Overview

Course:	CS181: Machine Learning
Course Level:	Upper-level undergraduate
Course Description:	"CS 181 provides a broad and rigorous introduction to machine learning, probabilistic reasoning and decision making in uncertain environments. We will discuss the motivations behind common machine learning algorithms, and the properties that determine whether or not they will work well for a particular task. You will derive the mathematical underpinnings for many common methods, as well as apply machine learning to challenges with real data. In doing so, our goal is that you gain a strong conceptual understanding of machine learning methods that can empower you to pursue future theoretical and practical directions. 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, reinforcement learning."[1]
Module Topic:	Discrimination
Module Author:	Lyndal Grant
Semesters Taught:	Spring 2021
Tags:	discrimination [phil], disparate treatment and disparate impact [phil], procedural and substantive fairness [phil], harm [phil], fairness [both], bias [both], algorithms [CS], machine learning [CS], predictive accuracy [both]
Module Overview:	Machine learning systems have incredible potential to both overcome biases in human decision-making, and to reinforce and further entrench such biases. In this module, we discuss what discrimination is, focusing on two kinds of discrimination recognized under United States law (disparate treatment and disparate impact discrimination). We then explore how discrimination can arise at various stages of the machine learning process. Finally, we explore the difference between accuracy and fairness, and consider the possibility that there are requirements of fairness that go beyond epistemic requirements of accuracy.
Connection to Course Material:	In this course, students learn how to tackle real-world decision-making problems using machine learning. This module builds on that material by asking students to consider how the decisions they make in the design of machine learning systems might involve different forms of discrimination. We consider what "biased data" might be, and how the outputs of the machine learning process (predictions and the decisions made on their basis) might be discriminatory.

Goals

[1] https://harvard-ml-courses.github.io/cs181-web/

Module Goals:	1. Familiarize students with the distinction between disparate treatment and disparate impact discrimination, along with the difficulties of applying disparate treatment standards in the context of machine learning. 2. Have students identify decision-making contexts where taking protected social group membership into account seems morally unobjectionable and those where it is discriminatory. 3. Have students think deeply about the connection between accuracy in predictions and fairness in decisions. 4. Have students reflect on the roles and responsibilities of individuals involved in each stage of the machine learning process.	
Key Philosophical Questions:	1. What is discrimination, and why is it wrong? 2. When is it wrong to make predictions/decisions on the basis of (protected) social group membership? 3. Are decisions made on the basis of accurate predictions necessarily fair?	The question "What's wrong with discrimination?" can appear so abstract as to be unhelpful. It is therefore helpful to start the class with a few clear cases of discrimination, and a few cases where it is less clear that wrongful discrimination has occurred. This helps students begin to get a feel for the theoretical question.

Materials

Key Philosophical Concepts:	● Disparate treatment and disparate impact ● Substantive and procedural fairness ● Predictive accuracy	The distinction between substantive and procedural fairness is used to categorize different kinds of explanations for what is wrong with wrongful discrimination. Students will likely be already familiar with the notion of predictive accuracy, but may assume that our reasons to avoid discrimination are just reasons to avoid subjecting people to the results of inaccurate predictions. It is therefore helpful to provide students with an example where an accurate ML prediction produces a seemingly unfair decision.
Assigned Readings:	● Salon Barocas and Andrew D. Selbst, "Big Data's Disparate Impact" (*California Law Review.*) Sections 1&2 only.	Barocas and Selbst's article explains the difference between disparate treatment and disparate impact discrimination, and gives a detailed

		overview of the stages in the machine learning process in which discrimination can arise. This makes the article well-suited to both the topic and the expertise of the students, who are already well-versed in machine learning processes. The article is very long and technical, so only sections 1 and 2 were assigned. A simple pre-class reading assignment allowed the Embedded EthiCS TA to determine that students had understood the main differences between disparate treatment and disparate impact discrimination. By showing how disparate impact and disparate treatment discrimination may arise in various stages of the machine learning process, the article helps students to see the practical utility of these theoretical tools.

Implementation		
Class Agenda:	1. What is discrimination? Defining discrimination in a morally neutral sense vs. wrongful discrimination. 2. Discrimination: Disparate treatment vs. disparate impact 3. What is wrong with discrimination?: Procedural vs Substantive Unfairness 4. Can we avoid discrimination by eliminating social group membership from data? Case study: map of distribution of Amazon same-day delivery service in Boston. 5. Accuracy and discrimination: what do we mean by "garbage in, garbage out?" 6. What is the relationship between accurate predictions and fair decisions?	It is helpful to distinguish early in the class between discrimination in a morally neutral sense (in which all statistical reasoning is discriminatory) vs. wrongful discrimination. In the final sections of the class (5 and 6), students are asked to consider whether fairness requires anything over and above decisions made on the basis of accurate predictions. Students are presented with the idea that even accurate predictions might yield unfair decisions. An example: workplaces that are hostile to minority employees may utilize recruiting tools that accurately predict that minority employees will perform worse at their jobs. However, the use of this tool seems nonetheless unfair because

		these employees perform worse because of their hostile workplace conditions.
Sample Class Activity:	At the beginning of the module, students are presented with the following case, and divided into groups of 2-3 for discussion. "Suppose that the age at which someone starts computer programming is strongly correlated with future success as a software engineer at Google. On average, boys tend to start computer programming earlier than girls. Would it be discriminatory for Google to use the age at which someone starts computer programming as part of their basis for deciding which software engineers to hire? Would it be illegal? Would it be unfair?"	This activity introduces a core puzzle in debates about algorithmic discrimination – when is basing decisions on a feature that is *correlated* with social group membership unfair to members of groups that are thereby disadvantaged? Including an activity early in the module helps stimulate student engagement and draws more of the class into the discussion. The assigned reading ("Big Data's Disparate Impact") and the reading questions students completed prior to the module prepared students to think about questions of legality.
Module Assignment:	Students were given a short article about the distribution of Amazon's same-day delivery service in various cities across the US and asked to complete the following assignment. https://www.bloomberg.com/graphics/2016-amazon-same-day/ *Please respond to the following two questions with answers of 1-2 paragraphs each. We do not expect you to do any outside research, though we encourage you to connect to lecture materials and the reading for the module where relevant.* *Question 1.* Some people think that Amazon's process for determining which neighborhoods would receive same-day delivery was wrongfully discriminatory, but others disagree. What do you think? Explain your reasoning. *Question 2.* Basing decisions about how to treat others on social group membership often strikes us as wrongfully discriminatory. For example, most people would say that refusing to hire someone because they are a woman is wrongful discrimination, at least under normal circumstances. However, there are cases in which deciding how to treat people on the basis of social group membership does *not* strike most people as unfair, even if it ends up disadvantaging members of one group relative to another. Describe one such case – it can be a real-world case, or a hypothetical scenario of your own devising – and explain why	The assignment is designed to encourage students to grapple with the question of whether we can altogether avoid discrimination (at least understood as disparate impact) in our decision-making by simply ignoring social group membership. It also requires students to consider what might differentiate cases of wrongful discrimination from those where taking social group membership into account is not wrongfully discriminatory.

	someone might think the treatment in question is not discriminatory.	
Lessons Learned:	This module stimulated broad-based engagement, in part because it focused on an issue – algorithmic discrimination – that most students are already familiar with and care about. Framing the last half of the module around the oft-cited (and deceptively simple) slogan "garbage in, garbage out" seemed to particularly resonate with students. We unpacked various things the slogan might mean, and considered practical implications for system design.	Marginal notes