Predict Responsibly: Fairness in Machine Learning

David Madras

University of Toronto, Vector Institute

October 23, 2018

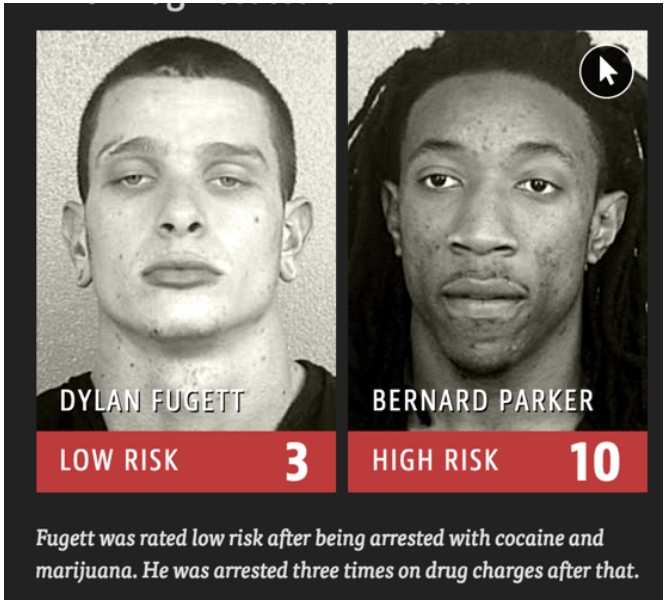
Overview

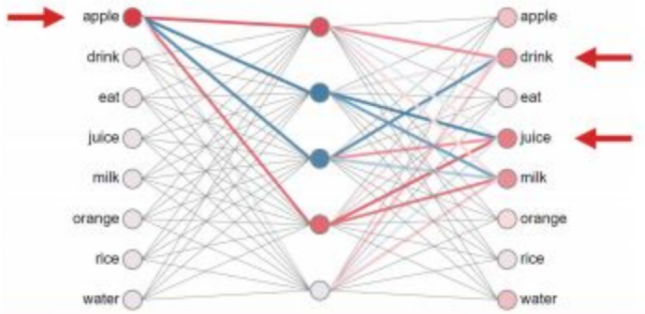
1. Intro to Fairness in Machine Learning
2. Defining Fairness
3. "Fairness" is Complicated
4. Predict Responsibly: NIPS 2018 paper (with Toni Pitassi & Rich Zemel)

Algorithmic Unfairness

- Algorithms are pervasive, high-stakes, high-impact
- Need more than just "accuracy"
- Eg. fairness
  - and interpretability
  - and accountability
  - and security
  - and safety
  - and robustness
  - ...

Where Can Unfairness Arise?





Defining Fairness

- Can we define fairness? Should we try?
- To attack these problems using machine learning, we need to **define** them mathematically
- If we can't, then maybe machine learning is the wrong tool for the problem
- Probably no perfect definitions, but maybe some useful ones

Fair Classification

Fair classification is the most common setup, involving:

- X, some data
- Y, a label to predict
- $\hat{Y}$, the model prediction
- A, a **sensitive attribute** (race, gender, age, socio-economic status)

We want to learn a classifier which is:

1. accurate
2. fair with respect to A

Fair Classification: Definitions

Most common way to define fair classification is to require some invariance with respect to the sensitive attribute

- Demographic parity: $\hat{Y} \perp A$
- Equalized Odds: $\hat{Y} \perp A | Y$
- Equal Opportunity: $\hat{Y} \perp A | Y = y$, for some y
- Equal Calibration: $Y \perp A | \hat{Y}$
- Fair Subgroup Accuracy: $\mathbb{1}[Y = \hat{Y}] \perp A$

Note: Many of these definitions are incompatible!

Question: Will it work to just remove A from our data?

Fair Representations

Learn a representation of X which removes all information about A

- Can't just erase A from the data, since other variables are correlated

If our representations achieve some conditional independencies, so will classifiers learned from those representations

- Adversarial approach: LAFTR (Madras et al.)
- VAE approach: VFAE (Louizos et al.)

