ECON5181: Machine Learning Methods

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 #918
Lectures: Thursday, 3:30pm – 6:15pm
Venue: TBC

Course Description

Large-scale data set has become increasingly available in many fields of economics. This presents challenges to statistical inference and even merely "understanding" the data. Meanwhile, it offers abundant opportunities for new inquiries and answers. In this course, we introduce the core statistical methods to work with big data (structured and unstructured), and show how these techniques can be combined with econometric tools in economics research. While we cover major machine learning tools, including supervised learning methods, dimensionality reduction, and unsupervised learning, we will focus on their concrete applications in current empirical research. Examples will be drawn from various lines of research, including text as data, relevant prediction problems in economics, and causal inference.

Required Materials

In addition to the journal articles, the lecture will draw heavily from the following two textbooks:

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, p. 18). New York: springer.
A free copy of the book is available here https://www.statlearning.com/

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1, No. 10). New York: Springer series in statistics.
Available here https://web.stanford.edu/~hastie/ElemStatLearn/

A great reference on working with textual data is:

Schütze, H., Manning, C. D., & Raghavan, P. (2008). *Introduction to information retrieval* (Vol. 39, pp. 234-265). Cambridge: Cambridge University Press.
Available here https://nlp.stanford.edu/IR-book/information-retrieval-book.html

Prerequisites

Students should be familiar with undergrad-level statistics and econometrics. Prior knowledge of programming is not required, although students are strongly encouraged to learn the basics of R beforehand.

Learning Outcome

Students will i) understand the key ideas behind the design of machine learning algorithms, ii) understand the assumptions for each method to work, iii) learn to code simulation and statistical methods in R and/or python, and iv) identify fruitful cases to deploy ML methods and evaluate the associated merits and constraints.

Grading

There will be problem sets (three or four) throughout the term, accounting for 30% of the grade. Students are encouraged to discuss with fellow classmates, however, the final submission must be written up individually. A mid-term exam will account for another 20% of the grade. In the second half of the course, students will work in groups to complete a final project. A short written paper detailing the question, empirical model, method, data, and results will be due at the end of the term. This will account for the remaining 50% of the grade.

Course Outline

***Note: Applications are subject to change

1. Introduction: Machine learning vs economics
    (a) Two cultures: estimation and prediction
    (b) Overview of ML in economics

2. Linear regression
    (a) The challenge with many covariates
    (b) Covariate selection: economic and automated methods
3. Penalized regression
    (a) Sparsity: Lasso
    (b) Other shrinkage methods (e.g., elastic net)
    (c) Bias/variance trade-off, overfitting, regularization, cross validation
    (d) Application: demand estimation, network analysis
4. Non-linear models
    (a) Trees and partitioning method
    (b) Ensemble: bagging, boosting, random forest
    (c) Neural network and stochastic gradient descent
    (d) Application: predicting housing prices
5. Dimension reduction and discovery
    (a) PCA and factor models
    (b) k-means clustering
    (c) Application: grouped patterns of heterogeneity, forecasts in macro and finance
6. Text as data
    (a) Latent Dirichlet Allocation
    (b) Generative language model
    (c) Application: FOMC, measuring political polarization
7. Causal inference (optional)

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/.