ID2222 Data Mining (Datautvinning)

2024/2025, period 2 (HT24)

The course studies the fundamentals of data mining, data stream processing, graph processing, and machine learning algorithms for analyzing large amounts of data. You will practice using big data processing platforms, such as Apache Spark and Apache Flink, to implement data mining algorithms.

Intended learning outcomes

After this course, students will be able to:

- to mine different types of data, e.g., high-dimensional data, graph data, and infinite/never-ending data (data streams);
- to program and build data-mining applications;
- to know how to solve problems in real-world applications, e.g., recommender systems, association rules, link analysis, and duplicate detection;
- to master various mathematical techniques, e.g., linear algebra, optimization, and dynamic programming.

Examination

- LAB1 - Programming Assignments, 3.0 credits, grading scale: P, F
- TEN1 - Examination, 4.5 credits, grading scale: A, B, C, D, E, FX, F

The examination consists of a computer-based exam (TEN1) at campus and programming assignments (LAB1). Assignments can be done in groups of two students. Assignments must be submitted to Canvas on the due dates and presented in person at specially appointed reporting sessions at the KTH Kista campus. The final grade is based on the performance of the computer-based campus exam and the programming assignments.

- The ID2222 exam is a proctored on-campus computer-based closed-book exam in Canvas. The exam consists of questions of different types, e.g., Multiple Choice, Multiple Answer, True/False, and Numeric, to be answered in Canvas on a KTH computer or your laptop at the KTH Kista campus.
- Each homework programming assignment is to be uploaded to Canvas by the due date and presented to a teaching assistant during the reporting session. If submitted on time and successfully presented, you will earn bonus points added to your exam score.
- The exam passing grade (E) is 60 points out of 100 without a bonus (120 with a total bonus).
    - Grade A: > 92
    - Grade B: 85-92
    - Grade C: 77-84
    - Grade D: 69-76
    - Grade E: 60-68
    - Grade FX (Fail; eligible for completion): 55-59
    - Grade F (Fail): < 55

Recommended Prerequisites

Knowledge of concepts and terminology associated with statistics, database systems, and machine learning; a course on data structures, algorithms, and discrete mathematics (such as ID1020 Algorithms and Data Structures); a course in software systems, software engineering, and programming languages; a course on processing, storing and analyzing massive data sets (such as ID2221 Data-Intensive Computing).

Course Book

Mining of massive datasets, by Jure Leskovec, Anand Rajaraman, and Jeffrey D. Ullman, 3-rd edition, Cambridge University Press, 2020, ISBN: 978-1108476348 (http://www.mmds.org/

Tentative Lecture Layout

Eighteen lectures mainly follow the course book. Lectures are not compulsory but highly recommended.

1. Introduction (reading: Chapter 1 Data Mining)
2. Finding Similar Items: LSH (reading: Chapter 3 Finding Similar Items)
3. Frequent Itemsets (reading: Chapter 6 Frequent Itemsets)
4. Introduction to Apache Spark (reading: Spark Overview)
5. Mining Data Streams 1 - Models, Sampling, and filtering (reading: Chapter 4)
6. Mining Data Streams 2 - Counting and estimating on streams (reading: Chapter 4)
7. Data Stream Processing and Analytics: Introduction to Apache Spark Streaming
8. Eleven Lectures on Graph Fundamentals and Information Network Mining, Clustering, and Recommender Systems (Reading instructions are provided in Canvas).

Supplementary reading

Some articles give a student the background or better insight into the topics covered in the course. Some additional readings (articles) are recommended in Canvas during the course.