Higher School of Economics
National Research University
Faculty of Economic Sciences

Course: Data Mining

Approved by
the academic council of the program
Report № 2.9.5-12/11 from 25 May 2018

Author	Maria Alexandrovna Veretennikova
Credits	3
Contact Hours	32
Self-study Hours	82
Course	4
Course Format	Full-time

Course Description

Course title: Data Mining autumn 2018

Course pre-requisites: Mathematical statistics, basics of probability theory

Course type: Elective

Abstract: Data Mining (DM) is closely related to Computer Science and Artficial Intelligence. This field focuses on recognizing patterns in datasets. The main purpose of DM is to produce and study learning algorithms, that may be implemented with computers and imitate intelligent behaviour in dealing with data, for example in classifying objects into groups.

Learning objectives and expected outcomes

The aim of this course is to provide the basic skills for analysis of statistical data, in particular for regression and classification purposes. An important objective is operational knowledge of the studied techniques, hence there will be a practical side to the course as well as the theoretical side. We will use the statistical software R for the practical part, but it is not compulsory to be acquainted with R in advance.

Learning Outcomes:

1. Understanding and ability to present/describe the methods studied in this course in mathematical terms
2. Application of these methods to simple problems/examples
3. Understanding the limitations and benefits when using each of these DM methods
4. Using R software for implementing methods studied in the course

Topics

1. Data preprocessing and feature selection,
2. Linear regression, regularization, splines,
3. Model selection,
4. Classification: kNN, Naive Bayes, Support Vector Machines, logistic regression,
5. Model comparison

Reading list and other information resources

Course literature includes:

- (Required) G. James, D. Witten, T. Hastie, R. Tibshirani, An Introduction to Statistical Learning: with Applications in R, Springer, 2015, (ISL/ISLR).

You may be advised to look at specific paragraphs of these and/or other books/papers as we progress with the course.

Among useful websites:

- Optional: A Russian open source machine learning platform http://www.machinelearning.ru
- Optional: An open source learning platform about Support Vector Machines: http://www.svms.org

Grading System and knowledge assessment guidelines

Homework and grading criteria

Theoretical questions will be available from the beginning of the course, one by one, as we move on from one topic to another. Practical questions will be announced once certain blocks of topics have been covered. There will be 3 homeworks and all deadlines will be announced in advance. All will count towards your course grade.

Homework should be submitted before class. Late work will not be accepted. There will be 2 in-class tests, all of which will count towards your course grade.

All work should be written clearly in complete sentences and in a reasonable font size, preferably on A4 paper. Your thinking should be explained as well as it is possible.

For the midterm and for the term final you will be allowed to use a single-sided formula sheet that is handwritten/typed, a calculator will be allowed

too. Note that you are only allowed to write formulas on the formula sheet, there should be no words on it.

Your overall grade will be determined as follows:

H = homework (theoretical and practical parts), S = seminar/practical class activity,
T = term final (note, this is not an exam, but a term final),
C = in-class tests,
P = announced online platform learning: any online learning resources on the topic of decision trees, ensembles of trees, such as random forests and boosting

For the independent learning part P you are required to select a paper about an application of decision trees and write a mathematical essay about the method and the findings of the authors. The submission date is the same as the day of the final. You need to read about what kind of problems decision tree algorithms and their modifications/alterations may be useful for, very carefully explain what optimization problems this involves, e.g minimizing specific loss functions at every step of the algorithm. Explain how this is applied to a specific problem of your choice, referencing the articles, books and/or lectures you have used for the essay. The criteria for assessment are as follows: the overall clarity, mathematical depth, attention to detail and specificity of the method you select, presence of references. You should write at least 2 pages of A4, excluding references. There will be no upper bound for the number of pages. The essay should be concise, neat, with a balance between mathematical detail and specificity of the problem. Ask me if you have further questions about this task.

Extra credit may be given for a presentation on an advanced topic such as genetic algorithms for feature selection and clustering techniques.

Overall grade = 5*T + 2*H + 1*S + 1*C +1*P.

Grading scale
8 - 10 points: Great understanding of the theory and practical skills, correct solutions
6-7 points: One solution is not complete, theory and practical skills are not presented fluently
4-5 points: Solutions are not complete, there are gaps in knowledge/skills
1 - 3 points: Unsatisfactory work

Academic Integrity

The Higher School of Economics strictly adheres to the principle of academic integrity and honesty. Accordingly, in this course there will be a zero-tolerance policy toward academic dishonesty. This includes, but is not limited to, cheating, plagiarism (including failure to properly cite sources), fabricating

citations or information, tampering with other students' work, and presenting a part of or the entirety of another person's work as your own. HSE uses an automated plagiarism-detection system to ensure the originality of students' work. Students who violate university rules on academic honesty will face disciplinary consequences, which, depending on the severity of the offense, may include having points deducted on a specific assignment, receiving a failing grade for the course, being expelled from the university, or other measures specified in HSE's Internal Regulations.

Methods of Instruction and Expectations in regard to student activities

Classes: Attendance and keeping track on class activities is your responsibility. All teaching is during the lectures and seminars.

Seminars/practicals: Students are expected to actively participate in the practicals/seminars. This implies solving proposed problems, explaining solutions to the class and answering other questions on the topic discussed. At the beginning of several classes some time (approximately 15 minutes) will be allocated to answering short questions, which test your understanding of concepts we had just covered.

Missing tests/many classes/the final: Do not hesitate to contact me regarding unusual circumstances, preventing you from course evaluation and general attendance.

Disclosure: The instructor reserves the right to make any changes that she deems academically advisable. Changes will be announced in class. It is your responsibility to keep up with any changed policies.