Programming and Statistical Analysis

Intro and Course overview

Majid Sohrabi
National Research University Higher School of Economics

April 04, 2025

Teaching Team Contacts

Majid Sohrabi (lectures & seminars)

- Email: msohrabi@hse.ru
- Telegram: @MSohrabi_CS

Timetable

Lectures	Seminars	Location	Duration
Fridays 14:40 pm	Fridays 16:20 pm	Online	Modules 4th
Link	Check your timetable.		

Course content

Introduction to Programming and Statistics

- Overview of the field
- Importance and applications
- Types of data, structured vs. unstructured

Data Processing and Cleaning

- Handling missing data, outliers, noise
- Data transformation, normalization, standardization, encoding
- Feature selection and dimensionality reduction

Exploratory Data Analysis (EDA)

- Descriptive Statistics and Data Summarization
- Data Visualization Techniques (Histograms, Scatter Plots, Box Plots)
- Identifying Patterns, Trends, and Anomalies

Course content

Statistical Foundations for Data Analysis

- Probability Theory and Distributions
- Correlation and Causation, introduction to Regression Analysis

Machine Learning Fundamentals

- Supervised vs. Unsupervised Learning
- Linear and Logistic Regression
- Support Vector Machines
- Model Evaluation (accuracy, precision, recall, F1-score, ROC AUC), Cross-Validation

Data Mining Techniques

- Classification, decision tree, Naïve Bayes, k-Nearest Neighbors
- Clustering, k-Means, Hierarchical Clustering, DBSCAN

Course content

Advanced-Data Mining and Machine Learning

- Ensemble Methods: Random Forests, Gradient Boosting

Data Visualization and Reporting

- Advanced Visualization Techniques: Heatmaps, Geospatial Data Visualization

Overview

Elective for 1st year PhD in Cognitive Science

Duration: 1/4 of the academic year (module 4)

Assessment elements:

- Homework assignments (30% weight)
- Exam (70% weight), in the form of a project, with **progress tracked during the module** (topic choice deadline, preliminary results deadline, final result deadline)

Format:

- Online (lectures & seminars)

Grade Formula

Grade Component	Percentage	Evaluation Criteria
Homework	30%	Homework with deadline, each homework consists of several tasks and a single homework is a 10-point scale.
Final Project	70%	The final project is a 10-point scale (in groups), choose a dataset, make relevant analysis, write a report, and present their works.

The formula

$$\text{Final grade} = 0.3 \cdot \text{Homework} + 0.7 \cdot \text{Final Project}$$

$0 \leq$ Homework score ≤ 10

$0 \leq$ Final Project score ≤ 10

Rounding to the closest integer

Arithmetic rounding. E.g. 3.5 is rounded to 4, 3.49 is rounded to 3.

Homework

A small set of tasks each week (or 2 weeks) (jupyter notebooks)

Solve tasks to earn points

Deadline: 1-2 weeks per homework

Homework grade $= 10 \cdot \min\left(1, \frac{\sum \text{points}}{\text{total}}\right)$

Exam

Exam in the form of project defense

The project is either:

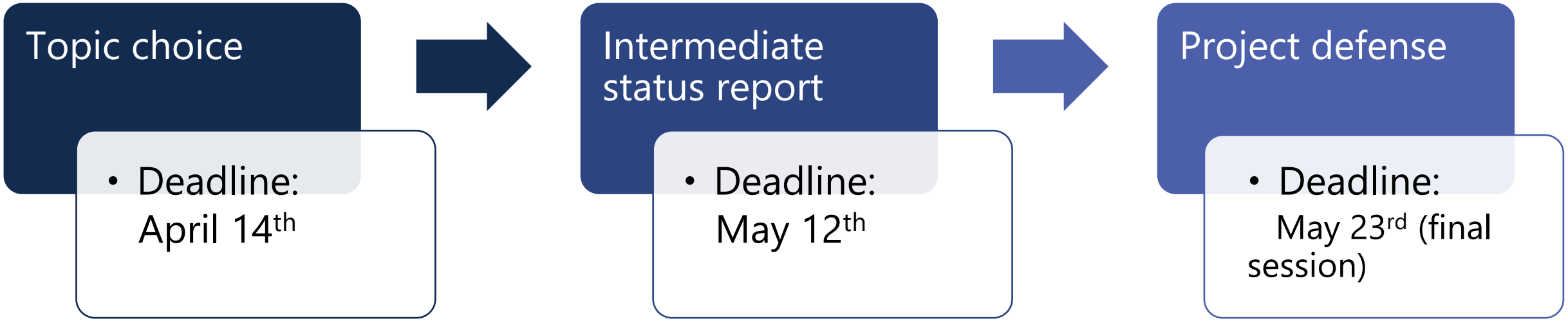
- Participation in a competition (on www.kaggle.com or similar)
  - Teams of up to 2 people are OK (roles of all members of a team should be clear and significant)

Or:

- Implementation of some technique or study from an advanced machine learning or data analytics paper
  - I'll provide some suggestions later,
  - or you can find something that interests you by yourself
  - Individual work

Please discuss your choice with me

Exam project timeline



Missing any of the 3 stages has a negative point (-2 points) on the final mark for the project.

The intermediate status report is a “MUST” to be eligible for final defense.

All members of a group need to attend the final defense.

Motivation

Programming, Statistics and data analysis underlie machine learning and artificial intelligence (AI) technologies

Comprehension of its basic principles helps to understand the world we live in

https://www.youtube.com/watch?v=RNnZwvklwa8&ab_channel=AdamEubanks

Thank you!

Majid Sohrabi

msohrabi@hse.ru

@MSOHRABI_CS