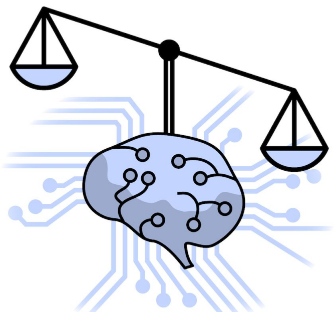


MoFlipTest: Multi-Objective FlipTest

Rifat Mehreen Amin

Ludwig-Maximilians-Universität

25th June, 2021

Outline

- Motivation
- Counterfactuals in ML
- The problem we want to tackle
- Related works
  - FlipTest
  - MOC
  - QII & MCS
- Notations
- Proposed Idea

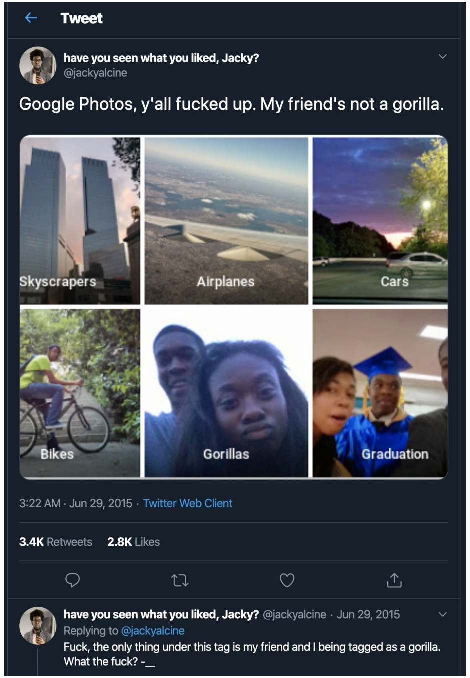
Motivation

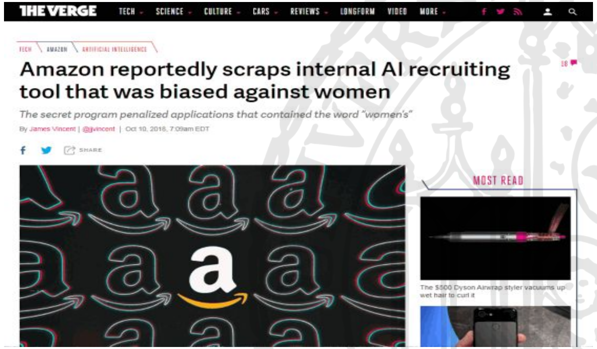
Erik Carter

China Is Collecting DNA From Tens of Millions of Men and Boys, Using U.S. Equipment

Even children are pressed into giving blood samples to build a sweeping genetic database that will add to Beijing's growing surveillance capabilities, raising questions about abuse and privacy.

Published June 17, 2020
By Sui-Lee Wee



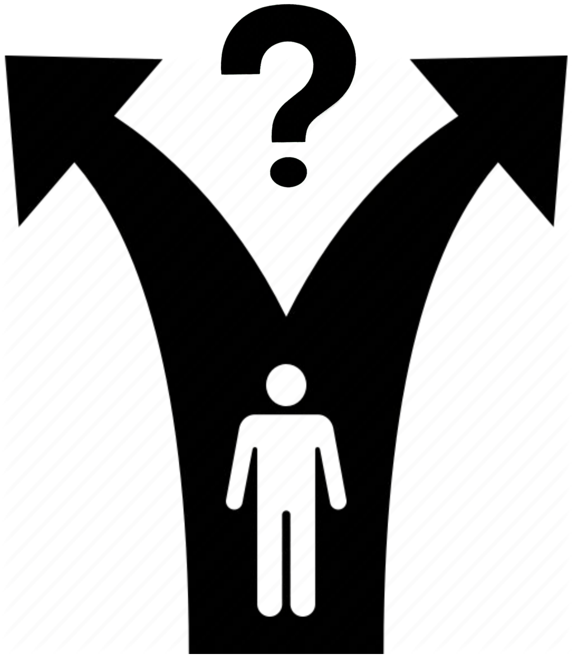


Motivation: bias against protected classes

Legally recognized ‘protected classes’

Race; Color; Sex; Religion; National origin; Citizenship; Age; Pregnancy; Familial status; Disability status; Veteran status; Genetic information [4]

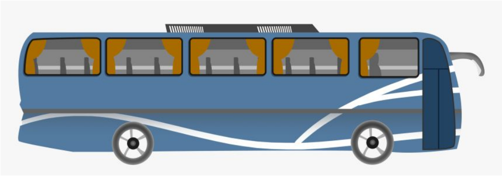
Counterfactuals



Counterfactuals cntd.

