Overview

Course:	CS 181: Machine Learning
Course Level:	Upper-level Undergraduate
Course Description:	"Introduction to machine learning, providing a probabilistic view on artificial intelligence and reasoning under uncertainty. Topics include: supervised learning, ensemble methods and boosting, neural networks, support vector machines, kernel methods, clustering and unsupervised learning, maximum likelihood, graphical models, hidden Markov models, inference methods, and computational learning theory. Students should feel comfortable with multivariate calculus, linear algebra, probability theory, and complexity theory. Students will be required to produce non-trivial programs in Python."[1]
Module Topic:	Bias in Machine Learning Design
Module Author:	Michael Pope
Semesters Taught:	Spring 2023
Tags:	Bias [Phil], Trust [Phil], Fairness [Phil], Design [CS], Machine learning [CS], Calibration [CS], Prediction [CS]
Module Overview:	Through a case of racial bias in healthcare, this module investigates how seemingly innocuous technical decisions and straightforward data labels can result in biased outcomes. Students discuss important choice points in design and generate possible alternatives, before examining how systems can be improved once bias is discovered.
Connection to Course Material:	This module connects to course material in two ways. First, prior to the module, students examine prediction problems in machine learning. Likewise, discussion activities included choice points similar to those found in problem sets. Second, the module interfaced with the courses focus on ML's broader impact, showing how technical design decisions can manifest those impacts. The healthcare case presents a system exactly like those studied and designed in this course. It also provides opportunities to discuss conceptual, practical, and ethical considerations for deploying ML systems in complex contexts.

Goals

Module Goals:	1. Investigate how data labels and reasonable design choices can result in biased outcomes. 2. Identify potential sources of bias and brainstorm alternative design choices, focusing on alternative predictor variables. 3. Examine how technical choices can improve outcomes and address bias.
Key Philosophical Questions:	1. How do developers' intentions in designing a system relate to negative impacts of that system? 2. Can developers be morally responsible for outcomes, even if design choices are not the sole cause of those outcomes? Q1: A crucial part of this module's primary case study is to show students how reasonable design decisions and the best intentions

[1] Link to Harvard course catalog. Link to course site.

		can still result in undesirable ethical and social impacts. Q2: Having discussed the impacts of design decisions, the module examines how designs can (and should be) sensitive to contextual factors, including the populations and communities impacted, their histories, and base rates of public trust.

Materials

Key Philosophical Concepts:	● Bias ● Prediction ● Trust ● Disparate treatment and disparate impact	*Bias*: The module focuses on the introduction of bias through system design and deployment. Bias, in the sense of discrimination, is discussed in connection with fairness and statistical bias. *Prediction*: In connection with bias, this module investigates the ways model predictions can mask salient ethical and social issues. *Trust*: Uptake is a crucial component of a system's success and depends on trust. The module emphasizes the vulnerabilities that accompany trust and trust's role in deploying and improving system design.
Assigned Readings:	● Ledford, H.. (2019). "Millions of Black People Affected by Racial Bias in Health-Care Algorithms." Nature NEWS.	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.

Implementation

Class Agenda:	1. Introduction: Discussion of US healthcare system and the promise of ML 2. Case Study 1: Bias in Healthcare Algorithms 3. Discussion: Improving performance and ethical design 4. Case Study 2: Identifying potential impacts through PredPol 5. Debrief and final discussion	Framing how the biased outcome arises in the first case study is crucial for success in this module. Discussions and engagement with other case studies relies on students seeing how innocuous and reasonable design decisions can fail to meet *technical* and *ethical* goals.

Sample Class Activity:	Prior to discussing the outcome of the healthcare case, students engage in a think-pair-share exercise to assess the information sources that could be useful for distributing health resources to those with the greatest need. Specifically, they discuss patient claims history, electronic health records, and the US Department of Health and Human Services list of determinants of health, which include economic stability, education, health care access, housing, and social context. Given the scope of these factors, discussion prompts ask students to consider *why* certain information is relevant for predicting need.	This exercise is designed to achieve two goals. First, the discussion invites students to reflect on possibilities at the design stage *before* knowing the outcomes of deployment. Second, reflecting on possible information sources for predicting health need allows students to formulate reasons for or against using a particular information source, revealing potential impacts and tradeoffs (for example, between patient privacy and predictive accuracy).
Module Assignment:	In a problem set following the module, students are asked to assess model-assisted decision making in high-stakes situations. Questions in the section relate to a fictional case study involving a pharmaceutical company who has requested a model for testing a drug intended to treat a devastating disease. Students consider possible predictors and features of different possible regression models, including efficiency and uncertainty. To conclude, students revisit the data collection process during clinical trials and consider possible confounding factors in participant recruitment protocols.	Every problem set in this course includes a section on the broader impacts of the relevant topic. This allows the problem set for this module to cover more material than would have been possible if students were not accustomed with formulating explanations for design decisions and considering possible alternatives.
Lessons Learned:	Feedback on this module was overwhelmingly positive. In particular, students reported appreciating the links in the module between technical decisions and ethically-relevant outcomes. Since the first case study fostered engaged and lengthy discussions, one area that deserves particular care is ensuring enough time to examine the final case study, PredPol, in a sufficiently sensitive way. An alternative format for this module, when time is restricted, would be to focus on the first case study only.	