Government of Russian Federation

Federal State Autonomous Educational Institution of High Professional Education

«National Research University Higher School of Economics»

National Research University
High School of Economics
Faculty of Computer Science

Syllabus for the course
«Machine Learning and Data Mining»

(Машинное обучение и майнинг данных)

010402.68 «Applied Mathematics and Informatics»,
«**Data Sciences**» Master program

Authors:
Dmitry I. Ignatov, Candidate of sciences (Ph.D), associate professor, dignatov@hse.ru

Approved by: Head of Data Analysis and Artificial Intelligence Department, Sergey O. Kuznetsov

Recommended by:

Moscow, 2014

1. Teachers

Author, lecturer: Dmitry Ignatov, National Research University Higher School of Economics, Department of Data Analysis and Artificial Intelligence, associate professor

2. Scope of Use

The present program establishes minimum demands of students' knowledge and skills, and determines content of the course.

The present syllabus is aimed at department teaching the course, their teaching assistants, and students of the Master of Science program 010402.68 «Data Sciences»,

This syllabus meets the standards required by:

- Educational standards of National Research University Higher School of Economics;
- Educational program «Data Sciences» of Federal Master's Degree Program 010402.68 «Applied Mathematics and Informatics», 2014;
- University curriculum of the Master's program in «Data Sciences» for 2014.

3. Summary

The course "**Machine Learning and Data Mining**" introduces students to new and actively evolving interdisciplinary field of modern data analysis. Started as a branch of Artificial Intelligence, it attracted attention of physicists, computer scientists, economists, computational biologists, linguists and others and become a truly interdisciplinary field of study. In spite of the variety of data sources that could be analyzed, objects and attributes that from a particular dataset poses common statistical and structural properties. The interplay between known data and unknown ones give rise to complex pattern structures and machine learning methods that are the focus of the study. In the course we will consider methods of Machine Learning and Data Mining. Special attention will be given to the hands-on practical analysis of the real world datasets using available software tools and modern programming languages and libraries.

4. Learning Objectives

Learning objectives of the course "**Machine Learning and Data Mining**" (MLDM) are to familiarize students with a new rapidly evolving filed of machine learning and mining, and provide practical knowledge experience in analysis of real world data.

5. Learning outcomes

After completing the study of the discipline "Machine Learning and Data Mining", the student should:

- Know basic notions and terminology used in MLDM
- Understand fundamental principles of modern data analysis
- Learn to develop mathematical models of MLDM
- Be capable of analyzing real world data

After completing the study of the discipline "Machine Learning and Data Mining" the student should have the following competences:

Competence	**Code**	**Code (UC)**	**Descriptors (indicators of achievement of the result)**	**Educative forms and methods aimed at generation and development of the competence**
The ability to reflect developed methods of activity.	SC-1	SC-M1	The student is able to reflect developed mathematical methods for machine learning and data mining (data sciences)	Lectures and tutorials, group discussions, presentations, paper reviews.
The ability to propose a model to invent and test methods and tools of professional activity	SC-2	SC-M2	The student is able to improve and develop research methods as applicable to machine learning and data mining (data sciences)	Classes, home works.
Capability of development of new research methods, change of scientific and industrial profile of self-activities	SC-3	SC-M3	The student obtains necessary knowledge in machine learning and data mining, which is sufficient to develop new methods	Home tasks, paper reviews
The ability to describe problems and situations of professional activity in terms of humanitarian, economic and social sciences to solve problems which occur across sciences, in allied professional fields.	PC-5	IC-M5.3_5.4_5.6_2.4.1	The student is able to describe data analysis problems in terms of computational mathematics.	Lectures and tutorials, group discussions, presentations, paper reviews.
The ability to detect, transmit common goals in the professional and social activities	PC-8	SPC-M3	The student is able to identify mathematical aspects in machine learning and data mining tasks, evaluate correctness of the used methods, and their applicability in each current situation	Discussion of paper reviews; cross discipline lectures

6. Place of the discipline in the Master's program structure

The course "Machine Learning and Data Mining" is a course taught in the first year of the Master's program 010402.68 "Data Sciences" and is a base course for specialization "Intelligent Systems and Structural Analysis"

Prerequisites

The course is based on knowledge and understanding of

- Discrete mathematics
- Algorithms and data structures
- Linear algebra
- Theory of probability and statistical analysis

It also requires some programming experience in one of the languages:

- Python
- Matlab

7. Schedule

One pair consists of 1 academic hour for lecture and 1 academic hour for classes after lecture.

