USC Marshall School of Business

DSO 699: Casual inference with modern machine learning methods

Instructor: *Yingying Fan*
Office: *BRI 307B*
Office Hours: *Thursday 4-5pm via zoom*
Email: fanyingy@marshall.usc.edu

COURSE DESCRIPTION

Across disciplines, causal inference is a cornerstone of science, engineering, economics, and public policy. In medicine, we would like to understand how a patient might have responded if we offered a different treatment. In engineering system design and optimization, we would like to understand how the system would behave if we made different design choices. In public policy, we are constantly asking if different taxes, laws, regulations, or programs might improve (or hurt) society at large. Correctly answer such questions can help us make more informed and better decisions. Data are frequently collected and analyzed in the process of seeking statistically quantified answers to these questions. In contemporary applications, these data are decidedly large-scale, complex, and high-dimensional. These call for urgent need in designing modern statistical machine learning methods for causal inference.

Recently, an exciting set of tools at the intersection of causal inference and machine learning has emerged to tackle these types of questions in these settings. This course is a doctoral-level introduction to these tools. Our emphasis is primarily on developing the statistical tools to formally analyze these types of methods, with a goal of empowering students to both understand cutting-edge research in this area and contribute their own new (theoretically justified) methods to the field.

The first one-third of the course (approximately the first 5 weeks) will focus on basic concepts and methods in causal inference at the level of the Imbens and Rubin (2015) book. In the second two-thirds, we will move to more recent developments for causal inference using modern machine learning methods. This second part will primarily be based on recent research papers.

COURSE Topics (tentative and subject to change)

1. Potential Outcomes Approach
2. Randomized Experiments
3. Unconfounded Treatment Assignment and Propensity Score
4. Matching with Nearest-Neighbors Methods, Doubly Robust Methods
5. Tree, Random Forests, and Causal Forests
6. Causal Effects Estimation with Deep Learning Methods
7. Knockoffs Framework and Causal Inference
8. Causal Inference for Panel Data: Difference-in-Differences and the Matrix Completion Viewpoint
9. Double Machine Learning Method for Treatment and Structural Parameters

For each topic in items 4-9 listed above, I will first start with introducing the machine learning tools and their theoretical justification, and then move to their applications in causal inference.

COURSE PRE-REQUISITES

This course will be theoretically rigorous. Students are expected to have a strong background in proof-based mathematics, probability theory and statistics at a graduate level. Students are also expected to be able to code in either R or Python for homework assignments and real data applications. If you have concerns about whether this course is suitable for you, reach out to the instructor to discuss before registration.

COURSE EXPECTATIONS

I will assign readings before each class. Students are expected to read the assigned readings – slowly and carefully – and come to class prepared to participate in class discussions.

COURSE MATERIALS

Required Text:
Imbens, Guido W., and Donald B. Rubin. 2015. Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. New York: Cambridge UP.

Reading list *(tentative):*

1) Susan Athey, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens & Khashayar Khosravi (2021) Matrix Completion Methods for Causal Panel Data Models, Journal of the American Statistical Association, 116:536, 1716-1730
2) Arkhangelsky, Dmitry, Susan Athey, David A. Hirshberg, Guido W. Imbens, and Stefan Wager. 2021. "Synthetic Difference-in-Differences." American Economic Review, 111 (12): 4088-4118.DOI: 10.1257/aer.20190159
3) Ramachandra, Vikas & Athey, Susan & Wager, Stefan. (2015). Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. Journal of the American Statistical Association. 113. 10.1080/01621459.2017.1319839.
4) Athey S, Imbens G. Recursive partitioning for heterogeneous causal effects. Proc Natl Acad Sci U S A. 2016 Jul 5;113(27):7353-60.
5) Athey, Susan and Stefan Wager. Estimating Treatment Effects with Causal Forests: An Application. Observational Studies, 5, 2019.
6) Imbens G, Abadie A. Large Sample Properties of Matching Estimators for Average Treatment Effects. Econometrica. 2006;74 (1) :235-267.
7) Lin, Z., Ding, P., and Han, F. (2021+). Estimation based on nearest neighbor matching: from density ratio to average treatment effect. Manuscript.
8) Demirkaya, E., Fan, Y., Gao, L., Lv, J., Vossler, P. and Wang, J. (2022). Optimal nonparametric inference with two-scale distributional nearest neighbors. *Manuscript*.
9) Kallus, N. (2020). Generalized optimal matching methods for causal inference. J. Mach. Learn. Res., 21, 62-1.
10) Chi, C.-M., Vossler, P., Fan, Y. and Lv, J. (2022). Asymptotic properties of high-dimensional random forests. *Manuscript*.

