Econ 224 - Statistical Learning and Causal Inference
Spring, 2021

Course intructor: Karun Adusumilli

Email: akarun@sas.upenn.edu

Office: Perelman building, 631

Office hours: Tuesday 11am-12noon & 5-6pm

Recitation Instructors: Lucie L'Heude, Office hours: TBA

Course description: This course will teach you how to apply modern machine/statistical learning methods - potentially including (but not limited to) penalized regression, Random Forests, k-means clustering, and LDA - for analyzing economic data. While emphasis will be placed on employing machine learning methods for estimating causal effects, we will also discuss other economically relevant applications such as forecasting, text analysis and analyzing individual heterogeneity. After completing this course, you should be able to carry out machine learning analyses yourself for addressing economically relevant questions.

Prerequisites: The prerequisite for this course is Econ 104 or by permission from the instructor. To do well in this course you will need to be comfortable with Econometric methods typically covered in Econ 104 such as regression, instrumental variables and limited dependent variables (logit regression).

You are also expected to be proficient with the basics of R programming.

Learning: Econ 224 will be based around active learning. Each lecture is assosciated with an R Notebook called R Labs covering computing and other details for the material discussed in the lecture. I expect you to read this before the lecture. The workload for Econ 224 will be fairly high, approximately 6–7 hours of time spent outside of class per week. The lectures themselves will be a mix of traditional ones, where I discuss machine learning methods, and R Lab discussions, where I discuss programming aspects.

The RNotebooks also describe the reading material expected of you before each lecture. So you should go over these carefully. I will not be discussing these Notebooks in great detail in all the lectures. The ones I discuss in the lecture are marked as such in the schedule of classes (see last page). Some will be discussed by your RI in recitations. For the rest, you should go over them yourself, and ask questions on Piazza or in ofice hours if you have any questions.

Required Text: There are three required texts for this course:

- “An Introduction to Statistical Learning” (ISL) by James, Witten, Hastie, & Tibshirani: https://statlearning.com/ISLR%20Seventh%20Printing.pdf
- “Pattern Recognition and Machine Learning” (PRML) by Christopher M Bishop: http://users.isr.ist.utl.pt/~wurmd/Livros/school/Bishop%20-%20Pattern%20Recognition%20And%
%20Springer%20%202006.pdf
- “R for Data Science” (RDS) by Wickham & Grolemund: https://r4ds.had.co.nz/

All books are freely available from urls listed above. Printed versions are available on Amazon. ISL is the main textbook reference. PRML will be used to cover only a subset of the material.

Required Software: We will use the statistical package R via a front-end called RStudio throughout the course. Both R and RStudio are free and open source. Installation instructions appear on the last page of this syllabus. You will be taught machine learning methods through a series of R Notebooks, some of which are discussed in lecture as well. Additional R resources are listed on the last page of this syllabus. It is important that you download and update to the latest version of R.

Optional Texts: For students who want a deeper theoretical grounding in the material covered in Econ 224, I will assign optional readings from:

- “The Elements of Statistical Learning” (ESL) by Hastie, Tibshirani, and Friedman http://www.web.stanford.edu/~hastie/ElemStatLearn/

Note that this is purely optional and will not appear on problem sets or exams. Like its counterpart ISL, ESL is available as a free download from the authors’ website.

Course Policies

Class participation: All lectures and recitations will be recorded and posted to Canvas. However, I would urge you to try and attend as many lectures as possible. Since Econ 224 is an active learning course, you should make sure you are prepared before each lecture. To encourage participation, I will be awarding 5% of the final grade in participation credits. Participation can take many forms: Showing up in the lectures or recitations, asking and answering questions on Piazza, coming to the office hours. This does not mean that you absollutely need to attend lectures in you are in very different time zone: What I want to see is evidence that you are actively participating and working on the course and keeping up with the material. Participation points are discretionary and given in consultation with the RI.

Academic Integrity: All suspected violations of the code of academic integrity as set forth in the Pennbook will be reported to the Office of Student Conduct. Confirmed violations will result in a failing grade for the course.

Department Policies: All Economics Department course policies are in force in Econ 224 even if they are not explicitly listed on this syllabus. See: https://economics.sas.upenn.edu/undergraduate/course-information/course-policies for full details.

Piazza: We will be using an online discussion forum called Piazza, accessible via Canvas, for all written communication in this course. We will use Piazza to make course announcements, answer questions about course material and respond to private messages from individual students regarding personal issues. By asking your question and getting an answer on Piazza, you create a positive externality: other students benefit from your questions and you benefit from theirs. You can even post anonymously if you like. The RI and I will actively moderate Piazza both to answer questions and approve (or correct) answers written by your fellow-students. All written communication for Econ 224 should be directed to Piazza, not to the instructors' personal email accounts.

Grading and Assignments

Grades for this course will be determined based on 5 homeworks, 2 take home mid-terms, participation, and most importantly, a final project. Specifically,

$$\begin{aligned}\text{Overall score} &= (20\times\text{Homeworks}) + (12\times\text{midterm1}) + (13 \times \text{midterm2})\\ &\quad + (50 \times \text{Final project}) + (5 \times \text{Participation}).\end{aligned}$$

Course Curve: There will be no curve in Econ 224. This course will demand a larger amount of work than other courses, but provided that you put in the time and effort, you will do well.

