Machine learning methods for economics and finance applications

Course outline

JS

August, 2023

Course title	:	Machine learning methods for economics and finance applications
Lecturer	:	Jonas Striaukas. Email: js.fi@cbs.dk or jonas.striaukas@gmail.com
Lecture time	:	TBA
Auditorium	:	TBA
Course website	:	https://jstriaukas.github.io/teaching.html

Prerequisites : introduction to statistics, linear regression and some basic familiarity with the statistical software (e.g., R, Python) is expected, but otherwise it is a self-contained course. Each topic will cover an extensive set of real data examples where students will be shown step by step how to implement and apply methods thought during lectures. I will try to switch from R & Python so that all students get exposure to both languages.

Books : The course is self-contained. Nonetheless, I suggest the following books (in particular the first book in the list):[1]

i) relevant for the course:

- (main book for the course) James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, p. 18). Springer (New York).
  ▶ online copy: pdf.
- (more advanced book that covers similar topics as the main book) Hastie, T., & Friedman, J. H., Tibshirani, R. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2, pp. 1-758). Springer (New York). ▶ online copy: pdf.

ii) relevant for more curious students:

- (graduate level book – covers more recent topics) Fan, J., Li, R., Zhang, C. H., & Zou, H. (2020). *Statistical foundations of data science.* Chapman and Hall/CRC.
- (graduate level book – financial data examples) Nagel, S. (2021). *Machine learning in asset pricing.* Princeton University Press.

Exam : TBA

[1] I list an additional set of books for students who want to learn more about statistical learning methods, statistical theory, introductory probability theory for high-dimensional problems, etc., at the end of this outline.

Information about the course

Topics

The course will cover the following main topics:

- **Week 1:** Introduction to statistical/machine learning
  - Material: slides shinyapps slides pdf tablet friendly slides pdf
- **Week 2:** Linear regression
  - Material: slides shinyapps slides pdf tablet friendly slides pdf
- **Week 3:** Multiple linear regression
  - Material: slides shinyapps slides pdf tablet friendly slides pdf
- **Week 4:** Loss function, classification, logistic, & quantile regressions
  - Material: slides shinyapps slides pdf tablet friendly slides pdf
- **Week 5:** Guest lecture
  - Material: TBA
- **Week 6:** Principal component analysis
  - Material: slides shinyapps slides pdf tablet friendly slides pdf
- **Week 7:** Time series
  - Material: slides shinyapps slides pdf tablet friendly slides pdf
- **Week 8:** Resampling methods
  - Material: slides shinyapps slides pdf tablet friendly slides pdf
- **Week 9:** Optimization methods
  - Material: slides shinyapps slides pdf tablet friendly slides pdf
- **Week 10:** Introduction to advanced topics in machine learning
  - Material: slides shinyapps slides pdf tablet friendly slides pdf

Details on each topic

For each topic, I will discuss real and simulated data examples to show how theory is applied in practice. The amount of material for theory/methods vs empirics/practical implementation will be split to 50/50.

Week 1: *Introduction to statistical/machine learning.*
I will introduce the notation and contents of the course. I will provide an overview of machine/statistical learning, and different data types, with some empirical examples. I will then discuss parametric vs nonparametric methods linking them to machine learning. I will introduce the basic trade-offs such as the bias-variance trade-off, discuss regression versus classification, and the assessment of the model's accuracy.

Topic 2: *Linear regression.*
I will introduce the linear regression with a single independent variable. I will discuss the statistical properties, Gauss-Markov theorem, model miss-specification. Lastly, I will show how to estimate, predict and construct confidence intervals using linear regression model.

Topic 3: *Multiple linear regression.*
I will introduce the multiple linear regression (MLR) with several independent variables. I will discuss how to estimate and predict using MLR. I will then discuss multicollinearity issue in MLR and ways to overcome it. The topic will conclude with the potential issues of linear regression and some comparison with the basic nonparametric method such as K-nearest neighbor.

Topic 4: *Loss function, classification, logistic & quantile regressions.*
I will introduce the concept of the loss function. I will then discuss two important cases besides the least squares, i.e., logistic and quantile losses. I will discuss the properties of both cases and will focus on various applications for which these methods should be applied. I will then cover the classification methods with greater detail and an extensive set of data examples.

Topic 5: *Guest lecture.* (If this is not possible, I will add a topic on panel data after time series.)

Topic 6: *Principal component analysis.*
I will introduce the principal component analysis (PCA) as unsupervised learning method. I will show how to estimate principal components and discuss the basic properties of the method. I will discuss numerous important examples in economics and finance with various applications.

Topic 7: *Time series.*
I will introduce the concept of dependence and discuss the time series data. I will review some of the basic properties and models. I will then discuss the implications of the dependence for applying ML methods to such data.

Topic 8: *Resampling methods.*
I will introduce the resampling methods for ML: i) cross-validation, ii) bootstrap. I will discuss the implementation of these methods aiming to show the advantages and disadvantages when applied to certain type of data.

Topic 9: *Optimization methods.*
I will introduce the convexity as the main theoretical concept for optimization methods. I will discuss the gradient descent, and some of its extensions relevant for the implementation of the ML methods.

Topic 10: *Introduction to advanced topics in machine learning.*
I will introduce the high-dimensional regression problems and the tools to analyze such data (e.g., ridge, LASSO). I will briefly discuss about the neural networks and the causal machine learning. In this topic, I will not go into any technical details.

Instructions to install R

The main R software is available at https://cran.r-project.org. Once on this website, you need to select an file to download for your operating system (OS), so if you work on Windows you need to download and install Window `.exe` file. Please install the most recent version of R. I also strongly advise to install RStudio (free version of it) from https://rstudio.com. You need to install R prior to RStudio.

Install R packages: suppose you need to install an R package called '`forecast`'. You should write in RStudio console:

```
install.packages("forecast")
```

Additional material

List of useful books

The list is meant for students who want to get to know more about ML and want to keep studying different methods at the graduate level.

- (introductory book on high-dimensional stats) Giraud, C. (2021). *Introduction to High-Dimensional Statistics* (Vol. 112, p. 18). Chapman & Hall/CRC Monographs on Statistics and Applied Probability.
- (great book for LASSO methods) Bühlmann, P., & Van De Geer, S. (2011). *Statistics for high-dimensional data: Methods, theory and applications.* Springer Science & Business Media.
- (more advanced book on high-dimensional stats) Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint* (Vol. 48). Cambridge University Press.
- (introductory probability theory for high-dimensional problems) Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science* (Vol. 47). Cambridge university press.
- (reference book for concentration inequalities) Boucheron, S., Lugosi, G., & Massart, P. (2013). *Concentration inequalities: A non-asymptotic theory of independence.* Oxford university press.
- (fun book on learning methods) Cesa-Bianchi, N., & Lugosi, G. (2006). *Prediction, learning, and games.* Cambridge university press.