11) Farrell, M.H., Liang, T. and Misra, S., 2020. Deep learning for individual heterogeneity: an automatic inference framework. arXiv preprint arXiv:2010.14694.
12) Xiong, Ruoxuan & Pelger, Markus. (2019). Large Dimensional Latent Factor Modeling with Missing Observations and Applications to Causal Inference. SSRN Electronic Journal. 10.2139/ssrn.3465357.
13) Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, James Robins, Double/debiased machine learning for treatment and structural parameters, The Econometrics Journal, Volume 21, Issue 1, 1 February 2018, Pages C1–C68
14) Candès, E. J., Fan, Y., Janson, L. and Lv, J. (2018). Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society Series B* 80, 551-577.
15) Ge, X., Chen, Y.E., Song, D. et al. (2021) Clipper: p-value-free FDR control on high-throughput data from two conditions. Genome Biol 22, 288.
16) Künzel, Sören & Sekhon, Jasjeet & Bickel, Peter & Yu, Bin. (2017). Meta-learners for Estimating Heterogeneous Treatment Effects using Machine Learning. Proceedings of the National Academy of Sciences. 116. 10.1073/pnas.1804597116.
17) Fan, J., Imai, K., Lee, I., Liu, H., Ning, Y. and Yang, X., 2021. Optimal covariate balancing conditions in propensity score estimation. Journal of Business & Economic Statistics, pp.1-14.
18) Athey, Susan & Wager, Stefan. (2021). Policy Learning With Observational Data. Econometrica. 89. 133-161. 10.3982/ECTA15732.

If you have any questions or need assistance with the Blackboard Course Pages, please contact the Marshall HelpDesk at 213-740-3000 or HelpDesk@marshall.usc.edu.

GRADING

Assignments (these are all tentative and we can discuss later)	**Grade share**
Homework	25%
Midterm	30%
Class presentations	20%
Class project	25%
TOTAL	100%

HOMEWORK

There will be approximately bi-weekly homework assignments. These assignments will be a combination of theoretical/empirical exercises. I strongly encourage students to discuss the assignments with whomever is willing. But the final submitted work must represent the students own work. Short of extraordinary circumstances as defined by university policy, late homework will not be accepted.

EXAM

We will have one in-class midterm in Week 9 (tentatively).

In Class Presentation

The in-class presentations will spread out throughout the semester (after I finish introducing the basic concepts from the book). Roughly, for each topic I discuss there will be one or two in-class presentations. Each student is required to make at least one in-class presentation. The presentations should be based on selected research papers on causal inference and should be individual work. It will be evaluated by the quality of slides, presentation, and discussions.

Final project

Students are responsible for finding a causal inference application and a suitable dataset for implementing the machine learning methods we learn in this class. The final project can be done in groups, but the group size cannot exceed 3. If you have difficulty deciding what to work on for your final project, please come up with a list of potential topics and discuss with me during my office hours.

STATEMENT OF ACADEMIC CONDUCT AND SUPPORT SYSTEMS

USC seeks to maintain an optimal learning environment. Students are expected to submit original work. They have an obligation both to protect their own work from misuse and to avoid using another's work as their own. All students are expected to understand and abide by the principles of academic honesty outlined in the University Student Conduct Code (see University Governance, Section 11.00) of SCampus (www.usc.edu/scampus or http://scampus.usc.edu). The recommended sanctions for academic integrity violations can be found in Appendix A of the Student Conduct Code.

Students with Disabilities:
USC is committed to making reasonable accommodations to assist individuals with disabilities in reaching their academic potential. If you have a disability which may impact your performance, attendance, or grades in this course and require accommodations, you must first register with the Office of Disability Services and Programs (www.usc.edu/disability). DSP provides certification for students with disabilities and helps arrange the relevant accommodations. Any student requesting academic accommodations based on a disability is required to register with Disability Services and Programs (DSP) each semester. A letter of verification for approved accommodations can be obtained from DSP. Please be sure the letter is delivered to me (or to your TA) as early in the semester as possible. DSP is located in GFS (Grace Ford Salvatori Hall) 120 and is open 8:30 a.m.–5:00 p.m., Monday through Friday. The phone number for DSP is (213) 740-0776. Email: ability@usc.edu.

Support Systems:
Student Counseling Services (SCS) - (213) 740-7711 – 24/7 on call
Free and confidential mental health treatment for students, including short-term psychotherapy, group counseling, stress fitness workshops, and crisis intervention. https://engemannshc.usc.edu/counseling/

National Suicide Prevention Lifeline - 1-800-273-8255
Provides free and confidential emotional support to people in suicidal crisis or emotional distress 24 hours a day, 7 days a week. http://www.suicidepreventionlifeline.org

Relationship & Sexual Violence Prevention Services (RSVP) - (213) 740-4900 - 24/7 on call
Free and confidential therapy services, workshops, and training for situations related to gender-based harm. https://engemannshc.usc.edu/rsvp/

Sexual Assault Resource Center
For more information about how to get help or help a survivor, rights, reporting options, and additional resources, visit the website: http://sarc.usc.edu/

Office of Equity and Diversity (OED)/Title IX compliance – (213) 740-5086
Works with faculty, staff, visitors, applicants, and students around issues of protected class. https://equity.usc.edu/

Bias Assessment Response and Support
Incidents of bias, hate crimes and microaggressions need to be reported allowing for appropriate investigation and response. https://studentaffairs.usc.edu/bias-assessment-response-support/

Student Support & Advocacy – (213) 821-4710
Assists students and families in resolving complex issues adversely affecting their success as a student EX: personal, financial, and academic. https://studentaffairs.usc.edu/ssa/

Diversity at USC – https://diversity.usc.edu/
Tabs for Events, Programs and Training, Task Force (including representatives for each school), Chronology, Participate, Resources for Students

Emergency Preparations
In case of an emergency if travel to campus is not feasible, the USC Emergency Information web site (http://emergency.usc.edu/) will provide relevant information, such as the electronic means the instructors might use to conduct their lectures through a combination of USC's Blackboard learning management system (blackboard.usc.edu), teleconferencing, and other technologies.