Machine Learning and Econometrics, Part I

Aims and objectives As it has been shown that the standard econometric tools are intrinsically not predictive, a new strand of economic literature increasingly uses Machine Learning procedures.

This course aims at making participants familiar with the potential of Machine Learning procedures for social scientists and aims to provide participants the tools for using Machine Learning in their analysis. The teaching will provide the basic knowledge and the way to implement the Machine Learning algorithms mostly used in the recent economic literature: **Lasso, Ridge, Elastic Net regressions, Random Forests, Gradient Boosting Machine, and Neural Networks**. The training will consist in replicating a few scientific papers using Machine Learning in economics.

After the course, participants are expected to have an improved understanding of the potential to perform Machine Learning, thus becoming able to master Machine Learning tasks both from a data mining and a predictive perspective.

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.

This course meets the **"Research and Discovery"** objective. Students immerse themselves in a research project and experience the reflection and revision involved in producing and disseminating original scholarship or creative works.

Questions for Students:

1. How do I establish my point of view, take intellectual risks, and begin producing original scholarship or creative works?
2. How do I narrow my topic, critique current scholarship, and gather evidence in systematic and responsible ways?
3. How do I evaluate my findings and communicate my conclusions?

Learning outcomes:

1. Frame a topic, develop an original research question or creative goal, and establish a point of view, creative approach, or hypothesis.
2. Obtain a procedural understanding of how conclusions can be reached in a field and gather appropriate evidence.
3. Evaluate the quality of the arguments and/or evidence in support of the emerging product.
4. Communicate findings in a clear and compelling ways.
5. Critique and identify the limits of the conclusions of the project and generate ideas for future work.

Textbooks We will be using a mixture of the following textbook.

1. (APM) Max Kuhn. Kjell Johnson, Applied Predictive Modeling. The book can be downloaded for free fom this link: Applied Predictive Modeling.

2. (ISL) Gareth, Witten, Hastie, and Tibshirani, An Introduction to Statistical Learning with Applications in R, Second edition. The book can be downloaded for free from An Introduction to Statistical Learning.

Problem sets There will be approximately 6 problem sets (Assignments) over the course of the two-weeks course program. Only 5 best will count towards the final grade. Problem sets are independent work and not a group project. You should submit your assignments on the appropriate due date.

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 courses. 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.

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 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. The grade will be based both on the oral presentation and the hard-copy of the paper.

Assessment Plan: **One Midterm exam**, a **Final exam** and a **Final class project** will be considered as the main assessment plan for this course. Each assignments including exams will have a computational assignments in which students will work intensively on data manipulation including their modeling aspect as well.

Grading Your final grade will be based on:

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

Prerequisites Knowledge of basic statistics and econometrics, a basic knowledge of R is preferable but not required.

Teaching methods This two-weeks course will be taught entirely online including office hours and exams. Material will be provided in course folder available on google drive or any other platform accessible to students.

- Complete lecture notes and class notes will be provided. Class notes do deviate from complete lecture notes and you are responsible for material as taught in the notes.
- Slides. Brief slides based off the class notes will be provided as well.
- Videos (If possible). New (pre-recorded) videos to be created for the course. Online lecture will be recorded as well and will be aviable to students if needed.

Outline of course content: The goal is to structure a course that must be completed within the two-weeks course program. The basic time frame will be as follows:

Lesson	**Main Topic**	**Readings/Assignments**
1	Introduction of the course A short boot camp in R/Rstudio Big Data and Statistical Learning I Big Data and Statistical Learning II	We will be using Markdown Making our first Markdown file
2&3	Regression I Regression II Regression III Uncertainty	Vietnam Shrimp Data. **Assignment 1 is due**
4&5	Resampling methods Model selection Model Tuning and Data Splitting/Recommendations Regularization I	**Assignment 2 is due**
6&7	Regularization II Classification I Classification II	**Assignment 3 is due**
–	**Midterm Exam**	
8&9	An overview of algorithms mostly used in the recent economic literature part 1 (Lasso, Ridge, Elastic Net regressions)	**Assignment 4 is due**
10&11	An overview of algorithms mostly used in the recent economic literature part 2 (Random Forests, Gradient Boosting Machine, and Neural Networks)	**Assignment 5 is due**
12&13	Comparing algorithms' performances Class Predictions Evaluating Predicted Classes	**Assignment 6 is due**
14&15	Replicating scientific papers using R part 1 (prediction tasks)	Reading Papers
16	Replicating scientific papers using R part 2 (counterfactual tasks) Rule Based Models —	Reading Papers
– –	**Research Project Presentation I** **Research Project Presentation II**	Oral Presentation Oral Presentation
–	**Final Exam**	Will be based on all topics covered

References Here are some useful references.

1. Susan Athey and Guido Imbens. Machine learning methods economists should know about, Annual Review of Economics, Vol. 11:685-725, 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.