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
«Modern Methods in Decision Making»

010402.68 «Applied Mathematics and Informatics»
«Data Sciences», Master Program

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

Approved by:

Recommended by:

Moscow, 2016

1. **Teachers**

Author, assistant professor: Geoffrey G. Decrouez, National Research University Higher School of Economics, Faculty of Computer Science, Department of Data Analysis and artificial Intelligence.

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 010402.68 «Applied Mathematics and Informatics». 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, 2014;
* University curriculum of the Master's program in «Data Science» (010402.68) for 2014.

⌈

Summary: The focus of this course is on statistical learning theory for regression and classification problems. We first describe the general framework of probabilistic approach to pattern recognition. We then introduce the multivariate linear regression model, regularization techniques including ridge regression and the lasso, splines and smoothing splines. Linear classification models, such as logistic regression and liner discriminant analysis are also discussed. Non-linear techniques include regression and classification trees, ensemble techniques (bagging, random forests, boosting), support vector machines, and the use of kernels. We also discuss model selection techniques, such as Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC) and cross-validation.

3. **Learning Objectives**

The learning objective of the core course «Modern Methods in Decision Making» is to provide students with essential theoretical and practical knowledge in modern statistical learning techniques, such as

- Linear regression model, regularization techniques, splines;
- Classification techniques, logistic regression, tree-based methods;
- Ensemble techniques, bagging, boosting;
- Elements of convex optimization, support vector machine, kernels;
- Validation techniques;
- Use of R environment;

4. **Learning outcomes**

After completing the study of the discipline « Modern Methods in Decision Making » the student should:

- Know modern regression and classification techniques.
- 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.

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

The course «Modern Methods in Decision Making» is a course taught in the first year of the Master's program «Data Science». It is compulsory for all students of the Master's program.

Prerequisites

The course is in the continuation of the core course «Modern methods of Data Analysis» proposed in Modules 1 and 2 in the Master`s program «Data Science». Students are expected to be already familiar with some statistical learning techniques, and have skills in analysis, linear algebra and probability theory. Students must have completed the course «Probability Theory and Mathematical Statistics».

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

* A good command of the English language, both orally and written.
* A sound knowledge in probability theory, statistics and linear algebra.

[

Main competences developed after completing the study this discipline can be used to learn the following disciplines:

* Theoretical aspects of statistical learning techniques.
* Implement and test the techniques on real data.

[

After completing the study of the discipline « Modern Methods in Decision Making » 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 models in statistical learning.	Lectures and tutorials.
The ability to propose a model to invent and test methods and tools of professional activity	C-2	SC-M2	The student is able to select a model using validation techniques and to test it on dataset from coming from real-life examples.	Examples covered during the lectures and tutorials. Assignments.
Capability of development of new research methods, change of scientific and industrial profile of self-activities	C-3	SC-M3	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.

6. **Schedule**

Two pairs consist of 2 academic hours for lecture followed by 2 academic hour for tutorial after lecture.

Module 3:

№	Topic	Total hours	Contact hours		Self-study
			Lectures	Seminars	
1.	Probabilistic Approach to Pattern Recognition	9	2	2	5
2.	Linear Regression	11	2	2	7
3.	Regularization Techniques	11	2	2	7
4.	Polynomial Regression and Splines.	11	2	2	7
5.	Validation techniques.	15	4	4	7
6.	Logistic regression and LDA	15	4	4	7
7.	Elements of Vapnik Chervonenkis Theory	15	4	4	7

Module 4:

№	Topic	Total hours	Contact hours		Self-study
			Lectures	Seminars	
8.	Tree-based Techniques	11	2	2	7
9.	Bagging, Random Forests	11	2	2	7
10.	Boosting	17	5	5	7
11.	SVM	17	5	5	7
12.	Reproducible Kernel Hilbert Spaces	19	6	6	7
Total:		**162**	**40**	**40**	**82**

Requirements and Grading

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

9. Assessment

The assessment consists of two homeworks and one mid-semester exam at the end of Module 3. The homework problems are based on each lecture topics and are handed out to the students throughout the semester.

Final assessment is the final exam. Students have to demonstrate knowledge of the material covered during both module 3 and 4.

The grade formula:

The exam is worth 50% of the final mark.

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

Final=0.25*(Homeworks)+ 0.25*(Mid-term exam)+0.5*(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	

10. Course Description

The following list describes main mathematical definitions which will be considered in the course in correspondence with lecture order.

Topic 1. Probabilistic Approach to Pattern Recognition.
Loss function, Risk, Bayes estimator, Empirical Risk Minimization, Bias-Variance Tradeoff, Approximation and Estimation Error.

Topic 2. Linear regression techniques
Multivariate Linear Regression, Ridge regression, Lasso, Elastic-net.

Topic 3. Polynomial regression and splines
Polynomial regression, splines, natural spline, smoothing splines.

Topic 4. Validation techniques
Akaike Information Criteria, Bayesian Information Criteria, Cross-Validation.

Topic 5. Linear Classifiers
Logistic regression, Linear Discriminant Analysis.

Topic 6. Elements of Vapnik-Chervonenkis Theory
Bounds on the estimation error, Vapnik-Chervonenkis inequality, Vapnik-Chervonenkis dimension, Structural Risk Minimization.

Topic 7. Tree-based models
Classification and Regression Trees (CART).

Topic 8. Ensemble Methods
Bagging, Random Forests, Convex Relaxation, Boosting, AdaBoost, Gradient Boosting.

Topic 9. Support Vector Machine
Elements of Convex Optimization, Kernels, Reproducible Kernel Hilbert Spaces.

11. Term Educational Technology

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 student;
* consultation time on Monday mornings.

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 usage of interdisciplinary papers to present connections between probability theory and statistics. The course is intended to be adaptive, but it is normal to differentiate tasks in a group if necessary, and direct fast learners to solve more complicated tasks.

13. Recommendations for students

The course is interactive. Lectures are combined with classes. Students are invited to ask questions and actively participate in group discussions. There will be special office hours for students, which would like to get more precise understanding of each topic. The lecturer is ready to answer your questions online by official e-mails that you can find in the “contacts” section**.** References found in section 15.1 are suggested to help students in their understanding of the material. This course is taught in English, and students can ask teaching assistants to help them with the language.

14. Final exam questions

The final exam will consist of a selection of problems equally weighted.

One two-sided hand-written A4 sheet prepared by the student herself/himself before the exam is allowed. Each question focuses on a particular topic presented during the lectures.

The problems consist in exercises and/or multiple choice questions on any topic seen during the lectures throughout module 3 and 4. To be prepared for the final exam, students must be able to solve questions from the problem sheets, and questions from the two assignments.

15. Reading and Materials

Recommended Reading: We do not follow a particular textbook, but the lecture notes are mainly inspired from the following books:

[1] L. Devroye, L. Gyorfl, G. Lugosi. A Probabilistic Theory of Pattern Recognition (1996). Springer
[2] T. Hastie, R. Tibshirani, J.Friedman. The Elements of Statistical Learning (2009). Springer

[3] G. James, D. Witten, T. Hastie, R. Tibshirani. An introduction to Statistical Learning (2013). Springer
[4] C. Bishop. Patter recognition and machine learning (2006). Springer.
[5] C. Giraud. Introduction to High-Dimensional Statistics (2015). Chapman & Hall/CRC Press

Course webpage: All material of the discipline are posted on the course webpage. Students are provided with links to the lecture notes, problem, their solutions and additional readings.

16.Equipment

The course requires a black board, laptop and projector.

Lecture materials, course structure and the syllabus are prepared by Geoffrey Decrouez.