Public Health 290 — Spring 2019 Syllabus
Targeted Learning with Biomedical Big Data

Lecture meets TuTh 11:00A–12:30P in Dwinelle 211
Discussion meets W 10:00–11:00A in BWW 1208

Section 011 | Course Control Number: 31997

Instructor: Mark van der Laan
E-mail: laan@berkeley.edu
Office Hours: Th 12:45P–1:45P in BWW 5311

GSI: Rachael Phillips
E-mail: rachaelvphillips@berkeley.edu
Office Hours: W 11:00A–12:00P in BWW 1208

Course Description

This course teaches students how to construct efficient estimators and obtain robust inference for parameters that utilize data-adaptive estimation strategies (i.e., machine learning). Students will perform hands-on implementation of novel estimators using high-dimensional data structures, providing students with a toolbox for analyzing complex longitudinal, observational, and randomized control trial data. Students will actively learn and apply the core principles of the Targeted Learning methodology, which (1) generalizes machine learning to any estimand of interest; (2) obtains an optimal estimator of the given estimand, grounded in theory; (3) integrates the state-of-the-art ensemble machine learning techniques; and (4) provides formal statistical inference in terms of confidence intervals and testing of specified null hypotheses of interest. It also integrates causal inference thereby allowing one to define estimands that represent the answer to causal questions of interest.

Instructional Strategy

Most pedagogical studies (i.e., those concerned with the methods and effectiveness of teaching) indicate that lectures by themselves are a poor way of engaging students and promoting learning. To address this problem, this course will use a Blended Learning/Hybrid Classroom format. This involves shifting the majority of the material presented in class to out of class. Instructional core content is delivered online, outside of the classroom. Class time is spent exploring topics in greater depth and creates meaningful learning opportunities. This rearrangement allows for more interactive, active learning opportunities during class time like group discussion, Q&A, problem solving activities, and labs where students will apply the methods presented to real data. It also allows for self-paced comprehension of complex core concepts. Video lectures give students the ability to pause, rewind, and even re-watch content delivery opposed to traditional lectures that require content delivery to occur in a fixed time and place. The video lectures are intended to serve as jumping off points to drive discussion, activities, and clarification during class time. Thus, it is recommended that students watch the video lecture *before* class. Under this instructional model, coming to class confused is welcome, but coming to class empty-headed is ineffective.

Prerequisites

It is highly recommended for students to have some background in statistics and mathematics. Also recommended is data analysis experience and/or familiarity with the `R` software.

Credit Hours: 4

Learning Outcomes

At the completion of this course, students should be able to

- Formulate the statistical estimation problem by defining the scientific question in terms of a structural causal model, type of intervention, and causal parameter of interest.
- Establish identifiability of the causal target parameter from the observed data.
- Understand the framework of the Super Learner approach for the estimation of parameters and oracle inequalities for the general cross-validation selector.
- Evaluate the conditions guaranteeing asymptotic linearity and efficiency of the TMLE.
- Use the `R` statistical software to do the following with complex, big data:
  - Apply the Super Learner algorithm to estimate various functional parameters.
  - Implement TMLE to construct an efficient estimator of a target parameter and construct statistical inference with confidence intervals.
  - Design data simulations to practically evaluate estimators and their inference.

Requirements and Materials

There will be assigned and optional readings from the following textbooks that serve as excellent references, encompassing all of the topics covered in this course and more:

- *Targeted Learning: Causal Inference for Observational and Experimental Data* by Mark J. van der Laan and Sherri Rose (2011) {abbreviated to vdL&R (2011)}
- *Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies* by Mark J. van der Laan and Sherri Rose (2018) {abbreviated to vdL&R (2018)}

These textbooks are freely available in PDF for students to download through SpringerLink. They are also posted on bCourses in Files > Readings.

Mark van der Laan's blog (`https://vanderlaan-lab.org/post/`) is encouraged for web-based discussion. At any time, students are invited to submit a question to the blog by sending an email to `vanderlaan.blog@berkeley.edu`. Course assignments will require submission of questions to the blog.

The following will be available on bCourses: links to access lecture videos; video lecture notes; weekly exercises and in-class worksheets; upcoming due dates, the syllabus, attendance, and grades; a sign-up sheet for course material presentations; the schedule for guest interviews and a document for proposing questions; and a research project sign-up sheet and project guidelines.

