Syllabus for HSE Machine Learning (ML) course

Course Schedule

Machine Learning, 2022-2023. This syllabus is shared by 2 programs (with differences pointed out):

1. DSBA (ПАД): ML1, Fall: 9/1-12/30, Spring: 1/10-TBD. Calendar
2. ICEF (МИЭФ): ML, Fall: 9/1-12/30, Spring: 1/10-April (module 3 only). Calendar

Teaching Team

Reserves: Maria Tikhonova, Mikhail Makarov

Lecturer	Alexey Boldyrev	Maxim Karpov	Oleg Melnikov	Sergey Chervontsev
DSBA (ПАД) SA	Maxim Karpov: 201	Mike Ovyan: 203 (Online)	Lika Lomidze: 202	Tatiana Iazykova: 204
ICEF (МИЭФ) SA	Maxim Karpov	~~Margarita Lazareva~~		
Head TA	Emil Akopyan			
- DSBA TA	Alexander Stepin (STACK designer)	Yaroslav Ruban	Nastya Tinyakova ~~Ilona Nadrus~~	Kseniia Lysanyuk
- ICEF TA	Evgenij Veselkov	Saraa Ali		

Course Description

This course introduces the students to the elements of machine learning, including supervised and unsupervised methods such as linear and logistic regressions, splines, decision trees, support vector machines, bootstrapping, random forests, boosting, regularized methods.

1. **The first two modules** (Sep-Dec`22) DSBA/ICEF/OOC students apply **Python** programming language and popular packages to investigate/visualize datasets and develop machine learning models that solve theoretical and data-driven problems.
2. **The next two modules** (Jan-Jun`23) DSBA/ICEF (only) students apply **R** programming language and dive deeper into mathematical, statistical, and algorithmic concepts.

Course Objectives

1. The course aims to help the students to develop an understanding of learning from data, to familiarize them with a wide variety of algorithmic and model based methods to extract information from data, teach to apply and evaluate suitable methods to various datasets by model selection and predictive performance evaluation.
2. DSBA and ICEF students: the course is designed to prepare DSBA/ICEF students for the upcoming University of London (UoL) examination.
3. OOC students: the course is designed to prepare OOC students (with diverse backgrounds) to career changes, job interviews, and applications of machine learning in their personal and company projects. We still expect OOC students to have minimal sufficient preparation.

Course prerequisites

1. English language proficiency: since all async/sync communication are offered entirely in **English**
2. Calculus, linear algebra, basic probability, statistics, programming in high level language (Python, R, or alike)

Course technical requirements

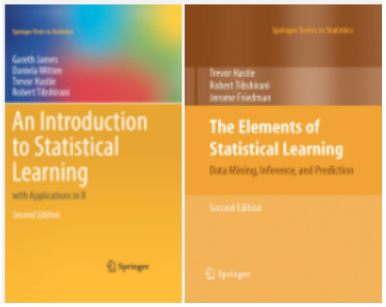
~~Connectivity is key in an online course. To properly connect, engage and participate, the students must have a PC with a mic/speakers and webcam and a reliable high speed internet connection suitable for video conferencing.~~ Chrome web browser, Google account with GDrive and Google Colab plugin

Textbooks

Required (free):

1. Fall: An Introduction to Statistical Learning (📖ISLR). **2ed**. J. Gareth, et al. Supplemental videos. Errata for 1ed.
2. Spring: Elements of Statistical Learning (📔ESL). **2ed**. Jan 2017. T. Hastie, et al. Errata

Optional (mostly free):



1. Dive into Deep Learning. Zhang, Lipton, Li, Smola. 2021
2. Pattern Recognition and ML , C.M. Bishop. ISBN 0387310738
3. Foundations of ML, M. Mohri, et. al.
4. Deep Learning, Ian Goodfellow, et. al. 2016. MIT Press
5. Hands-On ML (HOML). 2ed, A. Géron, ISBN 1492032646
6. Applications of DNN with Keras, Jeff Heaton. 2021

Prerequisite Refreshers:

1. Mathematics for ML. Deisenroth, Faisal, Ong. 2020
2. The Matrix Cookbook. Petersen, K.B. and Pedersen, M.S., 2008.
3. Computer Age Statistical Inference (CASI). Bradley Efron & Trevor Hastie. 2016
4. Bayesian Data Analysis (BDA). 3ed. Andrew Gelman, et al. 2020
5. Statistical Learning with Sparsity: Lasso and Generalizations (SLS). Trevor Hastie. 2015. Book site.

