Data Mining

Project 3

Due: At 23:59, 2 days before the appello of January or February 2016.

Instructions

You must hand in the homeworks electronically and before the due date and time.

Handing in: You must hand in the homeworks by the due date and time by an email to `aris@dis.uniroma1.it` that will contain as attachment (not links!) a .zip or .tar.gz file with all your answers and subject
`[Algorithmic Methods for Data Mining] Project 3`
After you submit, you will receive an acknowledgement email that your project has been received and at what date and time. If you have not received an acknowledgement email within 2 days after you submit then contact the instructor.

The solutions for the theoretical exercises must contain your answers either typed up or hand written clearly and scanned.

The solutions for the programming assignments must contain the source code, instructions to run it, the output of your solutions, and the corresponding RMSE values. Also, during the exam, you should be able to execute it (at your computer) to see how it performs online.

For information about collaboration, and about being late check the web page.

In this final project you have to implement a recommender system for movies. As a dataset we will use the `MovieLens 100K` dataset, available at `http://grouplens.org/datasets/movielens`.

Feel free to use whatever method you want and whatever libraries you can find online.

First evaluate the recommender system offline. Split the dataset (i.e. pairs of user $\times$ movie) into a training set (80% of pairs) and test set (20% of pairs). After training the system on the training set, evaluate it on the test set. We are interested in the RMSE.

Now consider a new user. Assume that you are given a list of movies and scores, corresponding to the preferences of this new users. These will be provided by a file. The user is also interested in some new movies. Recommend to him or her the movies he or she may like.

Some notes about the homework:

- On purpose we have left the project open ended. There is no technique that is better or worse for everything, so try to do your best.
- For extra credit you can try with some of the larger datasets.