Machine Learning

Course outline

JS

August, 2022

Course title	:	Machine learning
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

Important : if you have any questions regarding teaching material, you can write m them via email after each lecture (or during office hours) — I will allocate 10-15 minutes at the beginning of the following lecture to answer your questions so that everybody benefit from the discussion.

Prerequisites : introduction to statistics, linear regression and some basic computing in statistical software (e.g., R, Python) is assumed, but otherwise, it is a self-contained course. I recommend reading through the introductory book (the first book in the list of recommended books) before the course to have a rough idea of what we will cover during the course.

Books : I try to make material self-contained so that you don't need to buy any book for the course. However, if you like the course and want to have a deeper understanding of a particular topic or machine learning methods in general, I suggest the following books:[1]

- (great introductory book) James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, p. 18). Springer (New York).
  ▶ online copy: pdf.
- (moderate level overview 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.
- (main book for the course – covers more recent topics) Fan, J., Li, R., Zhang, C. H., & Zou, H. (2020). *Statistical foundations of data science*. Chapman and Hall/CRC.
- (financial data examples) Nagel, S. (2021). *Machine learning in asset pricing*. Princeton University Press.

Exam : TBA

Software : students need to either install statistical software R or some other software for different languages (Python, Matlab, Julia, Octave, ...). I will be showing examples mainly in R and Python. I advise you to install the software before the course starts (if you don't have it

[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.

already). See below for the instructions on how to install R and Python. You are free to use Matlab/Julia/C++/etc., programming language of your choice to code the examples we cover in the class as well as homework assignments.

Information about the course

Topics

The course will cover the following main topics — the links are provided for the slides:

- **Topic 1:** Introduction to learning, multiple and nonparametric regression
  ▶ Material: slides shinyapps slides pdf tablet friendly slides pdf
- **Topic 2:** High-dimensional linear regression
  ▶ Material: slides shinyapps slides pdf tablet friendly slides pdf
- **Topic 3:** High-dimensional regression properties and generalized linear models (GAMs)
  ▶ Material: slides shinyapps slides pdf tablet friendly slides pdf
- **Topic 4:** Prediction, loss functions and M-estimators
  ▶ Material: slides shinyapps slides pdf tablet friendly slides pdf
- **Topic 5:** Introduction to deep learning
  ▶ Material: slides shinyapps slides pdf tablet friendly slides pdf
- **Topic 6:** Introduction to causal machine learning
  ▶ Material: slides shinyapps slides pdf tablet friendly slides pdf

Details on each topic

Topic 1: *Introduction to learning, multiple and nonparametric regression.*
We start the section with a review of challenges and advantages of *Big data*. We then will look at the least squares estimator, discuss some of the basic properties. We will then cover multiple and nonparameteric regression models and estimation. We will conclude with empirical risk analysis and several new articles that analyze the estimator. A brief introduction to learning theory – empirical risk, loss function, concentration and other basic concepts will be introduced throughout the section.[2]
Topic 2: *Penalized least squares.*
In this section we will cover penalized least squares estimation. Variable selection, different penalization methods will be discussed (folded-concave penalized least squares, ℓ_1 penalization, aka LASSO, other alternatives). We will review some of the numerical algorithms and implementation

[2]**IMPORTANT:** Throughout the course all theoretical derivations will serve us the purpose of understanding certain learning techniques, how and why they work, and under which assumptions. I do not expect students to reproduce more complicated proofs, however, I expect that students understand and can interpret theory results, i.e., can argue why certain methods work for certain types of data, what properties should the data satisfy, etc.

for relevant minimization problems. We will conclude the section with structured nonparameteric models and structured penalization such as group LASSO.
Topic 3: *Penalized least squares: properties.*

Topic 4: *Prediction, loss functions and M-estimators.*

Topic 5: *Introduction to deep learning.*

Topic 6: *Introduction to causal machine learning.*

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

- (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.