There will be an anonymous mid-semester feedback survey provided through SurveyMonkey; the link will be sent to students' `berkeley.edu` email addresses.

Assessment

Research Project (25%)
Consists of 2 brief presentations and 1 comprehensive, final presentation. Involves the application of the methods learned to student-led research projects. Students will be asked to apply the roadmap of statistical learning to address a particular question of interest from their own research or that of their fellow students. Project guidelines with details on the required deliverables are posted on bCourses.

Weekly Exercises (20%)
Brief exercises will be distributed on bCourses on Tuesday's to check comprehension of key concepts and will include questions from Tuesday's in-class worksheet. Exercises will be due on bCourses on the following Monday.

Present Course Material (15%)
In pairs, students will present a piece of course material (a published paper or vdL&R chapter) and lead a discussion on it for 45 minutes during lecture. Presenters should read the content to be presented and prepare slides in advance of the class meeting. During the presentation, they should briefly summarize the main ideas and key results. Presenters should also prepare a few questions or other discussion points (such as concerns or possible extensions) to encourage discussion and critical thought by other class members. Students will have freedom to decide among themselves which dates they wish to present.

Interview Contribution (15%)
As a class, we will be hosting a series of 45-minute online interviews featuring experts in the field. A few days before each interview, students are expected to propose questions for it. These questions may be personalized to the interviewee or more general. Contribution will be evaluated based on the questions proposed for the interview. Students are also expected to arrive on time and stay for the duration of the interview. The door will be locked at 15 minutes past the hour to minimize disturbances.

Class Preparedness (10%)
Includes engagement during class and coming prepared to class. Preparedness for class will be assessed in the form of short, open-note worksheets that are straightforward for students who completed the "Before Week" activities before class. As mentioned above, questions from the in-class worksheets will be included in the weekly exercise.

Blog Question (5%)
In a group of 1 to 3, students will ask a question to the blog by sending an email to `vanderlaan.blog@berkeley.edu`. The question can be related to (but is not limited to) projects, research, or course content – the criteria is broad, just focus on a statistical question that is not easily "Google-able".

Attendance (5%)
Student attendance will be recorded at all meetings and will be posted on bCourses. If you cannot attend, email Prof. van der Laan in order to excuse your absence.

Course Feedback (5%)
Involves completing the anonymous mid-semester survey and the final course evaluation.

Course Policies

Late Assignments

Deducted 10% of the total number of points for every late day.

Accommodations

Please speak with the instructional staff as soon as possible if you require any particular accommodations, and we will work out the necessary arrangements.

Scheduling Conflicts

Notify the instructional staff by the second week of the term about any known or potential conflicts (e.g., religious observances, interviews, team activities, conferences).

Collaboration and Independence

All assignments should clearly list collaborators and references. Assignments will not be considered for credit if they are a replicate of another classmate's. With that in mind, you may work together but you should complete answers independently.

Honor Code

"*As a member of the UC Berkeley community, I act with honesty, integrity, and respect for others.*" The purpose of the Honor Code is to enhance awareness of the need for the highest possible levels of integrity and respect on campus, both within and outside the academic context. We hope and believe that the code will catalyze a series of ongoing conversations about our principles and practices. Together, through engagement, we can create a consistent message and ethos in our classrooms, labs, departments, and throughout the academic enterprise, to ensure that the core values of academic integrity and honesty are being embraced by both students and faculty.

Academic Integrity

One of the most important values of an academic community is the balance between the free flow of ideas and the respect for the intellectual property of others. Researchers don't use one another's research without permission; scholars and students always use proper citations in papers; and students may not circulate or post materials (handouts, exams, syllabi — any class materials) from their classes without the written permission of the instructor.

Any test, paper or report submitted by you and that bears your name is presumed to be your own original work that has not previously been submitted for credit in another course unless you obtain prior written approval to do so from your instructor. In all of your assignments, you may use words or ideas written by other individuals in publications, web sites, or other sources, but only with proper attribution. If you are not clear about the expectations for completing an assignment or taking a test or examination, be sure to seek clarification from your instructor or GSI beforehand. Finally, you should keep in mind that as a member of the campus community, you are expected to demonstrate integrity in all of your academic endeavors and will be evaluated on your own merits. The consequences of cheating and academic dishonestyincluding a formal discipline file, possible loss of future internship, scholarship, or employment opportunities, and denial of admission to graduate school are simply not worth it.

