Harvard CS 182: Artificial Intelligence November 16, 2022

Lecture 20: Fairness

Lecturer: Ariel Procaccia *Author: Amir Shanehsazzadeh*

1 Introduction

People originally thought that AI algorithms were fundamentally unbiased. However, people have realized that these algorithms and models learn and are trained on data that may encode existing biases in society.

1.1 Unfairness

1. AI algorithms thought to be unbiased.
2. AI algorithms are trained on data containing biases and so they amplify said biases.
3. Substantial evidence of discrimination by these algorithms.
4. For example, facial recognition algorithms from top companies like Microsoft and IBM were found to be much less accurate on darker females and most accurate on men with lighter skin.
5. Professor Cynthia Dwork (at Harvard) has played a key role in launching and shaping this field.

2 Fairness

2.1 Individual Fairness

First some notation:

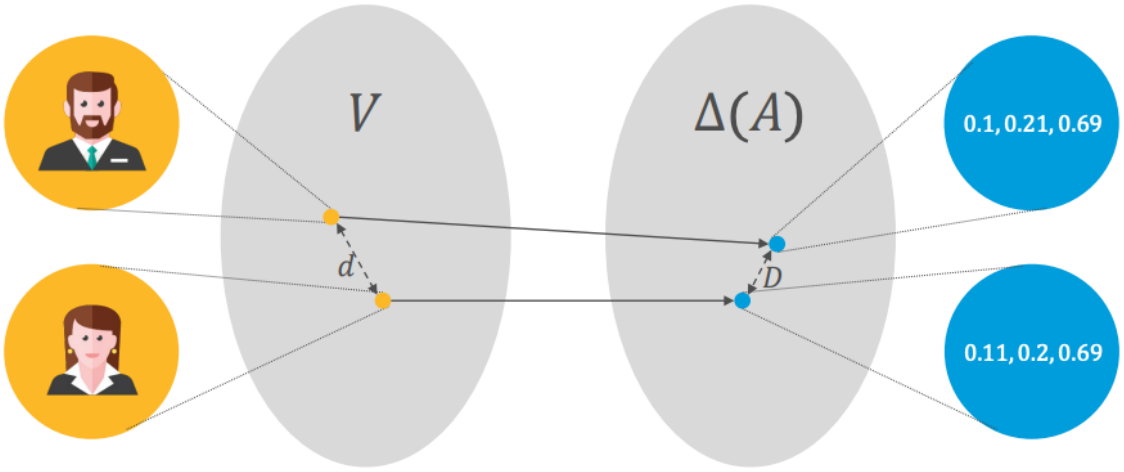
- Let V be a set of individuals and A a set of outcomes.
- Let $M : V \to \Delta(A)$ be a randomized classifier. Here $\Delta(A)$ is the set of distributions over outcomes.
- We require a metric on the individuals $d : V \times V \to \mathbb{R}^+$.

- We also require a metric on the distributions over outcomes $D : \Delta(A) \times \Delta(A) \to \mathbb{R}^+$.
- We say that M satisfies the *Lipschitz property* if for all $x, y \in V$ we have

$$D(M(x), M(y)) \leq d(x, y).$$

  Intuitively, the Lipschitz property implies that if two people are "close" then their distribution of outcomes will be "close" as well.

The image below demonstrates the distinction between the space of individuals and the space of their distributions of outcomes.



- An easy Lipschitz classifier is the constant classifier, M such that $M(x) = M(y)$ for all $x, y \in V$.
- The constant classifier is not interesting though because it results in $D \equiv 0$. Intuitively, our goal is not just to satisfy the Lipschitz property but to also optimize some objective or loss function of the form

$$L : V \times A \to \mathbb{R}^+.$$

  This function maps individuals and outcomes to a real number which is meant to have some real-world context. For example, Google's objective function is ad revenue (loss function is loss in ad revenue).

