Week 2

MORPH Algorithmic Fairness

Agenda:
Fairness in Classification

- Feedback
- This week in fairness
- Individual fairness
  - Aware & unaware
- Group fairness
- Looking ahead to week 3

This Week in Fairness


Bomze @tg_bomze · Jun 19

Face Depixelizer

Given a low-resolution input image, model generates high-resolution images that are perceptually realistic and downscale correctly.

😺GitHub: github.com/tg-bomze/Face-...
📙Colab: colab.research.google.com/github/tg-bomz...

P.S. Colab is based on the github.com/adamian98/pulse

547 4.4K 11.2K


Bomze @tg_bomze · Jun 20

With default settings, I got this result.

155 2.8K 12.9K

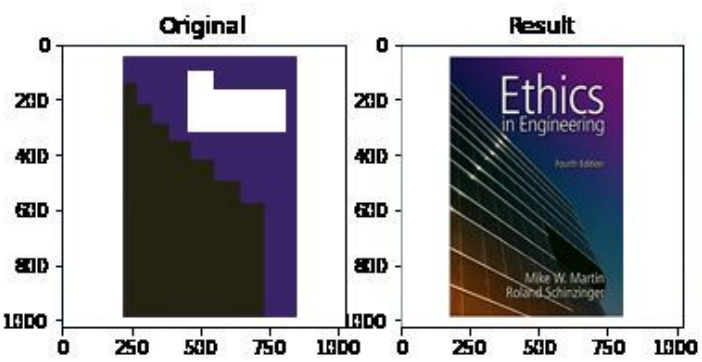
But then people realized...


Chicken3gg @Chicken3gg · Jun 20
Replying to @tg_bomze
🤔🤔🤔
Original
0
200
400
600
800
1000
0
250
500
750
1000
280
4.3K

Chicken3gg @Chicken3gg · Jun 20
Replying to @tg_bomze
🤔🤔🤔
Original
Result
0
200
400
600
800
1000
0
250
500
750
1000
280
4.3K
24.4K


Robert Osazuwa Ness @osazuwa · Jun 20
An image of @BarackObama getting upsampled into a white guy is floating around because it illustrates racial bias in #MachineLearning. Just in case you think it isn't real, it is, I got the code working locally. Here is me, and here is @AOC.
246
3.2K
7.1K


Robert Osazuwa Ness @osazuwa · Jun 20
This is @LucyLiu
Lucy Liu
22
266
1K


Original
Result
Ethics
in Engineering
Fourth Edition
Mike W. Martin
Roland Schinzinger
0
200
400
600
800
1000
0
250
500
750
1000

Someone tried a quick improvement(?)... so there are definitely many ways to improve and mitigate such cases.

What are your immediate reactions? thoughts?

If you’re interested in reading more:

News:
https://www.theverge.com/21298762/face-depixelizer-ai-machine-learning-tool-pulse-stylegan-obama-bias

Great piece summarizing the takeaways, discourse, and researchers’ response:

https://thegradient.pub/pulse-lessons/

Individual Fairness

What is “fairness”?
How do we define it?

Law

- Law is a natural starting point
- Two main notions:
  - **Disparate treatment**
    - Focuses on process
    - Unfair if decision was made w.r.t. to sensitive attribute
  - **Disparate impact**
    - Focuses on outcome
    - Unfair if outcome harms/benefits specific people

And many other definitions in…

- Philosophy — equality, distributive justice
- Economics — how do you divide and assign resources to people fairly
- etc.

→ What about in algorithms?

Fairness through Unawareness

- To not include sensitive attributes in consideration / data / algorithm at all
- Similar to idea of disparate treatment
- Problem: proxy features

- So let’s try fairness through awareness!

Key Idea:

Similar individuals should be treated similarly.

