Algorithmic Methods of Data Mining

Homework 3

Due: 12/2/2017, 23:59
If you want to do the exam in the January session, then it is due on **23/1/2017**.

Instructions

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

Groups: You have to do this homework in the groups of the previous two homeworks.

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 to some file-uploading server!)** a .zip or a .tar.gz file with your answers and subject
`[Algorithmic Methods for Data Mining] Homework 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 the deadline then contact the instructor.

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 *book-crossing* dataset, available at `http://www2.informatik.uni-freiburg.de/~cziegler/BX/`.

In class we have said that there exist multiple ways for peforming recommendations and what we have seen is just a the very basics. Therefore, when trying to choose what techniques to use in practice, there exists a long process for designing and testing the different approaches that we consider.

As with most supervised-learning tasks, we split the dataset into a *training set* (usulally around 80–90% of the data) and to a *test set*. You train your method (i.e., you learn the various parameters) using only the training set, and then you use the test set to evaluated it. This allows you to compare different approaches on data that they have not seen. One way to measure the performance of an approach is by computing the RMSE error (see, for example, Section 9.4 in the ZAL book, or Section 9.4 in the LRU book).

To obtain a higher confidence in the result we try to remove the dependence on the particular splitting by performing *k-fold cross validation*. There are different ways to do it, here is a simple one. Assume that we decide to use 80% of the data as training set. Then we partition the dataset into five equal groups. Then five times we take one of the groups, we use it as a test set, and the other four ones as a training set. We compute the RMSE error for the different approaches and at the end we take the average over the five independent runs.

The goal is to implement different recommender systems for books. The first part will be offline. Implement at least one of the recommender systems in the LRU book. Then perform cross validation and report the error of your approach(es). For the splitting of data into training and test set, use a 20% test set, that is, 20% of user–book pairs.

Next, consider an online recommender system. Build a system that accepts some books (as a list of ISBNs in a file, one ISBN per line) and returns some recommendations to the user. Present the recommendations that you performed. During the exam you have to demonstrate the online version of your system.

What we describe above is the very least you should do and 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.

There are several things you can do for a better grade. Some ideas:

- Implement and compare different approaches for recommendations, especially if they are of different type.
- Try to find some way to use content information from Amazon and use it in your recommendation.
- Be more intelligent when reading books. For example, allow the input file to give the title and/or author of the book, find some potential matches and let the user choose which one.