**Course Syllabus**

|  |  |
| --- | --- |
| Title of the course | Machine Learning and Data Mining |
| Title of the Academic Programme  | Advanced social analysis |
| Type of the course  | Elective |
| Prerequisites | Foundations of calculus, probability theory and mathematical statisticсs |
| ECTS workload | 2 |
| Total indicative study hours | Directed Study | Self-directed study  | Total |
| 28 | 48 | 76 |
| Course Overview | Rapid developments of social networking sites, online media and other internet-generated data are making machine learning an essential analytical tool of social scientists and industrial analysts of social data.Nowadays, social researchers should not only be able to work with different types of data, such as textual or relational data, but should also have skills to interpret results obtained with complex mathematical algorithms. In this course students will, first, get to know basic machine learning algorithms and their main advantages and limitations for social science goals. Second, they will obtain skills to work with machine learning software / codes. Third, by the end of the course all students will produce small-scale research project that may be used in their Master theses. Depending on the level of a student group, the course may be based on one of the following software tools:1. Orange
2. R
3. Python
 |
| Intended Learning Outcomes (ILO) | By the end of the course the students will be able:* To collect data from the Internet for social science research;
* To analyze them with machine learning tools;
* To visualize results of the analysis;
* To present the resulting project.
 |
| Teaching and Learning Methods | Seminars, team work, project work |
| Content and Structure of the Course |
| **№** | **Topic / Course Chapter** | **Total** | **Directed Study** | **Self-directed Study** |
| **Lectures** | **Tutorials** |
| 1 | Introduction to machine learning. | 1 | 1 |  |  |
| 2 | Overview of mathematical formalism necessary for understanding of machine learning | 2 | 1 | 1 |  |
| 3 | Data collection and existing databases for machine learning | 4 | 1 | 1 | 2 |
| 4 | Data preprocessing | 5 | 1 | 2 | 2 |
| 5 | Cluster analysis (Kmeans, Cmeans, Hierarchical clustering) | 9 | 1 | 2 | 6 |
| 6 | Linear models of classification and regressions | 7 | 1 | 2 | 4 |
| 7 | KNN and SVM classification | 11 | 1 | 2 | 8 |
| 8 | Naïve Bayes classifier | 11 | 1 | 2 | 8 |
| 9 | Topic Modeling | 15 | 1 | 4 | 10 |
| 10 | Decision trees | 11 | 1 | 2 | 8 |
| **Total study hours** | 76 | 10 | 18 | 48 |
| Indicative Assessment Methods and Strategy  | * 80% - class work
* 20% - Presentation of project
 |
| Readings / Indicative Learning Resources | Mandatory 1. Murphy K. P., Machine learning : a probabilistic perspective, The MIT Press, 2012, 004 M96, ISBN: 978-0-262-01802-9
2. Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshiran, An Introduction to Statistical Learning (with applications in R), 1st ed. 2013, Corr. 7th printing 2017 Edition
3. Orange Data Mining Library Documentation, Release 3, 2018, https://media.readthedocs.org/pdf/orange-data-mining-library/latest/orange-data-mining-library.pdf

Optional1. Orange: <https://orange.biolab.si/docs/>
2. Orange video course: <https://www.youtube.com/playlist?list=PLmNPvQr9Tf-ZSDLwOzxpvY-HrE0yv-8Fy&disable_polymer=true>
3. RStudio Cheat Sheets: <https://www.rstudio.com/resources/cheatsheets/>
4. Python Tools for Machine Learning: <https://www.cbinsights.com/research/team-blog/python-tools-machine-learning/>
 |
| Indicative Self- Study Strategies | **Type** | **+/–** | **Hours** |
| Reading for seminars / tutorials (lecture materials, mandatory and optional resources) |  | 25 |
| Assignments for seminars / tutorials / labs |  |  |
| E-learning / distance learning (MOOC / LMS) |  |  |
| Fieldwork |  |  |
| Project work |  |  |
| Other (please specify) |  |  |
| Preparation for the exam |  |  |
| Academic Support for the Course | Academic support for the course is provided via LMS, where students can find: guidelines and recommendations for doing the course; guidelines and recommendations for self-study; samples of assessment materials |
| Facilities, Equipment and Software | Orange - an open-source data visualization, machine learning and data mining toolkitR Studio - free and open-source integrated development environment (IDE) for RJupyter Notebook - open-source software for python |
| Course Instructor | Sergei Koltcov |