How to prepare for the exam Advances in Data Mining (2021)

Wojtek Kowalczyk
w.j.kowalczyk@liacs.leidenuniv.nl

14/12/2021

The exam will consist of several questions and small problems that will test your knowledge and understanding of the key concepts that have been covered during the course and the practical assignments. It will have a form of a multiple-choice quiz, where, for every question or problem you will have to find the correct answer. Additionally, you will have to briefly justify your answer on a separate sheet of paper.

Selecting the right answer might require knowledge of some formulas and an ability of applying them to specific values or situations. It will be a "closed book exam" so you will not be allowed to use any notes, textbooks, smartphones, calculators, etc. The best way of preparing for the exam is studying the relevant fragments of the MMDS textbook and slides, paying a special attention to examples and solving example problems (you may skip problems that are marked with ! or !!). You could also study example solutions of problems used in earlier editions of the course. Additionally, you should be able to answer specific questions related to the practical assignments (A0, A1, A2, and A3).

The exam will be on-site. You can find the details about the location and time on the university website. In exceptional situations, students who are not able to physically attend the on-site exam should contact me via email (wojtek@liacs.nl) as soon as possible.

During the exam you may expect questions that are related to:

Introduction (Chapter 1, Section 1.3 Things Useful to Know, subsections 1.3.1 – 1.3.3, 1.3.5)

Section 1.3.1(Importance of words in documents)

- The concept and the definition of TF-IDF.
- Possible uses of TF-IDF in text mining.

Section 1.3.2, 1.3.3 (Hash Functions and Indexes)

- The concept of a hash function.
- The concept of an index.
- Some examples and applications.

Section 1.3.5:

- The importance of the constant e and its approximation by the expression $(1+1/x)^x$ (or as the limit of the sequence $(1+1/n)^n$, with $n \to \infty$).
- ***Exercises 1.3.1, 1.3.2, 1.3.4***

Recommender Systems (Chapter 9, Lecture Slides, Assignment 1)

You should be familiar with all the recommender algorithms that have been discussed during the course.

Similarity Search (Chapter 3: Finding Similar Items)

You are supposed to know the material that has been covered during our lectures and the practicals. In particular, you should know:

Section 3.1:

- Example applications of similarity search, Jaccard similarity.

Section 3.2:

- The concept of shingles, k-shingles; ability of estimating the number of possible shingles and memory requirements for storing their hashes (section 3.2.3); all examples and exercises from this section.

Sections 3.3 and 3.4:

- Everything; you may skip the exercises marked as difficult ones (! or !!).

Sections 3.5, 3.7.1-3.7.3 and 3.7.6 (exercises):

- Definitions of: Jaccard-, Hamming-, cosine-, Euclidean- similarity measures (you may skip the edit distance), all examples and exercises 3.5.4, 3.5.5, 3.5.10.

Mining Data Streams (Chapter 4)

Sections 4.1, 4.2, 4.3

- Everything; special stress on Bloom filters, their properties and applications (section 4.3).

Section 4.4 plus slides (probabilistic counting); skip the exercise 4.4.5.

PageRank algorithm (Chapter 5)

Section 5.1

- The key idea behind PageRank, the definition of page rank, the transition matrix and Markov processes; calculation of PageRank by iterative matrix multiplication; spider traps, teleporting. All examples (of 5.1) and exercises 5.1.1 and 5.1.2.

Sections 5.2.1 and 5.2.; Section 5.2.6 (exercises)

- An efficient implementation of the PageRank algorithm on a single computer which has enough RAM to store the vector v_new and disk space to store the transition matrix M and the vector v_old (check slides!).

Hadoop and MapReduce (Chapter 2, *Sections 2.1, 2.2 (skip 2.2.5-2.2.7), slides used during the course)*

- The concept of HDFS (Hadoop Distributed File System) and of the MapReduce framework.
- All examples from the slides, esp. matrix-vector and matrix-matrix multiplication.
- Estimating the chance of losing data that is stored on the Hadoop distributed file system.

SVM (Chapter 12, *Sections 12.3.1-12.3.5)*

- The key concepts: margin, learning as a quadratic optimization problem; handling data sets that are not linearly separable; multi-class classification problems. Watch the presentations of Jure Leskovec (parts 2, 3, 4, 5, 6) from the *mmds.org* site (Chapter 12).

PCA, LLE, t-SNE, RandomForest, XGBoost

- The key ideas behind dimensionality reduction techniques: linear projections that preserve/explain of the original variance (PCA), preserving the local topology - distances to the nearest neighbors (LLE), or local density estimates (t-SNE).

- The overall idea behind the RandomForest algorithm. The most important hyper-parameters used by this algorithm. Out-of-Bag (OOB) accuracy estimate. Measuring importance of attributes. The concepts of Bagging and Boosting. Study Sections 15.1-15.3 of the ESLII book: https://web.stanford.edu/~hastie/Papers/ESLII.pdf

- The overall idea behind the XGBoost algorithm. It is sufficient that you study the simplified presentation of this algorithm, https://xgboost.readthedocs.io/en/latest/tutorials/model.html .

Example Questions

- To get an idea of what to expect, study the questions/problems (with answers/solutions) that have been used in the AiDM exams in 2018 and 2019. You may skip questions related to topics that have not been covered in 2021 (e.g., estimating moments, counting 1's, etc.).