ECON5180: Economics and Data Science

Term 2, 2023-2024
Department of Economics
The Chinese University of Hong Kong

Prof. CAO Siying
Email: s.cao189@cuhk.edu.hk
Office Hours: By appointment
Office: Dept. of Economics #907
Lectures: Tuesday 3:30-6:15pm
Venue: Wong Foo Yuan #403

Course Description

This is a weekly reading group discussing recent literature that combines novel data with the new generation of tools from data science to answer questions in political economy, law and economics, and other topics in economics (e.g., economic history) and broader social sciences. We will discuss how and why machine learning works, in what way the new method and potentially new forms of data contribute to our knowledge, and new venues where deployment of such tools can be fruitful.

Prerequisites

This course is aimed at MPhil/PhD students interested in expanding their research toolbox. Students should have a solid background in statistical techniques, such as comes from the equivalent of a first year PhD econometrics sequence.

Learning Outcome

The key objective of this class it to bring students to the frontier of research that find new ways to explore and work with data in the age of “Big Data". Students should develop a solid understanding of what makes machine learning work conceptually, how they

relate to and complement standard econometric tools, and be able to write a chapter in the thesis that incorporate these tools.

While we will cover specific algorithms developed in machine learning as we go along, from supervised to unsupervised learning, this is NOT a course on ML theory or the mechanics and nitty-gritty implementation of these tools. Nor will the computational aspects of the methods be treated in great depth. There are courses offered by the Computer Science Department that better serve these purposes. Students are expected to learn the techniques and practical implementation of methods on their own. Plenty of online tutorials are available. I can also give pointers if necessary.

Grading

This is a non-standard course in Economics. Here is how we will organize the reading group. I will provide an overview of the subject and background lectures in the first two class meetings. For each session in the rest of the term, I will prepare a short primer on the topic we discuss that week. One or two students will give a 45 minute presentation of the paper chosen from the reading list. All other students coming to the meeting are expected to have read the paper and activitely contribute to paper discussion [30% of grade]. Starting from the mid-term, each student will pitch 1-2 ideas to the class for feedback [30% of grade]. Students will then submit a final project proposal at the end of the term, which ideally build on the those ideas.

The proposal should flesh out in detail the research question, why we should care (a brief discussion of the literature can help), and with what data and methodology it would be executed [40% of grade]. A few graphs or tables with descriptive facts can also be quite effective in motivating your research question[1]. Students are welcome to propose a project related to their thesis in progress.

Final proposal grading. You will be graded on clarity, the novelty and relevance (i.e, importance) of the problem you address, and the feasibility of your empirical strategy. Besides the research question itself being interesting, relevance here also entails the suitability and creativity in applying the machine learning methods to help you answer the question. The proposal does not need to include actual implementation or results, but the project should be promising. Actual results are neither sufficient nor necessary for a good grade. The proposal should be an approriate length, meaning that you are neither too brief nor ramling on with irrelevant stuff. With that, 3-6 pages should do the work. The paper is **due via email absolutely no later than noon on Wednesday, May 15**.

[1] I think it's generally true that for empirical work, If you can't show anything interesting from some simple statistics or charts, the project is not likely to be that interesting.

Reading List

General Reference

- Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. Springer, Berlin: Springer series in statistics, 2001.
- Murphy, Kevin P. *Machine learning: a probabilistic perspective*. MIT press, 2012.
- Schütze, Hinrich, Christopher D. Manning, and Prabhakar Raghavan. *Introduction to information retrieval*. Vol. 39. Cambridge: Cambridge University Press, 2008.
- Mullainathan, S. and Spiess, J. (2017). Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives*, 31(2):87-106
- Athey, Susan, and Guido W. Imbens. "Machine learning methods that economists should know about." *Annual Review of Economics* 11 (2019): 685-725.

Week 1. Background Lecture: Introduction to machine learning

- Lecture notes, general reference

Week 2. Background Lecture: Key principals

- Lecture notes, general reference

Week 3. Cookbook, a menu of methods

- Lecture notes, general reference

Week 4. Prediction Problem

- *** Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human decisions and machine predictions. The Quarterly Journal of Economics, 133(1), 237-293.
- Kleinberg, J., Ludwig, J., Mullainathan, S., & Sunstein, C. R. (2020). Algorithms as discrimination detectors. Proceedings of the National Academy of Sciences, 117(48), 30096-30100.
- Arnold, D., Dobbie, W., & Yang, C. S. (2018). Racial bias in bail decisions. The Quarterly Journal of Economics, 133(4), 1885-1932.
- Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., & Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. Science, 353(6301), 790-794.

Week 5. ML for Causal Inference

- Belloni, A., Chen, D., Chernozhukov, V., & Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. Econometrica, 80(6), 2369-2429.
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., & Newey, W. (2017). Double/debiased/neyman machine learning of treatment effects. American Economic Review, 107(5), 261-65.
- *** Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association, 113(523), 1228-1242.

Week 6. Debrief on prediction challenge + Natural Language Processing Intro

***NO REQUIRED READING

- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. the Journal of machine Learning research, 3, 993-1022.
- Rudolph, M., & Blei, D. (2018, April). Dynamic embeddings for language evolution. In Proceedings of the 2018 World Wide Web Conference (pp. 1003-1011).
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).