Treating Similar Individuals Similarly

- Binary classification algorithm
  - Positive or negative / 1 or 0 / accept or reject
- Any 2 individuals who are similar <u>with respect to a particular task</u> should be classified similarly

Dwork et al. (2006)

Similarity

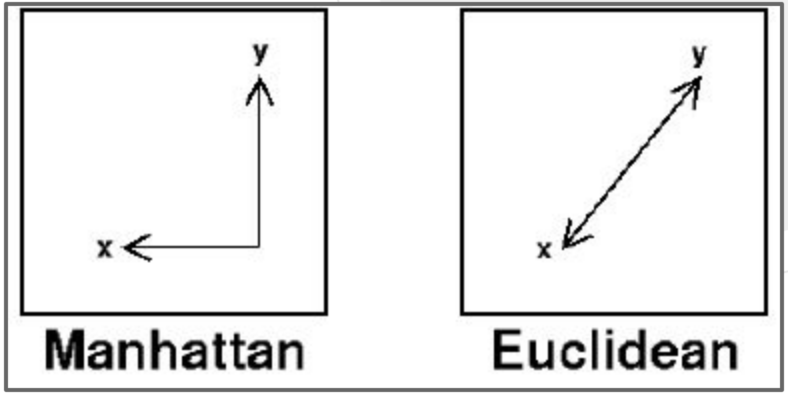
- How do we define **similarity**?
- We assume a **distance metric**
  - Similarity metric between individuals $d(x, y)$
    - # of features
    - Graphical distance (like word embeddings paper)
    - Many more ideas
  - Similarity measure between distributions of outcomes $D(x, y)$

Any 2 individuals who are similar with respect to a task should be classified similarly

Dwork et al. (2006)

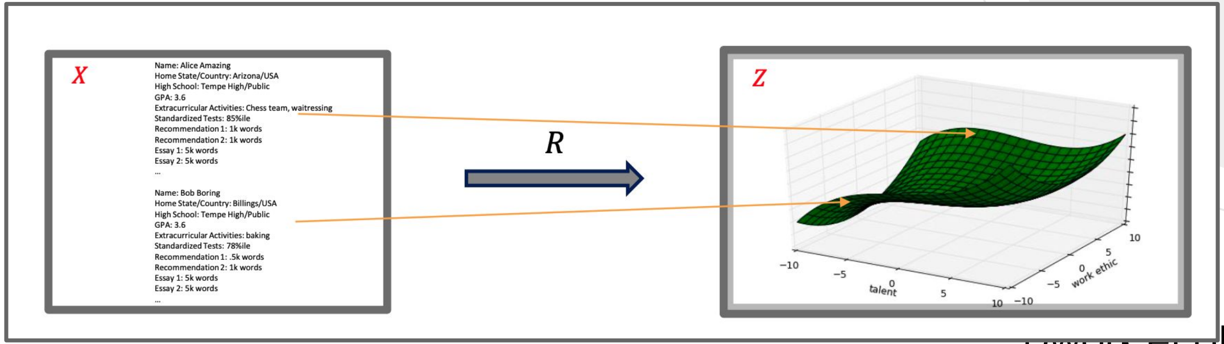
Distance Metric

- In math, a **distance function/metric** is a function that defines the distance between each pair of elements in a set
- Properties to uphold:
  - $d(x, y) = 0$ iff $x = y$
  - $d(x, y) = d(y, x)$
  - $d(x, y) \leq d(x, z) + d(z, y)$
- Many ways to choose or define a metric



Parameters

- Universe
- Outcome space
- A **representation mapping** between them:
  - Map from individuals to distributions over outcomes



Dwork et al. (2006)

Lipschitz Condition

"Any two individuals x, y that are at distance $d(x, y) \in [0, 1]$ map to distributions $M(x)$ and $M(y)$, respectively, such that the statistical distance between $M(x)$ and $M(y)$ is at most $d(x, y)$."

