Advanced Data Mining

Piotr Lipiński

List of Assignments 4 – recommendation systems

Assignment 1. *(5 points)*
See the MovieLens datasets (https://grouplens.org/datasets/movielens/).
a) Run the simple collaborative filtering algorithms for preparing recommendations on the MovieLens 100k dataset as well as on a few larger MovieLens datasets (depending on your computational platform). You may use either your own implementation or some implementations available, such as the scikit-surprise package.
b) Use also the matrix factorization algorithms, such as SVD and SVD++.
c) Propose a methodology of comparing the results of different recommendation algorithms. Why the regular cross-validation is not a best choice?
d) Try to compare the recommendation accuracy for experienced users (with many products rated) and for less active users (with only few products rated).
e) Propose a method of evaluating and comparing the coverage and the diversity of the recommendation generated.

Assignment 2. *(optional - 5 bonus points)*
The Yelp Open Dataset (https://www.yelp.com/dataset) contains data businesses, their users and ratings of businesses by their users as well as some rich meta-data concerning the users and the businesses.
a) Try to use the regular recommendation techniques to generate recommendations of the businesses to the users and evaluate them (similarly to Assignment 4).
b) Propose a method of clustering similar businesses (based on some selected meta-data from the dataset) and generate recommendations of the business clusters to the users (e.g. recommend clusters of businesses instead of individual businesses) and evaluate them.
c) For each user *u*, for each business cluster *c*, evaluate the mean rating $\mu_{u,c}$ of the businesses from the cluster *c* by the user *u*. Next, normalize the ratings $r_{u,b}$ of the businesses *b* by the user *u* by subtracting $\mu_{u,c}$ from $r_{u,b}$ (for *c* being the cluster containing the business *b*). Run the regular recommendation algorithms on the normalized ratings and compare the results with a).
In the case of large computational requirements and long computation times, you may focus on a selected part of the entire dataset.