Systems

Ensure that your name & email matches exactly to those in Moodle LMS or we can lose you in grade matching. Update all profiles with your presentable photo to help us authenticate you and credit your effort (i.e. participation, test proctoring).

1. **Moodle LMS** (a.k.a. Smart LMS): for posting weekly material, additional videos, calendar, and for posting/collecting/grading quizzes, tests, and HW, etc.
    a. Due to some bugs, HSE Moodle Admins are recommending to use Moodle interface, not Smart LMS. Our team strongly encourages use of Moodle interface (see lecture 1 video recording).
2. **Discourse Forum**: for nearly all written communication in our course. We are working on embedding Discourse into Moodle LMS
    a. We'll invite everyone by Sep 1
    b. All posts must be in English. Please do not post code for HW (whether correct or not)
    c. Exception: specific HW grading feedback is easier to discuss in Moodle's gradebook
3. **Telemost.Yandex.ru**: tentative; for video meetings and recorded videos
4. **Google Drive**: for release of seminar Colab notebooks, lecture presentation slides, Starter Colab files for Kaggle
5. **Google Colab** plugin for Google Drive: for Kaggle programming assignments.
    a. We require the use of LaTeX and Markdown syntax for all write ups
6. **DataCamp.com**: for additional preparatory courses in the first few weeks of each semester.
    a. Invitations were sent out on Sep 2
7. **Kaggle.com**: for weekly data science competitions in teams of ~3-4 (randomly grouped) students
8. **GitHub**: cleaned up data files are saved in this GitHub. These are primarily used in seminars, not in HW.

Datasets

1. Unless stated otherwise, we'll use this Github repo for our datasets
2. All data must load over the internet from a URL. Please do not submit local data files (to avoid penalties).
3. We value **small** data files demonstrating student's understanding of core concepts, not file management. Avoid files over a few megabytes.

Communication

1. All written course communication is via Discourse forum (not email, not Telegram, etc.) in **English**. We may not reply otherwise.
2. Add a presentable **profile photo with your face clearly visible** to LMS, Yandex Telemost, DataCamp, Kaggle, ... We and proctors use these to authenticate you.
3. Private messages: send to "**Instructors**" (to lower confusion, faster reply), tag an individual, if needed
4. Assignment grading feedback can be discussed directly via Moodle's Gradebook

5. Unless you have a strictly private matter, **all posts should be in the public forum,** where we can scale our assistance to everyone.
6. Course assistance: If you feel the TAs did not resolve your issue, please escalate to the head TA, then to the lecturer.
7. Complaints of inappropriate activity: make a post to the Teaching Team in a private forum.
8. Note: the forum answers may not always be correct (typos, question misunderstandings, etc.).
9. Always keep the communication professional, respectful and cordial.
10. Our response time is within 2 business days, but we'll try to assist sooner. Please start early on all assignments

Seminar/teaching assistants (SA/TA) and program details

Program:	**DSBA**				**ICEF**	
Lectures, 80 min	Regular: Mon 16:20, R204					
Lecturer's Office Hour (OH)	See announces in Discourse forum					
Group/Sections	201	202	203	204	1	2
SA	Maxim Karpov	Mike Ovyan	Lika Lomidze	Tatiana Iazykova	Maxim Karpov	Margarita Lazareva
Seminar date/time duration: 80 min	Thu 16:20, D210	Tue 11:10 via Zoom	Thu 16:20 D507	Sa 14:40, D208	Wed 13:00, R614	Mon 19:40, R612
Office hours:	Thu, 09:30, S925	DM me in telegram @th3luck	DM me in telegram @lomlika	DM me in telegram @winteryfox	Thu, 09:30, S925	DM me in telegram @margo_lazareva

Discourse Monitoring:

Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
Tatiana Iazykova Yaroslav Ruban	Kseniia Lysanuyk Lika Lomidze	Emil Akopyan Saraa Ali	Maxim Karpov Nastia Tinyakova Mike Ovyan	Evgeniy Veselkov Alexey Boldyrev	Margarita Lazareva Mike Ovyan	Oleg Melnikov

Course Grading Formula:

Grading formula is stated in the official course home page. Moodle's gradebook shows your up to date performance, including your current constituent and aggregate grades. If you suspect an error in grade calculation, please let us know ASAP. We use natural grade aggregation in LMS. Rounding (to the nearest integer) is used to report 0-10 scale grades to HSE. There are no blocking grading components.

