Government of Russian Federation

Federal State Autonomous Educational Institution of High Professional Education

"National Research University Higher School of Economics"

Syllabus for the course "Modern Methods in Statistical Learning"

for postgraduate program in 09.06.01 Computer Science and Computer Engineering / 05.13.01 "Systems Analysis, Control Theory, and Information Processing", 05.13.11 "Mathematical Theory and Software for Computing Machinery, Systems, and Networks", 05.13.17 "Theoretical Foundations of Computer Science", 05.13.18 "Mathematical Modeling, Numerical Methods, and Software Systems"

Authors:
Geoffrey G. Decrouez, assistant professor, ggdecrouez@hse.ru

Approved by the Academic Council of the School for Postgraduate Studies in Computer Science on October 26, 2014

Moscow - 2014

1. Scope of Use

This program establishes the minimal requirements to postgraduate students' knowledge and skills for 09.06.01 Computer Science and Computer Engineering / 05.13.01 "Systems Analysis, Control Theory, and Information Processing", 05.13.11 "Mathematical Theory and Software for Computing Machinery, Systems, and Networks", 05.13.17 "Theoretical Foundations of Computer Science", "05.13.18 Mathematical Modeling, Numerical Methods, and Software Systems" and determines the content of the course and educational techniques used in teaching the course.

The present syllabus is aimed at faculty teaching the course and postgraduate students studying 09.06.01 Computer Science and Computer Engineering / 05.13.01 "Systems Analysis, Control Theory, and Information Processing", 05.13.11 "Mathematical Theory and Software for Computing Machinery, Systems, and Networks", 05.13.17 "Theoretical Foundations of Computer Science", 05.13.18 "Mathematical Modeling, Numerical Methods, and Software Systems".

This syllabus meets the standards required by:

- Educational standards of National Research University Higher School of Economics;
- Postgraduate educational program for 09.06.01 Computer Science and Computer Engineering.
- University curriculum of the postgraduate program for 09.06.01 Computer Science and Computer Engineering / 05.13.01 "Systems Analysis, Control Theory, and Information Processing", 05.13.11 "Mathematical Theory and Software for Computing Machinery, Systems, and Networks", 05.13.17 "Theoretical Foundations of Computer Science", 05.13.18 Mathematical Modeling, Numerical Methods, and Software Systems", approved in 2014.

2. Learning Objectives

The learning objective of the course "Modern methods in statistical learning" is to provide PhD students with essential theoretical and practical knowledge in modern statistical learning techniques, such as:

- Linear models, regularization techniques, splines;
- Classification techniques, logistic regression, tree-based methods;
- Validation techniques;
- Neural networks, deep learning;
- Time-series modeling;
- Sampling algorithms;
- Elements of Bayesian learning;
- Use of R environment.

3. Main Competences Developed after Completing the Study of This Discipline

After completing the study of the discipline the PhD student should:

- Know modern regression and classification techniques in supervised learning.
- Know how to implement these models using a programming language such as R.
- Understand the theory behind the most widely used statistical learning models.
- Use validation techniques to select a candidate model for the purpose of prediction.
- Think critically with real data.
- Learn to develop complex mathematical reasoning.

After completing the study of the discipline the PhD student should have developed the following competences:

Competence	Code	Descriptors (indicators of achievement of the result)	Educative forms and methods aimed at generation and development of the competence
the ability to carry out theoretical and experimental research in the field of professional activity	ОПК-1	PhD students obtain necessary knowledge in statistical learning sufficient to develop and understand new methods in closely related disciplines such as in Machine Learning.	Assignments, additional material/reading provided
the ability to develop new research methods and apply them in research in one's professional field	ОПК-2	The PhD student is able to select a model using validation techniques and to test it on dataset coming from real-life examples.	Examples covered during the lectures and tutorials. Assignments.
the ability to objectively evaluate the outcomes of research and development carried out by other specialists in other scientific institutions	ОПК-4	The PhD student is able carry out comparative testing of competing models or methods.	Examples covered during the lectures and tutorials. Assignments.
the ability to do research in transformation of information into data and knowledge, models of data and knowledge representation, methods for knowledge processing, machine learning and knowledge discovery methods, principles of building and operating software for automation of these processes	ПК-4	The PhD student is able to develop and analyze statistical models, implement them in a programming language, and select the best model using validation techniques.	Lectures, tutorials, and assignments.