"What if I had taken the public bus instead of a taxi, I might have reached office earlier"

- Idea of two fictitious world
- Different Interpretations



Counterfactuals in ML

- Used to explain predictions
- Used to check if an ML model is fair
- Used for local explanation of a model (individual instance)
- Counterfactual explanations for reversing unfavorable outcome

The problem we want to tackle

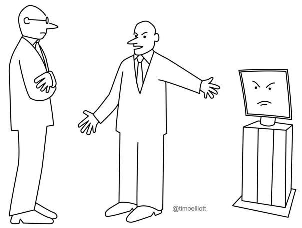
“Given a fitted model *f*, we want to test whether it is fair or not with respect to the definition of counterfactual fairness:

A predictor is fair if it would give the same prediction in a world where we were different [4]”

Related works

FlipTest [1]:

- A fairness testing approach
- Generates counterfactuals by optimal transport mapping of the predicted probabilities of the two groups
- Idea of Flipsets
    - Positive Flipset (F+)
    - Negative Flipset (F-)
- Transparency report
- Works for binary classifiers

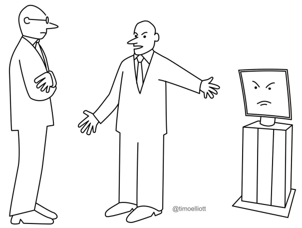


His decisions aren't any better than yours — but they're WAY faster...

Related works cntd.

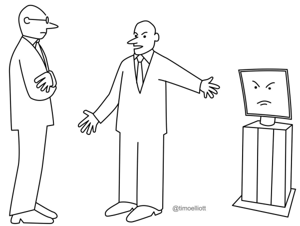
MOC (Multi-Objective Counterfactual method) [2]:

- Model-agnostic
- Focuses on four objectives
  - Counterfactual prediction close to desired prediction
  - Counterfactual close to instance
  - Sparse feature changes
  - Counterfactuals having likely feature values or combinations
- Works for numerical and categorical features
- Works for classifiers and regression



His decisions aren't any better than yours — but they're WAY faster...

Related works (for extension) cntd.



His decisions aren't any better than yours — but they're WAY faster...

QII (Quantitative input influence) [3]:

- Captures the degree of influence of inputs on outputs of systems
- Takes into account the correlated inputs

MCS (Model-based Counterfactual Synthesizer) [6]:

- Synthesizes model-based counterfactuals for prediction reasoning

Notations

x: original instance

y: original prediction

x': counterfactual instance

y': counterfactual prediction

Proposed idea (MoFlipTest)

- Multi-Objective Counterfactual method
- Model-agnostic
- Transparency report
- Flipsets
  - Positive Flipset (F+)
  - Negative Flipset (F-)
- Will work for numerical and categorical features

Proposed idea (MoFlipTest) cntd.

- Focuses on three objectives of MOC:
  - Counterfactual close to instance x'→x
  - Sparse feature changes
  - Counterfactuals having likely feature values or combinations (hard to achieve)

- Observe the difference between y'→y
  - Comparing outcomes with flipsets

Proposed idea (An extension) cntd.

An extension to our idea:

- We would like to use QII for finding out the exact features that are responsible for discrimination

Proposed idea (example)

A loan application example

Sex	Income	Height	Education	Age	Y
female	50k	5’4”	Bachelors	28	✗
male	60k	5’9”	Masters	30	✓

✗ = Loan denied

✓ = Loan accepted

Proposed idea (example) cntd.

