Medical Data Privacy and Ethics in the Age of Artificial Intelligence

Lecture 6: Algorithmic Fairness

Zhiyu Wan, PhD (wanzhy@shanghaitech.edu.cn)
Assistant Professor of Biomedical Engineering
ShanghaiTech University

March 14, 2025

Learning Objectives of This Lecture

- Understand 6 algorithmic fairness metrics
- Know 3 types of bias mitigation methods

Biases in Computational Medicine Studies

- Examples
  - Associations between Framingham risk factors and cardiovascular events are significantly different across ethnic groups.
  - Video stream analysis algorithms are challenging for Asian individuals.
  - Undiagnosed silent hypoxemia, detected from pulse oximetry, occurred three times in Black people due to their dark skin.

Xu J, Xiao Y, Wang WH, Ning Y, Shenkman EA, Bian J, Wang F. Algorithmic fairness in computational medicine. EBioMedicine. 2022 Oct 1;84.

Computational bias

- Data bias
  - Patients of low socioeconomic status may have limited access to health care
  - Sampling bias (Selection bias)
    - Melanoma detection algorithms based on classification of skin lesion images may perform poorly on dark-pigmented skin if the training images contain predominantly lighter skin.
    - Face2Gene, a machine learning algorithm to recognize Down syndrome based on facial images, performed much better in Caucasian than in African.
  - Allocation bias
    - Emulate clinical trials with real world data such as EHRs

Computational bias

- Data bias
  - Attrition bias
    - It can occur if there are systematic differences in the way different groups of participants are recruited or are dropped from a study.
  - Publication bias
    - It occurs when the decision to publish a study depends on its own results.
    - It makes people overestimate the effectiveness of specific treatments or models.
- Measurement bias
  - When the data are labeled inconsistently
  - When Diseases are collected or measured inaccurately

Computational bias

- Measurement bias
  - When the data are labeled inconsistently
  - When Diseases are collected or measured inaccurately
  - Response bias
    - When respondents tend to give inaccurate or even wrong answers on self-reported questions.
    - Example 1: People might tend to always rate themselves favorably or feel pressured to provide socially acceptable answers.
    - Example 2: Misleading questions can lead to biased answers.
    - Example 3: Demographic groups who are willing to answer survey questions are sometimes different from those who are not.
  - Algorithmic bias

A case study

- Build an alerting algorithm in ICU setting (e.g., for developing sepsis)
- Machine learning algorithm based on the patient's EHR and the patient's race.
- Consider only two demographic groups (e.g., Black or white)
  - A in {0, 1}: Protected attribute
  - X: Observable attributes
  - U: Relevant latent attributes not observed
  - Y in {0, 1}: Outcome to be predicted
  - $\hat{Y}$ in {0,1}: Prediction

Fairness metrics

- Unawareness
  - No protected attribute A is explicitly used in the decision-making
  - A: Protected attribute (e.g., race)
  - $\hat{Y}=f(X, A)=f(X)$

- Demographic Parity
  - The outcomes must be equal
  - $P(\hat{Y}=y|A=0) = P(\hat{Y}=y|A=1)$, y in {0,1}
  - P: Proportion or Percentage

Fairness metrics

- Equalized Odds
  - Different groups deal with similar odds
  - $P(\hat{Y}=1|A=0, Y=y) = P(\hat{Y}=1|A=1, Y=y)$, y in {0,1}
  - The true positive rates (of those who actually developed sepsis, how many were correctly predicted to be positive) and false positive rates in both demographic groups are equal

- Equal Opportunity
  - The true positive rates in both groups are equal.
  - $P(\hat{Y}=1|A=0, Y=1) = P(\hat{Y}=1|A=1, Y=1)$

Fairness metrics

- Individual Fairness
  - Similar individuals have similar predictions.
  - Individuals *i* and *j*, if distance $d(i, j)$ is small, then $|\hat{Y}(i) - \hat{Y}(j)|$ is small.