Tentative Course Outline

Week	Content
1: 22–24 Jan.	• *Topics:* The roadmap of statistical learning. • *Before:* Read Forewards and Preface of vdL&R (2011); watch Week 1 video lecture.
2: 29–31 Jan.	• *Topics:* Data-generating experiments and traditional data analysis. • *Before:* Read Ch. 1 of vdL&R (2011); watch Week 2 video lectures. • *Deliverables:* Week 1 Exercise due Mon. 28 Jan. on bCourses.
3: 05–07 Feb.	• *Topics:* Structural causal models, causal quantities, identification. • *Before:* Read Ch. 2 of vdL&R (2011); watch Week 3 video lecture. • *Deliverables:* Week 2 Exercise due Mon. 04 Feb. on bCourses.
4: 12–14 Feb.	• *Topics:* Interventions and identifiability results. • *Before:* Watch Week 4 video lecture. • *Deliverables:* Week 3 Exercise due Mon. 11 Feb. on bCourses.
5: 19–21 Feb.	• *Topics:* Understanding the challenges of nonparametric density estimation: Super Learning of a density. • *Before:* Watch Week 5 video lecture. • *Deliverables:* Week 4 Exercise due Mon. 18 Feb. on bCourses. • **Project Presentations I – Proposal on Wed. 20 Feb.**
6: 26–28 Feb.	• *Topics:* Super Learner and an oracle inequality for the general cross-validation selector, Super Learning in prediction, Super Learning of optimal individualized treatment rules. • *Before:* Read Ch. 3 of vdL&R (2011); Watch Week 6 video lecture. • *Deliverables:* Week 5 Exercise due Mon. 25 Feb. on bCourses.
7: 05–07 Mar.	• *Topics:* Super Learning of conditional multinomial distributions or densities. • *Before:* Watch Week 7 video lecture. • *Deliverables:* Week 6 Exercise due Mon. 04 Mar. on bCourses; Mid-semester feedback survey due Fri. 08 Mar.
8: 12–14 Mar.	• *Topics:* Measure-theoretic Integration, Highly Adaptive Lasso (HAL). • *Before:* Optional, see bCourses homepage. • *Deliverables:* Week 7 Exercise due Mon. 11 Mar. on bCourses.

Week	**Content**
9: 19–21 Mar.	• *Topics:* Asymptotic linearity, influence curves, and statistical inference based on influence curves. • *Before:* Watch Week 9 video lecture. • *Deliverables:* Week 8 Exercise due Mon. 18 Mar. on bCourses.
26–28 Mar.	• *Topics:* None — it's Spring Break • *Before:* N/A
10: 02–04 Apr.	• *Topics:* Pathwise differentiable target parameters, gradients and the canonical gradient of infinite-dimensional models. • *Before:* Watch Week 10 video lectures. • *Deliverables:* Week 9 Exercise due Mon. 01 Apr. on bCourses. • **Project Presentations II – Progress on Wed. 02 Apr.**
11: 09–11 Apr.	• *Topics:* Definition of MLEs and NP-MLEs, efficient influence curves, theorems of efficiency. • *Before:* Watch Week 11 video lectures. • *Deliverables:* Week 10 Exercise due Mon. 08 Apr. on bCourses.
12: 16–18 Apr.	• *Topics:* Efficient one-step estimators, online one-step estimators. • *Before:* Watch Week 12 video lecture. • *Deliverables:* Week 11 Exercise due Mon. 15 Apr. on bCourses; Blog question due Fri. 19 Apr. by email to `vanderlaan.blog@berkeley.edu`
13: 23–25 Apr.	• *Topics:* Targeted Maximum Likelihood Estimation (TMLE). • *Before:* Read Ch.4–6 of vdL&R (2011); watch Week 13 video lecture. • *Deliverables:* Week 12 Exercise due Mon. 22 Apr. on bCourses.
14: 30 Apr. – 02 May	• *Topics:* TMLEs of causal effects of multiple time-point interventions based on longitudinal data. • *Before:* Optional, see bCourses homepage. • *Deliverables:* Week 13 Exercise due Mon. 29 Apr. on bCourses.
15: 07–09 May	• **Project Presentations III – Comprehensive**