1st objective and 2nd objective: x' close to x and small #feature changes

Sex	Income	Income'	Height	Education	Age	Y	Y'
female	50k	60k	5’4”	Bachelors	28	✗	✓
male	60k	60k	5’9”	Masters	30	✓	✓

Proposed idea (example) cntd.

3rd objective: counterfactuals should have likely feature values or combinations

Sex	Income	Height	Education	Education'	Age	Age'	Y	Y'
female	50k	5’4”	Bachelors	Masters	28	32	✗	✓
male	60k	5’9”	Masters	Masters	30	30	✓	✓

Proposed idea (example) cntd.

Suppose our method generated a counterfactual where the sensitive attribute is changed from female to male and changed the outcome:

Sex	Sex'	Income	Height	Education	Age	Y	Y'
female →	male	50k	5'4"	Bachelors	28	✗	✓
male →	female	60k	5'9"	Masters	30	✓	✓

Proposed idea (example) cntd.

It means our model might be biased based on an individual’s sex or gender

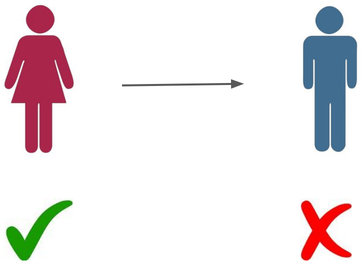
Sex	Sex'	Income	Height	Education	Age	Y	Y'
female	male	50k	5’4”	Bachelors	28	✗	✓
male	female	60k	5’9”	Masters	30	✓	✓

Proposed idea (example) cntd.

Flipset: the set of women who had a different model outcome post translation

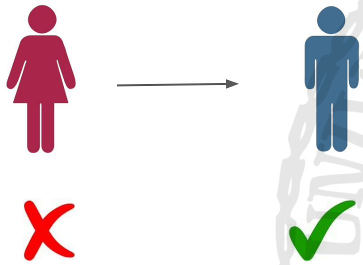
F(+)

Positive Flipset



F(-)

Negative Flipset



Proposed idea (example) cntd.

Information from flipset:

- Compare flipset distribution to overall female population to uncover potential discrimination

Proposed idea (example) cntd.

Transparency Report: it gives the insight about how the model discriminates.

- Here, $h : X \rightarrow \{0, 1\}$ is a binary classifier, $G(x) \rightarrow$ function for MOC counterfactual generation. $F(h, G)$ is the corresponding flipset

$$\frac{1}{|F^{\star}(h,G)|} \sum_{\boldsymbol{x} \in F^{\star}(h,G)} \boldsymbol{x} - G(\boldsymbol{x}), \ \ and$$

$$\frac{1}{|F^{\star}(h,G)|} \sum_{\boldsymbol{x} \in F^{\star}(h,G)} \operatorname{sign}(\boldsymbol{x} - G(\boldsymbol{x}))$$

$$Here, \star \in \{+, -\}$$

Proposed idea (challenges to tackle)

- Generalize to regression
- Making it multi-objective

References

1. Emily Black, Samuel Yeom and Matt Fredrikson. “FlipTest: Fairness Auditing via Optimal transport”. In: *CoRR* abs/1906.09218 (2019). arXiv: 1906.09218. URL : http://arxiv.org/abs/1906.09218.
2. Susanne Dandl, Christoph Molnar, Martin Binder, and Bernd Bischl. “Multi-Objective Counterfactual Explanations”. In: Lecture Notes in Computer Science (2020), pp. 448–469. ISSN : 1611-3349. DOI : 10.1007/978-3-030-58112-1_31. URL : http://dx.doi.org/10.1007/978-3-030-58112-1_31.
3. Anupam Datta, Shayak Sen, and Yair Zick. “Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems”. In: 2016 IEEE Symposium on Security and Privacy (SP). 2016, pp. 598–617. DOI : 10.1109/SP.2016.42.
4. Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. “Counterfactual Fairness”. In: Advances in Neural Information Processing Systems 30. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Curran Associates, Inc., 2017, pp. 4066-4076.
5. https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb
6. Fan Yang, Sahan Suresh Alva, Jiahao Chen and Xia Hu. “Model-Based Counterfactual Synthesizer for Interpretation”.