4. Place of the Discipline in the Postgraduate Program Structure

This is an elective course for 05.13.01 "Systems Analysis, Control Theory, and Information Processing", 05.13.11 "Mathematical Theory and Software for Computing Machinery, Systems, and Networks", 05.13.17 "Theoretical Foundations of Computer Science", 05.13.18 "Mathematical Modeling, Numerical Methods, and Software Systems".

Postgraduate students are expected to be already familiar with some statistical learning techniques, and have skills in analysis, linear algebra, and probability theory.

The following knowledge and competences are needed to study the discipline:

- A good command of the English language, both oral and written.
- A sound knowledge of probability theory and linear algebra.

5. Schedule

№	Topic	Total hours	Contact hours			Self-study
			Lectures	Seminars	Practice lessons	
1.	Decision Theory. Linear Regression.	20	2	2		16
2.	Shrinkage methods.	30	4	4		22
3.	Polynomial regression and splines.	26	2	2		22
4.	Validation techniques.	26	2	2		22
5.	Classification.	86	10	10		66
6.	Deep learning.	30	4	4		22
7.	Time-series modeling.	90	12	12		66
8.	Sampling algorithms.	34	6	6		22
	Total	342	42	42		258

6. Requirements and Grading

Mid-Term Exam	1	Mid-semester test.
Homework	2	Solving 2 homework tasks and examples.
Exam	1	Written exam. Preparation time – 180 min.

7. Assessment

The assessment consists of two homeworks and one mid-semester exam. The homework problems are based on each lecture topics and are handed out to the PhD students throughout the semester.

Final assessment is the final exam. Postgraduate students have to demonstrate knowledge of the material covered during the entire course.

8. The grade formula

The exam is worth 60% of the final mark.

Final course mark *is obtained from the following formula*: Final=0.2*(Homeworks)+ 0.2*(Mid-term exam)+0.6*(Exam).

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	

9. Course description.

Topic 1. Decision Theory and Linear regression

Loss function, Conditional expectation, Bayesian vs frequentist, Expected loss, Linear regression, Gram-Schmidt orthogonalization, confidence and predictive intervals. Introduction to Bayesian linear regression.

Topic 2. Shrinkage methods

Ridge regression, lasso, elastic-net, singular value decomposition, effective degrees of freedom.

Topic 3. Polynomial regression and splines

Polynomial regression, splines, natural spline, smoothing splines, singular value decomposition.

Topic 4. Validation techniques

AIC, BIC, cross-validation, bootstrap.

Topic 5. Classification

Logistic regression, tree-based methods, bagging, boosting, AdaBoost, support vector machine, elements of convex optimization.

Topic 6. Deep learning

Neural networks, back-propagation, deep learning.

Topic 7. Time-series modeling

Hidden-Markov models, Wiener filtering, Kalman filtering.

Topic 8. Sampling algorithms

Rejection sampling, importance sampling, MCMC, Gibbs sampling.

10. Educational technologies

The following educational technologies are used in the study process:

- discussion and analysis of the results during the tutorials;
- regular assignments to test the progress of the PhD student;
- consultation time on Monday mornings.

11. Final exam questions

The final exam will consist of a selection of problems equally weighted. No material is allowed for the exam. Each question will focus on a particular topic presented during the lectures. The first question of the exam will ask the PhD students to prove a result or a theorem proved during the class.

The questions consist in exercises on any topic seen during the lectures. To be prepared for the final exam, PhD students must be able to solve questions from the problem sheets and questions from the two assignments.

12. Reading and Materials

Literature:

1. T. Hastie, R. Tibshirani, J.Friedman. Elements of Statistical Learning: Data Mining, Inference, and Prediction (2009). Springer
2. G. James, D. Witten, T. Hastie, R. Tibshirani. An introduction to Statistical Learning (2013). Springer

Literature for self-study:

1. C. Bishop. Pattern recognition and machine learning (2006). Springer.

13. Equipment.

The course requires a laptop and a projector.