National Research University Higher School of Economics

The International College of Economics and Finance

Syllabus for *Data Science For Economists*, 2019-2020.

Lecturer

Vitalijs Jascisens

Course Pre-requisites

- Statistics;
- Mathematics for Economists.

In the second part of the course I will present and derive statistical properties of various estimators. **To be able to follow this part of the course students should have a certain level of mathematical maturity**. In practice, this means that students should have done some simple mathematical proofs before taking this class (for example, understand "epsilon-delta" arguments in the context of limits of sequences).

Course description

The objective of this course is to provide students with a hands on introduction to data science in economics (or more broadly to data science in the social sciences).

The course consists of three parts:

1. Introduction to programming;
2. Overview of the most commonly used machine learning algorithms;
3. Time permitting, an introduction to causal inference and applications of machine learning algorithms to causal inference.

In the first part of the course students will learn basic programming using computing language R. Obtained skills will allow to implement all methods taught subsequently. Additionally, students will learn how to explore and analyse structured and un-structured data sets. Finally, provided introduction to programming will also be useful in subsequent courses in econometrics and economics. At first, to gain intuition, we will study how to solve the problem in a brute force manner and then explore R packages and built in functions to deal with a problem in the most efficient manner.

In the second part of the course we will focus on most commonly used machine learning algorithms. We will cover regression techniques (parametric, nonparametric and high-dimensional), classification methods, resampling methods, model selection, unsupervised learning and text analysis (time permitting). Finally, in the last part of the course we will cover research papers which have recently applied machine learning methods to causal inference in economics.

Learning Objectives and Learning Outcomes

At the end of the course students should have developed the following skills:

- Ability to write simple computer programs using computing language R;
- Implement basic machine learning algorithms;
- Understand assumptions and statistical properties of machine learning algorithms;
- Be able to use machine learning algorithms to solve real world business problems.

The course contributes to the development of the following general competencies:

- Ability to work with information: to find, evaluate and use information from various sources, necessary to solve scientific and professional problems;
- Ability to do research, including problem analysis, setting goals and objectives, identifying the object and subject of research, choosing the means and methods of research, assessing its quality;
- Ability to collect and analyse the data;
- Able to solve problems in professional sphere based on analysis and synthesis;
- Capability to work in a team;

Methods of Instruction

The following methods and forms of study are used in the course:

- Lectures (2 hours a week);
- Classes (2 hours a week). During classes class-teachers will cover a selection of both applied and theory problems;
- Teachers' consultations;
- Self study.

In total the course includes: 32 hours of lectures, 32 hours of classes.

Main Textbooks

First Part Of The Course:

1. G1 - Grolemund, Garrett, Hands-On Programming with R., Sebastopol, CA: O'Reilly, 2014;
2. G2 - Grolemund, G., Wickham, H., 2015. R for Data Science. O'Reilly, Sebastopol, CA;

Second Part Of The Course:

1. JWHT - James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani, An Introduction to Statistical Learning: With Applications in R, 1st ed. 2013, corr. 7th printing 2017 ed., New York: Springer-Verlag New York Inc., 2013; JWHT will be the main text for the second part. **In order to be able to follow lectures and ask questions students are expected to read the relevant chapters before the lecture.**

2. C - Shalizi, Cosma, Advanced Data Analysis from an Elementary Point of View [WWW Document], n.d. URL http://www.stat.cmu.edu/~cshalizi/ADAfaEPoV/ (accessed 08.08.19); This is a somewhat more mathematically advanced book. Where possible I will complement JWHT with materials from C (I will announce readings from C at least one week before the lecture).

Third Part of The Course:

Remark: This list is somewhat preliminary, depending on our progress more papers will be announced.

1. Athey, S., 2018. The Impact of Machine Learning on Economics (NBER Chapters). National Bureau of Economic Research, Inc.;
2. Belloni, A., Chernozhukov, V., Hansen, C., 2014. High-Dimensional Methods and Inference on Structural and Treatment Effects. Journal of Economic Perspectives 28, 29–50;
3. Einav, L., Levin, J., 2014. Economics in the age of big data. Science 346, 1243089;
4. Gentzkow, M., Kelly, B.T., Taddy, M., 2017. Text as Data (NBER Working Paper No. 23276). National Bureau of Economic Research, Inc;
5. Mullainathan, S., Spiess, J., 2017. Machine Learning: An Applied Econometric Approach. Journal of Economic Perspectives 31, 87–106.

Supplementary Textbooks

First Part Of The Course:

1. M1 - McKinney, W., 2017. Python for Data Analysis: Data Wrangling with Pandas, Numpy, and IPython, 2nd ed. O'Reilly Media, Inc, USA, Beijing;
2. W - Wickham, Hadley, Advanced R, Boca Raton, FL: CRC Press Inc, October 2014.

