Matrix Completion

Défi Big Data – Mines St-Étienne – 2018/2019

This work will be evaluated. You must provide a **report in pdf**, your **R script** and the file containing your **predictions** on Campus before Monday, 17^{th} of December at 6 p.m.
N.B. Your report should not only comment the results, but must give interpretation and analyze them.

During this lab, you will investigate completion matrix techniques, using the package softImpute of R. In particular, the following functions may be helpful:

- 'Incomplete' reconstructs the incomplete matrix from the raw data.
- 'softImpute' fits a low-rank matrix approximation to a matrix with missing values via nuclear-norm regularization.
- 'impute' produces predictions from the low-rank solution of softImpute.

Dataset

The data set consists of: 3141793 ratings $\in \{1, \ldots, 10\}$ from 39189 users on 4365 movies (anime). Each user has rated at least 20 movies. The data can be found on the data science platform Kaggle (www.kaggle.com).

The purpose of the session is to predict, for a set of users, the ratings of movies they have not seen yet, and, based on these predictions, to suggest a new movie to an user.

Objectives

First, you need to examine the data, provided in the file 'data.csv'. Is it relevant to apply a low-rank based completion technique in this context?

Consider the softImpute variants proposed by the corresponding package. Which parameters will you tune?

In order to realize this tuning, you need to separate the dataset into a training set and a validation set. Which pitfall(s) should be avoided?

To compare the different parametrizations, you can use the root mean square error (RMSE), given by:

$$\sqrt{\frac{\sum_{i=1}^{N} (x_i^{pred} - x_i^{obs})^2}{N}},$$

where x_i^{pred} is the i^{th} prediction, x_i^{obs} is the corresponding realization and N is the total number of predictions.

After finding the most relevant set of parameters, you will predict the ratings corresponding to users and movies given in the files 'test.csv'. **Please save your prediction in a csv file and return it within your report and your code.**

Last, you are invited to recommend **ONE** movie to users number 29686, 22761 and 16132. Movies' titles given their id are provided in the file 'names.csv'.

Bonus: Can you suggest a prediction method using the NMF?