# 机器学习+大数据计量经济学
# Machine Learning+Econometrics of Big Data

Fall 2023

Instructor: 马骏

- Instructor: 马骏
- Email: jun.ma@ruc.edu.cn
- Office: 北校区一号楼西配楼106
- Time: 周一18:00-20:25
- Classroom: 2407
- Lecture slides, homework and answers will be posted on the course website: https://ruc-econ.github.io/Statistical_learning/.

# Course Description

- This course provides an introduction to modern statistical machine learning (ML) techniques and the intersection of ML and econometrics.

- Two parts: 1. introduction to selected topics from modern statistical ML theory; 2. applications of two ML techniques, LASSO and random forests in econometric estimation and inference of structural and causal effects.

- This course focuses on understanding of basic concepts, applications of the ML methods in both prediction and econometric (causal inference) contexts, instead of mathematical details. Basic statistical properties of the ML methods will be discussed in class.

- There will be conceptual homework questions that involve proofs, as well as applied homework questions that involve real-life data and applications.

# Textbooks and References

- James, G., Witten, D., Hastie, T. and Tibshirani, R.: An Introduction to Statistical Learning with Applications (ISL) in R 2nd edition, freely available: https://web.stanford.edu/~hastie/ISLR2/ISLRv2_website.pdf

- Wooldridge J.M.: Introductory Econometrics: A Modern Approach

- Wasserman, L.: All of Statistics: A Concise Course in Statistical Inference

- ISL is the required textbook for the first part of the course. The second part uses lecture slides that are based on academic papers. Wooldridge and Wasserman cover the econometric and statistical knowledge required for this course.

- Other useful references will be mentioned in class.

# Prerequisite

- Students are expected to have good knowledge about calculus, probability and statistics and linear algebra.

- Basic mathematical tools (e.g., conditional expectation, convergence in probability) will be reviewed in class.

- Prior knowledge in undergraduate econometrics (e.g., algebraic and asymptotic properties of OLS, instrumental variables, maximum likelihood, binary choice models) is required. We will review these concepts in class.

# Grading

- 20%*homework + 40%*mid-term exam + 40%*final project.
- Students are allowed to take a “cheat sheet” to the exam. The cheat sheet is a two-sided, handwritten, A4 paper where you can write any information.
- No final exam.
- Homework should be handed in before class.
- Late homework will not be accepted.
- Mid-term exam will be held after we finish the first part of the course.

# Software

- We will learn to use a free and popular statistical software R: https://www.r-project.org/.

- We use RStudio which provides a user interface for R: https://www.rstudio.com/ and RMarkdown for homework and final project: https://rmarkdown.rstudio.com/.

- For writing the final project and other academic papers, you may find it worthwhile to learn to use LYX: https://www.lyx.org and JabRef: https://www.jabref.org/.

- Tutorials on using R will be held in class.

- Applied homework questions will involve R applications.

# Final Project

- Students will be asked to find data from a published economics paper, re-do the econometric analysis using one of the ML methods taught in the course, discuss the differences and finally write the report as a short academic article.
- Detailed instructions for the final project will be provided later.
- The project should be finished individually. A proposal including which paper you are looking at and where to find the data should be sent to the instructor before the end of the semester.
- The submission deadline will be set as late as possible (maybe after the spring festival) and announced later before the end of the semester.

# Syllabus

1. Introduction to statistical ML (ISL ch. 1, 2)
2. Linear regression (ISL ch. 3)
3. Classification (ISL ch. 4)
4. Resampling methods: cross-validation and bootstrap (ISL ch. 5)
5. Regularized linear regression: subset selection, ridge and LASSO regressions (ISL ch. 6)
6. Nonparametric regression techniques: series regression and smoothing splines (ISL ch. 7)
7. Tree-based methods: regression/classification trees and random forests (ISL ch. 8)
8. Deep learning (ISL ch.10)
9. ML and econometrics
10. Adaptive LASSO
11. Post and double LASSO
12. Instrumental variable model with many controls and instruments
13. Treatment effects framework
14. Causal forests
15. Double ML