DS-GA 3001.003

Special Topics in DS - Causal Inference in Machine Learning

CSCI-GA 3033.108 Special Topics in CS - Causal Inference in Machine Learning

Overview

Recent advances and successes in machine learning have had tremendous impacts on society. Despite such impact and successes, the currently dominant paradigm of machine learning, broadly represented by deep learning, relies almost entirely on capturing statistical correlations among features from data. As the famous phrase “correlation does not imply causation” suggests, such correlation-driven approaches cannot uncover causal relationships among variables of interest. In this introductory course, we aim to provide early-year M.Sc. and Ph.D students, who are already familiar with machine learning, with the foundational ideas and techniques in causal inference so that they can expand their knowledge and expertise beyond correlation-driven machine learning.

Target Audience

M.Sc. students and early-year PhD students in computer science and data science.

- Prerequisites: students should have taken the following courses.
    - Computer Science
        - MATH-UA.235 Probability and Statistics
        - CSCI-GA.2565 Machine Learning
        - CSCI-GA.2569 Inference and Representation (optional; encouraged)
        - CSCI-GA.2572 Deep Learning (optional; encouraged)
    - Data Science
        - DS-GA 1002 Probability and Statistics for Data Science
        - DS-GA 1003 Machine Learning
        - DS-GA 1005 Inference and Representation (optional; encouraged)
        - DS-GA 1008 Deep Learning (optional; encouraged)

General Information

- **Lectures:** 4.55pm-6.35pm on Mondays
    - REDACTED
    - Lectures will be in person, although they will be livestreamed via Zoom as well.
- **Lab sessions**: 4.55pm-5.45pm on Tuesdays
    - REDACTED
    - Lab sessions will be in person.

- **Instructor**: Kyunghyun Cho
- **Assistant**: Daniel Jiwoong Im, Taro Makino and Divyam Madaan
- **Office Hours** (starts on Jan 29 2024)
    - **Instructor**

        - Kyunghyun Cho: 12-1pm on Mondays at REDACTED
    - **Assistant**
        - Divyam Madaan: 4-4.30pm on Tuesdays at REDACTED
        - Daniel Jiwoong Im: 4-4.30pm on Wednesdays REDACTED
        - Taro Makino: 4-4.30pm on Thursdays at REDACTED

- **Grading**
    - Paper presentation 50%
    - Remote final exam 50%

- **Course Site**:
    - **Campuswire**: REDACTED
        - For discussion and announcements
        - Registration code: REDACTED
    - **Google Calendar**: REDACTED
        - Add this to your calendar so that you get the up-to-date course schedule.

- **Lecture Note**: REDACTED
    - This lecture note will be continuously updated throughout the semester.

Schedule

Note that the schedule below is only a guideline. The content of each lecture will be decided as the course progresses.

wk	Lecture (Monday)	Lab (Tuesday)
01/22	Logistics & Recap 1. Probabilistic graphical models 2. Structural causal models	
01/29	Confounders, colliders and mediators 1. Definitions 2. Causal vs. spurious correlation	(Makino) * Build structural causal models (SCM) that correspond to the basic setup. * Compute various conditional distributions. * Estimate these conditional distributions from data.
02/05	Confounders, colliders and mediators (2) 1. Conditional distributions are not interventional distributions.	(Makino) * Continue from last week's SCM. * Show the difference between the interventional distribution and the conditional distribution.

