Machine learning and data mining

University of Toronto, Winter 2022

ECO480H1

Course information

Classes	Wednesday, 11 AM – 2 PM
Delivery mode	In person
Instructor	Clement Gorin
Office hours	GE351, Wednesdays, 3 PM – 5 PM
Email	clement.gorin@utoronto.ca
Zoom	https://utoronto.zoom.us/j/2198654082
Teaching assistant	Saba Ale Ebrahim
Office hours	GB448, Thursdays, 2 PM – 4 PM
Email	saba.aleebrahim@mail.utoronto.ca
Zoom	https://utoronto.zoom.us/j/89931453400

1 Overview

Motivation

Many important economic questions remain unanswered, partly because the data necessary to address them is encoded into high-dimensional data structures such as text or images. Applied economists have become increasingly interested in using machine learning models to transform these data into simpler representations, which can be used as inputs for subsequent economic analysis. After introducing predictive modelling, this course provides a comprehensive understanding of some of the most capable supervised learning models including random forest, gradient boosted trees, or specialised neural network structures to make sense of these complex forms of data and perform original economic analysis. From a strong base in theory and mathematical formalisation, focus is kept on intuition and effective implementation using Python.

Objective

This course aims to equip you with the necessary knowledge and tools to pursue independent projects. This includes acquiring a sound conceptual understanding understanding of a range of empirical models, as well as the ability to implement them on actual datasets. You are encouraged to think critically about each model, its assumptions, the problems to which they can be usefully applied, and their limitations.

Spirit

I believe that learning is an interactive process. I strive to create an interactive and stimulating learning environment where ideas can be exchanged freely and easily. In exchange, you are expected to engage actively by participating in discussions, preparing homework, reading the required papers, as well as familiarising regularly with the previously covered material.

2 Organisation

Schedule (tentative) & readings

#	Date	Topic	Reading
1	2023.01.11	Predictive modelling	–
2	2023.01.18	Linear predictions	Breiman (2001)
3	2023.01.25	Generalised additive models	Mullainathan et al. (2017)
4	2023.02.01	Prediction trees	Kleinberg, Ludwig, et al. (2015)
5	2023.02.08	Ensembles methods	Kleinberg, Lakkaraju, et al. (2017)
6	2023.02.15	Neural networks	Athey et al. (2017)
7	2023.03.01	Backpropagation	LeCun et al. (2015)
8	2023.03.08	Better optimisation	Lones (2021)
9	2023.03.15	Image modelling	Naik et al. (2017)
10	2023.03.22	Convolutional networks	Mueller et al. (2021)
11	2023.03.29	Embedding spaces	Gentzkow et al. (2019)
12	2023.04.05	Recurrent networks	Iyyer et al. (2014)

Digital tools

The course material including lectures slides, exercises, datasets and research papers are regularly uploaded on Quercus. The material is usually updated before every session so make sure to check the platform regularly. For questions related to the course content, I strongly suggest that you use the Piazza discussion channel, so that everyone can benefit from the additional explanation. You are encouraged to address the questions if you feel confident about your answer, which Saba can endorse or clarify afterwards.

Computer resources

You should come to class with a computer with administrator rights and a working internet connection. For practical exercises and assignments, we use Jupyter notebooks containing your written answers as well as the code to replicate your analysis. These notebooks will be introduced during the practical sessions. To write and execute notebooks, you can use the Google Colab remote environment (requires a Google account). Otherwise, you can install Python on your computer. Make sure to install an Anaconda or Miniconda distribution of Pyhton 3.10, which provides the basic modules for scientific computing and simplify module management. In addition, install a development environment that supports notebooks, such as Visual Studio Code. Installation instructions, along with a Conda environment file will be provided on Quercus. Installation issues will be addressed during the practical sessions.

3 Assessments

Overview

You are assessed on the basis of participation, three homework, a term paper and a proposal. Every assignment can be done in groups of two. Groups should submit a single document and each member will receive the same grade. Obviously, groups must remain the same for the term paper proposal and the term paper. Note that your grades may include adjustments to the raw scores, such as adding points to everyone's score or not counting an unduly difficult question. Your grades, rather than your raw score, best reflects the quality of your work.

Summary of assessments

Assessment	Share	Start date	Deadline
Participation	12.5%	Throughout the semester	
Homework 1	12.5%	23.02.08	23.02.14 (23:59)
Homework 2	12.5%	23.03.08	23.03.14 (23:59)
Homework 3	12.5%	23.03.22	23.03.28 (23:59)
Term paper proposal	15%	–	23.02.28 (23:59)
Term paper	35%	–	TBA

Participation

Your participation grade depend on your degree of engagement throughout the semester and reflect your attendance, the quality of your questions and responses either in class or on Piazza, as well as your ability to discuss the previously covered material and readings. Each class will start with a short group discussion about the required reading. As an example, consider these different levels of involvement. (1) Generally absent, tries to respond when called. (2) Demonstrates adequate preparation, offers straightforward information without elaboration, demonstrates sporadic involvement in discussions. (3) Demonstrates good preparation, offers interpretation and analysis, responds to other students constructively, demonstrates consistent ongoing involvement. (4) Demonstrates excellent preparation, offers analysis and synthesis, puts together pieces of the discussion to develop new approaches.

Homework

Homework consists in performing empirical analysis using the methods studied in class, or programming a statistical procedure from scratch using Python. Homework should be submitted as a Jupyter notebook. Details specific to each homework will be provided on Quercus.