Homework (HW) Assignment:

1. HW is released weekly with about a 2 week deadline, due on Sundays.
2. **Theory part** (individual): HW are released via Moodle/Smart LMS. These are individualized (not in groups!) and auto-graded assignments (thanks to STACK plugin with Maxima CAS backend). Some written responses will be selectively hand-graded by TAs.
    1. Quick Maxima syntax is provided below and in each assignment.

3. **Practical part** (group): students are randomly assigned into groups to compete on Kaggle.com. Teamwork is evaluated by instructors and team partners (via regular surveys). Team's grade is awarded to everyone on a team, but individual grades may be adjusted up/down according to the team's feedback about one's team effort.

Tests and Exams - per module:

1. We will have a cumulative online exam/test at the end of each module (during the HSE examination sessions). Do not book travel that conflicts with test dates.
2. Tests are open book/internet, but individual, i.e. no collaboration.
3. Test questions are drawn from quiz banks, not HW. HW deepens your understanding, but a test measures it.

University of London (UoL), Course ST3189 (ML):

1. DSBA students: ~~UoL grade (exam & coursework project) replaces the exam grade in module 4~~
    a. Since not all students take UoL exam, we will release our typical Exam in Module 4
2. Coursework Project in R programming language is for DSBA/ICEF students only and is administered by LSE/UoL. It is released around January and is due around April 1. Although students are given a 3-4 months window, this exercise is meant to be completed in a few days. Typically, students work on it in Feb/Mar. More details on the UoL site.

Quizzes in LMS - weekly:

1. Quizzes are based on lectures, seminars, textbooks, posted videos, and other material delivered via our course.
2. Quizzes individualized (shuffled and sampled from question banks) for each student. Most questions are auto-generated. Students can only navigate sequentially (no jumping around questions - anti cheating measure).
3. Students have 3 (possibly delayed) quiz attempts with repeats penalized by 10-33%. Best score is used for the quiz. In some cases, students can immediately check and resubmit answers at a variable penalty (10-33%)
4. All choice questions are multi-answer (regardless of singular/plural formulation). Incorrect answers lower your score to prevent guessing.
5. Numeric answers are typically accepted with 0.01=1% of error, i.e. round to at least 4 decimal places, if needed. Please do not round any intermediate calculations. It's best to use as many decimals as fits in the answer box.
6. The quiz answers are released after its due date. In the meantime, students see hints, if they answer it incorrectly.
    a. We do this because we allow students to have multiple retries and because students take quizzes at different times. This policy minimizes cheating.
7. We always use natural logarithm (inverse of exp()) in this course.
8. Back button is disabled to minimize cheating. You can only move forward. Note: you have multiple attempts.
9. Quizzes have a single deadline (typically, on Sundays).

Participation:

An active participation in our course includes:

1. meaningful Q&A posts in the course forums
2. attendance and participation in lectures and seminars (with the webcam turned on for the online meetings)
3. inappropriate behavior is penalized. E.g. redundant or uninformative posts, abusive behavior, cheating, plagiarism and other violations of Honor Code, ethical behavior and student integrity.

Participation grade depends on objective and subjective metrics across Discourse discussion forum, Moodle, etc.

1. **Objective** part includes calculation of the proportion of logged online and offline time working on our course. For example, we may rank students according to likes (from peers and teaching team) they receive for their posts in the forum, as well as number of questions, number of answers, number of post views, visits, etc. In Yandex Telemost this might use the connection time, use of webcam, etc.
2. **Subjective** part includes instructor's judgment of the quality of posts, questions, answers, regular/active participation in lectures/seminars, help to peers and assistance in building a meaningful learning environment.

Deadline Extensions and Makeup:

1. Only valid verifiable excuses are accepted for 1-2 day extensions.
    a. DSBA: submit your doctor's note via student services.

