Project supervisor: Rodrigo Cabral Farias (I3S)

Tensor and matrix factorizations in python: application to recommender systems

With applications ranging from large scale recommender systems [1] to geoscience applications [2] and from chemometrics applications [3] to community structure retrieval [4], models such as nonnegative matrix factorization [5,6], collective matrix factorization [7] and tensors [8] became widespread data mining and prediction tools.

Example in recommender systems:

In Video on Demand (VOD) platforms, e.g. Netflix, one is interested in recommending movie titles to users. For each user the recommender system must generate a personalized set of proposals that is expected to be fit to the user profile. By doing so, the number of movie demands is expected to be maximized.

The recommender platform has data on the movie ratings for each user and the ratings are normally indexed by a time stamp. Since users only rate movies they have watched, the three dimensional block of data (User x Movie x Time) has multiple missing entries. The advantage of tensor and matrix models in this context is that, under some constraints, they can be fit to multidimensional blocks of data with only a few present entries. Thus, by querying the obtained model on the missing points, we can predict their values (the missing ratings). Clearly, two questions arise in this problem:

- What is the best temporal model? A tensor model? A collective matrix model? Independent matrix factorization models for the time slices?

- What do we do when the blocks of data are too big for processing with standard tensor decomposition algorithms?

Objective of the project:

In this project, the students are asked to develop a python library for fitting nonnegative matrix, collective matrix and tensor factorization models. The students will be first asked to code standard algorithms from the literature such as multiplicative algorithms [6, 9] and alternating methods [10], then adapt them to data with missing entries and finally to propose strategies which are scalable to datasets of increasing sizes. The different algorithms will be tested on the Movielens dataset [11]. A comparison between the different models should be carried out in terms of prediction accuracy of the missing entries.

Requirements and acquired knowledge:

This project will require some skills on optimization, linear algebra, machine learning and coding. After this project, the students will be able to grasp some of the difficulties of dealing with big data in a machine learning/optimization context and will have some experience on recommender systems.

References:

[1] Zhang, S., Wang, W., Ford, J., & Makedon, F. (2006, April). Learning from Incomplete Ratings Using Non-negative Matrix Factorization. In SDM (Vol. 6, pp. 548-552).

[2] Pauca, V. P., Piper, J., & Plemmons, R. J. (2006). Nonnegative matrix factorization for spectral data analysis. Linearalgebra and its applications, 416(1), 29-47.

[3] Smilde, A., Bro, R., & Geladi, P. (2005). Multi-way analysis: applications in the chemical sciences. John Wiley & Sons.

[4] Gauvin, L., Panisson, A., & Cattuto, C. (2014). Detecting the community structure and activity patterns of temporal networks: a non-negative tensor factorization approach. PloS one, 9(1).

[5] Paatero, P., & Tapper, U. (1994). Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. Environmetrics, 5 (2), 111-126.

[6] Lee, D. D., & Seung, H. S. (2001). Algorithms for non-negative matrix factorization. In Advances in neural information processing systems (pp. 556-562).

[7] Singh, A. P., & Gordon, G. J. (2008). Relational learning via collective matrix factorization. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 650-658). ACM.

[8] Kolda, T. G., & Bader, B. W. (2009). Tensor decompositions and applications. SIAM review, 51(3), 455-500.

[9] Cichocki, A., Zdunek, R., Phan, A. H., & Amari, S. I. (2009). Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation. John Wiley & Sons.

[10] Comon, P., Luciani, X., & De Almeida, A. L. (2009). Tensor decompositions, alternating least squares and other tales. Journal of chemometrics, 23(7-8), 393-405.

[11] https://grouplens.org/datasets/movielens/