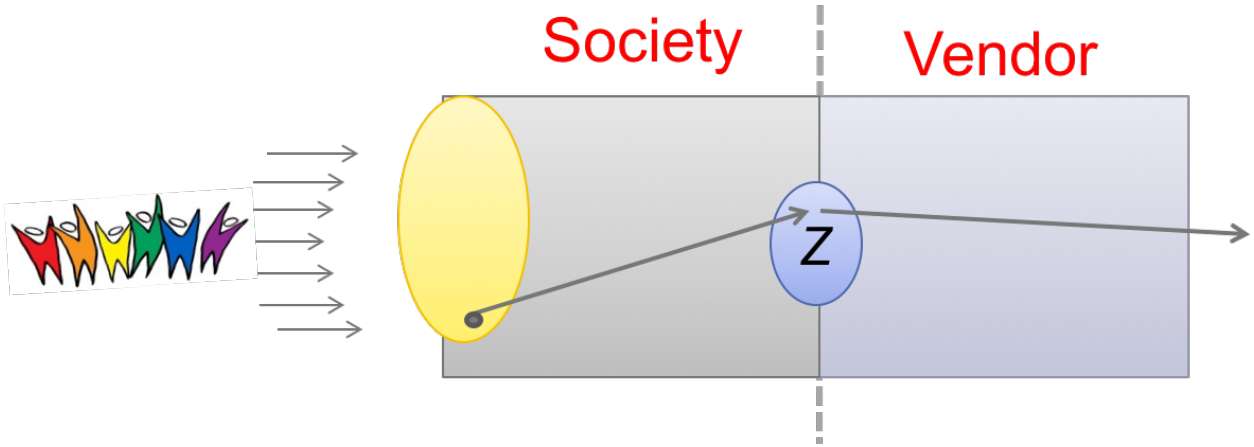


Figure 1: Learning a Fair Representation [Dwork et al.]

Fairness & Causality

Key idea: **counterfactual fairness**

- What would happen if the sensitive attribute for this person had been different, but everything else had been the same? What would our model do?

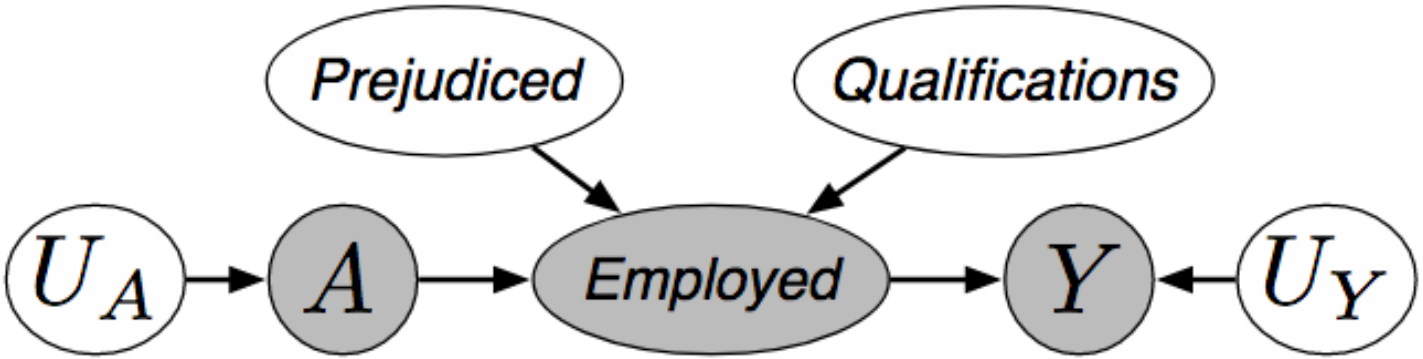


Figure 2: A causal model of the data-generation process [Kusner et al.]

- Additionally, causal inference allows us to model **actions in the world** (interventions), as opposed to just predictions

The Golden Rule of Fairness Definitions

The statistician George Box once wrote

> *All models are wrong, but some are useful.*

The Golden Rule of Fairness Definitions

The statistician George Box once wrote

> *All models are wrong, but some are useful.*

When talking about fairness in ML, I think we should similarly remember

> *All fairness definitions are wrong, but some are useful.*

The Golden Rule of Fairness Definitions

The statistician George Box once wrote

> *All models are wrong, but some are useful.*

When talking about fairness in ML, I think we should similarly remember

> *All fairness definitions are wrong, but some are useful.*

We've talked about useful. How are these definitions wrong?

Multi-attribute Fairness

We may be concerned with fairness for multiple sensitive attributes

- e.g. race *and* age, gender *and* SES
- Satisfying fairness definition for one attribute may make it unfair for another ("fairness gerrymandering")

Dataset Bias

We often define our model's fairness in terms of the labels Y. But what if these labels are themselves biased?