2. If you miss a deadline (with a valid/verifiable excuse), contact instructors ASAP in a private post to arrange a new deadline.
3. Submissions are not allowed after the solutions have been released.
4. Any test/exam retakes will be rescheduled as per university policy (see also dedicated rubric below)
5. Note: accommodating exceptions is difficult and time consuming. Typically, a verifiable medical emergency is a valid reason, but travel and conferences are not. It is the student's responsibility to start their work early, so as to hedge against any unforeseeable life event.

Regrading:

1. We aim to grade fairly, accurately, and timely. If you believe we made a crude grading error, make a post in a private forum in LMS ASAP (within 1 week of grading) & tag a grader.
2. To discourage frivolous appeals, we reserve the right to deduct a 2-5% of the grade, if your appeal lacks a strong justification or the benefit fails to exceed 2-5%. Be sure it is worth the mutual effort.

Source of Truth:

- If in doubt, always use the course textbook as the main source of truth. It has been well edited, provides full context, and has very few typos.

Retakes: (as of 25/09/2022)

DSBA students:

On the first retake the cumulative course grade is taken into consideration. On the second retake there are following rules:
If the cumulative grade + 2nd retake grade >= 4, then the student receives the sum of these marks, multiplied by their coefficients.
If the cumulative grade + 2nd retake grade < 4, but the 2nd retake grade >= 7, then the student is given a 4.
If the 2nd retake grade is <7 and the cumulative grade is <4, the student receives a failing grade.

ICEF students:

Retakes are organized in accordance with the HSE Interim and Ongoing Assessment Regulations (incl. Annex 8 for ICEF). Grade determination after retakes is done in accordance with ICEF Grading Regulations (par. 5) for BSc and MSc. Final grades are given initially out of 100. They are also transferred to 10- and 5-points grades in accordance with the ICEF Grading Regulations (par. 3).

Posting Guidelines:

- Do NOT post code/solutions and assignments on the Internet or outside of our class at any time.
- Do NOT post code/solutions within our course, unless explicitly permitted by instructors (e.g. ok to share within your assigned team or after instructors release the solutions).
    - Doing so may affect your grade. Give your peers a chance to discover and learn on their own. We do encourage high-level discussions of problems/solutions. E.g. you can describe the steps you take to reproduce the problem or you can post parts of the error message you observe. Use minimal reproducible example.
- Phrase your questions precisely and concisely. Identify URL of the referenced page or video (+time), or textbook page.
- Format your posts, i.e. use LaTeX and code formatting to make your messages easier to read.
- The instructor team tries to reply daily, but please be patient.
- Peer assistance is encouraged, just please avoid posting threads for the sake of posting (the quality does matter).
- Before asking a question, quick-search prior posts for similar concerns.

Assignment Guidelines & Rubric (unless otherwise instructed):

1. Rule 0: if ever in doubt, just ask. Better ask than be sorry. Asking the right question in the right way is a valuable skill one shall develop.
2. Every output (table, numbers, plots, ...) must have a purpose and an interpretation. We never produce outputs for the sake of filling the void.
3. Every output must be legible, readable, meaningfully organized, neat, clean, compact, precise, and concise. Avoid generic terms, such as data and model. Instead, be specific and use observations, features, linear regression, logistic regression, etc.
4. No overplotting. No redundant plots. No data dumps (pages of table outputs).
5. Annotate your programming code! Don't expect people to understand the full context.
6. Assure no errors and warnings in your code.
7. Always start your modeling with a simple baseline model (version 0), which you can do with minimal data munging.
    1. This is your Proof of Concept (PoC) and your benchmark for any future improvement.
8. Start your work early! Never procrastinate. We are busy on assignment due days and may be inaccessible to answer last minute questions.
9. Clearly label and order problems as assigned (to help us avoid missing your solutions)
10. Submit IPYNB file (for execution) and its PDF (for grading feedback).
    1. Try this converter or others. If we can't run your code as is, it earn a zero
    2. NEW: flatten your PDFs so that Moodle displays plots for graders
11. Name your files as HW[#]-[Student Name].[***]
12. Seed all random number generators (RNG) to ensure reproducibility of your work
13. ~~Quizzes/~~Kaggle assignments are due on Sunday night (MSK), soft deadline. 2 more days (till Tuesday) are allowed at a penalty of 10% per day to the hard deadline.

