CS529– Applied Artificial Intelligence

Instructor : Shashi Shekhar Jha (shashi@iitrpr.ac.in)

Lab Assignment - 1

Due on 10/09/2019 2400 Hrs

Total: 100 Points

Submissions Instructions:

The submission is through google classroom in a single zip file. In case you face any trouble with the submission, please contact the TA:

- Neeru Dubey, neerudubey@iitrpr.ac.in

You can comment in the google classroom with your queries as well which will be beneficial for the whole class.

Your submission must be your original work. **Do not indulge in any kind of plagiarism or copying.** Abide by the honour and integrity code to do your assignment.

Any late submissions after the due date will attract penalties.

You submission must include:

- A legible PDF document with all your answers to the assignment problems.
- A folder named as 'code' containing the scripts for the assignment along with the other necessary files to run your code.
- A README.txt file explaining how to execute your code.

Naming Convention:

Name the ZIP file submission as follows:

YourName_rollnumber.zip E.g. if your roll number is 2016csx1234 and Abcd,

then you should name the zip file as: Abcd_2016csx1234.zip

About the Dataset

The dataset to be used for this assignment is from GroupLens, a research lab in the Department of Computer Science and Engineering at the University of Minnesota. You are going to use the MovieLens Data (https://grouplens.org/datasets/movielens/). This dataset is quite rich and well documented and has been often used by various researchers to experiment with a variety of recommender system models and compare their performance.

As the authors claim that the online link of the dataset may keep on getting updated with new data, you will be using the dataset provided to you in a zipped file with this assignment.

You can go through the README.txt (provided along with dataset) to know about the dataset.

For the dataset, see other attachments in the assignment post in google classroom.

Programming Language

For this and all successive assignments, you will be using **Python** as the chosen programming language. All data wrangling operations should be done using ***Numpy*** and ***Pandas***.

In case you are new to python, you can use the following resource to get you started on Python https://wiki.python.org/moin/BeginnersGuide/Programmers

Questions

Understanding the dataset [20 points]

Q.1. Plot a histogram to show the variation of the number of user ratings per movie in the dataset. What do you observe from the graph? Also plot the cumulative distribution function (CDF) of number of ratings per movie and list down your observations. Does the plot approximates any well-known distribution? [5 points]

Q.2. Plot the number of movies rated by each user. Also, plot the CDF of number of movies rated by each user. What do you observe from these plots? Are there any outliers in the dataset? [5 points]

Q.3. Provide a list of most popular movies ($\geq$ 100) from the dataset. Mention the criteria that you used, to find the most popular movies. What are your observations from the data for the most popular movies? [10 points]

Recommendation modelling [80 points]

Before you start creating your recommendation engines, the first task is to split the dataset in two parts - Training and Test sets. This essentially means deleting a set of observed entries from the ratings matrix so that you can verify the predicted ratings against the observed ones forming the Test set. The usual ratio of split is 70:30 for Train and Test sets respectively.

It is upto you to create the training and test sets. Remember, a good test set should cover all variations (even some unseen ones) that are there in the complete dataset. Can you think of any sampling based approach to generate the test set? Mention your approach in the submission document. [5 points]

Note that the metrics to be used to evaluate the performance of the different recommender models will be :

1. Root Mean Squared Error (RMSE) and
2. Mean Absolute Error (MAE).

You must implement your own functions to evaluate these metrics. Do not use any built-in function (however, you can verify your implementation by checking the output against the built-in function). ***Submissions that use in-built functions for these metrics will not be evaluated.***

Q.4. Implement the user-based neighbourhood methods using the Pearson correlation coefficient and Cosine similarity measure for the raw rating matrix as well as for the mean-centred rating matrix. Later on, change the prediction function to include standard deviation of user ratings. Document and comment on the comparison of the results. Can you highlight the “Long Tail effect”? Also, take measures to counter the “Long Tail effect” and compare the new results with the earlier ones. Further, you must choose a value of **K** (for selecting the top-K neighbourhood of the target user). What is your approach to choose the best K value? Hint: use different K values to evaluate for the best K. [20 points]

Q.5. Implement the item-based neighbourhood method using the Pearson correlation coefficient and Cosine similarity for the mean-centred rating matrix. Which similarity method performs better in item-based collaborative filtering, comment? Compare your results against the user-based methods. You can use the same value of **K,** as in the previous question. [15 points]

Q.6. Propose an approach to combine the user-based and item-based neighbourhood methods in a single recommendation model. Does this new approach improves the performance of your predictions? If “Yes” or “No”, discuss the reasons with respect to the data for your answer. [10 + 5 = 15 points]

Q.7 Implement both the clustering-based models discussed in the lecture.
Compare your results with the user-based, item-based and combined approaches.
[10+10 = 20 points]

In general, compare the performances of all the recommendation systems implemented so far. The comparison should be based on the accuracy along-with the complexity of the offline and online phases. Give your final comments as to which method would be most preferred for this particular dataset and why? [5 points]

__________________-******________________******________________________