Term paper

The term paper involves choosing a research question of interest to economists, and providing an answer using some of the tools studied in class. Your paper will be graded using the following criteria: The relevance of the research question, the use of the related literature, the eventual construction and cleaning of databases, the model choice and

the optimisation, the interpretation of the results, and the quality of the writing (e.g. clarity, referencing). The paper should be submitted as a Jupyter notebook and must not exceed 3000 words, excluding the title page, references, appendix, figures and code. Additional details will be provided on Quercus.

Term paper proposal

The proposal must clearly state your research question, as well as the reasons you think this is important and interesting. You must include a short review of two or more related articles, as well as a description of your empirical strategy, including possible data sources and potential models. Since some models will be covered after the submission deadline, you can consult me or Saba for advice. The proposal should be submitted as a PDF or Jupyter notebook and must not exceed 1000 words. Make sure to use this assignment as an opportunity to get early feedback about your term paper.

Late submission

Late submissions are given a grace period of 15 minutes for technical difficulties. After this delay, your assignment will receive a 10% percentage point penalty out of the original grade for each calendar day of late submission. For example, if your grade is 95%, but you submitted it more than 48 hours after the deadline (but less than 72 hours), your penalised grade is 75%.

Regrade requests

You can submit a regrading request using the appropriate online form within two weeks after the assessment has been returned. Make sure to clearly articulate why your answers deserves additional grades. The entire assessment will be reassessed and your grade could increase, decrease, or remain unchanged.

4 Practical information

Contact policy

For general questions about the content of the course, I suggest that you use the Piazza discussion channel, so that everyone can benefit from the additional explanations. For questions related to your assignments, contact Saba via private message on Piazza, or discuss during office hours. Otherwise, prefer speaking with me after the lecture or

during office hours. For personal or urgent matters, you can contact me by email, Zoom, or come directly to my office.

Academic integrity

The University of Toronto treats cases of cheating and plagiarism very seriously. Make sure to familiarise yourself with the Code of Behaviour on Academic Matters, which outlines the behaviours that constitute academic dishonesty and the processes for addressing academic offences.

Plagiarism

Normally, students will be required to submit their course essays to the Universitys plagiarism detection tool for a review of textual similarity and detection of possible plagiarism. In doing so, students will allow their essays to be included as source documents in the tools reference database, where they will be used solely for the purpose of detecting plagiarism. The terms that apply to the Universitys use of this tool are described on the Centre for Teaching Support & Innovation website.

Copyright

The course material belong to your instructor and are protected by copyright. The course materials is for your academic use, but should not be copied, shared, or used for any other purpose without the explicit permission of the instructor.

Equity, diversity and inclusion

The University of Toronto is committed to equity, human rights and respect for diversity. All members of the learning environment in this course should strive to create an atmosphere of mutual respect where all members of our community can express themselves, engage with each other, and respect one anothers differences.

Students with disabilities

The University of Toronto welcomes students with diverse learning styles and needs. You may wish to approach me and the Accessibility Services if you have a disability or health consideration that may require accommodations.

Religious observances

The University of Toronto provides reasonable accommodation of the needs of students who observe religious holy days other than those already accommodated by ordinary scheduling and statutory holidays. Make sure to approach me as early as possible to communicate any anticipated absences related to religious observances.

Family care responsibilities

The University of Toronto provides a family-friendly environment. You may wish to approach me and the Family Care Office if you are a student with family responsibilities.

5 Bibliography

Textbooks

Goodfellow, I., Y. Bengio, and A. Courville (2016). *Deep Learning*. MIT Press.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The elements of statistical learning*. Springer.

Nielsen, M. A. (2015). *Neural networks and deep learning*. Determination Press.

Textbooks are provided for reference and are freely available online.

Papers

Athey, S. and G. W. Imbens (2017). "The state of applied econometrics: Causality and policy evaluation". In: *Journal of Economic Perspectives* 31.2, pp. 3–32.

Breiman, L. (2001). "Statistical modeling: The two cultures". In: *Statistical Science* 16.3, pp. 199–231.

Gentzkow, M., B. Kelly, and M. Taddy (2019). "Text as data". In: *Journal of Economic Literature* 57.3, pp. 535–574.

Iyyer, M. et al. (2014). "Political ideology detection using recursive neural networks". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Vol. 1, pp. 1113–1122.

Kleinberg, J., H. Lakkaraju, et al. (2017). "Human decisions and machine predictions". In: *The Quarterly Journal of Economics* 133.1, pp. 237–293.

Kleinberg, J., J. Ludwig, et al. (2015). "Prediction policy problems". In: *American Economic Review* 105.5, pp. 491–495.

LeCun, Y., Y. Bengio, and G. Hinton (2015). "Deep learning". In: *Nature* 521, pp. 436–444.

Lones, M. A. (2021). "How to avoid machine learning pitfalls: A guide for academic researchers". In: *CoRR* abs/2108.02497.

Mueller, H. et al. (2021). "Monitoring War destruction from space: A machine learning approach". In: *Proceedings of the National Academy of Sciences* 118.23.

Mullainathan, S. and J. Spiess (2017). "Machine learning: An applied econometric approach". In: *Journal of Economic Perspective* 31.2, pp. 87–106.

Naik, N. et al. (2017). "Computer vision uncovers predictors of physical urban change". In: *Proceedings of the National Academy of Sciences* 114.29, pp. 7571–7576.

Specific references will be given in class.