Repository Entry for CS 181
Embedded EthiCS @ Harvard Teaching Lab
Cat Wade

NOTE: this module should be labeled as "under development."

Overview	
Course: CS181: Machine Learning	
Course Level: Upper-level undergraduate	
Course Description: "This course provides a broad and rigorous introduction to machine learning, probabilistic reasoning and decision making in uncertain environments. We will cover provide an overview of three major areas in machine learning: supervised learning, unsupervised learning, and reinforcement learning. Our learning approach will be conceptual, theoretical, and practical. 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."[1]	
Module Topic: Machine learning and discrimination	
Module Author: Cat Wade	
Semesters Taught: Spring 2018, Spring 2019	
Tags: discrimination [phil], disparate treatment and disparate impact discrimination [phil], social goods [phil], harm [phil], racism [phil], algorithms [cs], machine learning [cs], predictive accuracy [cs], optimization [cs]	
Module Overview: Machine learning systems are powerful tools, but risk repeating existing patterns of discrimination if not developed carefully. In this module, we begin by discussing 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 result from the application of machine learning to organizational decision-making problems and consider the relationship between discrimination and predictive accuracy. Finally, we briefly introduce students to different techniques for preventing discrimination in the development of machine learning systems.	
Connection to Course Technical Material: In this course, students learn how to tackle real-world decision-making problems using machine	

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

learning. This module introduces students to important ethical questions raised by applications of machine learning "in the wild," and gives them tools to think through those questions as they apply the machine learning skills they have acquired.	
Goals	
Module Goals: ● Understand the concept of discrimination and its variations. ● Be able to identify the ways in which machine learning can both enable and prevent discrimination. ● Practice communicating about discriminatory algorithms. ● Evaluate the tradeoffs associated with different ways of optimizing for fairness in machine learning.	
Key Philosophical Questions: 1. What is discrimination? 2. What is disparate impact/disparate treatment? 3. What is the connection between discrimination and decision-making? 4. What is the connection between discrimination and accuracy?	
Materials	
Key Philosophical Concepts: ● Discrimination ● Disparate Impact ● Disparate Treatment ● Social goods/harms	
Assigned Readings: There were no assigned readings for this module.	
Implementation	

Class Outline:
 	
1. Social goods and uses of machine learning.
2. Discrimination.
3. Machine learning, accuracy and discrimination.
  • Understanding discrimination as a kind of inaccuracy.
  • Discrimination despite accuracy.
4. Discrimination beyond accuracy.
  • Evaluating machine learning in terms of performance tasks.
5. Small group activity.	

6. Optimizing machine learning for fairness. • Formalizing a non-discrimination criterion. • Demographic parity. • Equalizing odds. • Well-calibrated systems. 7. Concluding discussion.	
Sample Class Activity: Students are given the following scenario: *Hiring at Forever 28*. Forever 28 has hired a new computer science team to design an algorithm to classify various job applicants. You notice that African-American sales representatives have significantly fewer average sales than white sales representatives. The algorithm's output recommends hiring far fewer African-Americans than white applicants when the percentages of applications from people of various races are adjusted for. Students then discuss the following questions in small groups: Q1: Is this discrimination? (Would it be disparate impact or disparate treatment discrimination?) Q2: Does this meet our other two criteria? (Make sure there is: (1) a performance task that achieves some social good; and (2) an unbiased dataset.) Q3: You have to communicate the results to your employer and make a recommendation about what to do. What do you say? After each question, the Embedded EthiCS fellow asks groups to share their responses with the full class.	