- Hiring: past performance of similar candidates
- Bail: defendants who do not receive bail are more likely to plead guilty
- School acceptance: ability of past students to succeed once in school

Feedback Loops

What if our model acts in the world, and affects its own training data?

- Predictive policing: where to send police to patrol each day?
  - This affects where and how much crime we observe
- Recommender systems
  - If recommendations are worse for some group, they may stop using the system
- Retraining the model on past data will amplify biases

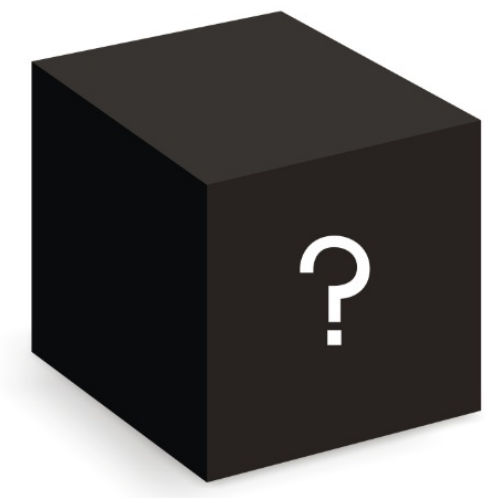
Multi-agent Systems

What if our model interacts with other decision-making agents?

- Doctors, judges, committees, other pieces of software ...
- Decision makers have biases in how they act and how they use assistive models

This is what we aim to address in our recent paper "Predict Responsibly: Improving Fairness and Accuracy by Learning to Defer" (NIPS 2018).

The Judge and the Black-Box

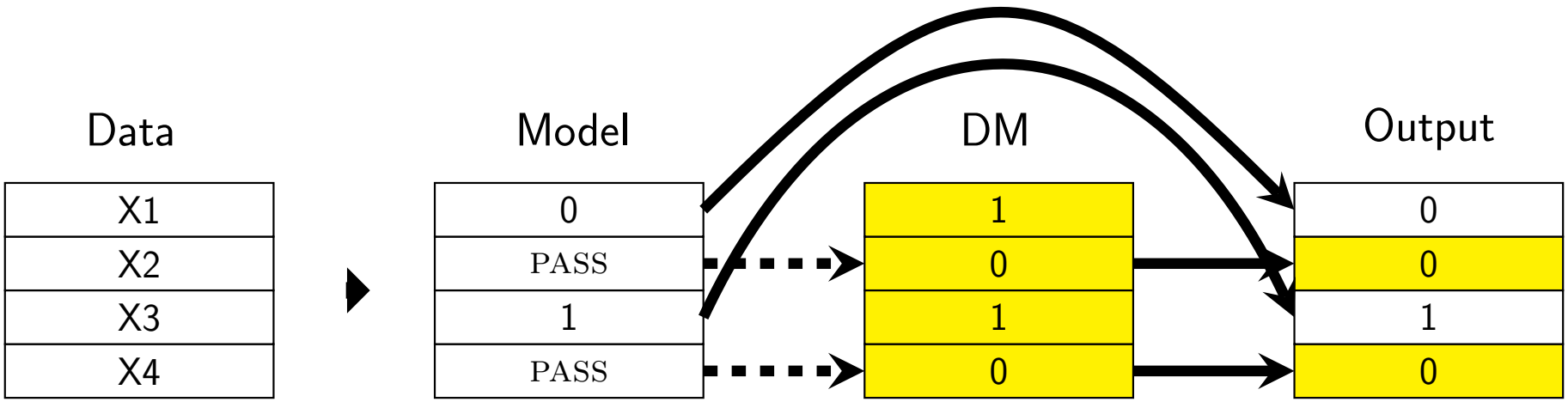


Decision

A Framework for Real-World Decision-Making Systems

Real-world decision systems are interactive processes with many agents

- Our framework is external decision-maker (DM) + ML model
- PASS or predict?
- PASSing can be: culling a large pool, flagging cases for review, etc.



