MOVIE RECOMMENDATION ENGINE

Mohammed Rizwan Rawani (11111030), Nitesh Bagmar (11111013)

Instructions and important points.

1. The folder named '11111030_11111013' has a folder inside it named 'programs'. This folder ('programs') contains the various programs and functions implemented. It also has a folder 'data', which contains the required data files.
2. As we had to perform some pre-processing in the data set taken from source, we are also submitting the data folder. The source of data is http://www.grouplens.org/node/73 (100k data set)
3. There are 27 programs (named as Program1 – Program27). The comment above a program describes what approach a program follows. Program1 – Program14 implement the approaches given in papers referred. Program15 and Program16 predict the ratings as just the mean user rating and mean item rating respectively and do not perform any type of filtering technique. Program17-Program27 implement our approaches (We kept on modifying the various steps one by one and we got best results when all the steps had been modified in Program26)
4. All the programs perform the steps from creating the rating matrix to prediction of ratings and evaluation of performance (calculating Mean Absolute Error). These programs differ in the approaches used for various steps.
5. Some programs are the extension of some previous programs (e.g. Program4 is an extension of Program2, Program14 is an extension of Program7). This means that some step has been modified in the new program. Comments have been given describing what step has been modified and referring to previous program as 'extension of previous program (say, Program2)'. In order to see the approaches for the remaining steps in such cases, refer to the comments of corresponding previous program.
6. There are total 19 functions for various steps from creating the rating matrix to calculation of error in the final prediction (These functions implement different approaches for these steps). The functions have been named in such a way that the name contains the step being implemented and the approach being used. Proper comments have also been given above every function. These functions are used by the Programs.
7. TAs need to just execute the programs. The functions are called within the programs. TAs may see the code of the functions if they wish.

8. To run the programs, change the current matlab folder to ‘programs’ folder and type the program name in the command window. Alternatively, right click on the program inside matlab and click run in the popup menu.
9. The results included in the report are of Program1 (Basic Algorithm), Program4 (rating matrix filled with mean user ratings), Program5(rating matrix filled with mean item ratings), Program7(rating matrix filled with mean user ratings along with a correction term), Program14(extension of Program7 using aggregate neighbourhood, i.e. centroid method), Program26(Our algorithm), Program27(Our algorithm using aggregate neighbourhood, i.e. centroid method). These are the complete programs.
10. We proceeded incrementally while shifting from one approach to another and other programs therefore were a combination of these approaches (e.g In Program2, we fill the initial rating matrix with corresponding mean user ratings and find the neighbourhood. But while predicting, we use the initial sparse matrix. This step was further modified in Program4 where we use the complete approximated matrix for all the steps). This was done to properly analyse the performance of various steps. Hence other programs that were just the intermediates while shifting form one approach to other have not been included in the report.
11. The programs may take 10 – 15 minutes for execution. This should not be thought of as a negative performance, because these programs implement all the steps required, over the complete data set. In actual recommendation systems that exist, only the last step (prediction of ratings) is performed online. Rest of the steps are performed offline. Even in the prediction step that is performed online, only one row (corresponding to the user logged in) of the prediction matrix is to be calculated, which would take very less time. Hence actual systems are able to generate predictions in real time.
12. There are 4 datasets of type I (u1.base & u1.test - u4.base & u4.test) and 2 datasets of type II (ua.base & ua.test, ub.base & ub.test) as mentioned in the report. We have averaged the performance of the algorithms over these datasets in the report.
13. The programs submitted use u1.base and u1.test. After running the programs for this data set, if TAs wish to check the programs for other data sets, then please perform the following steps:
    (a) Open the function CreateRatingMatrix.m
    (b) Change the data file being loaded to the desired file (say u2.base). Save changes.
    (c) Open the function error_calculation.m
    (d) Change the data file being loaded to the corresponding test file (say u2.test). Save changes.
    (e) Run the programs.