Web Intelligence

Assignment 3 – part 1&2 (of 2)

This assignment is about Collaborative Filtering (CF). We will be analyzing the MovieLens data (http://grouplens.org/datasets/movielens/). There are various data sets of different size. I suggest to use the smallest, but for interested parties, go ahead with using any version of the data set there. I further suggest to use Movie Lens 100K data set from 4/1998 as this one contains split to 5 test and training folds. In other data sets you will have to create the folds by using either their toolkit or other toolkit which will result in another time overhead. This particular data set contains over 100 000 ratings from 1000 anonymous users over 1700 movie titles.

1. Data manipulation

a) The 'README' file describes where you can find the "training data" and the "test data" that you will be using, and it describes their formats. As mentioned it contain 5 splits to training and test data. You have two options: computing it in one run only with one training and one testing file, or computing it 5 times and compute averages.

b) Load the two data sets (start by just loading subsets of the data by restricting to a much smaller set of user-id and movie-id, but still of a decent size to make it interesting). We will NOT be using the date/time information, so that information can be ignored!

c) Each user-movie pair in the test data should right now also be apparent in the training data. Take the rating that you see in the training data and add it to the user-movie pair in the test data. Now delete the user-movie pair from the training data. (It may be useful to have your data contained in some form of dictionaries data structure –or similar– for this step). *Why do we perform this data manipulation?*

d) You may want to save the smaller new training and test data, so you don't have to load all of the data and perform this manipulation every time you try something new in your program!

2. Learning

a) *We have talked about two basic paradigms of recommender system. Which one fits the above data? What would we need to build the other type?*

b) Subtract the movie and user means from the training data (the pre-processing step from the slides).

c) Construct a matrix factorization CF model (a.k.a. "Funk-SVD") for this training data. Use between 10 and 50 latent factors.

d) *Does matrix factorization seem as the better choice of CF technique for this data? Why/ Why not?*

3. Scoring

a) Evaluate RMSE = $\sqrt{\frac{\sum_{i=1}^{n}(\hat{r}_i - r_i)^2}{n}}$ on the test data, where $\hat{r}_i$ is the predicted value for a user-movie pair and r_i is the actual observed rating. (Remember to add in what you subtracted in the pre-processing step when you predict.) *What do you get?*

b) Now iterate the whole learning-scoring process (as many times as you like), where you experiment with either more latent factors, a bigger training set (more movie- and user-id), or both. *What are the results now? Any explanation?*