- Counterfactual Fairness
  - The predicted outcome does not change if a patient from one demographic group is assigned to the other demographic group
  - $P(\hat{Y} = y | A=0, X=x) = P(\hat{Y} = y | A=1, X=x)$ for all *x* and *y*
  - Counterfactual reasoning may negatively affect the process of causality identification (e.g., *Y* is dependent on *A*)

Fairness metrics (additional)

- Average odds difference (AOD)
  - AOD=1/2(Average TPR Difference + Average FPR Difference)

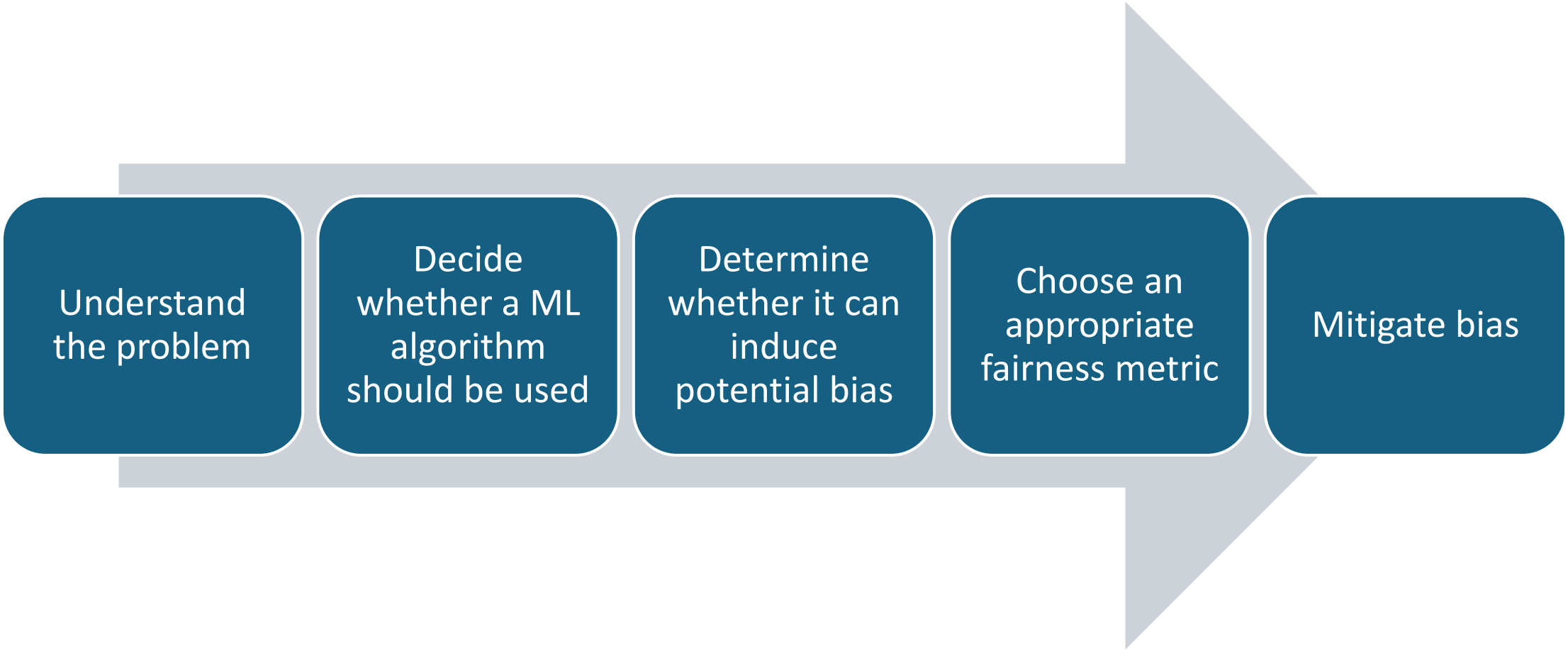
    = ½(|$P(\hat{Y}$ =1|A=0, Y=1) - P($\hat{Y}$ =1|A=1, Y=1)|

    +|$P(\hat{Y}$ =1|A=0, Y=0) - $P(\hat{Y}$ =1|A=1, Y=0)|)

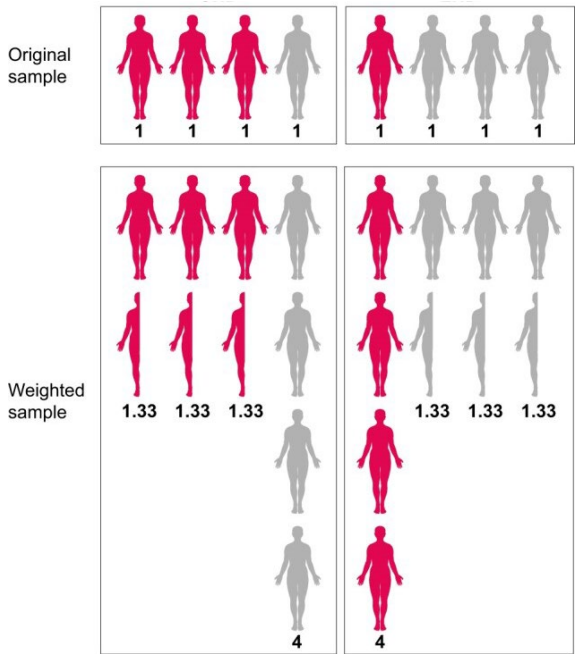
- Disparate impact (DI)
  - $DI_{ij} = \min\left(\frac{P\left(\hat{Y}=1\middle|Y=1,A=i\right)}{P\left(\hat{Y}=1\middle|Y=1,A=j\right)}, \frac{P\left(\hat{Y}=1\middle|Y=1,A=j\right)}{P\left(\hat{Y}=1\middle|Y=1,A=i\right)}\right), \mathrm{i},\mathrm{j}=0,1,\mathrm{i}\neq\mathrm{j}$
  - $DI = \max DI_{ij}$

