Syllabus
Data Analysis

Andrey Shestakov, avshestakov@hse.ru
Faculty of computer science, school of software engineering
Meeting Minute # ___dated _____ 20_

1. Course Description

a) Pre-requisites

Mathematical analysis, linear algebra, basic theory of probability, programming in Python

b) Abstract

This course presents the foundations of rapidly developing scientific field called intellectual data analysis or machine learning. This field is about algorithms that automatically adjust to data and extract valuable structure and dependencies from it. The automatic adjustment to data by machine learning algorithms makes it especially convenient tool for analysis of big volumes of data, having complicated and diverse structure which is a common case in modern "information era".

During this course most common problems of machine learning are considered, including classification, regression, dimensionality reduction, clustering, collaborative filtering and ranking. The most famous and widely used algorithms suited to solve these problems are presented. For each algorithm its data assumptions, advantages and disadvantages as well as connections with other algorithms are analyzed to provide an in-depth and critical understanding of the subject.

Much attention is given to developing practical skills during the course. Students are asked to apply studied algorithms to real data, critically analyze their output and solve theoretical problems highlighting important concepts of the course. Machine learning algorithms are applied using python programming language and its scientific extensions, which are also taught during the course.

The course is designed for students of the bachelor program "Software Engineering" at the Faculty of Computer Science, HSE.

2. Learning Objectives

The objective of the are:

- make students familiar with the major problems of data analysis, solved with machine learning (classification, regression, dimensionality reduction, clustering, collaborative filtering and ranking)
- make students acquainted with the major algorithms to solve stated problems

- give students a critical understanding of the subject, highlighting the limitations of each algorithm, data assumptions each algorithm relies upon, its strengths and weaknesses.
- teach students one of the most commonly used tools for machine learning: python programming language together with its major data analysis libraries - numpy, scipy, pandas, matplotlib and machine learning library scikit-learn.
- give students practical experience from application of studied methods to real datasets.

3. Learning Outcomes

- to know major problems of data analysis, solved with machine learning
- to know major algorithms to solve stated problems
- to understand dependencies between algorithms, their advantages and disadvantages
- to know python programming language together with its major data analysis libraries - numpy, scipy, pandas, matplotlib and machine learning library scikit-learn.
- to understand, which kinds of algorithms are more appropriate for what kinds of data
- to know the whole pipeline of research & development of machine learning methods
- to know, how to transform data to make it more suitable for machine learning algorithms
- to understand scientific articles about data analysis and machine learning.

4. Course Plan

	Number of lessons (1 lesson=90min)		**Self-study (astronomical hours)**
Section name	**Lectures**	**Seminars**	
Introduction to data science and machine learning.	1	0	12
K nearest neighbours method.	1	1	12
Decision trees.	1	1	12
Model evaluation	1	1	12
Linear classifier methods	1	1	12
Support vector machines	1	1	12
Regression	1	1	12
Boosting	1	1	12
Other ensemble methods: bagging, RandomForest, etc.	1	1	12
Feed Forward Neural networks	1	1	12
Convolutional Neural networks	1	1	12
Feature selection and dimentionality reduction	1	1	12

Introduction to NLP	1	1	12
Clustering	1	1	12
Introduction to recommendation systems	1	1	12
Total:	16	16	192

5. Reading List

a) Required

Lecture slides for the course on the course page

b) Optional

- Links to the additional materials on the course page.
- Machine learning in action / P. Harrington. – Shelter Island: Manning, 2012.
- Trevor Hastie, Robert Tibshirani, Jerome Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction., 2nd Edition, Springer, 2009
- Kevin P. Murphy. Machine Learning: A Probabilistic Perspective. Massachusetts Institute of Technology. 2012
- Mohri M. - Foundations of machine learning. 2012
- Christopher M. Bishop. Pattern Recognition and Machine Learning. Springer. 2006.

6. Grading System

Grade takes values 4,5,…10. Grades, corresponding to 1,2,3 are assumed unsatisfactory. Exact grades are calculated using the following rule:

[score]>0.35 => 4,

[score]>0.45 => 5,

…

[score]>0.95 =>10,

where [score] is calculated using the following rule:

[score]=0.7*[cumulative score] + 0.3*[exam2 score]

[cumulative score] = 0.8 * [homework score] + 0.2 * [exam1 score] + 0.2*[competition score]

[homework score] – total sum of obtained points divided by the total sum of maximum achievable points for all homeworks.

[exam1 score] – proportion of successfully answered theoretical and practical questions during exam after module 3

[exam2 score] – proportion of successfully answered theoretical and practical questions during exam after module 4

[competition score] – score for the competition in machine learning.

Participation in machine learning competition is aimed to give students an opportunity to get extra points and to get practical experience of application of studied methods to real data analysis task. The task is to make a prediction system and [competition score] is set according to the accuracy of the developed prediction system.

7. Examination Type

Knowledge of students is assessed by evaluation of their home assignments and exams. Home assignments divide into theoretical tasks and practical tasks. In theoretical tasks students are asked to answer questions and to prove mathematical statements. In practical tasks students are asked to program certain data processing and prediction methods, apply them to datasets and provide reports with their results and comments.

The course lasts during the 3rd and 4th modules. There are two exams during the course – after the 3rd module and after the 4th module respectively. Each of the exams evaluates theoretical knowledge and understanding of the material studied during the respective module.

8. Methods of Instruction

Lectures and seminars

9. Special Equipment and Software Support (if required)

- A projector for lectures and seminars
- Whiteboard with markers
- Computers with internet access and stable version of anaconda distribution