Second Part Of The Course:

1. EH - Efron, Bradley and Trevor Hastie, Computer Age Statistical Inference: Algorithms, Evidence, and Data Science, New York, NY: Cambridge University Press, 2016;
2. M2 - Murphy, K.P., 2012. Machine Learning: A Probabilistic Perspective. MIT Press;

Grading System and Examination Type

The Final grade is based on 4 problem sets (done in groups of two students) and the final exam. **No late submissions of problem sets will be accepted.** The final grade is determined as follows:

$$FG = 0.15 \times A1 + 0.15 \times A2 + 0.15 \times A3 + 0.15 \times A4 + 0.4 \times FE$$

Where:

- FG = Final course grade;
- A1-4 = Grade in the assignment 1-4;
- A2 = Grade in the assignment 2;
- A3 = Grade in the assignment 3;
- A4 = Grade in the assignment 4;
- FE = Final Exam

Remark: The final exam will based both on readings mentioned in the syllabus and on additional readings which will be announced in the class as the course progresses. Additionally, the final exam might include additional topics which are only covered in classes (and not in lectures). **Therefore**, **Students are strongly encouraged to attend all classes.**

Sample materials for knowledge assessment are available in ICEF Information system at https://icef-info.hse.ru.

All grades are given initially out of 100. The final grades are also transferred to 10- and 5-points grades in accordance with the ICEF Grading Regulations (par.3). Retakes are organized in accordance with the HSE Interim and Ongoing Assessment Regulations (incl. Annex 8 for ICEF). Grade determination after retakes is done in accordance with ICEF Grading Regulations (par. 5).

Course Outline.

Remark: Additional topics (without altering drastically the structure of the course) might be announced as the course progresses;

Remark: For the second part of the course I am listing only relevant chapters from JWHT. JWHT does not cover all topics which will be covered during the class. **Thus, it is essential (to fully benefit from the class) to read additional readings assigned during the class (which also might be covered in the exam, classes and problem sets).** As mentioned before, these readings will be announced at least one week prior to the class.

1. Introduction to Data Science In Economics.

Overview of the course. Data science process. Overview of the usage of machine learning algorithms in economics. General introduction to R and Rstudio. Data types and classes in R. R objects. Subsetting of R objects.

Readings:

- G1: Chapters 1 – 5.

2. Control Structures and Functions.

Control structures in R: if – else, for loops, nested for loops, while loops, repeat loops, next, break. Functions in R.

Readings:

- G1: Chapters 1, 6, 7 and 9;
- G2: Chapters 19, 21.

3. Vectorized Computation and Data Aggregation.

Usage of "apply" family in R. Usage of a data.table package in R.

Readings:

- Slides.

4. Working With Text and the Web.

Usage of a stringr package in R. Regular expressions in R. The structure of HTML and XML documents. Writing simple XML and HTML parsers.

Readings:

- G2: Chapter 14.

5. Introduction to Statistical Learning

Classification and regression, bias – variance tradeoff, mean squared error, in sample and out of sample mean squared error, parametric and nonparametric statistical models. First look at ordinary least squares (OLS) estimator.

Readings:

- JWHT: Chapters 2 and 3.

6. Large Sample Properties of OLS

Review of "asymptotic tools". "Linear plus noise representation" of OLS. Derivation of asymptotic properties of OLS. OLS in the matrix form.

Readings:

- JWHT: Chapter 2.

7. Classification

Introduction to the logistic regression. Estimation of the logistic regression via MLE. Discussion of the statistical properties of MLE. Linear discriminant analysis, quadratic discriminant analysis.

Readings:

- JWHT: Chapter 4.

8. Resampling Methods

Cross – Validation, jacknife, bootstrap, subsampling.

Readings:

- JWHT: Chapter 5.

9. Linear Model Selection and Regularization

Best subset selection, stepwise forward selection, backward forward selection, dimension reduction methods. Ridge regression. Lasso regression.

Readings:

- JWHT: Chapter 6.

10. Nonparametric Estimation

Polynomial regression, step functions, basis functions, regression splines, smoothing splines, kernel – methods, generalized additive models.

Readings:

- JWHT: Chapter 7.

11. Tree Based Methods

Decision trees, bagging, random forests, boosting.

Readings:

- JWHT: Chapter 8.

12. Support Vector Machines

Readings:

- JWHT: Chapter 9.

13. Unsupervised Learning

Readings:

- JWHT: Chapter 10.

14. - 16. Topics in Causal Inference

Readings:

- To be confirmed as the course progresses

Distribution of hours for topics and types of work

No	Topics titles	**Contact hours**	
		Lectures	**Classes**
1.	Introduction to Data Science In Economics	2	2
2.	Control Structures and Functions.	2	2
3.	Vectorized Computation and Data Aggregation.	2	2
4.	Working With Text and the Web.	2	2
5.	Introduction to Statistical Learning and the First Look at OLS.	2	2
6.	Large Sample Properties of OLS.	2	2
7.	Classification.	2	2
8.	Resampling Methods.	2	2
9.	Liner Model Selection and Regularization.	2	2
10.	Nonparametric Estimation.	2	2
11.	Tree Based Methods	2	2
12.	Support Vector Machines	2	2
13.	Unsupervised Learning	2	2
14.	Additional Topics	6	6
	Total:	32	32