№	Topic	Total hours	Contact hours		Self-study
			Lectures	Seminars	
1	Introduction to Machine Learning and Data Mining	10	2	2	6
2	Clustering and basic techniques.	11	2	2	7
3	Classification and basic techniques.	11	2	2	7
4	Frequent Itemset Mining and Association Rules.	11	2	2	7
5	Feature Selection and Dimensionality Reduction. Outlier detection.	11	2	2	7
6	Recommender Systems and Algorithms.	11	2	2	7
7	Ensemble Clustering and Classification.	10	2	2	6
8	Multimodal relational clustering.	11	2	2	7
9	Artificial Neural Methods and Stochastic Optimization. Elements of Statistical Learning.	11	2	2	7
10	Machine Learning Tools and Big Data.	11	2	2	7
	Total	108	20	20	68

8. Requirements and Grading

	Type of work	Characteristics		
		1	2	
Type of grading	Homework	5		Solving homework tasks and examples.
	Special homework – research projects		2	Independent modelling and verification of research papers results
	Exam		1	Written exam
Final				

9. Assessment

The assessment consists of classwork and homework, assigned after each lecture. Students have to demonstrate their knowledge in each lecture topic concerning both theoretical facts, and practical tasks' solving. All tasks are connected through the discipline and have increasing complexity.

Final assessment is the final exam. Students have to demonstrate knowledge of theory facts, but the most of tasks would evaluate their ability to solve practical examples, present straight operation, and recognition skills to solve them.

The grade formula:

The exam will consist of 10 problems, giving 10 points each, total 100 points for the exam

Final course mark *is obtained from the following formula*:

$$Ofinal = 0{,}6*Ocumulative+0{,}4*Oexam.$$

The grades are rounded in favour of examiner/lecturer with respect to regularity of class and home works. All grades, having a fractional part greater than 0.5, are rounded up.

Table of Grade Accordance

Ten-point Grading Scale	**Five-point Grading Scale**	
1 - very bad 2 – bad 3 – no pass	Unsatisfactory - 2	**FAIL**
4 – pass 5 – highly pass	Satisfactory – 3	**PASS**
6 – good 7 – very good	Good – 4	
8 – almost excellent 9 – excellent 10 – perfect	Excellent – 5	

10. Course Description

The following list describes main topics covered by the course with lecture order.

Topic 1. Introduction to Machine Learning and Data Mining

Content: Introduction to modern data analysis. Machine Learning. Data Mining and Knowledge Discovery in Data Bases. Basic tasks and examples.

Topic 2. Clustering and basic techniques.

Content: The task of clusterization. K-means and its modifications (k-medoids and fuzzy c-means clustering). Density-based methods: DB-scan and Mean Shift. Hierarchical clustering. Criteria of quality.

Topic 3. Classification and basic techniques.

Content: The task of classification. 1-Rules. K-Nearest Neighbours approach. Naïve Bayes. Decision Trees. Logistic Regression. Quality assessment: loss-function, error-matrix, cross-validation and learning curves.

Topic 4. Frequent Itemset Mining and Association Rules.

Content: Frequent itemsets. Apriori algorithm. Association rules. Interestingness measures: support and confidence. Connection with Lattice Theory and Formal Concept Analysis.

Topic 5. Feature Selection and Dimensionality Reduction. Outlier detection.

Content: Feature selection versus feature extraction. Singular Value Decomposition, Latent Semantic Analysis and Principal Component Analysis. Boolean matrix factorization. Projections.

Topic 6. Recommender Systems and Algorithms

Content: Collaborative filtering. User-based and item-based methods. Slope one. Association rules based and bicluster-based techniques. Quality assessment: MAE, precision and recall. SVD-based approaches: SVD++ and time-SVD.

Topic 7. Ensemble Clustering and Classification.

Content: Ensemble methods of clusterization for k-means partitions' aggregation. Ensemble methods of classification: Bagging, Boosting, and Random Forest.