Our values

1. Knowledge sharing. Everyone has much to contribute and to learn.
2. Academic integrity. We treat every individual with respect. In our education, we maintain the highest standards, acknowledge mistakes. We demonstrate fairness and impartiality.
3. Commitment. We demonstrate a high level of motivation and dedication to the course. We respect deadlines and other people's work.

We expect students to know and practice Student Code of Conduct, Code of Academic Ethics, Statement of Values, and Code of Conduct for HSE University Stafft, Statement of Values, Internal Labour Regulations of National Research University Higher School of Economics. Clarification from Vadim Radayev.

Additional information

Copyright: Each disseminated document in this course is copyrighted with all rights reserved. These may not be reproduced or distributed without an explicit permission of the author(s). The HSE also has a policy that one should know before audio/video recording.
Recorded lectures might be reused in other HSE courses, viewed by other HSE students, faculty, or staff, or used for other education and research purposes. Note that your image or voice might be incidentally captured. If you have questions, please contact a member of the teaching team.
Methods of Instruction:

1. Lectures are taught in the auditorium.
2. Seminars are also in the classroom.
3. HSE Smart LMS (based on Moodle LMS) is the platform for the course management, incl. quizzes, assignments, grading, forum, etc.

Soft Skills Development

In accordance with the recommendations of BSc and MSc Subcommittees, we are placing a greater focus on the development of soft skills. This is done in two ways. Specific with graded elements:

1. Students are encouraged to participate in lectures, seminars, and online discussions. ICEF and DSBA students will be groped for HW assignments (teams of 2 people). Students are also graded for joining lectures, seminars and turning on their webcams. We strongly believe that success in communication begins with being present, visible and actively involved in the sync and async element of the course. Objective assessment is done in the form of fraction of participation from the log files. Subjective assessment is done by a consensus of all instructors on the quality of questions asked, answers provided, quality of participation, etc.
2. Students will be taught ethics, integrity and honor code in order to clearly perceive and analyze boundaries of constrained communication flow while working on assignments. Assessment is done in the form of weekly auto-graded quizzes in LMS.

Organization of Studies for Persons with Limited Mobility and Disabilities:

1. If necessary, learners with limited mobility or a disability (as per his/her application), as well as per his/her individual rehabilitation program, may be offered the following options for receiving learning information with due consideration of his/her individual psycho-physical needs (e.g., via eLearning studies or distance technologies):
2. 7.1.1. for persons with impaired vision: enhanced fonts in hard copy documents; e-documents; audio files (transfer of study materials to an audio-format); hard copy documents with the use of Braille; individual consultation with a facilitated communicator; individual assignments and mentoring;
3. 7.1.2. for persons with hearing impairments: in hard copy; e-documents; video materials with subtitles; individual consultation with a facilitated communicator; individual assignments and mentoring;
4. 7.1.3. for persons with a muscular-skeleton disorder: in hard copy; e-documents; audio-files, individual assignments and mentoring.

Course Acronyms and Emojis:

We leverage acronyms/emojis in our course for brevity (and fun). Many of these are specific to our course and cannot be found on the internet or in course textbooks. So we compiled the list below. If any acronym or emoji is unclear, please let us know.

1. Textbooks:
    a. 📖=ISLR textbook (in Fall). E.g.: 📖1.1=Chapter 1.1 from ISLR,
    b. 📔=ESL textbook (in Spring)
    c. CASI=Computer Age Statistical Inference, BDA=Bayesian Data Analysis, HOML=Hands-on Machine Learning
2. Modules: M1=Module 1, M2=Module 2, M3=Module 3, M4=Module 4
3. Gradebook categories:
    a. 💼=Homework (HW) assignment
    b. ❓=quiz type of assignment
    c. 🎭=Participation (in our course)
    d. 💯=Exams
    e. Gradebook Totals:
        i. ∑🎭M1=total grade for Module 1 participation
        ii. ∑❓M1=total grade for quizzes in Module 1
        iii. ∑💯=total grade for exams
        iv. ∑💼M1=total grade for Module 1 homework assignments
4. Content type:
    a. 📄=document content
    b. 📹=video content or URL link
    c. 🐍=Python coding problem
    d. 📐=Math or calculation problem
5. Assignment types:
    a. 🛡️=ethics, honor and integrity category. E.g. ❓🛡️📄=ethics quiz based on a document
    b. 🏆=Kaggle competition (or well done congrats)
    c. 👩‍💻=DataCamp (DC) assignment
    d. 📣=student's Feedback & Partner Evaluation (within teams)
    e. ⚖️=peer assessment of Collab Notebooks (between teams)
    f. 🤖=autograded by STACK/Maxima
    g. 💬=textual answer expected

