CS 109a Repository Entry
Embedded EthiCS @ Harvard Teaching Lab

Overview

Course:	CS 109A: Data Science 1: Introduction to Data Science
Course Level:	Upper-level undergraduate
Course Description:	"Data Science 1 is the first half of a one-year introduction to data science. The course will focus on the analysis of messy, real life data to perform predictions using statistical and machine learning methods. Material covered will integrate the five key facets of an investigation using data: (1) data collection - data wrangling, cleaning, and sampling to get a suitable data set; (2) data management - accessing data quickly and reliably; (3) exploratory data analysis – generating hypotheses and building intuition; (4) prediction or statistical learning; and (5) communication – summarizing results through visualization, stories, and interpretable summaries. Part one of a two-part series. The curriculum for this course builds throughout the academic year. Students are strongly encouraged to enroll in both the fall and spring course within the same academic year."[1]
Module Topic:	Bias in Machine Learning
Module Author:	Michael Pope
Semesters Taught:	Fall 2022
Tags:	Algorithms [CS], machine learning [CS], algorithmic design [CS], prediction [CS], calibration [CS], proxy variable [CS], label [CS], training data [CS], model cards [CS] bias [phil], discrimination [phil], fairness [phil], feedback loop [phil], luck-egalitarianism [phil]
Module Overview:	This module explores the ways unfair bias enters into machine-learning (ML) and how we can prevent it. Following an overview of possible ways bias can enter ML algorithms, the module introduces a case study involving a racially biased healthcare risk-prediction algorithm. In discussion groups, students practice identifying the ways in which the healthcare algorithm is unfair. To assist this exercise, they are introduced to the promises and limitations of luck-egalitarian approaches to equality. Next, the module explores how relabelling data can produce fairer outcomes. To close, the module provides ethical questions that students should ask during the development, design, and deployment of ML algorithms.
Connection to Course Material:	Students in this course learn to analyze real data and perform predictions using statistical and machine learning methods. This module shows how ethical questions about unfair bias can arise in using those methods.

The module's topic directly connects with the course's technical material. The topic introduces students to an active and growing research program on algorithmic bias. Focusing on racial bias in ML algorithms is especially timely, since the healthcare algorithm and others like it are widely used in public and private sectors today.

[1] Harvard course catalog listing: Link.

Goals

Module Goals:

By the end of the module, students should be able to:

1. Identify ways unfair bias can enter the data, design, and deployment of ML algorithms.
2. Describe how features of a major healthcare risk-prediction algorithm result in discrimination.
3. Critically assess luck-egalitarianism to formulate reasons this algorithm could be unfair.
4. Cultivate an awareness of ethical questions for designing and deploying ML algorithms.

Key Philosophical Questions:

1. How can unfair bias arise at different stages in the development of ML algorithms?
2. *Why* is it unfair that Black patients are sicker, on average, than White patients, despite having the same risk scores?
3. What ethical questions are relevant for addressing unfair bias in the development of ML algorithms?

Q1: Before focusing on a specific example and considerations of fairness, it is important to see that bias can enter ML algorithms in myriad ways. For instance, ML algorithms can be biased by unrepresentative data, or by predicting an inappropriate target variable.

Q2: Having suggested that the algorithm is unfairly biased, this question pushes students to explain why the bias is unfair. One plausible response is that unfairness arises because the algorithm disadvantages Black patients through factors that are beyond their control. Yet, by critically assessing this response, we see how some inequalities (including health inequalities) could be unfair and should be compensated, despite not resulting from anyone's choice.

Q3: Formulating ethical questions for a particular situation arises from recognizing ethical features of data, design, and deployment in machine learning. Only after identifying the questions can we develop solutions.