Topic 8. Multimodal relational clustering

Content:
Biclustering. Spectral co-clustering. Triclustering. Two-mode networks. Folksonomies and resource-sharing systems. Multi-modal approaches.

Topic 9. Artificial Neural Methods and Stochastic Optimization. Elements of Statistical Learning.

Content:
Artificial Neural Networks. Basic idea of Deep Learning. Gradient descent. Statistical (Bayesian) view on Machine learning.

Topic 10. Machine Learning Tools and Big Data.

Content:
Orange, Weka, Knime, and Scikit Learn. Machine Learning for Big Data: Mahout and MALLET.

11. Term Educational Technology

The following educational technologies are used in the study process:
- discussion and analysis of the results of the home task in the group;
- individual education methods, which depend on the progress of each student;
- group projects on analysis of real data.

12. Recommendations for course lecturer

Course lecturer is advised to use interactive learning methods, which allow participation of the majority of students, such as slide presentations, combined with writing materials on board, and MLDM software tools for demonstration and practicing purposes. The course is intended to be introductory, that is rather broad in nature, but it is normal to differentiate tasks and projects in a group if necessary, and direct fast or more dedicated learners to solve more complicated tasks. The final group project and computational homework tasks are inevitable constituents of the cource.

13. Recommendations for students

Lectures are combined with classes. Students are invited to ask questions and actively participate in-group discussions and projects. There will be special office hours for students, which would like to get more precise understanding of each topic. Teaching assistant will also help you. All tutors are ready to answer your questions online by official e-mails that you can find in the “contacts” section. Note that the final mark is a cumulative value of your term activity and final results.

14. Sample final exam questions

1. For the given dataset define unknown labels of objects by three different machine learning techniques (Decision Tress, Naïve Bayes, 1-Rule).

   Gender Age Education Salary Grant credit?
   1 M young higher high +
   2 F young special high +
   3 F middle higher moderate +
   4 M old higher high +
   5 M young higher low -
   6 F middle high school moderate -
   7 F old special moderate -

8 F young special high t
9 F old higher moderate t
10 M middle special moderate t

2. For the given dataset and minimal values of support=1/3 and confidence=1/2 find a) all frequent itemsets and b) generate corresponding associations rules.
3. For the given dataset obtain different partitions into clusters and compare the results (k-means, k-medoids, hierarchical clustering) using one of the appropriate distance measures (Hamming, Euclid, or Manhattan distance).

15. Reading and Materials

15.1. Required Reading

1. Mohammed J. Zaki and Wagner Meira, Jr., Data Mining and Analysis: Fundamental Concepts and Algorithms, Cambridge University Press, 2014
2. Peter Flach Machine Learning: The Art and Science of Algorithms that Make Sense of Data, Cambridge University Press, 2012

15.2. Recommended Reading

1. C.M. Bishop. Pattern Recognition and Machine Learning Springer (2006) http://research.microsoft.com/en-us/um/people/cmbishop/PRML/index.htm
2. D. Barber Bayesian Reasoning and Machine Learning, Cambridge University Press, 2012
3. http://www-stat.stanford.edu/~tibs/ElemStatLearn/
4. J. Han, M. Kamber, J. Pei. Data Mining: Concepts and Techniques, Third Edition. — Morgan Kaufmann Publishers, 2012. — 703 p.
5. T. Hastie, R.Tibshirani, J. Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second Edition, Springer 2009
6. T. Mitchel. Machine Learning, McGraw Hill, 1997. http://www.cs.cmu.edu/~tom/mlbook.html
7. Witten, E. Frank, M. Hall. Data Mining: Practical Machine Learning Tools and Techniques, 2011, Morgan Kaufmann Publishers http://www.cs.waikato.ac.nz/ml/weka/book.html

Supplementary reading:

8. B. G. Mirkin. Core Concepts in Data Analysis: Summarization, Correlation, Visualization, Springer, 2011, 388 p.
9. B.G. Mirkin Clustering: A Data Recovery Approach. CRC Press, 2012.

