PH 196: Artificial intelligence for medicine and health policy

Fall 2019

W 3-6 PM | B5 Hearst Field Annex

Faculty Instructor

Ziad Obermeyer
zobermeyer@berkeley.edu
Office Hours: Book here

Teaching Staff

★ *GSI*: Nolan Pokpongkiat
nolanpokpongkiat@berkeley.edu
★ *Teaching fellow*: Scott Lee
scott.lee.3898@berkeley.edu
Office Hours: M 10-12
1204 Berkeley Way West

Course Overview

Over the coming decades, data and algorithms will transform medicine and our health care system. Whether you plan to be a doctor, an algorithm developer, or work elsewhere in the health sector, this course will help you understand the tremendous upside of artificial intelligence for health: what the tools of machine learning can do in this important sector, and where they can do harm. The course will focus on teaching concepts, not the mechanics of specific algorithms. But genuine conceptual understanding requires engagement with technical content (e.g., readings from computer science and statistics, problem sets requiring analysis of real datasets with statistical software). As a result, it is designed for students who are already comfortable with basic data analysis, thanks to coursework in data science/computer science, biostatistics/statistics, or economics (e.g., you should already know how to load and manipulate datasets in statistical software).

Learning Objectives

The focus of this course will not be on learning the mechanics of particular algorithms; there are many excellent courses for that in computer science, statistics, and elsewhere. Rather, it will teach you how to think about the role of artificial intelligence in the health sector, which is both a particularly important use case and also a 'model system' in which to study the application of machine learning to social problems. By the end of the course, you will have learned two major categories of things:

1. *How to identify unsolved problems* in health that artificial intelligence can help solve.
    a. A central idea is learning to see 'prediction problems' (can I predict *y* with *x*)—which are the kinds of problems machine learning solves—and distinguish them from 'estimation problems' (does *x* cause *y*)—where machine learning alone is not enough.
    b. To do this, we will need to understand both the basis of causal inference tools—which solve estimation problems—and the basics of prediction tools. We'll do this by learning about specific real-world problems in medicine and policy.

2. *The unique challenges* of problems in health, which complicate efforts to apply successful methods from other fields. Specifically, you will learn:
    a. How the observational nature of health datasets complicates the evaluation of algorithmic performance, just as it complicates causal inference, and how some lessons from causal inference can help understand these issues.
    b. How hard it is to measure health outcomes, and how these measurement challenges can lead machine learning to automate human error and bias

    This will allow you be an 'educated consumer' of health algorithms wherever you encounter them in your future: in a career in health, as a patient/consumer of health data, or simply in newspaper headlines every day.

Requirements and responsibilities

Your responsibilities are to attend lecture three hours per week, participate in discussions and answer questions in class, take the midterm exam, and complete assignments and a final group project on time. You should look at the assigned readings before class each week—it's likely they will make more sense after lecture, but you will want to know what to pay attention to and identify key concepts *before* you walk into class. Since we want to encourage some collaboration on all non-exam assignments, as long as each student writes up their assignment in their own words, discussion with other students is ok, and even encouraged.

Evaluation

1. *Problem sets: 5%, 15%, 15% (35% total).* To complete these, you will need to be comfortable with basic statistical concepts and the notation that accompanies them (e.g., conditional expectation and probability). The first problem set will be focused on testing these skills. If you don't find yourself very comfortable with this, the rest of the course will be difficult. The second and third problem sets will involve two basic skills.
    a. Reasoning through toy problems
        i. Analytical: e.g., you have a dataset describing medical decision D affecting outcome Y, with contextual variables X.
            1. Use potential outcomes notation to write the outcomes we observe in our dataset in terms of conditional expectation and potential outcomes (hint: expand out $E[Y|D=0]$ and $E[Y|D=1]$).
            2. Imagine fitting a predictor of Y using X in those for whom $D=0$ (and imagine you've done a good job, so predictions approach expected values). Now imagine applying this to those for whom $D=1$. Can you take predictions at face value? when you compare expected values, what are the two key differences (hint: literally, take the difference)?
        ii. Substantive: e.g., you are a program officer from a foundation. A co-founder of a company that predicts depression, using social media data, is pitching you for funding. What is the hardest question you could ask and what would she likely answer?
    b. Running your own analyses in a health dataset

        i. Data will be posted on bCourses. The analyses will be about concepts, not specific coding practices, so you can use any language you like. We will talk through a 'software toolbox' that should prepare you to do the problem sets in whatever language you choose. However, course staff will be providing support in Python, which we recommend if you are on the fence.
        ii. Analytical: e.g., using the dataset documentation provided, create a variable measuring whether the patient has had a heart attack during the follow up period, and predict it using data available at the start of the follow up period. Pick a measure of accuracy and report your results.
        iii. Substantive: e.g., what are sources of measurement error in the outcome? How would these affect performance?

