Final report

Introduction

First of all, in this assignment, I want to show a research way to find a high-quality recommendation system based on already known solutions, since the mathematical component of the recommendation allows us to operate on matrix factorizations, within which I implemented the task.

Data analysis

The dataset on the basis of which the solutions were made is a relatively small data set (similar amounts of data can be run on the CPU), having a very high discharge. Despite this, it was decided to use the most effective approach and consider the data as a large data set in the future, which will be constantly expanded.

Model implementation

In my work, I have considered non-DL approaches, as they allow the problem to converge faster if we consider the problem of ranking rather than predicting the rating.

Thus, I considered two models based on matrix factorization: classical SVD and hybrid (experimental) SVD. The main difference between the methods is that in the hybrid version, I use more features and preprocessing with data based on several articles about hybrid methods from references.

Model Advantages and Disadvantages

After studying materials from well-known recommendation system competitions (for example, from Netflix or Amazon), I made a number of judgments on the topic that, basically, the "PureSVD" models win. They are much easier to Fine-tune and their convergence of solution complexity drops much faster than that of neural networks

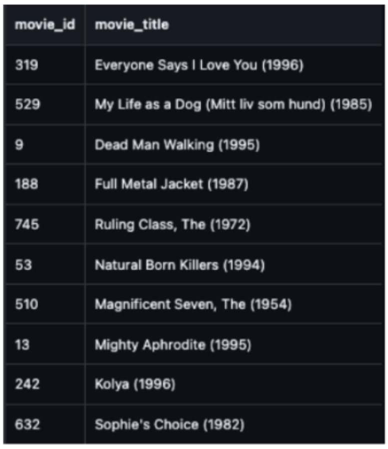
Also, based on the material of the articles, such models show better accuracy than multilayer neural networks, since large models tend to retrain on the "user-item" principles.

Training process

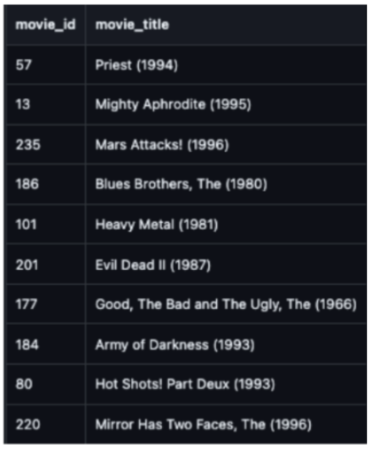
The process of training the solution did not take long to wait, since it is a multiplication of matrices. As a learning optimization, a CSR-matrix was developed, which allows to reduce the occupied volume of the "user-item" matrix by about 40%.

Evaluation

The evaluation of the trained models was carried out by setting the "main title of the film", for which a selection of films consisting of 10 sets was selected. Then, the models were compared using the RMSE metric, which ensures the reliability of information in verifying information for non-network models. A graph was built reflecting the results. Below are film selections (“Schindler's List” and “Snow White and the Seven Dwarfs”, respectively):



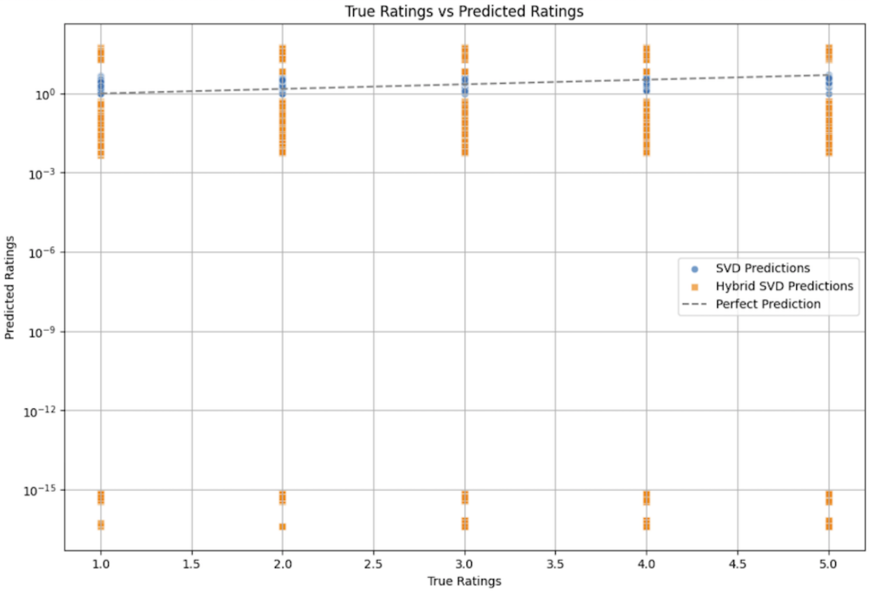
movie_id	movie_title
319	Everyone Says I Love You (1996)
529	My Life as a Dog (Mitt liv som hund) (1985)
9	Dead Man Walking (1995)
188	Full Metal Jacket (1987)
745	Ruling Class, The (1972)
53	Natural Born Killers (1994)
510	Magnificent Seven, The (1954)
13	Mighty Aphrodite (1995)
242	Kolya (1996)
632	Sophie's Choice (1982)



movie_id	movie_title
57	Priest (1994)
13	Mighty Aphrodite (1995)
235	Mars Attacks! (1996)
186	Blues Brothers, The (1980)
101	Heavy Metal (1981)
201	Evil Dead II (1987)
177	Good, The Bad and The Ugly, The (1966)
184	Army of Darkness (1993)
80	Hot Shots! Part Deux (1993)
220	Mirror Has Two Faces, The (1996)

Results

According to the final data and indicators of deviation from the 100% correct decision, it can be noted that additional signs do not always fully reflect the validity and increase accuracy. Thus, in my example, the Hybrid version may even be more accurate than the classic SVD, but it adds high instability due to the multitude of offsets.



In the future, you can experiment with data normalization and better feature extraction, since this was not considered in this paper, but if we talk about the quality of the result, then it shows its place here.