15.3. List of review papers

1. Rakesh Agrawal, Ramakrishnan Srikant: Fast Algorithms for Mining Association Rules in Large Databases. VLDB 1994: 487-499
2. Rakesh Agrawal, Maria Christoforaki, Sreenivas Gollapudi, Anitha Kannan, Krishnaram Kenthapadi, Adith Swaminathan: Mining Videos from the Web for Electronic Textbooks. ICFCA 2014: 219-234

3. Inderjit S. Dhillon. 2001. Co-clustering documents and words using bipartite spectral graph partitioning. In Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '01). ACM, New York, NY, USA, 269-274.
4. Dmitry I. Ignatov, Sergei O. Kuznetsov, Jonas Poelmans: Concept-Based Biclustering for Internet Advertisement. ICDM Workshops 2012: 123-130
5. Leonid Zhukov Spectral Clustering of Large Advertiser Datasets. Technical report. Overture R&D, 2003.
6. Leandro Balby Marinho, Alexandros Nanopoulos, Lars Schmidt-Thieme, Robert Jäschke, Andreas Hotho, Gerd Stumme, Panagiotis Symeonidis: Social Tagging Recommender Systems. Recommender Systems Handbook 2011: 615-644
7. Dmitry I. Ignatov, Sergei O. Kuznetsov, Jonas Poelmans, Leonid E. Zhukov: Can triconcepts become triclusters? Int. J. General Systems 42(6): 572-593 (2013)
8. Dmitry Gnatyshak, Dmitry I. Ignatov, Sergei O. Kuznetsov: From Triadic FCA to Triclustering: Experimental Comparison of Some Triclustering Algorithms. CLA 2013: 249-260
9. Bing Liu, Lei Zhang: A Survey of Opinion Mining and Sentiment Analysis. Mining Text Data 2012: 415-463
10. Charu C. Aggarwal, ChengXiang Zhai: A Survey of Text Classification Algorithms. Mining Text Data 2012: 163-222
11. Bo Pang, Lillian Lee, Shivakumar Vaithyanathan: Thumbs up? Sentiment Classification using Machine Learning Techniques. CoRR cs.CL/0205070 (2002)
12. Thorsten Joachims: Text Categorization with Suport Vector Machines: Learning with Many Relevant Features. ECML 1998: 137-142
13. Claudio Carpineto, Carla Michini, Raffaele Nicolussi: A Concept Lattice-Based Kernel for SVM Text Classification. ICFCA 2009:237-250
14. Steven P. Crain, Ke Zhou, Shuang-Hong Yang, Hongyuan Zha: Dimensionality Reduction and Topic Modeling: From Latent Semantic Indexing to Latent Dirichlet Allocation and Beyond. Mining Text Data 2012:129-161
15. David M. Blei, K. Franks, Michael I. Jordan, I. Saira Mian: Statistical modeling of biomedical corpora: mining the Caenorhabditis Genetic Center Bibliography for genes related to life span. BMC Bioinformatics 7: 250 (2006)
16. David M. Blei, Andrew Y. Ng, Michael I. Jordan: Latent Dirichlet Allocation. Journal of Machine Learning Research 3: 993-1022 (2003)
17. David M. Blei: Probabilistic topic models. Commun. ACM 55(4): 77-84 (2012)
18. Thomas Hofmann: Probabilistic Latent Semantic Analysis. UAI 1999:289-296
19. Thomas Hofmann: Unsupervised Learning by Probabilistic Latent Semantic Analysis. Machine Learning (ML) 42(1/2):177-196 (2001)
20. Deerwester S, Dumains ST, Furnas G, Landauer TK, Harshman R. Indexing by latent semantic analysis. J Am Soc Inf Sci 1990, 41:391–407.
21. Nicholas E. Evangelopoulos. Latent semantic analysis WIREs Cogn Sci 2013, 4:683–692

15.4. Course telemaintenance

All material of the discipline are posted in informational educational site at NRU HSE portal www.ami.hse.ru. Students are provided with links to research papers, electronic books, data and software.

16. Equipment

The course requires a laptop, projector, and acoustic systems.
It also requires opportunity to install programming software, such as:

- Python
- Matlab
- Orange, Weka or their analogs.

on student personal computers.

Lecture materials, course structure and syllabus are prepared by Dmitry Ignatov