Machine Learning in Econometrics: Lecture 1

Instructor: Chaoyi Chen
NJE & MNB

September 5, 2023

General information

- The syllabus is the main source of information for the course. Please check the syllabus before asking questions.
- Assignments:
  - ASSN 1 Due Date: Tuesday, 31 October, In-class
  - ASSN 2 Due Date: TBA
  - Grace Period: Hand-in within two weeks, Max 80%.
- Exams:
  - Final: Date TBA, In-class
- Grades:
  - 20% Each Assignment, 20% Group Project, 40% Final Exam
- Lecture attendance:
  - Strongly encourage you to take all lectures.
  - Absent for more than three lectures: 0 mark will be given.

General information continue

- Book:
  - James, G., Witten, D., Hastie, T., & Tibshirani, R. (2022). *An introduction to statistical learning: With applications in R*. Springer.
- Lecture room: Info park campus, Lámfalussy lecture room, 4th floor.
- Office Hours: 1:30-2:30pm Tuesday, Info park campus, 1st floor.
- Email: chaoyi.chen@nje.hu

Term project

- Group number: you will form groups of 3-4 people and each group will turn in one project. (If there is no group available for you by the end of next week, please let me know).
- The paper should consist of **a simple empirical analysis** using **panel data**, and it should be at most **15 pages** long (incl. tables, figures, and bibliography). You can use either *word* or *LaTeX* to write your essay. Formal requirements will be strictly enforced. a Harvard style is recommended.
- All authors should have **same** contributions to the project formulation and paper preparation
- Project Due Date: 15 December. **NO EXTENSION**

Term project schedule suggestion

- Choose your group topic on forecasting by the end of September (e.g. stock return forecasting)
- A complete literature review with respect to your topic by the end of October
- After literature review, start to work on your topic. This includes
  - Introduce your motivation and outline the literature review.
  - Construct your econometric model and choose the empirical machine learning method
  - Obtain data
  - Estimate the model and summarize the findings
  - Conclude your paper
- Your first version of the paper should be available by the end of November.
- The final project deadline is 10 December.

Learning machines: AT-AT Walkers?

How supervised ML works in econometrics

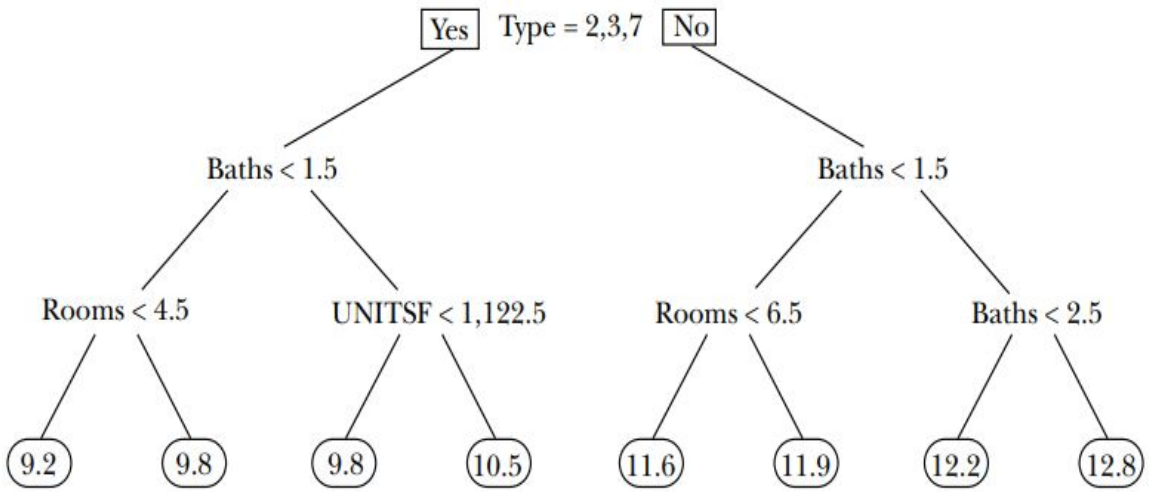
- Supervised machine learning (ML) methods include data-driven algorithms to predict **y** given **x**.
  - Pros: ML revolves around the problem of prediction: produce predictions of y from x. Comparing to the traditional methods (e.g. OLS), ML manages to fit complex and very flexible functional forms to the data without simply overfitting. Usually, it finds functions that work well out-of-sample.
  - Cons: Many economic applications, instead, revolve around parameter estimation: produce good estimates of parameters that underlie the relationship between y and x. It is important to recognize that machine learning algorithms are not built for this purpose. A black box!