Feldman M, Friedler SA, Moeller J, Scheidegger C, Venkatasubramanian S. Certifying and removing disparate impact. In proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining 2015 Aug 10 (pp. 259-268).

Fairness-aware problem solving



Bias mitigation



- Pre-processing
  - Choice of sampling (Resampling)
    - Ensure that all demographic groups are properly and proportionately represented in the training dataset
    - Under-sample the majority group or oversample the minority group
    - Collecting more data is better
  - Reweighting
    - Inverse propensity score weighting
    - w(1,1)=1/P(A=1|Y=1)=1.25
    - w(1,0)=1/P(A=1|Y=0)=1.5
    - w(0,1)=1/P(A=0|Y=1)=5
    - w(0,0)=1/P(A=0|Y=0)=3

	Case(1)	Control(0)
White(1)	80	200
Black(0)	20	100

⇩

	Case(1)	Control(0)
White(1)	100	300
Black(0)	100	300

Bias mitigation

- In-processing
  - Prejudice remover
    - Make predictions be independent from the protected attribute
  - Adversarial learning

Loss function: prediction error Loss function: equalized odds bias

  - Interpretable models: reveals biased decision-making process
  - Independent learning
    - Trains a model for each protected group → Reduces the performance
    - Transfer learning → Align the sample distributions

Bias mitigation

- Post-processing
  - Equalized odds post-processing
    - Changing output labels to achieve the equalized odds objective
  - Adjust the risk scores of the instances in the disadvantaged group
  - Adjust the ranking order of the samples across different protected groups
  - Causal analysis approach

Popular software libraries

Project	Developer	Year	Description	Publication
FairMLHealth	KenSci	2020	Tools and tutorials for evaluating bias in healthcare machine learning.	GitHub
AIF360	IBM	2019	Fairness metrics for datasets and machine learning algorithms, interpretation of the metrics, and approaches for reducing bias in datasets and models. It is available in both Python and R.	IBM Journal of Research and Development
Fairlean	Microsoft	2020	A Python package to evaluate fairness and mitigate any observed inequities.	Microsoft Tech
Fairness-comparison	Sorelle et al.	2019	Compare fairness-aware machine learning techniques. It aims to facilitate benchmarking of fairness-aware machine learning algorithms.	ACM FAccT
MEASURES	Cardoso et al.	2019	A benchmark framework for assessing discrimination-aware models.	AAAI/ACM CAES
Fairness Indicators	Google	2024	A suite of tools built on top of TensorFlow Model Analysis that enable regular evaluation of fairness metrics in product pipelines.	Google Colab
ML-fairness-gym	Google	2020	A general framework for studying and exploring long-term equity effects in carefully constructed simulation scenarios where learning subjects interact with the environment over time.	Google Blog
Themis-ml	Niels Bantilan	2017	A Python library built on top of pandas and sklearn that implements fairness-aware machine learning algorithms.	J. of Technology in Human Services
FairML	Julius Adebayo	2017	A Python toolkit for auditing machine learning model deviations.	Github

Readings for the Next Week

- None
- Optional
  - ❑ 1. Molnar, Christoph. Interpretable machine learning. 2020. (Ch. 5)
    - https://christophm.github.io/interpretable-ml-book/
  - ❑ 2. Lundberg, S. M., & Lee, S. I. A unified approach to interpreting model predictions. NeurIPS. 2017 (Original SHAP paper)

Feedback Survey

- One thing you learned or felt was valuable from today's class & reading

- Muddiest point: what, if anything, feels unclear, confusing or "muddy"

- https://www.wjx.cn/vm/hX0mIro.aspx