Big picture idea:

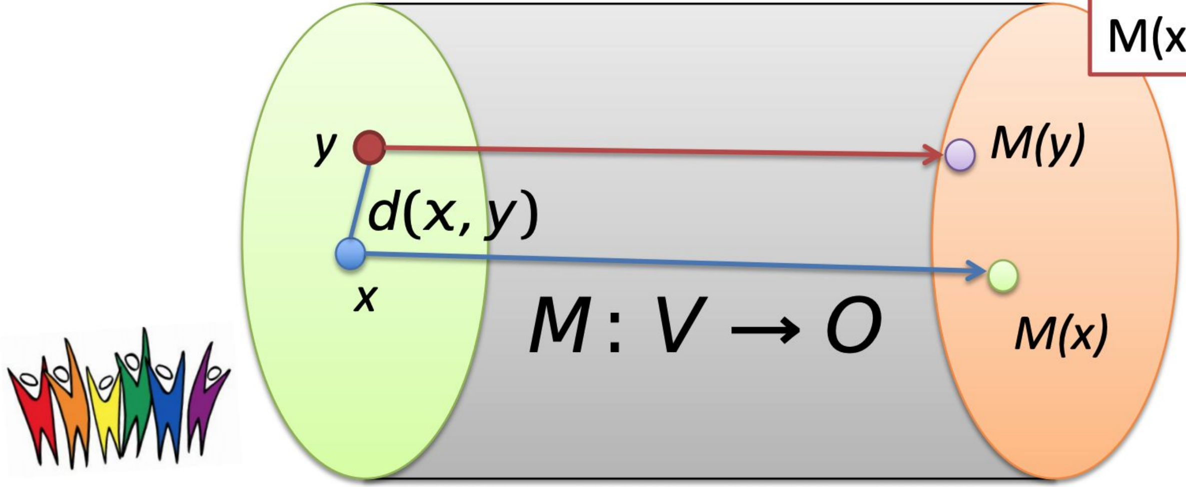
difference in outputs ≤ difference in inputs

$$\mathbf{D(M(x), M(y)) \leq d(x, y)}$$

Dwork et al. (2006)

Think of V as space
with metric d(x,y)
similar = small d(x,y)

How can we compare M(x) with M(y)?



V: Individuals

O: outcomes

Slide taken from [here]

Loss Function

- Some classifiers are more fair than others
- To capture this, we have the idea of **loss**
  - Minimizing loss = better and fairer!

- Thus, our goal becomes:
  Find a mapping from individuals to distributions over outcomes that minimizes expected loss subject to the Lipschitz condition.

Dwork et al. (2006)

Linear Program

A way to achieve best outcome (ex: max profit, min cost) with requirements or constraints represented by linear relationships

→ **Optimization**

Linear Program

$$
\begin{aligned}
\underset{x_1,x_2,x_3,x_4}{\text{maximize}}\ 3x_1 + 2x_2 - x_3 + x_4 & & s.t. \\
x_1 + 2x_2 + x_3 - x_4 &\leq 5 \\
x_1 + x_2 + x_3 + x_4 &= 1 \\
2x_1 + 4x_2 - x_3 - x_4 &\geq -1 \\
x_2, x_3, x_4 &\geq 0
\end{aligned}
$$

Linear Program

$$
\begin{aligned}
\max_{x_1,x_2,x_3,x_4} \quad & \boxed{3x_1 + 2x_2 - x_3 + x_4} && s.t. \\
x_1 + 2x_2 + x_3 - x_4 &\leq 5 \\
x_1 + x_2 + x_3 + x_4 &= 1 \\
2x_1 + 4x_2 - x_3 - x_4 &\geq -1 \\
x_2, x_3, x_4 &\geq 0
\end{aligned}
$$

Optimization Objective

Linear Program

$$\begin{aligned}
\underset{x_1,x_2,x_3,x_4}{\text{maximize}} \quad 3x_1 + 2x_2 - x_3 + x_4 & & & \quad s.t. \\
x_1 + 2x_2 + x_3 - x_4 & \leq & 5 & \\
x_1 + x_2 + x_3 + x_4 & = & 1 & \\
2x_1 + 4x_2 - x_3 - x_4 & \geq & -1 & \\
x_2, x_3, x_4 & \geq & 0 &
\end{aligned}$$