- This leads to the following optimization problem:

$$\min \sum_{x \in V} \sum_{a \in A} \mu_x(a) \cdot L(x, a) \tag{1}$$

$$\text{s.t. } \forall x, y \in V,\ D(\mu_x, \mu_y) \leq d(x, y), \tag{2}$$

$$\forall x \in V,\ \mu_x \in \Delta(A). \tag{3}$$

Metrics

- Various options for the metric D.
- Example: *total variation distance*, defined for distributions P and Q as

$$D_{tv}(P, Q) = \frac{1}{2} \sum_{a \in A} |P(a) - Q(a)|.$$

- $D = D_{tv}$ makes the optimization problem a linear program.
- Where does the similarity metric d come from? It's quite difficult to construct one and this is actually an obstacle. We ran a poll in class to discuss some ideas which included the L2 norms and cosine or PCA similarity. These ideas *might* work.

Envy-Freeness, Revisited

- Each $x \in V$ has a utility u_{xa} for each outcome $a \in A$. This is now an individual utility, for example the utility a user gets from seeing a particular ad.
- A randomized classifier M is envy free if and only if for all $x, y \in V$,

$$\mathbb{E}_{a \sim M(x)} [u_{xa}] \geq \mathbb{E}_{a \sim M(y)} [u_{xa}] .$$

- A completely different way of thinking about individual fairness.
- Envy-freeness isn't useful in situations where there are desirable and undesirable outcomes, like with bail decisions (where everyone wants low bail) or loans (where everyone wants a favorable loan). This is because envy-freeness in these cases amounts to the very stringent case of everyone getting the desired outcome with the same probability.

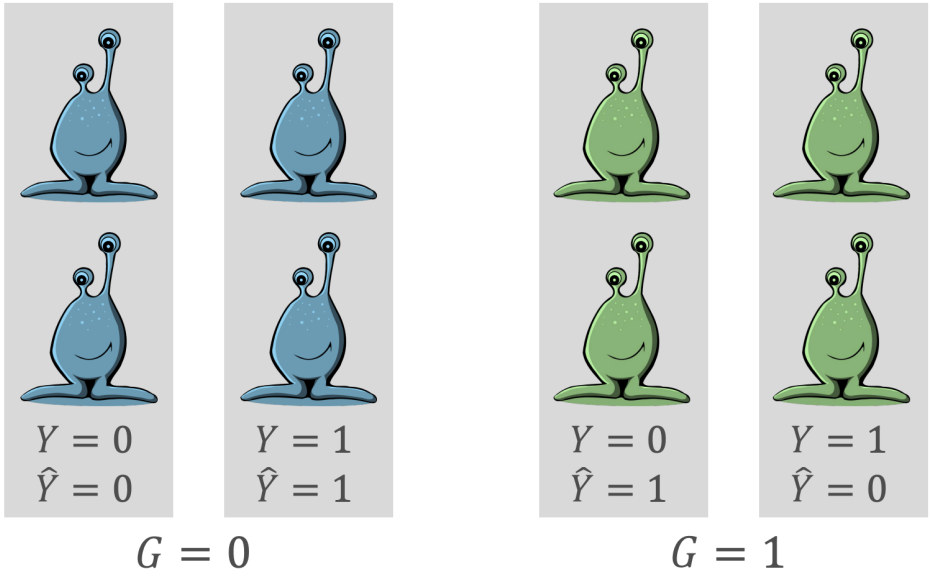
2.2 Group Fairness

- Suppose we are making a binary decision $\hat{Y} \in \{0, 1\}$ and there is a legally protected attributed $G \in \{0, 1\}$.
- *Demographic parity* is present if

$$\mathbb{P}(\hat{Y} = 1 | G = 0) = \mathbb{P}(\hat{Y} = 1 | G = 1).$$

- An issue with this is that to achieve demographic parity you could accept unqualified individuals when $G = 0$ and qualified individuals when $G = 1$. To see this consider the example below. Demographic parity is present between both groups but the accuracy

