Targeted Learning in the *tlverse*: Techniques and Tools for Causal Machine Learning

Course Information

Course Number: CE_17C
Date & Time: Monday, August 7th, 2023, from 8:30 AM – 5:00 PM Eastern Time
Location: Metro Toronto Convention Centre (North Building) Room 104C

Course Description

Great care is required when disentangling intricate relationships for causal and statistical inference in medicine, public health, marketing, political science, and myriad other fields. However, traditional statistical practice ignores complexities that exist in real-world problems, for example, by avoiding interaction terms in regression analysis because such terms complicate and obfuscate the interpretation of results. The field of Targeted Learning (TL) presents a solution to such practices by outlining a modern statistical framework that unifies semiparametric theory, machine learning, and causal inference. This workshop provides a comprehensive introduction to TL and its accompanying free and open-source software ecosystem, the *tlverse* (https://github.com/tlverse). It will be of interest to statisticians and data scientists who wish to apply cutting-edge statistical and causal inference approaches to rigorously formalize and answer substantive scientific questions. This workshop incorporates discussion and hands-on R programming exercises, allowing attendees to familiarize themselves with methodology and tools that translate to improvements in real-world data analytic practice.

Learning Objectives

As a result of this workshop, we aim for attendees to be able to:

1. Follow the roadmap of statistical learning / TL to define estimation problems in realistic statistical models and obtain valid inferences.
2. Apply standard cross-validation schemes using the *origami* R package, including V-fold, stratified, and clustered cross-validation.
3. Train a super learner using the *sl3* R package by selecting an appropriate cross-validation scheme, library of candidate machine learning algorithms, and loss function and meta-learner algorithm.
4. Estimate the effect of a static intervention, as defined in the *tmle3* R package, and apply the delta method to estimate transformations of existing parameters.

Course Prerequisites

Attendees are highly recommended to have had prior training in basic statistical concepts, such as confounding, probability distributions, (linear and logistic) regression, hypothesis testing and confidence intervals. Advanced knowledge of mathematical statistics may be useful but is not necessary. Familiarity with the R programming language is essential.

We request that attendees ***download the necessary software in advance and bring the laptop (and charger) with the downloaded software to the workshop***. The Wi-Fi at the conference may be unstable and the software download requires internet connection, so we highly recommend doing this in advance. If you are experiencing installation issues, please reach out via email or let us know during the workshop. The software download instructions are as follows:

1. Install R version 4.2 (https://cloud.r-project.org/), the most recent major version, and see the instructions for setting up R and RStudio (http://tlverse.org/jsm2023-workshop/setting-up-r-and-rstudio.html).
2. After updating R to version 4.2, please follow the software installation instructions (http://tlverse.org/jsm2023-workshop/install-software-packages.html#install-software-package). There, you will also find instructions for resolving the most common installation-related error, so please look there as a first step for troubleshooting.

The worksheets are on pages 4–7 of this handout. We will ask attendees to complete these worksheets during lecture, and we will review answers following each lecture. Attendees should therefore ***print this 7-page handout in advance and bring this with a pen/pencil to the workshop***.

Lastly, please ***complete the brief pre-workshop survey*** (https://forms.gle/4M4Dh6r31ehdsXpDA) at least a few days before the workshop.

Instructional Methods

This workshop was developed based on the learning objectives (LOs). There is time in the schedule for lecture and discussion, with opportunities for attendees to engage with the material via coding exercises and worksheets (WS).

Materials

Instructional materials are freely available in the GitHub repository for this course (https://github.com/tlverse/jsm2023-workshop) and the corresponding online vignette for this course (https://tlverse.org/jsm2023-workshop/), which is based on the *tlverse* handbook (https://tlverse.org/tlverse-handbook/). The coding exercises for *origami*, *sl3*, and *tmle3* R packages are located at the end of those packages' online vignette chapters.