02/12	Randomized controlled trials (RCT) 1. When do conditional and interventional distributions coincide? 2. Actively collecting data 3. Limitations 4. Outcome maximization	(Im) * Draw samples from a synthetic basic setup. * Show that we cannot estimate the potential outcome correctly if the covariate is not observed. * Show that we can estimate the potential outcome correctly if we run an interventional study (RCT).
02/19	President's Day	
02/26	Causal inference from observational data 1. Inverse probability weighting 2. Taste of doubly robust estimation 3. Matching	(Im) * Continue from last week's RCT and Introduce the bandit problem. * Show how to tackle a bandit problem; does solving a bandit problem necessitate precisely inferring a causal effect?
03/04	Causal inference from observational data 1. Instrumental variables 2. Difference-in-difference 3. Regression discontinuity	(Im) * Draw samples from a synthetic basic setup. * Show that we can estimate the potential outcome correctly with matching/IPW if the covariate is also observed. * Empirically demonstrate the potentially high variance of matching/IPW.
03/11	Paper presentations **4: Sex Bias in Graduate Admissions: Data from Berkeley.** **3: Collider bias undermines our understanding of COVID-19 disease risk and severity.** **12: Randomized Controlled Trials and real life studies. Approaches and methodologies: a clinical point of view.** **9: Evidence for Health Decision Making — Beyond Randomized, Controlled Trials.** **14: Comparison of Approaches to Advertising Measurement: Evidence from Big Field Experiments at Facebook.**	(Madaan) * Create a synthetic example that includes an instrument variable. * Run two-stage linear regression and show that the potential outcome can be correctly estimated even if the covariate is not observed.
03/18	Spring break	

03/25	Paper presentations **13 A contextual-bandit approach to personalized news article recommendation.** **15 A contextual-bandit approach to personalized news article recommendation.** **11 Mastering the game of Go without human knowledge.** **2 Can higher cigarette taxes improve birth outcomes?**	(Madaan) Numpyro tutorial
04/01	Distributional shift in machine learning 1. Recap of causal inference 2. Independent and identically distributed 3. Distributional shift and impossibility	(Madaan) Double machine learning
04/08	Distributional shift in machine learning 1. Invariance: stable correlations 2. Meta-learning	(Makino) Colored MNIST * invariance to the environment
04/15	Distributional shift in machine learning 3. Reinforcement learning with binary preference	(Madaan) * Build an SCM where the covariate is not a confounder but a collider. * Show the effect of explaining away.
04/22	Guest lecture 1 on real-world causal inference (Aahlad Puli: https://aahladmanas.github.io/)	Review and Q&A
04/29	Guest lecture 2 on real-world causal inference (Bryant Moy: https://bryantjmoy.com/)	Review and Q&A
05/06	Paper Presentations **8 The Blessings of Multiple Causes.** **7 Causal Inference by using Invariant Prediction: Identification and Confidence Intervals.** **6 Invariant Risk Minimization.** **5 When Training and Test Sets Are Different: Characterizing Learning**	Reading Day

	Transfer. **1 On causal and anticausal learning.** **10 Out-of-Distribution Generalization in the Presence of Nuisance-Induced Spurious Correlations.**	
05/13	Final Exam at 4-5.30pm	

Paper presentation

Students will form teams of three or four at the beginning of the semester by random assignment. Each team will be randomly assigned to a group of papers from the Paper List below, and present the selected papers, or a subset, as a group at a designated lecture slot (Teams 1-7 on March 11, teams 8-14 on March 25 and teams 15-21 on April 29). Each team will be given 15 minutes for presentation (12 minutes of presentation + 3 minutes of question-answering.) Presentation slides (in pdf) must be submitted by email a day before the presentation to the instructor.

Paper list

For some of the papers in the list below, use https://library.nyu.edu/services/computing/off-campus/ to access the full text articles if they are behind the paywall.

1. Confounder bias
    a. P. J. Bickel et al. Sex Bias in Graduate Admissions: Data from Berkeley. Science 187, 398-404 (1975).
        i. https://www.science.org/doi/10.1126/science.187.4175.398
2. Collider bias
    a. Griffith, G.J., Morris, T.T., Tudball, M.J. et al. Collider bias undermines our understanding of COVID-19 disease risk and severity. Nat Commun 11, 5749 (2020).
        i. https://www.nature.com/articles/s41467-020-19478-2
3. Randomized controlled trials in healthcare
    a. Frieden. Evidence for Health Decision Making — Beyond Randomized, Controlled Trials. The New England Journal of Medicine. 2017.
        i. https://www.nejm.org/doi/10.1056/NEJMra1614394
    b. Saturni S, Bellini F, Braido F, Paggiaro P, Sanduzzi A, Scichilone N, Santus PA, Morandi L, Papi A. Randomized Controlled Trials and real life studies. Approaches and methodologies: a clinical point of view. Pulm Pharmacol Ther. 2014 Apr;27(2):129-38. doi: 10.1016/j.pupt.2014.01.005. Epub 2014 Jan 24. PMID: 24468677.
        i. https://pubmed.ncbi.nlm.nih.gov/24468677/
4. Causal inference in marketing (A/B testing)

