ECON 590
Big Data and Machine Learning in Econometrics
Spring 2020

Instructor: Andrii Babii

Time and Location: T and Th, 2pm-3:15pm, Bingham 208

E-mail: andrii@email.unc.edu

Office Hours: T and Th, 1pm-2pm, Gardner 208A

Prerequisites

Econ 400 (Statistics), Econ 410 and 420 (Intermediate Microeconomics and Macroeconomics), and at least one semester of differential calculus.

Course description

Students will learn how to explore, visualize, and analyze high-dimensional datasets, build predictive models, and estimate causal effects. The course introduces key concepts and tools demanded in the business environment.

Examples of techniques include an advanced overview of linear and logistic regression, model selection and regularization, LASSO, cross-validation, experiments, and causal inference, estimation of treatment effects with high-dimensional controls, networks, classification and clustering, latent variable models, bagging and the bootstrap, decision trees and random forests, text analysis.

Students will learn basic underlying concepts and will build practical programming skills in R. Heavy emphasis is placed on the analysis of actual datasets, and on applications of specific methodologies. Examples may include consumer choice data, housing prices, asset pricing, network data, internet and social media data, sports analytics.

Textbooks

1. (BDS) Matt Taddy, *Business Data Science: Combining Machine Learning and Economics to Optimize, Automate, and Accelerate Business Decisions*

2. (ISL) Gareth, Witten, Hastie, and Tibshirani, *An Introduction to Statistical Learning with Applications in R*. The book can be downloaded for free from `http://faculty.marshall.usc.edu/gareth-james/ISL/`.

I strongly encourage you to read the relevant material **before** the class.

Problem sets

There will be approximately 6 problem sets over the course of the semester. Only 5 best will count towards the final grade. Problem sets are independent work and not a group project. You should hand in your assignments at the beginning of class the day they are due. Late problem sets (but before graded problem sets are given back) will be marked down by 50% with no exceptions.

Programming

Problem sets will involve data analysis using R. R is a very flexible, powerful, and popular language and environment for statistical computing and graphics. You can download and install it from https://www.r-project.org/. You may also want to check the R Studio GUI from https://rstudio.com/products/rstudio/.

I don't assume that you have used R in a previous class. I will provide in-class demonstrations, some limited statistical instructions, and code to accompany lectures and assignments. However, this is not a class on R. Like any language, R is only learned by doing. You should install it as soon as possible and familiarize yourself with basic operations.

Some useful R resources:

1. R in Action: https://livebook.manning.com/book/r-in-action-third-edition/welcome/v-2/
2. Tutorials: https://data.princeton.edu/R
3. YouTube tutorials, e.g., from Google Developers: https://www.youtube.com/playlist?list=PLOU2XLYxmsIK9qQfztXeybpHvru-TrqAP
4. Me and your classmates.
5. Web search: "do X in R". Try variations of X until you find an answer. You will find many answers on https://stackoverflow.com/.

Classroom etiquette

To maintain a good learning environment for everyone, you must turn off all cell phones, laptops, and other electronic devices during class.

Exam dates

1. Midterm: 2/27, Th (in class)
2. Research project hard copy: 4/18, Th
3. Research project presentations: 4/21 T and 4/23, Th
4. Final: 5/4, T, 12:00pm

Tentative Schedule

#	Date	Topics	Readings, Ch.	Problem Sets Posted	Problem Sets Due
1	1/9	Big Data and statistical learning I	BDS 0, ISL 1-2	PS1	
2	1/14	Big Data and statistical learning II	BDS 0, ISL 1-2		
3	1/16	Regression I	BDS 2, ISL 3		
4	1/21	Regression II	BDS 2, ISL 3	PS2	PS1
5	1/23	Regression III	BDS 2, ISL 3		
6	1/28	Uncertainty	BDS 1, ISL 5		
7	1/30	Resampling methods	BDS 1, ISL 5		
8	2/4	Model selection	BDS 3, ISL 6		
9	2/6	Regularization I	BDS 3, ISL 6	PS3	PS2
10	2/11	Regularization II	BDS 3, ISL 6		
11	2/13	Classification I	ISL 4		
12	2/18	Classification II	ISL 4		
13	2/20	Regression splines	ISL 7		
14	2/25	Local regressions	ISL 7	PS4	PS3
15	2/27	**Midterm exam**			
16	3/3	Regression trees I	BDS 9, ISL 8		
17	3/5	Regression trees II	BDS 9, ISL 8		
		Spring Break			
18	3/17	Random forests	BDS 9, ISL 8		
19	3/19	Factors	BDS 7	PS5	PS4
20	3/24	Clustering I	BDS 7, ISL 10		
21	3/26	Clustering II	BDS 7, ISL 10		
22	3/31	Treatment effects I	BDS 5-6		
23	4/2	Treatment effects II	BDS 5-6		
24	4/7	Treatment effects III	BDS 5-6		
25	4/9	Network data	Slides	PS6	PS5
26	4/14	Text as data I	BDS 8		
27	4/16	Text as data II	BDS 8		
28	4/21	Research Project Presentation I			
29	4/23	Research Project Presentation II		PS6	

Research project

For the research project, you should use data to analyze a prediction or a causal inference question using methods from this course. You should write a paper, approximately 15 pages long, and explain the research question, data, methodology, and results. One possibility is to focus on a prediction problem trying various techniques learned in this class. Another possibility is to focus on estimating the causal effect. For the former numerous datasets can be found at https://www.kaggle.com/datasets, which is an online community of data scientists and machine learners. There will be oral presentations at the end of the course. The grade will be based both on the oral presentation and the hard-copy of the paper.

Grading

Your final grade will be based on:

- 15% problem sets (5 best)
- 15% research project
- 30% midterm
- 40% final

There will be no make-up exam for the midterm. If you miss a midterm exam because of a medical or family emergency, the final exam score will be 70% of your final grade.

References

[1] Susan Athey and Guido Imbens. *Machine learning methods economists should know about*. https://arxiv.org/abs/1903.10075.

[2] Leo Breiman. *Statistical modeling: the two cultures (with comments and a rejoinder by the author)*. Statist. Sci., 16(3):199-231, 2001.

[3] Matthew Getzkow, Bryan Kelly, and Matt Taddy. *Text as data*. Journal of Economic Literature, 57(3):535-574, 2019.

[4] Sendhil Mullainathan and Jann Spiess. *Machine learning: an applied econometric approach*. Journal of Economic Perspectives, 31(2):87-106, 2017.