Data Analytics and Mining

Intro and Course overview

Majid Sohrabi

National Research University Higher School of Economics

October 04, 2024

Teaching Team Contacts

▶ **Majid Sohrabi (lectures)**

– Email: msohrabi@hse.ru

– Telegram: @MSohrabi_CS

▶ **Alexandra Kogan (seminars)**

– Email: akogan@hse.ru

– Telegram: @horror_in_black

Repository with course material

https://github.com/Majid-Sohrabi/DAM-2024

Timetable

Lectures	Seminars	Location	Duration
Fridays 6:10 pm	Fridays 7:40 pm	Online	Modules 1-2
Link	Check your timetable.		

Course content

▶ **Introduction to Data Analytics and Mining**

    – Overview of data analytics and mining

    – 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

▶ Compulsory course for year 2

▶ Duration: 1ˢᵗ half of the academic year (modules 1 and 2)

▶ Assessment elements:

  – Homework assignments (40% weight)

  – Tests (20% weight)

  – Exam (40% weight), in the form of a project, with **progress tracked during the semester** (topic choice deadline, preliminary results deadline, final result deadline)

▶ Format:

  – Online (lectures & seminars)

Grade Formula

Grade Component	Percentage	Evaluation Criteria
Homework (module 1)	20%	Weekly homework with deadline, each homework consists of several tasks and a single homework is a 10-point scale.
Test (module 1)	10%	In-class mini tests once a week or once in two weeks, each test consists of several questions and a single test is 10-point scale.

Grade Formula

Grade Component	Percentage	Evaluation Criteria
Homework (module 2)	20%	Weekly homework with deadline, each homework consists of several tasks and a single homework is a 10-point scale.
Test (module 2)	10%	In-class mini tests once a week or once in two weeks, each test consists of several questions and a single test is a 10-point scale.
Final Project	40%	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

Final grade = $0.2 \cdot$ Homework_1 + $0.1 \cdot$ Test_1 + $0.2 \cdot$ Homework_2 + $0.1 \cdot$ Test_2 + $0.4 \cdot$ Final Project

$0 \leq$ Homework score ≤ 10

$0 \leq$ Test score ≤ 10

$0 \leq$ Final Project score ≤ 10

Rounding to the closest integer

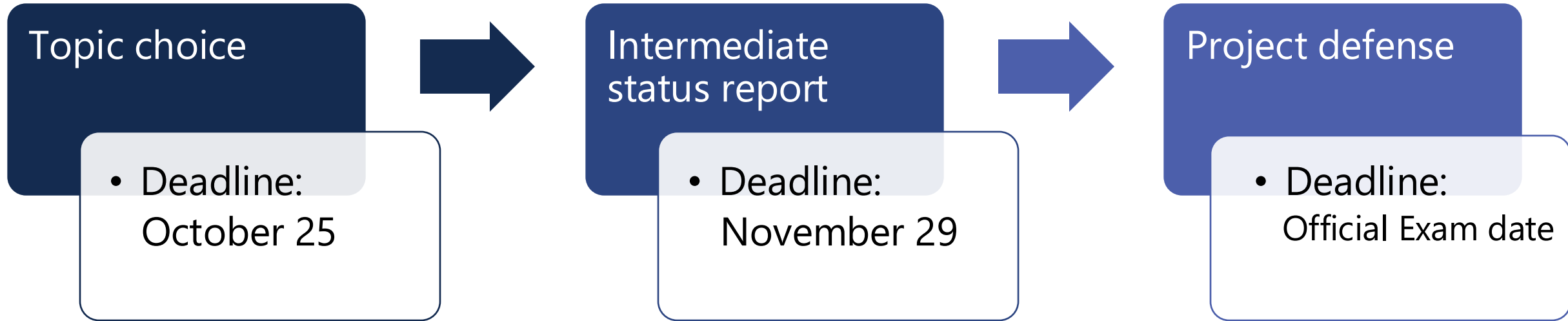Arithmetic rounding. E.g. 3.5 is rounded to 4, 3.49 is rounded to 3.

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 3 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

Topic choice	Intermediate status report	Project defense
• Deadline: October 25	• Deadline: November 29	• Deadline: Official Exam date

▶ Missing any of the 3 stages has a negative point 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

▶ 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