Course Tentative Outline

Time	Topic	Format	LO
08:30 – 10:15	Targeted Learning: the bridge from machine learning to statistical and causal inference	Introductory lecture by Mark & Alan, during which attendees complete WS 1 and ask questions	1
10:15 – 10:30	Morning Break		
10:30 – 10:40	The *tlverse* software ecosystem of R packages for Targeted Learning	Introductory lecture by Ivana, during which attendees ask questions	
10:40 – 11:40	Cross-validation and the *origami* R package	Lecture and live coding with Ivana, during which attendees complete WS 2 and ask questions	2
11:40 – 12:00	Coding exercise with the *origami* R package	Attendees break out into groups to complete exercise; instructors assist as needed	2
12:00 – 12:30	Super Learning and the *sl3* R package	Lecture and live coding by Rachael, during which attendees complete WS 3 and ask questions	3
12:30 – 14:00	Lunch Break		
14:00 – 14:20	Super learning and the *sl3* R package	Lecture and live coding by Rachael, during which attendees complete WS 3 and ask questions	3
14:20 – 14:45	Coding exercise with the *sl3* R package	Attendees break out into groups to complete exercise; instructors assist as needed	3
14:45 – 15:15	Targeted minimum loss-based estimation (TMLE) and the *tmle3* R package	Lecture and live coding by Nima, during which attendees complete WS 3 and ask questions	4
15:15 – 15:30	Afternoon Break		
15:30 – 16:00	Targeted minimum loss-based estimation (TMLE) and the *tmle3* R package	Lecture and live coding by Nima, during which attendees complete WS 4 and ask questions	4
16:00 – 16:30	Coding exercise with the *tmle3* R package	Attendees break out into groups to complete exercise; instructors assist as needed	4
16:30 – 17:00	Concluding remarks	Discussion and Q&A	all

Instructor Information (listed by presentation order)

Mark van der Laan (laan@berkeley.edu) is the Jiann-Ping Hsu/Karl E. Peace Professor of Biostatistics and Statistics at the University of California, Berkeley. He has made contributions to survival analysis, semiparametric statistics, multiple testing, and causal inference. He also developed the targeted maximum likelihood methodology and general theory for super learning. He is a founding editor of the Journal of Causal Inference and International Journal of Biostatistics. He has authored 4 books on targeted learning, censored data, and multiple testing, authored over 300 publications, and graduated over 50 Ph.D. students. He received the COPSS Presidents' Award in 2005, the Mortimer Spiegelman Award in 2004, and the van Dantzig Award in 2005.

Alan Hubbard (hubbard@berkeley.edu), Professor and Head of Biostatistics, University of California, Berkeley, co-director of the Center of Targeted Learning, and head of the computational biology core of the SuperFund Center at UC Berkeley (NIH/EPA), as well a consulting statistician on several federally funded and foundation projects. He has worked as well on projects ranging from molecular biology of aging, epidemiology, and infectious disease modeling, but most of his work has focused on semi-parametric estimation in high-dimensional data. His current methods-research focuses on precision medicine, variable importance, statistical inference for data-adaptive parameters, and statistical software implementing targeted learning methods. Currently working in several areas of applied research, including early childhood development in developing countries, environmental genomics, and comparative effectiveness research. He has most recently concentrated on using complex patient data for better prediction for acute trauma patients.

Ivana Malenica (imalenica@berkeley.edu) is a Postdoctoral Researcher in the Department of Statistics at Harvard University and a Wojcicki and Troper Data Science Fellow at the Harvard Data Science Initiative. She obtained her PhD in Biostatistics at UC Berkeley working with Mark van der Laan, where she was a Berkeley Institute for Data Science Fellow and a NIH Biomedical Big Data Fellow. Her research interests span non/semi-parametric theory, causal inference, and machine learning, with emphasis on personalized health and dependent settings. Most of her current work involves causal inference with time and network dependence, online learning, optimal individualized treatment, reinforcement learning, and adaptive sequential designs.

Rachael Phillips (rachaelvphillips@berkeley.edu) is a Senior Data Analyst for the Center for Targeted Machine Learning and Causal Inference (CTML) at UC Berkeley. She has a PhD in Biostatistics, MA in Biostatistics, BS in Biology, and BA in Mathematics. Motivated by issues arising in healthcare, the projects she's pursued include the development of clinical algorithm frameworks and guidelines; real-world data analysis methodologies for generating and evaluating real-world evidence; and biostatistics graduate-level courses and other educational material for Targeted Learning and causal inference. Rachael collaborates with clinicians and statisticians at UC San Francisco and Novo Nordisk, and during her PhD studies, she worked closely with researchers at the U.S. FDA under Drs. Susan Gruber and Mark van der Laan.

Nima Hejazi (nhejazi@hsph.harvard.edu) is an Assistant Professor of Biostatistics at the Harvard T.H. Chan School of Public Health. His research interests concentrate in causal inference and statistical machine learning (or "causal machine learning"), focusing on the development of efficient, model-agnostic statistical inference methods. Nima is often motivated by topics from non- and semi-parametric inference and efficiency theory; high-dimensional inference; targeted minimum loss-based estimation; biased sampling designs, especially outcome-dependent two-phase designs (e.g., case-control studies); and sequentially adaptive trials. He studies these topics through the lens of statistical parameters motivated by causal inference (e.g., heterogeneous treatment effects, dose-response curves, mediational direct/indirect effects). Nima is also deeply interested in high-performance statistical computing and is a passionate advocate for open-source software and the critical role it plays in the promotion of transparency, reproducibility, and "data analytic hygiene" in the practice of applied statistics and statistical data science. Recently, he has been captivated by the rich statistical issues and pressing public health challenges common in clinical trials and/or observational studies evaluating the efficacy of preventive vaccines or curatives/therapeutics for high-burden infectious diseases (HIV/AIDS, COVID-19), in infectious disease epidemiology, and in computational immunology.

