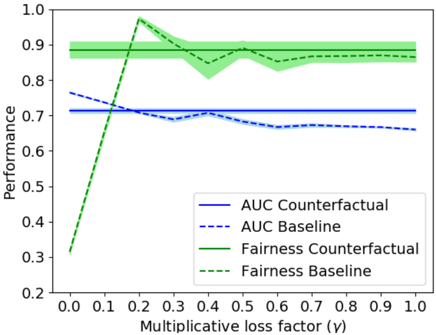
$\rightarrow$ value of 1 satisfies equalized odds

Performance Results



k is the right amount of samples to penalize for maximum fairness

Baseline Comparison



our model is in-line with the baseline method

Summary & Conclusions

Summary & Conclusions

- Fairness-aware learning is essential for having more **responsible** AI and **trustworthy** decision-support systems

Summary & Conclusions

- Fairness-aware learning is essential for having more **responsible** AI and **trustworthy** decision-support systems
- Counterfactual methods are reliable techniques to remove decision biases *from logged data* and learn impartial policies

Summary & Conclusions

- Fairness-aware learning is essential for having more **responsible** AI and **trustworthy** decision-support systems
- Counterfactual methods are reliable techniques to remove decision biases *from logged data* and learn impartial policies
- Biased classifiers can be modeled as sampling policy in counterfactuals and a fairness measure shapes the feedback

Summary & Conclusions

- Fairness-aware learning is essential for having more **responsible** AI and **trustworthy** decision-support systems
- Counterfactual methods are reliable techniques to remove decision biases *from logged data* and learn impartial policies
- Biased classifiers can be modeled as sampling policy in counterfactuals and a fairness measure shapes the feedback
- Our model effectively increases a measure of fairness while maintains an acceptable classification performance

Limitations and Future Direction

- Focus on **individual fairness** measures
  - Individual metrics are more reliable
  - Literature shows that group-based measures are not necessarily compatible with individual fairness
- Model the **long-term** effects of fairness
  - Decisions made by AI systems have time-varying consequences
  - Literature shows that modeling the static effect of bias does not guarantee fairness in long-term

Questions?

Thanks for your attention

Maryam Tavakol

m.tavakol@tue.nl

Technische Universiteit
Eindhoven
University of Technology