Constraints we must follow

Fairness LP

$$\begin{aligned} \mathrm{opt}(\mathcal{I}) \stackrel{\mathrm{def}}{=} \quad & \min_{\{\mu_x\}_{x\in V}} \quad \mathop{\mathbb{E}}_{x\sim V} \mathop{\mathbb{E}}_{a\sim\mu_x} L(x,a) \\ \text{subject to} \quad & \forall x, y \in V, : \quad D(\mu_x, \mu_y) \le d(x,y) \\ & \forall x \in V: \quad \mu_x \in \Delta(A) \end{aligned}$$

Dwork et al. (2006)

Fairness LP

$$\mathrm{opt}(\mathcal{I}) \stackrel{\text{def}}{=} \boxed{\min_{\{\mu_x\}_{x\in V}} \mathop{\mathbb{E}}_{x\sim V} \mathop{\mathbb{E}}_{a\sim\mu_x} L(x,a)}$$

$$\begin{aligned} \text{subject to} \quad & \forall x, y \in V, : \quad D(\mu_x, \mu_y) \leq d(x,y) \\ & \forall x \in V : \quad \mu_x \in \Delta(A) \end{aligned}$$

Optimization objective:
minimize expected loss over
all individuals & outcomes

Dwork et al. (2006)

Fairness LP

$$\mathrm{opt}(\mathcal{I}) \stackrel{\text{def}}{=} \min_{\{\mu_x\}_{x \in V}} \mathop{\mathbb{E}}_{x \sim V} \mathop{\mathbb{E}}_{a \sim \mu_x} L(x, a)$$

$$\text{subject to} \quad \boxed{\begin{aligned} &\forall x, y \in V, : \quad D(\mu_x, \mu_y) \le d(x, y) \\ &\forall x \in V : \quad \mu_x \in \Delta(A) \end{aligned}}$$

Constraints:

1) Lipschitz condition
2) Mapping outcomes are valid

Dwork et al. (2006)

To Sum Up: Individual Fairness

Definition 2.1. (Individual Fairness) A randomized classifier $C : U \to \Delta(O)$ is individually fair with respect to $D : \Delta(O) \times \Delta(O) \to [0, 1]$ and $d : U \times U \to [0, 1]$ if for every $u, v \in U$,

$$D(C(u), C(v)) \leq d(u, v).$$

Here, U is the universe of individuals being classified, O is the space of outcomes (which is simply $\{0, 1\}$ in binary classification tasks), D is a measure of similarity between distributions (eg. D_{TV}), and d is a given similarity metric between individuals.

Dwork et al. (2006)

Group Fairness

Statistical Parity

- % of people classified positive/negative matches the % demographic of general population
- Hire the same % of people in both groups

$$\boxed{\Pr[\text{outcome} \mid \text{person in } S] = \Pr[\text{outcome} \mid \text{person in } T]}$$

Statistical Parity

- Sometimes desirable, but can be abused! (How so?)
- Outcome might be “fair” but individuals can still be discriminated
  - In S, most skilled students study CS
  - In T, most skilled students study finance
  - Company hiring students interested in economics would hire wrong subset of S
- Self-fulfilling prophecy
  - When unqualified members of S are hired to justify future discrimination against S

Dwork et al. (2006)

Equalized Odds (Equality of Opportunity)

- Hiring analogy:
  - C = the decision made (hire or reject)
  - Y = the true standard of whether a person was qualified enough or not to be hired.
  - Ex: One can be rejected (C=0) but be capable enough for the job (Y=1).
- Hire equal % of individuals from the qualified subset of each group

$$\boxed{Pr_1[C = c \mid Y = y] = Pr_2[C = c \mid Y = y]}$$

Predictive Rate Parity

- Equal chance of success given hired/accepted
- Hiring analogy:
  - C = the decision made (hire or reject)
  - Y = the true standard of whether a person was qualified enough or not to be hired.
  - Ex: One can be rejected (C=0) but be capable enough for the job (Y=1).

$$\boxed{Pr_1[Y = y \mid C = c] = Pr_2[Y = y \mid C = c]}$$

Which group fairness notion do you agree with the most?