ECON 590
Big Data and Machine Learning in Econometrics
Fall 2020

Instructor: Andrii Babii

Time and Location: T and Th, 3pm-4:15pm, Hanes Art Center - Rm 0121

E-mail: andrii@email.unc.edu

Office Hours: Th, 12:30pm-2:30pm, via Zoom

Prerequisites

Econ 400, 410, and 420.

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, textual 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 do not 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/.

Exam dates

1. Midterm: 9/29, Th
2. Research project hard copy: 11/5, Th
3. Research project presentations: 11/12 Th and 11/17, T
4. Final: TBA (see Registrar)

Research project

For the research project, you will analyze a prediction or a causal inference question using methods learned in the course. You will write a paper, approximately 15 pages long, where you will 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

Tentative Schedule

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

possibility is to estimate causal effects. 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. For the latter, you can find more examples on Sakai. 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.

Classroom etiquette

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

Community Standards in Our Course and Mask Use

This fall semester, while we are in the midst of a global pandemic, all enrolled students are required to wear a mask covering your mouth and nose at all times in our classroom. This requirement is to protect our educational community your classmates and me as we learn together. If you choose not to wear a mask, or wear it improperly, I will ask you to leave immediately, and I will submit a report to the Office of Student Conduct. At that point you will be disenrolled from this course for the protection of our educational community. An exemption to the mask wearing community standard will not typically be considered to be a reasonable accommodation. Individuals with a disability or health condition that prevents them from safelywearing a facemask must seek alternative accommodations through the Accessibility Resources and Service. For additional information, see https://carolinatogether.unc.edu/university-guidelines-for-facemasks/.

Title IX Resources

Any student who is impacted by discrimination, harassment, interpersonal (relationship) violence, sexual violence, sexual exploitation, or stalking is encouraged to seek resources on campus or in the community. Please contact the Director of Title IX Compliance (Adrienne Allison Adrienne.allison@unc.edu), Report and Response Coordinators in the Equal Opportunity and Compliance Office (reportandresponse@unc.edu), Counseling and Psychological Services (confidential), or the Gender Violence Services Coordinators (gvsc@unc.edu; confidential) to discuss your specific needs. Additional resources are available at safe.unc.edu.

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.