How supervised ML works? An example

- We look to predict the value y of a house from its observed characteristics x based on a sample of n houses (y_i, x_i).
  - Algorithm: Take a loss function $L(y, \widehat{y})$ as an input and search for a function $\widehat{y}$ that has low expected prediction loss $E_{y,x}[L(y, \widehat{y})]$ on a *new* data point from the same distribution.
- For OLS, the $L(y, \widehat{y})$ would be the sum of squared errors and we may include all explanatory variables, x, in the regression. But why include all? Why not include interactions between variables?
  - Question? Could we choose the x automatically (e.g. which variables and interactions etc.)?

From regression to regression tree

- Machine learning searches for these automatically! Take an example of a typical learning function class: regression trees - mapping each vector of house characteristics to a predicted value.



The secret sauce

- How can a tree even be fitted here?
  - *Regularization*: measuring the complexity of a function. The less regularize, the better for in-sample, but more likely to overfit the model. For the regression tree case, instead of choosing the "best" overall tree, we could choose the best tree among those of a certain depth. The shallower the tree, the worse the in-sample fit, the less overfit. Tree depth is an example of a regularizer.
  - *Empirical tuning*: choosing the optimal level of regularization ("tune the algirithm"). In empirical tuning, we create an out-of-sample experiment inside the original sample. We fit on one part of the data and ask which level of regularization leads to the best performance on the other part of the data. This procedure can be enhanced by using cross-validation (CV). Finally, we pick the parameter with the best estimated average performance.

Most of supervised ML in one expression

- We can generalize this structure—regularization and empirical choice of tuning parameters—in the following central step.
  - **Step 1**: Conditional on a level of complexity, pick the best in-sample loss-minimizing function.

$$\text{minimize} \underbrace{\sum_{i=1}^{n} L(f(x_i), y_i)}_{\text{in-sample loss}}, \text{ over } \overbrace{f \in \mathcal{F}}^{\text{function class}}$$

$$\text{subject to } \underbrace{R(f) < c}_{\text{complexity restriction}}$$

  - **Step 2**: Estimate the optimal level of complexity using empirical tuning (normally choosing c as $R(f)$ is determined by the used ML method).

Supervised ML algorithms we will study

Function class $\mathcal{F}$ (and its parametrization)	*Regularizer $R(f)$*
Global/parametric predictors	
Linear $x\beta$ (and generalizations)	Ridge $\|\beta\|_2^2 = \sum_{j=1}^k \beta_j^2$ LASSO $\|\beta\|_1 = \sum_{j=1}^k \|\beta_j\|$
Local/parametric predictors	
Decision/regression trees	Depth, number of nodes
Random forest	Number of trees, number of variables used in each tree, size of bootstrap sample, complexity of trees
Nearest neighbors	Number of neighbors
Kernel regression	Kernel bandwidth