Materials

Key Philosophical Concepts:

- Bias
- Fairness/unfairness
- Luck-egalitarianism

Bias: The module distinguishes bias from unfair bias at the outset. Bias, in the sense of discrimination, can

		be further distinguished from statistical bias. *Fairness*: Having considered some ways that bias can enter machine learning, fairness is introduced with the case study. Students are asked to identify bias in the case study and assess whether it is fair or unfair. *Luck-egalitarianism*: To assist students' evaluation of fairness in the case study, the module presents luck-egalitarianism as an account of what renders some inequalities unfair. The view is then problematized by considering how contextual factors can influence choice, including influential insights from Elizabeth Anderson's article, "What Is the Point of Equality?"
Assigned Readings:	• Ledford, H.. (2019). "Millions of Black People Affected by Racial Bias in Health-Care Algorithms." *Nature NEWS.* • Barocas, S., Hardt, M., and Narayanan, A. "Introduction" to *Fairness and Machine Learning*.	Ledford offers a succinct and accessible overview of the study published in *Science* that uncovered racial bias in the healthcare algorithm discussed in the module. Barocas, Hard, and Narayanan provide an overview of machine learning, including clear examples wherein algorithms are unfairly biased. They discuss how bias can arise in machine learning and review some influential measures of fairness.

Implementation

Class Agenda:

1. Introduction: overview of ML algorithms, why we use them to improve human decision making, and terminology for bias and fairness.
2. Discussion of how bias arises in ML algorithms, namely problem formulation, dataset construction, and deployment.
3. Case Study: review of the ways an influential healthcare risk-prediction algorithm is racially biased.
4. Discussion of fairness.

	5. Luck egalitarianism as possible explanation and objections. 6. Data science intervention to address bias: relabelling. 7. Conclusion: identifying ethical dimensions of ML algorithms through questions at each stage of development.	
Sample Class Activity:	In small groups, students discussed the following questions, before sharing insights from their discussion in a subsequent large-group discussion: 1. Why might someone think that the outcome of the algorithm—namely that Black patients receive lower risk scores than White patients—is *fair* or *unfair*? 2. Consider that the algorithm predicts healthcare costs as a proxy for need. How does this technical choice affect the fairness of the algorithm?	This course is a large, undergraduate course, limiting the types of plausible discussion formats. Moreover, the case study material is sensitive. With that in mind, students are not asked to state their own opinions about the fairness of the algorithm. Rather, they are asked to think about the reasons and arguments one could have for thinking that the algorithm is fair or unfair. The second question turns students' attention to the ways that technical choices result in more or less fair outcomes.
Module Assignment:	Following the class session, the module content was incorporated into the ethics section of a quiz. Students were asked the following questions: 1. What is the relevance of free choice for evaluating the fairness of inequalities in predicting healthcare needs? 2. In the healthcare algorithm we discussed in the Embedded EthiCS module, Black patients have significantly lower health outcomes when compared with White patients, despite similar risk scores. How can relabelling data produce less biased outcomes? *Why* would those outcomes be fair or unfair? 3. Imagine that the healthcare algorithm discussed in class is used to determine health insurance premiums. Suppose that patients that receive higher risk scores are charged higher premiums. What are two *ethical* questions you would ask to assess whether this use of the algorithm is fair?	Q1: This question draws on the discussion of luck-egalitarianism in class. It is designed to invite students to reflect on the relationship of stakeholder choice, structural/contextual features that constrain choices, and the impacts of ML algorithms. Q2: This question has two parts. The first part asks students to review the ways that interventions on data use can have ethical impacts on ML systems. The second part asks students to formulate reasons for the fairness or unfairness of those impacts. Q3: This question challenges students to reflect on how the ML algorithm discussed in class could be used in a new context and, in turn, to formulate two questions about the ethical impacts of that use.
Lessons Learned:	Student responses to this module were overwhelmingly positive, especially the ways luck-egalitarianism oriented discussions of fairness.	

Students also appreciated the overview of the ways that bias can affect the fairness of ML algorithms throughout their development.

Pedagogical lessons:

- When situated in technical conversations, students appreciate the ways philosophical concepts can clarify fuzzy, but related ethical questions (e.g., “Why is this algorithm fair or unfair?”).
- A real-world case study, especially one that is currently in use, provides engaging and memorable content for student discussions. Moreover, looking at the ways the case study prompted hospitals to update the algorithm provides an example of how to address concerns about unfair bias in the real world.
- This module was designed for a class of approximately 250 students in a lecture-style classroom, with approximately 85 additional students participating online. While small-group discussions facilitated large-group participation, it is useful to have students determine their groups at the outset of the class meeting.