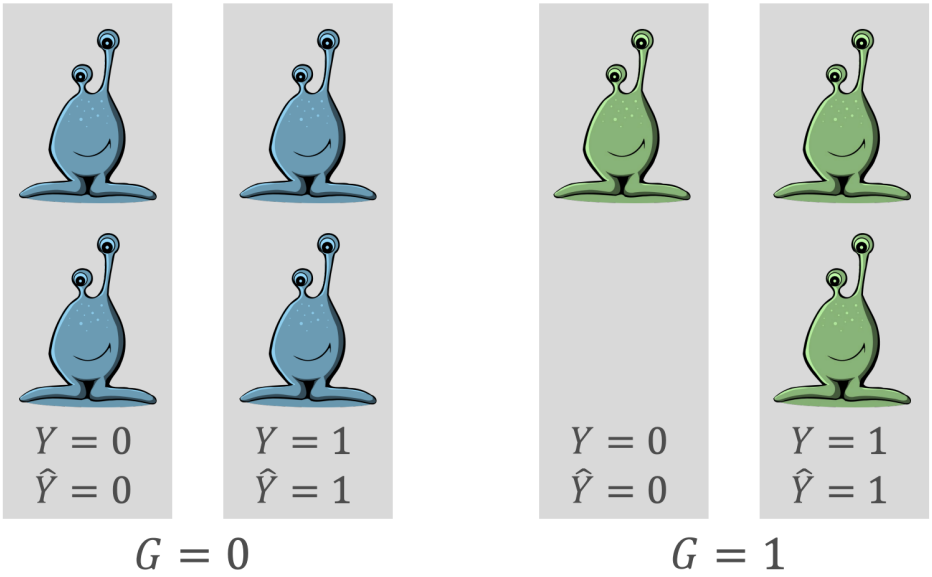
of the classifier is 1 for the blue aliens ($G = 0$) but 0 for the green aliens ($G = 1$).



- A more nuanced notion is *equalized odds*. $\hat{Y}$ satisfies equalized odds w.r.t. the protected attributed G if the groups have equal false positive rates and false negative rates, which is to say that for all $y, \hat{y} \in \{0, 1\}$ we have

$$\mathbb{P}(\hat{Y} = \hat{y} | G = 0, Y = y) = P(\hat{Y} = \hat{y} | G = 1, Y = y).$$

- Demographic parity and equalized odds are incomparable. The example above with the blue and green aliens shows that demographic parity does not imply equalized odds. Below is an example showing that equalized odds does not imply demographic parity.



2.3 Impossibility for Risk Scores

- Each person has a feature vector σ.
- p_σ denotes the fraction of people with feature vector σ and a true positive label.
- A person in group $G \in \{0, 1\}$ has a given probability of exhibiting σ.
- A *risk assignment* is as an assignment of people to bins, where each bin b is labeled with a score v_b seen as the probability of a positive label.
- *Calibration within groups* is achieved when for each group G and each bin b, the expected number of people from group G in b who belong to the positive class is a v_b fraction of the expected number of people from group G assigned to b.
- Can we have both calibration and equalized odds?
  - *Perfect prediction:* For each feature vector σ have $p_\sigma \in \{0, 1\}$.
    - Create two bins: one for people labeled as 0 probability $b = 0, v_0 = 0$ and another for people labeled with probability 1 $b = 1, v_1 = 1$. This is calibrated because any member of bin 0 has probability $v_0 = 0$ of being positively labeled and any member of bin 1 has probability $v_1 = 1$ of being positively labeled. Perfect prediction never makes a mistake so the false positive rates and false negative rates are all 0, meaning equalized odds is satisfied.
  - *Equal base rates:* The two groups have the same fraction of members in the positive class.
    - Let p be the equal base rate (the fraction of members in the positive class for any group). Create one bin with $v_b = p$. This is then calibrated. Using this mean probability prediction as a classifier creates equal false positive rates and equal false negatives rates and so equalized odds is satisfied. Note that this is a bad classifier because it just predicts a mean probability without considering feature vectors.
- *Theorem:* If a risk assignment satisfies calibration and equalized odds, the instance must allow for perfect prediction or have equal base rates.