Week 7. Text as Data

- Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. Journal of Economic Literature, 57(3), 535-74.
- *** Gentzkow, M., & Shapiro, J. M. (2010). What drives media slant? Evidence from US daily newspapers. Econometrica, 78(1), 35-71.
- Gentzkow, M., Shapiro, J. M., & Taddy, M. (2019). Measuring group differences in high-dimensional choices: method and application to congressional speech. Econometrica, 87(4), 1307-1340.
- Mueller, H., & Rauh, C. (2018). Reading between the lines: Prediction of political violence using newspaper text. American Political Science Review, 112(2), 358-375.
- Hansen, S., McMahon, M., & Prat, A. (2018). Transparency and deliberation within the FOMC: a computational linguistics approach. The Quarterly Journal of Economics, 133(2), 801-870.

- Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring economic policy uncertainty. The Quarterly Journal of Economics, 131(4), 1593-1636.
- Michalopoulos, S., & Xue, M. M. (2021). Folklore. The Quarterly Journal of Economics, 136(4), 1993-2046.
- Ash, E., & Chen, D. L. (2019). Case vectors: Spatial representations of the law using document embeddings. Law as Data, Santa Fe Institute Press, ed. M. Livermore and D. Rockmore.
- Grajzl, P., & Murrell, P. (2022). Did Caselaw Foster England's Economic Development during the Industrial Revolution? Data and Evidence. CESifo Working Paper No. 10088, Available at SSRN: https://ssrn.com/abstract=4286961

Week 8. Student Idea Pitch

- Each student will present 1-2 ideas to the class for feedback. 10 mins talk + 5 mins Q&A

Week 9. Unsupervised Learning

- Bonhomme, S., & Manresa, E. (2015). Grouped patterns of heterogeneity in panel data. Econometrica, 83(3), 1147-1184.
- *** Bonhomme, S., Lamadon, T., & Manresa, E. (2019). A distributional framework for matched employer employee data. Econometrica, 87(3), 699-739.
- Angrist, J., Azoulay, P., Ellison, G., Hill, R., & Lu, S. F. (2017). Economic research evolves: Fields and styles. American Economic Review, 107(5), 293-97.

Week 10. Other New Data

- Autor, D., Chin, C., Salomons, A. M., & Seegmiller, B. (2022). New Frontiers: The Origins and Content of New Work, 1940–2018 (No. w30389). National Bureau of Economic Research.
- Blumenstock, J., Cadamuro, G., & On, R. (2015). Predicting poverty and wealth from mobile phone metadata. Science, 350(6264), 1073-1076.
- *** Feigenbaum, J. J. (2016). Automated census record linking: A machine learning approach.
- Donaldson, D., & Storeygard, A. (2016). The view from above: Applications of satellite data in economics. Journal of Economic Perspectives, 30(4), 171-98.

- Glaeser, E. L., Kominers, S. D., Luca, M., & Naik, N. (2018). Big data and big cities: The promises and limitations of improved measures of urban life. Economic Inquiry, 56(1), 114-137.

- Kang, J. S., Choi, Y., Kuznetsova, P., & Luca, M. (2013). Using text analysis to target government inspections: Evidence from restaurant hygiene inspections and online reviews (No. 14-007). Harvard Business School.

Week 11. Neural Network, Deep Learning

- *** Chen, D. L., & Ornaghi, A. (2023). Gender Attitudes in the Judiciary: Evidence from US Circuit Courts. American Economic Journal: Applied Economics.

- Igami, M. (2020). Artificial intelligence as structural estimation: Deep Blue, Bonanza, and AlphaGo. The Econometrics Journal, 23(3), S1-S24.

- Khachiyan, A., Thomas, A., Zhou, H., Hanson, G., Cloninger, A., Rosing, T., & Khandelwal, A. K. (2022). Using Neural Networks to Predict Microspatial Economic Growth. American Economic Review: Insights, 4(4), 491-506.

- Nakamura, E. (2005). Inflation forecasting using a neural network. Economics Letters, 86(3), 373-378.

- Shen, Z., Zhang, R., Dell, M., Lee, B. C. G., Carlson, J., & Li, W. (2021). LayoutParser: A Unified Toolkit for Deep Learning Based Document Image Analysis. arXiv preprint arXiv:2103.15348.

Week 12. Images, Computer vision

- *** Adukia, A., Eble, A., Harrison, E., Runesha, H. B., & Szasz, T. (2021). What we teach about race and gender: Representation in images and text of children's books (No. w29123). National Bureau of Economic Research.

- Gebru, T., Krause, J., Wang, Y., Chen, D., Deng, J., Aiden, E. L., & Fei-Fei, L. (2017). Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States. Proceedings of the National Academy of Sciences, 114(50), 13108-13113.

Week 13. Student Presentations

- Presentation of final project proposal. This is meant to build on the ideas presented in the idea presentations. 20 minutes each student (15 mins presentation + 5 min Q&A).

Academic Honesty

Attention is drawn to University policy and regulations on honesty in academic work to the disciplinary guidelines and procedures applicable to breaches of such policy regulations. Details may be found at http://www.cuhk.edu.hk/policy/academichonesty/.