

Homework 1: Matrix Factorization

Form:	Zip file including Jupyter notebook file and images.
Language:	English
Requirements:	The report should be clear, readable and include all code documented
Submission:	Zip file via Moodle. The file name should include the students' ids
Contact:	
Deadline for submission:	December 6, 2020

Students will form teams of two people each, and submit a single homework for each team. The same score for the homework will be given to each member of the team.

Submit your solution in the form of an [Jupyter notebook file](#) (with extension ipynb). Images of graphs or tables should be submitted as PNG or JPG files. The code used to answer the questions should be included, runnable and documented in the notebook. Python 3.6 should be used.

The goal of this homework is to let you understand the details of matrix factorization algorithm and practice recommender system training and evaluation. You should implement the matrix factorization algorithm by yourself using only basic Python libraries (such as numpy).

Submission: Submission of the homework will be done via Moodle by uploading a zip file containing a Jupyter notebook and images. The homework needs to be entirely in English. The deadline for submission of Homework 1 is set to December 6, 2020 end of day Israel.

Question 1: Data exploration (15 points)

Download the [MovieLens 100K rating dataset](#). Calculate the dataset sparsity, distribution of number of ratings as well as the average rating value per user\item. Include additional exploration you find relevant to questions 2 and 3.

Discuss your insights and possible challenges related to the prediction task described in question 2.

Question 2: Matrix factorization model implementation and evaluation (60 points)

Split the data to 5 folds. Implement explicit matrix factorization algorithm to predict the rating a user will provide to an item and recommendation of top N items.

Use the RMSE, MRR and NDCG metrics to evaluate your recommender system. For the MRR and NDCG metrics use cutoff values of 5 and 10. Produce a learning curve graph and a table which compares the results of the following algorithms:

- Different choices of the latent factor dimension, regularization, learning rate.
- Gradient Decent vs. ALS
- Bias only based model

- Simple popularity based model

Describe your challenges and findings, including comparison between the algorithms as well as your recommendations for how to select the latent factor dimension, the contribution of regularization and pros\cons of the different algorithms. In your comparison refer to the evaluation metrics, training and inference duration, learning convergence and additional aspects you find relevant.

Question 3: Matrix factorization – item similarity and model explainability (25 points)

- a) Suggest a method to find similarity between items. Demonstrate the approach on a set of 5 item pairs of your choice based on item representation by your best matrix factorization model. Leverage the items side information available in the dataset to demonstrate the effectiveness of the suggested approach.
- b) Select two latent dimensions of your matrix factorization model and try to estimate the latent 'meaning' of the learnt matrix factorization model. Demonstrate your estimation, using few items from the dataset and the items side information available in the dataset and\or external information.

Good luck