    a. Brett R. Gordon, Florian Zettelmeyer, Neha Bhargava, Dan Chapsky. Comparison of Approaches to Advertising Measurement: Evidence from Big Field Experiments at Facebook. Marketing Sciences. 2019.
        i. https://pubsonline.informs.org/doi/epdf/10.1287/mksc.2018.1135
        ii. You can skip Section 8.
5. Contextual bandit in online recommendations
    a. Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. 2010. A contextual-bandit approach to personalized news article recommendation. In Proceedings of the 19th international conference on World wide web (WWW '10). Association for Computing Machinery, New York, NY, USA, 661–670.
        i. https://dl.acm.org/doi/abs/10.1145/1772690.1772758
6. Reinforcement learning
    a. Silver, D., Schrittwieser, J., Simonyan, K. et al. Mastering the game of Go without human knowledge. Nature 550, 354–359 (2017).
        i. https://www.nature.com/articles/nature24270
7. Causal inference from observational data
    a. William N. Evans, Jeanne S. Ringel. Can higher cigarette taxes improve birth outcomes? Journal of Public Economics, Volume 72, Issue 1, 1999.
        i. https://www.sciencedirect.com/science/article/pii/S0047272798000905
    b. Yixin Wang & David M. Blei (2019) The Blessings of Multiple Causes, Journal of the American Statistical Association, 114:528, 1574-1596.
        i. https://arxiv.org/abs/1805.06826
8. The principle of invariant prediction
    a. Jonas Peters, Peter Bühlmann, Nicolai Meinshausen, Causal Inference by using Invariant Prediction: Identification and Confidence Intervals, Journal of the Royal Statistical Society Series B: Statistical Methodology, Volume 78, Issue 5, November 2016, Pages 947–1012.
        i. Sections 1, 2, 7 and 8.
        ii. https://academic.oup.com/jrsssb/article-abstract/78/5/947/7040653
    b. Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, David Lopez-Paz. Invariant Risk Minimization. arXiv:1907.02893. 2019.
        i. Sections 1, 2, 4.3 and 5
        ii. https://arxiv.org/abs/1907.02893
9. Distribution shifts
    a. Amos, Storkey, 'When Training and Test Sets Are Different: Characterizing Learning Transfer', in Joaquin Quiñonero-Candela, and others (eds), Dataset Shift in Machine Learning (Cambridge, MA, 2008; online edn, MIT Press Scholarship Online, 22 Aug. 2013)
        i. https://academic.oup.com/mit-press-scholarship-online/book/13447/chapter-abstract/166918395
        ii. https://homepages.inf.ed.ac.uk/amos/publications/Storkey2009TrainingTestDifferent.pdf
    b. Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. 2012. On causal and anticausal learning. In Proceedings of the 29th International

Coference on International Conference on Machine Learning (ICML'12). Omnipress, Madison, WI, USA, 459–466.
    i. https://icml.cc/2012/papers/625.pdf
10. Causality in machine learning
    a. Aahlad Puli, Lily H. Zhang, Eric Oermann, Rajesh Ranganath. Out-of-Distribution Generalization in the Presence of Nuisance-Induced Spurious Correlations. International Conference on Learning Representations (ICLR), 2022.
        i. https://openreview.net/pdf?id=12RoR2o32T

Remarks

- A student in this course is expected to act professionally. Please follow the GSAS regulations on academic integrity found here: http://gsas.nyu.edu/page/academic.integrity
- Academic accommodations are available for students with disabilities. Please contact the Moses Center for Students with Disabilities (212-998-4980 or mosescsd@nyu.edu) for further information. Students who are requesting academic accommodations are advised to reach out to the Moses Center as early as possible in the semester for assistance.