Home Assignment 2

This assignment consists of a theoretical and a practical task. The description of the theoretical task can be found in the current document. The descriptions and template for the practical task can be found in the corresponding ipynb file. Your task is to complete the code and answer questions where necessary. Please don't remove any of the code we provide. For your convenience, we will also lunch a Kaggle competition on the dataset used in the practical task, so you can check your solutions.

You must submit one pdf file with the answer to the theoretical problem and an ipynb file with the solution to the practical problem. Please use LaTex to type the answer to the theoretical question. For the practical tasks, develop a solution yourself and provide a reproducible and readable code. Your code should be reproduced without any mistakes with "Run all" mode.

You will receive a total of 100 points for this assignment, if you complete all the tasks. Due to the upcoming final of our course the deadline is strict, so if you hand in the assignment the day after the deadline, you will receive 0 points.

1 Problem 1 (20 pts)

Theoretical Task

In the coupled factorization, assume you have item features Y of dimensionality $n_y < d$, where d is the optimal rank of decomposition.

Which representation of coupled factorization form will be more economic in terms of memory usage for the item cold-start scenario:

- when mapping is applied to the latent item features;

$$\mathcal{L}_{coupled}(A, \Theta) = \|A - PQ^\top\|_{\text{F}}^2 + \|X - PG^\top\|_{\text{F}}^2 + \|Y - QW^\top\|_{\text{F}}^2$$

- when mapping is applied to the initial item features?

$$\mathcal{L}_{coupled}(A, \Theta) = \|A - PQ^\top\|_{\text{F}}^2 + \|X - PG^\top\|_{\text{F}}^2 + \|YW - Q\|_{\text{F}}^2$$

Where $X \in \mathbb{R}^{M \times m_x}$ is the matrix of user attributes;

$Y \in \mathbb{R}^{N \times n_y}$ is the matrix of item features;

$\Theta = \{P, Q, G, W\}$;

$P \in \mathbb{R}^{M \times d}$ is the matrix of latent user features;

$Q \in \mathbb{R}^{N \times d}$ is the matrix of latent item features;

$G \in \mathbb{R}^{m_y \times d}$ is the mapping from user features to user latent features;

$W \in \mathbb{R}^{n_y \times d}$ is the mapping from item features to item latent features.

2 Problem 2 (80 pts)

Solving cold-start problem for antiviral drug discovery

You can find a code template in the corresponding notebook. You need to fill the gaps and answer questions where needed.

Description

Several previously unknown viruses have suddenly emerged in different regions of the planet. There's no time to analyze whether it was a sabotage or not. It is critical to prevent the virus outbreak from turning into a new global pandemic. Prominent research labs around the world have joined their efforts in analyzing the new viruses, but they have only been able to identify the virus families the new strains belong to. No effective anti-viral drugs were discovered so far. It is now essential to find chemical compounds that will have a strong anti-viral effect against the new viruses so that new drugs can be manufactured as soon as possible. However, despite multiple attempts with some well-studied compounds selected by experts for trials, all experiments have demonstrated resistance of the new strains to them.

Your goal

To aid humanity in the fight against the virus outbreaks, you and a little girl with immunity to virus need to reach another city you have been hired to build an accurate hybrid recommender system for viruses. Your task is to suggest the top five known chemical compounds for each new virus that maximize the probability of creating effective anti-viral drugs.

Evaluation metric

Solutions are evaluated using Recall@5 metric.

Kaggle competition For your convenience, we've setup Kaggle leaderboard for this home assignment. You can upload your submission using special function from the notebook and see if it gives sensible results. This will help you to immediately get an understanding of whether you're on the right way and track your progress. If you submission gets you 0 score, something is wrong with your implementation.

Note that the task is about exploration. You don't have to focus on achieving the best possible scores. Once you get reasonable scores, move to the next part of the assignment. The grading for this assignment is based on the correctness of your implementation, not the position on the leaderboard.