Worksheet 1: Traditional Data Analysis and the Roadmap of Targeted Learning

Please complete this worksheet during the first lecture led by Mark and Alan. Answers will be reviewed afterwards.

1. Common data science practice encourages users to "check" models after they have been fit to the data, so that if one of the checks fail, then a new model can be fit to the data. Why might this approach be problematic?

2. Common data science practice lets the type of data at hand dictate the scientific question of interest and the statistical model. Why is this problematic?

3. What is the purpose of the Roadmap of Targeted Learning?

4. Enumerate and briefly describe each step in the Roadmap of Targeted Learning.

5. Distinguish the field of causal inference from that of statistical estimation and inference and explain how causal inference can be integrated into the Roadmap of Targeted Learning.

Worksheet 2: Cross-validation

Please complete this worksheet during the cross-validation lecture led by Ivana. Answers will be reviewed afterwards.

1. Compare and contrast V-fold cross-validation with re-substitution cross-validation. What are some of the differences between the two methods? How are they similar? Describe a scenario when you would use one over the other.

2. Why is V-fold cross-validation inappropriate for use with time-series data?

3. What are the advantages and disadvantages of V-fold cross-validation relative to:
    a. holdout cross-validation?

    b. leave-one-out cross-validation?

4. Consider a classification problem with many predictors and a binary outcome, and the analyst proposes the following procedure:
    (i) First, screen the predictors, isolating only those covariates that are strongly correlated with the binary outcome labels.
    (ii) Next, train a learning algorithm using only this subset of covariates that are highly correlated with the outcome.
    (iii) Finally, use cross-validation to estimate the tuning parameters and the performance of the trained algorithm.

    Is this application of cross-validation correct? Why or why not?

Worksheet 3: Super Learning

Please complete this worksheet during the Super Learner lecture led by Rachael. Answers will be reviewed afterwards.

1. What is the Super Learner (SL) algorithm? When and why might one consider using it?

2. How does the SL use cross-validation?

3. [Fill in the blank] The discrete SL (dSL) chooses the candidate estimator that optimizes the

   _____.

4. What is a meta-learner? How is it different from other learners/algorithms and how is it similar to them?

5. What is the purpose of screeners in the SL?

6. What should the analyst consider when they specify SL? Provide a few examples and justify each of them.

Worksheet 4: Targeted Minimum Loss-based Estimation

Please complete this worksheet during the TMLE lecture led by Nima. Answers will be reviewed afterwards.

1. Describe how one might implement a substitution estimator for the average treatment effect (ATE). What is the nuisance function (there's only one) necessary for the construction of this estimator? Might there be restrictions on the nuisance estimation procedure to ensure compatibility with this substitution approach?

2. Often, we may find ourselves in situations in which we have imperfect or little knowledge about the best way to estimate nuisance functions in a real-world data analysis – this is where the Super Learner (SL) comes in.
    a. What might go wrong when SL-based nuisance function estimates are directly inputted into the substitution estimator from (1) above?

    b. What are the consequences of this issue for statistical inference?

3. The targeted maximum likelihood estimation (or, more generally, targeted minimum loss-based estimation) approach outlines a path for modifying nuisance parameter estimates for compatibility with the substitution formula in (1). This approach does so by introducing a parametric fluctuation (so-called "targeting") step.
    a. Describe this fluctuation step in words (no formulas!) and explain how it incorporates information about the efficient influence function (EIF).

    b. When implemented correctly, the step above yields a targeted minimum loss-based estimator (TMLE) that is *asymptotically linear* and *asymptotically efficient*. Give simple definitions (that is, in layman's terms) of what these properties are and why they may be desirable.

    c. Like TMLE, estimation of the ATE with augmented inverse probability weighting (A-IPW) is a double robust procedure and yields an asymptotically linear and asymptotically efficient estimator. How are these two estimators different from each other?

4. Cross-validated TMLE (CV-TMLE) is an augmentation that pairs targeted minimum loss-based estimation with cross-validation.
    a. How does this type of algorithm introduce cross-validation into the process of constructing a TMLE, and does this differ from the role that cross-validation plays in the Super Learner algorithm?

    b. What is one advantage of CV-TMLE over TMLE (without this augmentation)? Can you think of any disadvantages, whether theoretical or practical?