2. *Exam: Midterm (20%)*. In addition to the kinds of concepts tested in the problem sets, the midterm will also have substantive questions about the readings (e.g., choose an algorithm described in one of the technical readings, and describe its inputs and outputs. Choose one of the pitfalls for algorithms we discussed and describe how it applies, using specific examples from the readings).
3. *Final project* (*35% total*)
    a. Halfway through the course, you will be asked to form groups of between 2-4 to work on a final project. The project will consist of one of the following, at the discretion of the group:
        i. An independent analysis of a dataset where you fit an algorithm to answer a prediction problem related to health (broadly construed).
        ii. A detailed analysis of a for-profit company or non-profit group that purports to use machine learning to solve a health problem that answers the question: how much money would I bet on success?
        iii. An in-depth analysis of a major paper applying machine learning to health, where you identify in detail all the potential pitfalls with the approach, and reach a decision on whether you would apply the algorithm to a real patient.

    We will post a list of publicly available datasets, a list of companies/non-profits, and a list of major papers early in the class so you can plan ahead. You are welcome to choose your own, subject to the guidelines we will provide.

    b. You'll be asked to submit a very short (½ page) group proposal describing what you plan to do, and will receive approval and feedback on that to make sure you are on the right track.
    c. Grading will be based on the understanding demonstrated by the group of major concepts from the class, which will be measured in three ways:
        i. The *short proposal* (2.5%) described in 3b above
        ii. An *in-class presentation (10%)* in which each group will be assigned a time to present in the last few weeks of class
        iii. A *final paper (22.5%)*, to be submitted by Tuesday, December 17 at 11:59pm.

iv. You will be graded as a group. You'll also be asked to fill out a peer evaluation about your group members. This feedback may, in some cases, affect your group assignment grades.

4. *Class participation (10%)*
    a. Some of this will be measured by simply showing up to lecture
    b. Some of it will be based on answering questions in class (there will be many opportunities), asking good questions in class (questions answered in readings do not constitute good questions), contributing constructively to your classmates' learning and group presentations, and pointing out errors and inconsistencies in the faculty instructor's lectures and slides.

Prerequisites and software

You will struggle in this course if you are not already comfortable with basic statistical concepts and notation, and statistical software (i.e., loading a dataset and running a regression should be easy for you). Functionally, this means having an upper division level of computational and inferential background (e.g. DATA 100), which could also be satisfied with biostatistics (PH 142), statistics (100), economics (100A/B), computer science (88), or an equivalent course in another department.

While this is not a lab-based course, we will use statistical exercises with health data to teach concepts. You are free to use whatever statistical software you are most comfortable with. We will not be evaluating your code (or giving you code), but we will provide some support for students who need it, primarily in Python. If you don't have previous experience in Python, don't be alarmed: as long as you've had previous programming experience in some language, you should be fine in this course. We won't be implementing any tools ourselves, but rather using existing libraries widely used in the field. The web provides many great tutorials and resources to learn Python. For a quick crash course in Python, we recommend lab00, disc00, lab01 from CS61A. In addition, the GSI and Teaching Fellow will provide extra office hours in the first couple weeks to support students who are less comfortable with Python to get caught up.

Lectures and readings

How to think about the readings.

- *Technical papers*: Focus on the following four things when reading: (i) what is the output of the algorithm, i.e., what is the parameter of interest; (ii) what are the inputs to the model, i.e., what data are used to predict the outcome; (iii) what is the basic structure (not the nitty gritty details) of the function linking inputs to output; (iv) as we progress in the course, you will need to start paying attention to what could go wrong, e.g., is the algorithm's problem well-specified, is there measurement error or bias, can I meaningfully evaluate the algorithm's performance in the dataset used?

