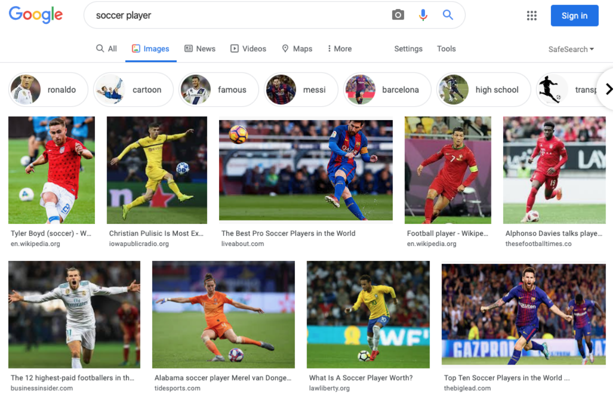
July 2020

Fair Classification with Counterfactual Learning

Dr. Maryam Tavakol

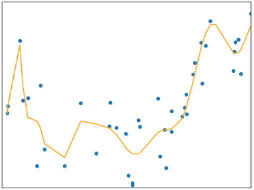
What is Fairness



ML/DM Basics

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

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.



The Role of Biases

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

Fairness-aware 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-aware Learning

How:

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

Definition of Fairness

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

Learning Framework

- ML/DM methods often depend of **factual** reasoning
- *Alternatively:* counterfactual methods learn unbiased policies from logged bandit data via **counterfactual** reasoning

Learning Framework

- ML/DM methods often depend of **factual** reasoning
- *Alternatively:* counterfactual methods learn unbiased policies from logged bandit data via **counterfactual** reasoning

Connect two concepts:

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

Counterfactual Bandits

treatments

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

Counterfactual Bandits

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

Counterfactual Bandits

treatments

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

Goal: learn a policy to optimize the outcome

Counterfactual Learning (cont.)

Goal:

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

1. **Evaluation:** estimate the loss of any policy (unbiased)

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

2. **Learning:** optimize the objective

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

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		×

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		×

extendable to multi-class classification

Sampling 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^*

Sampling 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

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

Summary of the Approach

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: our model effectively increases a measure of fairness while maintains an acceptable classification performance