Societal impact & approaches

ex: medical diagnosis

"Weapons of maths destruction" by Cathy O'Neil

- Black-box software (at large scale) for important matters
  - hiring
  - hiring
  - (big school)
  - loans (banks)
  - Schuying arbitrarily / stochastic
  - no feedback / questioning possible
  - large scale => arbitrariness nightmare
  - illegal criteria / proxy

ex: COMPAS: redlining prediction
  - high bias
  - self-fulfilling prophecy: police patrols

think twice about the impact of your algorithms before deploying them

be responsible & careful

⇒ AI ethics = Modified declaration for responsible AI

"FAT"-ML: Fairness, Accountability, Transparency

Interpretability by design: "X-AI"

Explainable

Causality

\[
\text{age} \rightarrow \text{gender} \rightarrow \text{sickle}
\]

\[
x = \Delta \text{File}
\]

\[
x^2 + y^2 = 1
\]

\[
y^2 = (\sqrt{x})^2
\]

\[
\text{independent}
\]

\[
\text{dependent but not correlated}
\]

\[
E[(x-Y)(y-y)] \Rightarrow \text{Correlated}
\]
Issues related to datasets

Dataset poisons

car  

bag  

Equation for car

dog  

easy to detect for NLP pipeline  

caus  

depth  

Fairness

Introduction

Words/view:

words ∈ dictionary

Definition 1:
- Simplicity: unawareness
  - do not include sensitive features
- Awareness
  [Cynthia Dwork et al., 2012]

F: stochastic

\( \phi_D(D(x), D(x')) \leq d(x, x') \)

\( S \) need relevant metric

MMD

Kullback-Leibler (KL) - optimal transport
Def 2: Equal opportunity / e-fairness

- Input: (x, A)
- Sensitive attribute (e.g., ethnicity)
- Output: \( \hat{y} \): predicted
- y: being hired

Equal opportunity:

\[ P(\hat{y} = 1 | A = a, Y = 1) = P(\hat{y} = 1 | A = a', Y = 1) \]

Group-based def:

Statistics

Relax: e-fairness:

\[ |P(\hat{y} = 1 | A = a, Y = 1) - P(\hat{y} = 1 | A = a', Y = 1)| < \epsilon \]

Def 3: Group-based

Same distibution of outputs | errors

Algorithms

Fairness / accuracy trade-off

[After training]

+a 

Bias 

decies at post-processing

(easy for group-based fairness)

1 threshold per group

Need to know the group

i.e., the sensitive attribute

[During training] while optimizing

\[ \min_{\theta} \mathcal{L}(D, \theta) \]

Adversarial training

- \( g \): sensitive task

- \( g \): make \( g \) fail

- criterion
Information Bottleneck

\[ f(x;a) \rightarrow (x,z) \rightarrow \max I(x;2) \text{ mutual information} \]

Differential privacy

Why care about privacy?

- Netflix prize 2007 movies
- Cross
- JMYB

Electricity consumption

- 67% of US citizens: identifiable from birthdate, gender, zip code

Group insurance is voter roll database

If no data sharing

Queries on a database

\[ \text{ask statistical questions} \]

Ex: wages in company

\[ \begin{array}{c}
\text{from join} \\
N \text{ persons} \\
\text{N+1 persons} \\
\text{average } s' \end{array} \]

\[ (s') x N \]

Principle:

1) add noise
2) bound the # of requests
\[ \text{Algorithm: } A \]
\[ \text{Dataset: } D_1 \]
\[ \epsilon = D_1 + \text{one element} = D_2 \]

A has \((\epsilon, \delta)\)-privacy if:

For subsets \(S\) of \(\text{Im}(A)\), for datasets \(D_1, D_2\) differing only by one element,

\[ p(A(D_1) \in S) \leq e^\epsilon p(A(D_2) \in S) + \delta \]

\[ (\forall \delta > 0) \]

---

Federated learning

any hospitals & medical task

No data transfer
- transfer parameters (encrypted)

\( g_1, g_2, \ldots, g_n \) from a generator model

\[ \text{DNA} \]

\[ \text{AATTS} \]

\[ \text{min}_x d(x', x) \in \mathbb{R}^d \text{ real} \]

\[ \text{Generated} \]