Boston College, Department of Economics
ECON 3389 Machine Learning In Economics
Spring 2021

Class format: Online, Asynchronous
Instructor: Anatoly Arlashin, anatoly.arlashin@bc.edu
Office: Maloney Hall 339
Office Hours: online via Zoom appointment

Course Overview

One of the defining features of the world around us today is the ever-increasing amounts of data that describe our daily lives. This "big data" phenomenon, as it became known, has led to developing of new methods, called "machine learning", that allow high-dimensional statistical analysis in ways that were either impossible or infeasible for classical statistical methods, such as regression analysis.

The goal of this course is to provide students with an introduction to modern data-driven learning in a framework that makes it applicable to causal economic analysis. While we will cover the necessary theoretical foundations, the emphasis will be placed on application and learning how and when to use these methods, as well as when and how these methods can fail.

The coursework will include homework assignments with simulated and real-world data, bi-weekly online discussions on real-life data analysis issues, case study presentations and group project. We will use R as our main data analysis software, with significant amount of class time devoted to learning how to efficiently code various analytical models. Prior coding experience is welcomed, but not required.

Important: this syllabus proves only the basic information about the course. The details on course structure, assignments, policies and grading are provided in the introductory lecture, and all students are required to familiarize themselves with both the syllabus and the introductory lecture.

Prerequisites

For students majoring in Economics, Economic Statistics (ECON1151) and Econometric Methods (ECON2228) are required prerequisites. Students coming from different departments should have similar command of statistical methods. A solid knowledge of differential calculus at the level of MATH 1102 (the ”preferred” co-requisite for ECON 2228) is highly recommended. Prior knowledge of programming is not a prerequisite, but student should expect to learn a lot about coding in R, which may prove to be the most challenging aspect of this course.

Literature

Main textbook (required):

1. *An Introduction to Statistical Learning, with Applications in R*. James, G., Witten, D., Hastie, T., and Tibshirani, R (2013), available for free at: https://www.statlearning.com/

Additional introductory references (recommended):

2. *Machine learning with R Cookbook.* Chiu, Yu-Wei. (2015).
3. *Data Mining and Business Analytics with R.* Ledolter, J. (2013)
4. *A Gentle Introduction to Effective Computing in Quantitative Research*. Paarsch, H.J., Golyaev, K. (2016)
5. *Fundamentals of Machine Learning for Predictive Data Analytics*. Kelleher, J, Mac Namee, B. and D'Arcy, A. (2015)
6. *Learning from Data: A Short Course.* Abu-Mostafa, Y. , Magdon-Ismail, M. and Lin, H. (2012)

Additional advanced references:

7. *The elements of statistical learning: data mining, inference and prediction.* Hastie, T., Tibshirani, R., Friedman, J., & Franklin, J. (2009), available for free at https://web.stanford.edu/~hastie/ElemStatLearn/
8. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Hastie, T., Tibshirani, R.,Wainwright, M. (2015)
9. *Deep Learning.* Goodfellow, I., Bengio, Y., Courville, A., Bach, F. (2016)
10. *Machine learning: a probabilistic perspective.* Murphy, K. P. (2012).
11. *Pattern recognition and machine learning*. Bishop, C. M. (2006).
12. *Introduction to high-dimensional statistics*. Giraud, C. (2014).
13. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Bühlmann, P. and Geer van de, S. (2011)

Required Software

The primary software environment is the R statistical programming language, which can be downloaded for free from http://www.r-project.org. RStudio is the recommended interface for the R statistical programming language software, which can also be downloaded for free at http://www.rstudio.org.

Canvas

Canvas is the Learning Management System (LMS) at Boston College, designed to help faculty and students share ideas, collaborate on assignments, discuss course readings and materials, submit assignments, and much more - all online. As a Boston College student, you should familiarize yourself with this important tool. For more information and training resources for using Canvas, click here.

All class materials will be available online via Canvas.

Course components

Your course grade will be determined using the following components:

Homework	60%
Canvas discussions	10%
Case study	10%
Course project	20%

There are no letter grades or curving per each individual part of the grade. The overall course score is calculated as a weighted sum of all components, and then translated from 0-100 scale into a letter grade using a curved distribution.

All students can access final grades through Agora after the grading deadline each semester. Transcripts are available through the Office of Student Services.

Deadlines, Late Work and Make Up Policy

All deadlines are strictly enforced. Late work is not accepted and no credit will be earned on late work unless the student has arranged an extension ahead of time with me (and that is quite possible, I am flexible with everyone's challenging circumstance and time constraints), with rare exceptions based on individual circumstances (e.g. inability to communicate with me ahead of time because of an emergency).

Please note that hardware ("WiFi was down") and software ("Canvas was glitching") issues are not considered valid excuses for late submissions.

Course outline and schedule

- Part I: Introduction
    - Course Overview
    - Introduction to R and RStudio
- Part II: Learning theory
    - Statistical models, loss functions, optimization
    - Supervised vs Non-supervised learning
    - Model selection, bias-variance trade-off, overfitting and underfitting
- Part III: Regression analysis
    - Simple Linear Regression
    - Multivariate linear regression
    - Beyond basic regression models: categorical variables, non-linear marginal effects, polynomial regressions.
- Part IV: Classification
    - Linear probability model
    - Logistic regression
- Part V: Cross-validation and bootstrap
- Part VI: Penalization methods
    - Regularization: Ridge regression
    - Sparsity: LASSO
    - Best of both: elastic net
- Part VII: Ensemble methods
    - Random forests
    - Boosted regression trees

Accommodation and Accessibility

Boston College is committed to providing accommodations to students, faculty, staff and visitors with disabilities. Specific documentation from the appropriate office is required for students seeking accommodation in this course. Advanced notice and formal registration with the appropriate office is required to facilitate this process. There are two separate offices at BC that coordinate services for students with disabilities:

- The Connors Family Learning Center (CFLC) coordinates services for students with LD and ADHD.
- The Disabilities Services Office (DSO) coordinates services for all other disabilities.

Find out more about BC's commitment to accessibility at www.bc.edu/sites/accessibility.

Academic Integrity

Boston College values the academic integrity of its students and faculty. It is your responsibility to familiarize yourself with the university's policy on academic integrity. If you have any questions, always consult your professor. Violations of academic integrity will be reported to your class dean and judged by the academic integrity committee in your school. If you are found responsible for violating the policy, penalties may include a failing grade as well as possible probation, suspension, or expulsion, depending on the seriousness and circumstances of the violation.

See this link for a full discussion of the university's policies and procedures regarding academic integrity.