Introduction to statistical/machine learning and R programming

Applied Machine Learning for Economics and Finance

Jonas Striaukas

Course details

Basic info:

- **My email**: jost.fi@cbs.dk or jonas.striaukas@gmail.com
- **Lecture time**: TBA
- **Auditorium**: TBA
- **Office hours**: TBA
- **Course website**: https://jstriaukas.github.io/teaching

Exam:

- **Structure**: TBA
- **When**: TBA

What I expect from you:

- Understand the concepts we learn in the class. In particular derivations of some simple theoretical results as well as full understanding of more complex theory.
- Be creative, active during class presentations and work hard! And try **not** to miss classes...

Machine learning, computing, etc.

"The purpose of computing is insight, not numbers."

Richard Hamming

Topics of the course

- Introduction to statistical/machine learning and R programming
    - *Brief introduction on statistical learning, data, some definitions. Basic programming in R statistical software.*
- Predictive linear regression
    - *Revisit linear regression, its basic properties. Estimation, assessment of the parameter estimates. Discussion on best linear predictor. Construction of predictions and prediction confidence intervals.*
- Multiple linear regression and regularization
    - *Estimation.*
- Loss function, classification. Logistic and quantile regression
    - BLAH BLAH
- Guest lecture
    - BLAH BLAH

Topics of the course

- Principal component analysis and factor models
    -
- Time series models
    -
- Resampling methods for ML
    -
- Optimization methods for ML
    -
- Introduction to advanced topics in machine learning
    -

Nowadays, Big Data are ubiquitous: from the internet, biology and medicine to government, business, economics, finance, ...

Some quotes:

- "*There were 5 exabytes of information created between the dawn of civilization through 2003, but that much information is now created every 2 days*", according to Eric Schmidt, the CEO of Google,in 2010.
- "*Big data is not about the data*", according to Gary King of Harvard University.

Do we need ML or even AI to understand economics and/or finance data?

- **Yes!** ML is not that different from classical statistics/econometrics...

Big data examples in economics and finance:

- high-frequency financial assets data (e.g., stocks, bonds, fx, derivatives, ...);
- large panels of economic data (e.g., 131 macroeconomics time series FRED MD database with monthly updates, McCracken and Ng (2016));
- spatial data (e.g., state-level data in the US, Euro area data);
- text-based data (e.g., newspaper articles, GDELT project; EC news data);
- ... .

What is statistical/machine learning?

Notation

Let $x_{i,j}$ represent the value of the j^{th} variable for the i^{th} observation, where $i = \{1, 2, \ldots, n\} \triangleq [n]$ and $j = \{1, 2, \ldots, p\} \triangleq [p]$.

We let $\mathbf{X}$ (typically called the *design* matrix) denote an $n \times p$ matrix whose (i, j)th element is $x_{i,j}$. That is,

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & x_{13} & \ldots & x_{1p} \\ x_{21} & x_{22} & x_{23} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & x_{n3} & \ldots & x_{np} \end{pmatrix}$$

At times we will instead be interested in the rows/columns of

$$\text{Rows: } x_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix} \text{ Columns: } \mathbf{x}_j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{pmatrix} \text{ Design matrix: } \mathbf{X} = \left(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p\right)$$

What is statistical/machine learning?

Supervised learning

Suppose that we observe a quantitative response $\mathbf{Y}$ and p different predictors, $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$.

Assume that there is some relationship between $\mathbf{Y}$ and $\mathbf{X} = (\mathbf{x}_i)_{i \in [p]}$, which can be written in the very general form:

$$\mathbf{Y} = f(\mathbf{X}) + \mathbf{e}$$

Questions:

- What is f? Can we infer/estimate it from the data?
- How assumptions on f after the estimation and/or prediction?
- Can we generalize f such that the prediction of $\mathbf{Y}_{\text{new}}$ is as accurate as possible, given that we have $\mathbf{X}_{\text{new}}$?
- Assumptions on $\mathbf{e}$, $\mathbf{X}$, $\mathbf{Y}$. Binary outcomes.
- Etc ...

What is statistical/machine learning?

Supervised learning

Suppose that we observe a quantitative response $\mathbf{Y} = (y_1, \ldots, y_n)^\top$ and p different predictors, $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$.

Assume that there is some relationship between $\mathbf{Y}$ and $\mathbf{X} = (\mathbf{x}_i)_{i \in [p]}$, which can be written in the very general form:

$$\mathbf{Y} = f(\mathbf{X}) + \mathbf{e}$$

Learning $f(.)$ from the data, i.e., the function that links $\mathbf{Y}$ and $\mathbf{X}$ is called *supervised learning*.

What is statistical/machine learning?

Supervised learning — data example

What is statistical/machine learning?

Unsupervised learning

Suppose that we observe a matrix of data points, i.e., $\mathbf{X} = (\mathbf{x}_i)_{i \in [p]}$. An example could be an image.

Questions:

- Suppose p is large. Can we summarize the data in compact way?
- Can we learn patterns in $\mathbf{X}$?
- Can we summarize the data in compact way and use this compact information in a predictive model?

Learning patterns in $\mathbf{X}$ is called *unsupervised learning*.

What is statistical/machine learning?

Unsupervised learning — data example

McCracken, M. W., and S. Ng (2016): "FRED-MD: A monthly database for macroeconomic research," Journal of Business & Economic Statistics, 34(4), 574–589.