Homeworks: I will assign 5 problem sets over the course of the semester. Each problem set is due at 11:59am on the Thursday morning after it was assigned. Problem set solutions should be submitted electronically to canvas and include both an .pdf file generated from RMarkdown and the .Rmd file used to create it. (In week 1 we will explain RMarkdown and how to use it to generate pdf documents.) Your grade for this component of the course will be calculated by averaging your 4 highest problem set scores. You may discuss problem set questions and how to solve them with your fellow students, but your code and write-up must be your own. Specifically, I expect you to adhere to an "empty hands policy." If you meet another student to discuss the problem set, you should leave the room with “empty hands,” i.e. no written or digital notes of your discussion. In particular, this means that you may not share code files with one another. If you discuss the problem set with your classmates, please indicate which students you discussed with at the top of your write-up. Failing to adhere to this collaboration policy constitutes a violation of academic integrity.

Exams: There will be one mid-term on the week of March 3, a second one around the week of April 14. The mid-term is a take-home one and you can work on it for up to 3 days. Unlike the homeworks, the mid-terms focus on the conceptual understanding of the material. You are free to use any resource, including the textbook, course notes, recordings, and even the internet to help you with your exam. However, you are not allowed to discuss the exam with any ‘human’: this implies talking about the exam with your fellow students is not allowed, nor is posting on online discussion forums. In particular, you are not to email, text, call, chat, or talk to anyone about the exam except with me and your TAs. There will be no makeup midterm. Sudden emergencies, of course, will be discussed and determined by the undergraduate chair.

Submitting exams: You should upload you solutions to Canvas as **PDF** files. It is fine to take pictures of your answers and upload them, but you should make sure that you convert all images into pdfs and collate them in the right order in which the questions were asked. You should also ensure that the pdfs are of good quality, and all your answers are legible. Failure to follow all these steps could result in your submission being declined.

Regrade Requests: Exam regrade requests must be made in writing within a week of receiving your graded exam. As we re-grade the entire exam, your score could rise or fall. You may not discuss your answers with an RI or the instructor before submitting a regrade request.

Final Project: This is the most important part of Econ 224. It is an opportunity for you to show what you have learned in the course by carrying out a substantive research project on a topic of your choice. Final projects will be carried out in groups of 3–4. You are welcome to form your own group; if you do not wish to form your own group, we will be happy to assign one for you. Four class meetings have been set aside for you to work on your group projects and get help from the instructor and RI: see the tentative semester plan on page 4 of this document. You will also be expected to work on the project outside of class. There will be no problem sets during week 13 giving you ample time to work on your projects. Projects are due at 11:59pm on Sunday, April 25th. In our final two class meetings of the semester, you and your group will give a short presentation sharing what you learned from your project. Full details and requirements for the final projects will be provided before Spring Break.

Installing R and RStudio

First, download and install R from https://cloud.r-project.org/. Second, download and install RStudio by visiting http://rstudio.org/download/desktop and clicking the link listed under "Recommended for Your System." If you have trouble, ask your RI or the instructor for help in office hours.

Here are links to some additional free resources to help you learn R:

- http://cran.r-project.org/other-docs.html
- http://www.twotorials.com/
- http://www.r-bloggers.com/google-developers-r-programming-video-lectures/
- http://cran.r-project.org/doc/contrib/Farnsworth-EconometricsInR.pdf
- http://www.ats.ucla.edu/stat/R/

Table 1. Schedule (Tentative)

Day	Date	Lecture	HW	R Lab	Exam
Wed	Jan 20	1. Introduction to ML			
Mon	Jan 25	2. R Lab 1: Data analysis in R		R Lab 1 (L)	
Wed	Jan 27	3. R Lab 2: Data visualisation in R	HW1	R Lab 2 (L)	
Mon	Feb 1	4. Statistical learning and regression		R Lab 3	
Wed	Feb 3	5. Resampling methods		R Lab 4 (L)	
Mon	Feb 8	6. High dimensional methods 1		R Lab 5	
Wed	Feb 110	7. High dimensional methods 2	HW2	R Lab 6 (L)	
Mon	Feb 15	8. Tree based methods I		R Lab 7 (L)	
Wed	Feb 17	9. Tree based methods II		R Lab 8 (L)	
Mon	Feb 22	10. Causal inference and ML		R Lab 9	
Wed	Feb 254	11. Doubly robust methods		R Lab 10 (L)	
Mon	Mar 1	12. Topics in causal inference		R Lab 11	
Wed	Mar 3	No class: Mid-term I			Mid-term I
Mon	Mar 8	13. Instrumental variables		R Lab 12 (R)	
Wed	Mar 10	No class: Spring break	HW3		
Mon	Mar 15	14. Classification methods		R Lab 13	
Wed	Mar 17	15. Unsupervised learning	HW4	R Lab 14	
Mon	Mar 29	16. Mixtures and EM algorithm		R Lab 15 (R)	
Wed	Mar 31	17. Introduction to Bayesian methods			
Mon	Apr 5	18. Approximate Bayesian inference		R Lab 16 (L)	
Wed	Apr 7	19. Latent Dirichlet Allocation	HW5	R Lab 17	
Mon	Apr 12	No class: Engagement day			
Wed	Apr 14	No class: Mid-term II			Mid-term II
Mon	Apr 19	20. Miscallaneous topics/Final Projects			
Wed	Apr 21	Final Projects			
Mon	Apr 26	Final Projects			
Wed	Apr 28	Final Projects			

Schedule

The schedule is tentative and will updated in the course of the semester. R Labs refer to R Notebooks. The letter L denotes that some parts of the Notebook will be discussed in lecture, R that it will be discussed in recitations. In all cases you are expected to go over these before the relevant lectures.