- *Other papers*: These are meant to give you a sense of how algorithms are used, what problems they can apply to, etc. You can use these to build up a general library of ways that algorithms can and should (not) be used.

Week 1 (8/28) Big medical problems and how big data might help

Medicine is about life and death problems. Solving them requires breaking them down into specific questions. We'll distinguish between two broad kinds: 'prediction problems' (can I predict *y* with *x*; these are good questions for machine learning) and distinguish them from 'estimation problems' (does *x* cause *y*; where machine learning alone is insufficient).

Readings

Athey, S., 2017. Beyond prediction: Using big data for policy problems. Science, 355(6324), pp.483-485.

Kleinberg, J., Ludwig, J., Mullainathan, S. and Obermeyer, Z., 2015. Prediction policy problems. American Economic Review P&P, 105(5), pp.491-95

Obermeyer, Z. and Lee, T.H., 2017. Lost in thought—the limits of the human mind and the future of medicine. New England Journal of Medicine, 377(13), pp.1209-1211.

Week 2 (9/4) Some key insights from the causal inference toolbox

We'll spend most of our time on prediction, but the causal inference toolkit (for answering 'estimation' problems) has a lot of tricks that will come in handy. These tools, particularly as they are applied to data that comes from complex systems (social systems like medicine, economics, and policy; biological systems like gene networks and cells) have critical lessons about the statistical pitfalls that distort and manipulate the bp data we use.

Problem Set 0 released Monday 9/2 (due 9/10)

Readings

Angrist, J.D. and Pischke, J.S., 2014. Mastering 'metrics: The path from cause to effect. Princeton University Press.

Chapters 1-2, *including appendices*.

Week 3 (9/11) How prediction works

We'll cover the basic science of predicting accurately, and how this gets translated into practice. We'll focus on concepts, so everything you need to know here you'll be able to think about using regression; you can build up from regression to fancier methods later. For those of you who want to get fancier now, you can read about how this extends to penalized regression (ridge, lasso), tree-based classifiers, and tricks like boosting and bagging.

Problem Set 0 (due Tues 9/10 at 11:59pm)

Readings

James, G., Witten, D., Hastie, T. and Tibshirani, R., 2013. An introduction to statistical learning. New York: Springer. Available here.

Ch. 1: 9-42, on the bias-variance tradeoff
Ch. 5: pp. 175-194, on cross-validation
Ch. 6: pp. 238-243, on more dimensions, more problems
Optional
Ch. 6: pp. 203-227, on ridge and lasso
Ch. 8: pp. 303-332, on trees and forests; boosting and bagging

Week 4 (9/18) The 'unstructured' data toolbox: Images and text

Tools like regression and trees are useful for well-behaved, 'structured' data, i.e., data made up of discrete fields. But a lot of medical data look very different -- x-rays, doctors' notes, etc. We'll spend this week learning how algorithms for analyzing these data work.

Problem Set 1 Part 1 released Monday 9/16 (due 10/3)

Readings

Gulshan, V., Peng, L., Coram, M., Stumpe, M.C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J. and Kim, R., 2016. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. Jama, 316(22), pp.2402-2410.

Ghassemi, M., Naumann, T., Doshi-Velez, F., Brimmer, N., Joshi, R., Rumshisky, A. and Szolovits, P., 2014, August. Unfolding physiological state: Mortality modelling in intensive care units. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 75-84). ACM.

Optional

Kang, J.S., Kuznetsova, P., Luca, M. and Choi, Y., 2013. Where not to eat? improving public policy by predicting hygiene inspections using online reviews. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (pp. 1443-1448).

Week 5 (9/25) The self driving doctor

This week we'll present the most optimistic case we can for machine learning in medicine, as it's currently made. We'll cover two major topics: *automation* (replacing part of all of a doctor's job with algorithms) and *decision support* (augmenting doctors' decision making capabilities with algorithms). Since medical algorithms are in very early stages, we'll illustrate these with examples from other fields that make people very optimistic to run this playbook in medicine ("it's like Uber for…").

Problem Set 1 Part 2 released Thursday 9/26 (due 10/3)

Readings

Khosla V. 20-percent doctor included: Speculations & musings of a technology optimist. Introduction.