6. Projects/Competitions:
   a. 🔢=(digit) Recogn17er
   b. 🚗=Auto
   c. 🧬=Genomics
   d. 🔮=Crystals
   e. 🏃=Human Activity Recognizer (HAR)
   f. 📈=Crypto currency
   g. 🪐=Venus-Mars
   h. 📍=Graphs or graphical models
   i. ☢️=Radon
7. Grading Rubric and Quiz Evaluation:
   a. 😱=poor performance
8. Other:
   a. 👥=teams, LB=leaderboard, 💂=UoL (University of London)
   b. SA=seminar assistant, TA=teaching assistant,
   c. SL=supervised learning, UL=unsupervised learning

Quick Maxima+STACK Rules

Moodle's STACK types of quiz questions allow entry of algebraic expressions and can even check algebraic equivalences. The entry format is closer to 🐍 syntax than to LaTeX. For example, to enter the answer $e^{-t}\sin(3t)$, you can type `e**(-t)*sin(3*t)`. Use care and triple-verify your formulas before submission! Typos and re-attempts are costly. In backend, STACK uses Computer Algebra System (CAS) Maxima to evaluate symbolic algebraic expressions and compare these to teacher's answers. If in doubt, ask in Forum. Basic entry rules are:

1. STACK's student reference for Answer Input
2. Keep consistent variable names and capitalization. If the question uses Y (not y), then type in `Y` (not `y`). If the question uses χ, then type (Greek) `chi`, not (Latin) `X` or `x`
3. Subscripted elements, such as y_i or y_1, should be entered as `y[i]` or `y[1]`, respectively
   1. don't type `yi`, `y1`, `y_i`, or `y_1`. These render similarly (!!!), but actually are Maxima's atomic names, not composite or array elements
4. Special constants (more): π =`pi` or `%pi`, $\exp()$ =`exp()`, e^x =`e^x` or `e**x` or `%e^x`
5. Sums and products: $x_1 + x_2$ =`x[1]+x[2]` or `sum(x[i],i,1,2)`; $z_1 z_2$ =`z[1]*z[2]` or `product(z[j],j,1,2)`
6. Use explicit algebraic symbols. E.g. $2a$ =`2*a`
7. Greeks are typically entered in English: μ =`mu`, Λ =`Lambda`
8. Use fractions instead of decimals, if possible. E.g. 0.5 =`1/2`, $0.875a$ =`a*7/8`
   1. Function `rat()` may convert a floats to a rational (if floats are enabled in the question)
9. A list is `[1,x,y]`, a set is `{1,x,y}` or `set(1,x,y)`
   1. FYI: A Maxima solver returns a list of solutions for an equality, but we expect (unordered) sets for mathematical correctness. Pay attention to the notation, which should be unequivocal!
10. $\propto$ is equivalence up to a multiplicative constant. Drop all multiplicative constants in your answer, i.e. $f(x) := a\sqrt{bx}\cdot e^{c+x}\cdot \ln x^d \propto \sqrt{x}\cdot e^x \cdot \ln x$
11. $\simeq$ is equivalence up to an additive constant. Drop all additive constants in your answer, i.e. $f(x) := a + (b+x)^2 + e^{c+x} \simeq x^2 + 2bx + e^{c+x}$.
12. Other functions: $x!$ is `factorial(x)` or `x!`, $\frac{N}{k}$ is `binomial(N,k)` or use factorials

Symbolic programming in Maxima is similar to

1. Free: in 🐍: SimPy, SymPy Live, SymPy Gamma; in **R**: rSymPy, Ryacas (based on YACAS)
2. Paid: Wolfram Mathematica, Wolfram Alpha, Matlab's Symbolic Toolbox, Maple

Maxima is free. It is worthwhile to learn its basics; and, it's more fun with wxMaxima (GUI). 🌄