Machine learning in applied economic analysis

Lecturers

- Thomas Heckelei
- Hugo Storm
- [Kathy Baylis]

Course Description

Machine learning (ML) now offers great potential for expanding the applied economist's toolbox. Data availability has dramatically increased and ML methods are well equipped to exploit large volumes of data more efficiently than traditional statistical methods. Researchers have developed and improved algorithms that push the boundaries of ML. The community has a strong open source tradition, including powerful DL libraries (e.g. tensorflow.org, pytorch.org) and pretrained models (e.g. VVGNet, ResNet), increasing the potential for adoption. In the past few years, economists have begun to realize that the predictive power of ML methods may not only be used as such, but can also improve causal identification. In this course, we introduce ML to applied economists by placing it in the context of standard econometric and simulation methods. We identify shortcomings of current methods used in agricultural and applied economics, and discuss both the opportunities and challenges afforded by ML to supplement our existing approaches.

The aim of this course is to allow students to understand when and where ML is appropriate, and what one can gain relative to standard econometric methods. We will focus on concepts, tuning and applications, not coding from scratch. Our aim is to enable students to learn specific applications relevant to their research. We use real-world data and Jupyter notebooks to allow students to collaborate with each other and run code remotely. Students will be exposed to ML models like LASSO and tree-based methods, as well as ensemble methods for prediction. They will also be introduced to neural networks. We will focus on interpretation as well as prediction accuracy. We will also cover the use of ML methods in causal estimation.

Competence to have / things to do for students before course starts:

- Requirement: Econometric background like in the course Advanced Applied Econometrics I or similar (with multiple regression, consistent and efficient estimation under Gauss-Markov and violations of it, endogeneity and IV estimation, static panels)
- Assignments to do before the course:
    - Reading Storm, Baylis, Heckelei 2020 (https://academic.oup.com/erae/article/47/3/849/5552525)
    - Prepare python and sklearn so that you can run a regression on sklearn (scikit-learn.org)
    - To be turned in: Access Jupyter notebook and run OLS using python (on sklearn)

General structure

The general daily organisation is such that there is a lecture in the morning (sometimes continues after lunch) and a lab session in the afternoon.

1. ML intro and prediction (2 days)

- **Lecture Day 1**: Introduction to ML
  - what matters for prediction? R-squared, prediction error, out-of-sample validation → ML with train/test split; bias vs variance, compare to econometrics (2 hrs)
  - regularization (LASSO, Ridge, ElasticNet) (1 hrs)

- **Lab Day 1: Introduction to notebooks; illustration with OLS and LASSO**
  - Prediction with simulated data - comparison with OLS to see non-linearities)
  - model selection via train/test split vs bias vs variance; regularization (LASSO; Trees/Forest)
  - compare results to OLS.

  Assignment 1: Fill out Jupyter notebook and summarize results on LASSO vs OLS comparison. Try two different train/test splits and compare the results

- **Lecture Day 2: Prediction cont'd and interpretation**
  - Trees, Random Forests, boosted trees and ensemble methods (2 hrs)
  - Interpretation of predictive models (1 hr)

- **Lab Day 2: Prediction and interpretation with forestry data using predefined methodology**
  - Split into groups to try different methods
  - Introduction of ensemble prediction
  - Preparation of the 1st stage or counterfactual for later causal analysis
  - Interpret errors in prediction

  Assignment 2: compare predictions and interpret

2. Prediction with Neural Networks (1 day)

- **Lecture Day 3**: **Introduction to NN and variations on NN**
  - General NN intro
  - autoencoders / unsupervised pre-training / feature extraction
  - transfer learning (if still fits)

- **Lab Day 3**: **Neural Network prediction of deforestation using sklearn**
  - Comparing versions with pre-training and without pre-training
  - binary classification (unchanged/decreased share of forest or pixel basis) and a continuous output (share of forest)

  Assignment 3: program own autoencoder

3. Causal identification (2 days)

- **Lecture Day 4**: **Introduction causal identification with ML**
  - Intro to the role of prediction in causal identification
  - Model selection for inference using LASSO double selection

- **Lab Day 4**: **NN for counterfactual and LASSO to discern true DGP**
    - Some further tweaking prediction of future forest cover (or forest cover change) for some aggregate area from baseline forest - same prediction objective as in day 2. Then use it to estimate the ATT.
    - LASSO to discern true DGP using simulated data

  Assignment 4 - Discuss which assumptions are needed for the predicted counterfactual to be used for inference.

- **Lecture Day 5**: **Double ML and deep IV**
    - Double ML
    - Deep IV
    - Matrix Completion for Panel Data

- **Lab Day 5: Causal ML and interpretation**
    - Double ML (Treatment is Protected Areas; outcome is deforestation)
    - some interpretation

Assignment 5 – To be done at home. Group project covering prediction task, counterfactual, causal identification and interpretation

Grading

Pass/fail based on handing in assignments