Modelling a Decision System

Given data X, auxiliary data Z, and labels Y, define system output $\hat{Y}$, model predictions $\hat{Y}_M$, model PASS decisions s, and DM predictions $\hat{Y}_D$:

$$\hat{Y} = (1 - s)\hat{Y}_M + s\hat{Y}_D$$

$$\hat{Y}_M = P_M(Y = 1|X); \quad s = g_s(X); \quad \hat{Y}_D = P_D(Y = 1|X, Z)$$

We can model the joint output of the system as

$$P_{defer}(Y|X, Z) = \prod_i [\hat{Y}_{M,i}^{Y_i}(1 - \hat{Y}_{M,i})^{1-Y_i}]^{1-s_i}[\hat{Y}_{D,i}^{Y_i}(1 - \hat{Y}_{D,i})^{1-Y_i}]^{s_i}$$

Learning to Defer

We can minimize the negative log-probability of this system. We call this **learning to defer**.

$$\mathcal{L}_{defer}(Y, \hat{Y}_M, \hat{Y}_D, s) = -\sum_i [(1 - s_i)\ell(Y_i, \hat{Y}_{M,i}) + s_i \ell(Y_i, \hat{Y}_{D,i})]$$

This bears similarity to the loss for **rejection learning**, which sets $\ell(Y_i, \hat{Y}_{D,i}) = \gamma_{reject}$:

$$\mathcal{L}_{reject}(Y, \hat{Y}_M, \hat{Y}_D, s) = -\sum_i [(1 - s_i)\ell(Y_i, \hat{Y}_{M,i}) + s_i \gamma_{reject}]$$

Key Takeaway: we include information about the decision-maker in training our model.

Experiments

- Datasets: COMPAS (predict recidivism fairly w.r.t. race), Health (predict co-morbidity fairly w.r.t. age)
- Compare learning to **defer** with learning to **reject**
- We simulated three types of decision-makers (DMs) to evaluate our model:
  - High-accuracy DM
  - Highly-biased DM
  - Inconsistent DM

Experiments - High-accuracy DM

- A **high-accuracy** DM may have useful auxiliary information Z
- We simulated a high-accuracy DM by training a classifier to predict Y from data X and Z
- Yielded a DM with higher accuracy than the ML model we trained

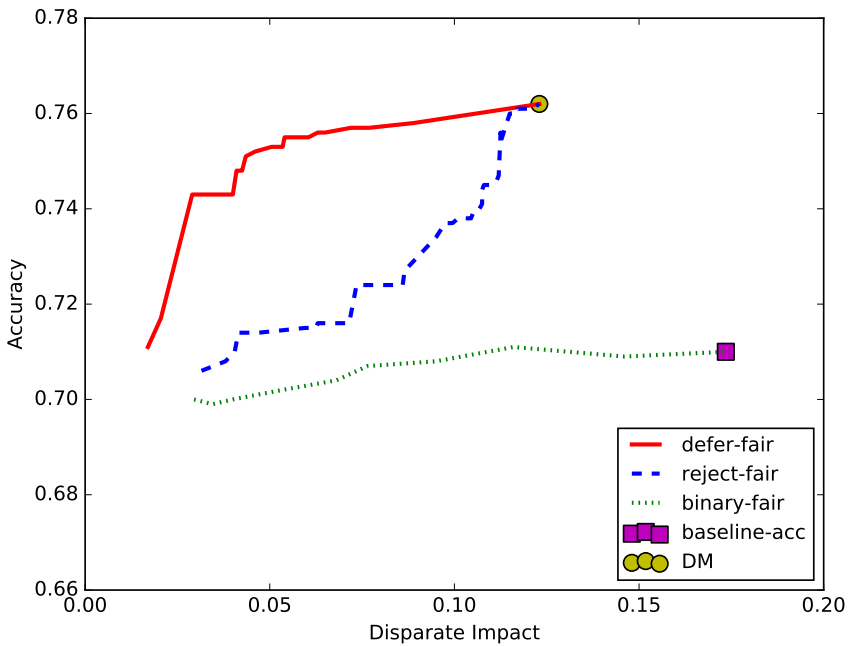


Figure 3: COMPAS dataset. Fairness-accuracy tradeoff. Top left corner is best.