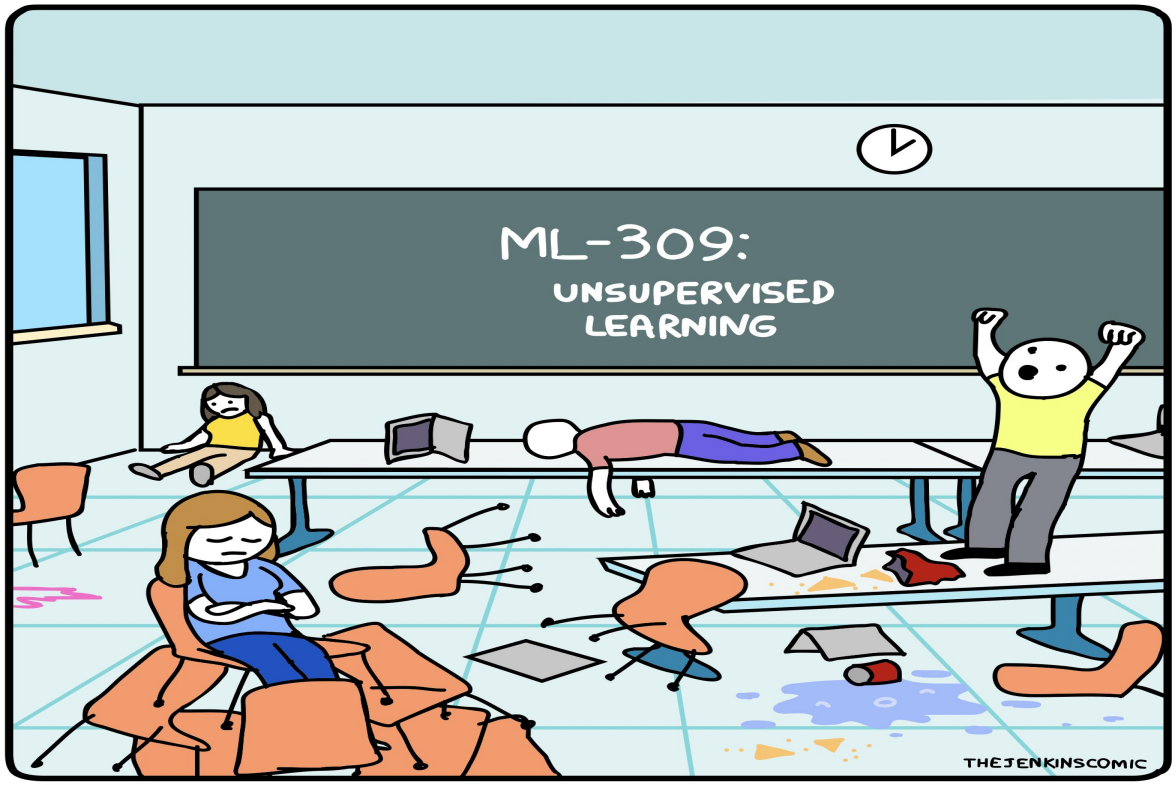
Supervised ML algorithms we will study cont.

Function class $\mathcal{F}$ (and its parametrization)	*Regularizer $R(f)$*
Mixed predictors	
Deep learning, neural networks	Number of levels, number of neurons per level, connectivity between neurons
Combined predictors	
Bagging	Number of draws, size of bootstrap samples (and individual regularization parameters)
Boosting	Learning rate, number of iterations (and individual regularization parameters)
Ensemble	Ensemble weights (and individual regularization parameters)

How unsupervised machine learning works in econometrics

- Unsupervised ML methods include data-driven algorithms to classify **x** (no **y**).
  - Pros: It can see what human minds cannot visualize. There is lesser complexity compared to the supervised learning task.
  - Cons: Hard to explain the economic intuition. A black box!
- Our course will focus on the supervised ML. However, we will also cover some of the unsupervised learning method (e.g. the principal component analysis).

Unsupervised learning



Learning outcomes

At the end of this course you should

- have knowledge of ML analysis.
- be able to understand the algorithms of various ML methods and apply ML methods on the prediction.
- be able to use R to perform an empirical analyses.
- be able to read and understand journal articles that make use of the methods introduced in this course.
- be able to make use of ML methods in your own academic/practical work.

Course outline

- Topic 1: Reviews
  - Linear regression and OLS
  - Cross-validation
  - Goodness-of-fit measures
- Topic 2: Global regression
  - Shrinkage methods: Ridge
  - Shrinkage methods: LASSO
  - Dimension reduction: PCA
- Topic 3: Local regression
  - Kernel regression
  - Regression trees
- Topic 4: Bootstrapping and model averaging
  - Bagging
  - Random forest
  - Boosting
  - Ensemble
- Topic 5: Deep learning
  - Single and multilayer layer neural networks
  - Convolutional layer neural networks