Bates DW, Saria S, Ohno-Machado L, Shah A, Escobar G. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. Health Affairs. 2014 Jul 1;33(7):1123-31.

Week 6 (10/2) Where medical data come from

Nearly all the medical data we have are produced as 'exhaust' -- a byproduct of two key activities in the health system: the business of diagnosing and treating illness (electronic health records) and the business of getting paid to do those things (insurance claims). Only a minority of health data sources are specifically assembled to do analysis (biobanks and longitudinal datasets). This week we'll learn about the messy process of making medical data, and the many distortions that it introduces into what we observe.

Problem Set 1 (due Thursday 10/3 at 11:59pm)

Readings

May, T. The fragmentation of health data.
https://medium.com/datavant/the-fragmentation-of-health-data-8fa708109e13

LeSueur, D. 5 Reasons Healthcare Data Is Unique and Difficult to Measure https://www.healthcatalyst.com/insights/5-reasons-healthcare-data-is-difficult-to-measure

Week 7 (10/9) **Midterm**

Week 8 (10/16) **Problems with prediction 1: If only there weren't counterfactuals**

If you thought running the tech playbook in medicine was going to be simple, you're not alone: many across the Bay are betting millions on it. But getting prediction right in medicine is harder than it looks, as we'll see this week and next. We'll start by covering a pitfall that will look familiar to you from week 2 (counterfactuals, selection bias). We often train an algorithm to help doctors make a decision, because we think doctors *would* make better decisions using the algorithm. But our dataset contains only the decisions doctors *currently* make. So if the doctor does A but the algorithm says B, how do we know who's right… when we only see A? We'll illustrate this with both medical examples, and examples from further afield -- tech companies have learned this lesson the hard way too.

Group formation and project topics posted, project proposal assigned (due 10/22)

Readings

Lakkaraju, H., Kleinberg, J., Leskovec, J., Ludwig, J. and Mullainathan, S., 2017, August. The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 275-284). ACM.

Week 9 (10/23) **Problems with prediction, part 2: Measurement error**

The second class of problems that distort medical (and other) algorithms is measurement error. You might think this is a simple problem in medicine: we'd like to predict heart attack and it seems like something doctors would have figured out how to measure by now. But this too turns out to be harder than it looks. We'll link back to our lecture from week 6 on medical data and see how seemingly subtle errors can have major consequences—which can be hard to diagnose, precisely because of measurement error. This leads to problems not just for predictive accuracy but also for algorithmic bias affecting disadvantaged groups.

Project Proposal due Tues, 10/22 at 11:59pm

Problem Set 2 released 10/21 (due 11/5)

Readings

Mullainathan, S. and Obermeyer, Z., 2017. Does machine learning automate moral hazard and error? American Economic Review P&P, 107(5), pp.476-80.

Inspecting algorithms for bias. MIT Tech Review.

Graber ML. The incidence of diagnostic error in medicine. BMJ Qual Saf. 2013 Jun 14:bmjqs-2012.

Week 10 (10/30) Putting it all together: A fully worked-through example

This week we'll spend the whole lecture working through one paper. It will cover nearly all the challenges we've discussed and tries to work through a set of solutions, all in an effort to understand one decision: how physicians test patients for heart attack. It ends on a fairly optimistic view about the profound changes we can expect from algorithms that help doctors make decisions. If you think spending 3 hours on one paper is long, imagine how the 5 years it took us to write it felt.

Mullainathan, S. and Obermeyer, Z., 2019. Who is Tested for Heart Attack and Who Should Be: Predicting Patient Risk and Physician Error. NBER Working Paper No. 26168.

Week 11 (11/6) No class

Problem Set 2 (due Tues 11/5 at 11:59pm)

Week 12 (11/13) Group presentations

Week 13 (11/20) Group presentations

Week 14 (11/27) No class (Thanksgiving break)

Week 15 (12/4) Medical mysteries: Heart disease

A few times in the class, we've touched on an uncomfortable possibility: that the decisions doctors make are less than perfect. It may be even less comforting to think that, in addition to doctors making mistakes, the underlying science on which medicine is based can be surprisingly uncertain. This week we'll walk through one example of this: heart disease, which despite being the leading cause of death in the world, is surprisingly poorly understood. We'll talk about what we know, what we don't, and how algorithms might help us know more.

12/17, 11:59pm **Final projects due**