Experiments - Highly-biased DM

- A **highly-biased** DM may have internal biases against some subgroups
- We simulated these biases by training a DM with a fairness regularization coefficient $\alpha < 0$

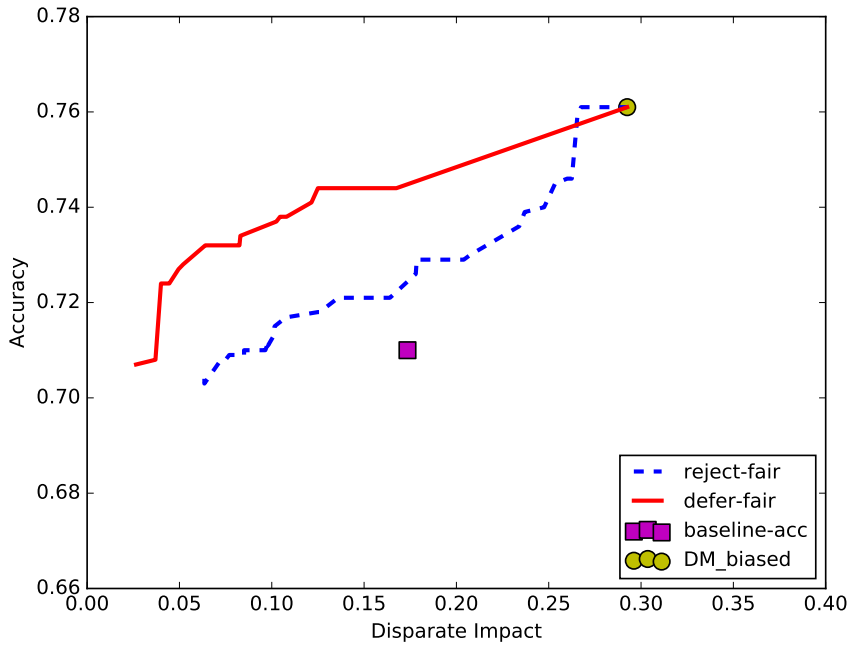


Figure 4: COMPAS dataset. Fairness-accuracy tradeoff. Top left corner is best.

Experiments - Inconsistent DM

- An **inconsistent** DM may have low accuracy, despite having auxiliary information Z
- Simulated this by post-hoc flipping DM's predictions on some "unreliable" subgroup

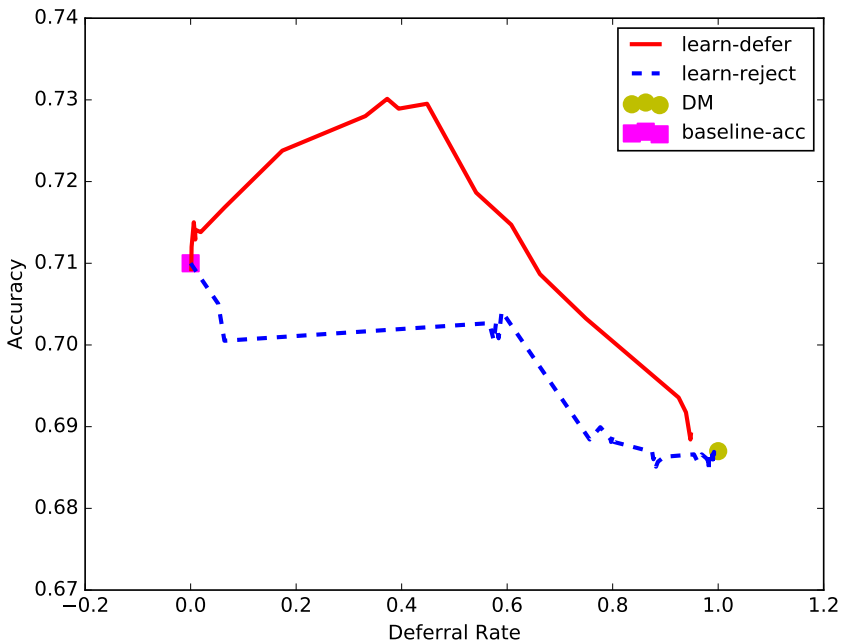


Figure 5: COMPAS dataset. Deferral rate-accuracy tradeoff. Higher is better.

Summary & Conclusions

"Learning to Defer" conclusions:

- Many ML models will be used as part of larger systems
- This should affect the way we train these models
- **Learning to defer** is a generalization of rejection learning; allows us to better optimize the behaviour of a system as a whole, for a wide range of objectives

Summary & Conclusions

"Learning to Defer" conclusions:

- Many ML models will be used as part of larger systems
- This should affect the way we train these models
- **Learning to defer** is a generalization of rejection learning; allows us to better optimize the behaviour of a system as a whole, for a wide range of objectives

General fair ML conclusions

- No fairness definition is perfect, but many are useful
- Understanding fairness in ML systems requires dealing with a lot of complexity
- In high-impact, high-stakes machine learning, we require different things from our models