**Федеральное государственное автономное образовательное учреждение**

**высшего образования**

**"Национальный исследовательский университет**

**"Высшая школа экономики"**

Факультет компьютерных наук

Департамент программной инженерии

**Рабочая программа дисциплины**

Анализ данных

 (на английском языке)

Data analysis

для образовательной программы «Программная инженерия»

направления подготовки 09.03.04 «Программная инженерия»

уровень – бакалавр

Разработчик программы: Китов В.В., v.v.kitov@yandex.ru

Одобрена на заседании департамента программной инженерии «\_\_\_»\_\_\_\_\_\_\_\_\_\_\_\_ 2016 г.

Руководитель департамента Авдошин С.М. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Рекомендована Академическим советом образовательной программы

«\_\_\_»\_\_\_\_\_\_\_\_\_\_\_\_ 2016 г., № протокола\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Утверждена «\_\_\_»\_\_\_\_\_\_\_\_\_\_\_\_ 2016 г.

Академический руководитель образовательной программы

Шилов В.В. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Москва, 2016

*Настоящая программа не может быть использована другими подразделениями*

*университета и другими вузами без разрешения подразделения-разработчика программы.*

**1. Course Description**

**1.1. Title of a Course**

Data analysis

**1.2. Pre-requisites**

Mathematical analysis, linear algebra, basic theory of probability.

**1.3. Course Type**

Compulsory

**1.4. 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 course is to make students familiar with

- 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.

The course contributes to the development of the following competencies: УК-2, УК-9, УК-10, ПК-1, ПК-2, ПК-3, ПК-6, ПК-11, ПК-12, ПК-13, ПК-15.

**4. Course Plan**

|  |  |  |  |
| --- | --- | --- | --- |
| **Section name** | **Number of lessons (1 lesson=45min)** |  | **Self-study (astronomical hours)** |
| **Lectures** | **Seminars** |  |
| Introduction to data science and machine learning. | 1 | 0 | 12 |
| K nearest neighbours method. | 2 | 2 | 12 |
| Decision trees. | 2 | 2 | 12 |
| Model evaluation | 1 | 1 | 12 |
| Bayes decision theory | 1 | 2 | 12 |
| Linear classifier methods | 1 | 1 | 12 |
| Support vector machines | 1 | 1 | 12 |
| Kernel trick | 1 | 1 | 12 |
| Regression | 1 | 1 | 12 |
| Boosting | 2 | 2 | 12 |
| Other ensemble methods: bagging, RandomForest, etc. | 1 | 1 | 12 |
| Neural networks | 2 | 2 | 12 |
| Feature selection | 1 | 1 | 12 |
| Feature extraction | 1 | 1 | 12 |
| Density estimation: parametric, mixture, KDE. | 1 | 1 | 12 |
| Clustering | 1 | 1 | 12 |
| Collaborative filtering | 1 | 1 | 12 |
| Ranking | 1 | 1 | 12 |
| **Total:** | 22 | 22 | 216 |

**5. Reading List**

**5.1. Required.**

Lecture slides for the course on the [course page](http://wiki.cs.hse.ru/%D0%98%D0%BD%D1%82%D0%B5%D0%BB%D0%BB%D0%B5%D0%BA%D1%82%D1%83%D0%B0%D0%BB%D1%8C%D0%BD%D1%8B%D0%B9_%D0%B0%D0%BD%D0%B0%D0%BB%D0%B8%D0%B7_%D0%B4%D0%B0%D0%BD%D0%BD%D1%8B%D1%85_%28%D0%BF%D1%80%D0%BE%D0%B3%D1%80%D0%B0%D0%BC%D0%BC%D0%BD%D0%B0%D1%8F_%D0%B8%D0%BD%D0%B6%D0%B5%D0%BD%D0%B5%D1%80%D0%B8%D1%8F%29).

**5.2. Optional.**

• Links to the additional materials on the [course page](http://wiki.cs.hse.ru/%D0%98%D0%BD%D1%82%D0%B5%D0%BB%D0%BB%D0%B5%D0%BA%D1%82%D1%83%D0%B0%D0%BB%D1%8C%D0%BD%D1%8B%D0%B9_%D0%B0%D0%BD%D0%B0%D0%BB%D0%B8%D0%B7_%D0%B4%D0%B0%D0%BD%D0%BD%D1%8B%D1%85_%28%D0%BF%D1%80%D0%BE%D0%B3%D1%80%D0%B0%D0%BC%D0%BC%D0%BD%D0%B0%D1%8F_%D0%B8%D0%BD%D0%B6%D0%B5%D0%BD%D0%B5%D1%80%D0%B8%D1%8F%29).

• Andrew R. Webb, Keith D. Copsey. Statistical Pattern Recognition. 3rd Edition, John Wiley & Sons Ltd., 2011.

• Trevor Hastie, Robert Tibshirani, Jerome Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction., 2nd Edition, Springer, 2009. <http://statweb.stanford.edu/~tibs/ElemStatLearn/>.

• Kevin P. Murphy. Machine Learning: A Probabilistic Perspective. Massachusetts Institute of Technology. 2012.

• Christopher M. Bishop. Pattern Recognition and Machine Learning. Springer. 2006.

**6. Grading System**

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.

**7. Guidelines for Knowledge Assessment**

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]=0.6\*[homework score]+0.2\*[exam1 score]+0.2\*[exam2 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 questions during exam after module 3

[exam2 score] – proportion of successfully answered theoretical 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.

**8. Special Equipment and Software Support**

• A projector for lectures

• Whiteboard with markers

• Internet access