Q1 Information and Honor Code

0 Points

In this assignment, you will work on the Colab 4 notebook and obtain results from it. If your answers are float values, round the decimal number to the **nearest 0.001**. For example, 0.2435 would become `0.244`.

You can submit as many times as you want, and the last submission will be graded. Only the fully correct answer will receive 1 point. No late day is allowed for any Colab assignment.

Please verify that you have read the above instructions and the Stanford Honor Code and that you have not given or received unpermitted aid while completing this assignment.

If you have any questions about how the Honor Code applies to Colab assignments or other parts of the course, please contact the teaching staff for clarification.

◉ I have read and understood the above information

Q2 Collaborative Filtering

7 Points

Q2.1 Train size

1 Point

How many ratings are in the training dataset? (Integer)

Q2.2 Test size

1 Point

How many ratings are in the test dataset? (Integer)

Q2.3 Prediction error - 1

1 Point

Use Spark collaborative filtering (maximal iteration 10, **rank 10**, **regularization 0.1**, and drop rows with NaN value in rating), and train your model on the training set.

What is the root-mean-square error (RMSE) for predicting the movie ratings on the test set? (Float. The answer might vary due to the training process, and we will accept answer +/- 0.004 from the reference.)

Q2.4 Prediction error - 2

1 Point

Now, you train a new model with different rank and regularization, while keeping the other parameters the same as Q2.3. With **rank 100** and **regularization 0.1**, what is the RMSE for predicting the movie ratings on the test set? (Float. The answer might vary due to the training process, and we will accept answer +/- 0.002 from the reference.)

Q2.5 Prediction error - 3

1 Point

You further test different regularizations (1, 0.3, 0.1, 0.03, 0.01) with **rank 100**. Which one of the regularizations gives the lowest RMSE for predicting the movie ratings on the test set?

Q2.6 Recommendation

2 Points

Use the collaborative filtering model trained with **rank 100** and **regularization 0.1**. Now you want to make some recommendations. Instead of examining individual users, we want to recommend a movie to all users.

First, we generate the top-1 recommendation for each user in the dataset.

Second, we count the number of times a movie is recommended, and identify the movie that is recommended to the largest number of users.

Which is the most recommended movie to all users output by your model?

- Someone Else's America (1995)
- Titanic (1997)
- Schindler's List (1993)
- Pather Panchali (1955)
- Angel Baby (1995)