Titre du projet : Gradient algorithms for co-clustering and higher-order co-clustering

Type de projet : (X) recherche ( ) développement

Résumé du projet :

Given a matrix of data containing values of different variables (the matrix columns) for different individuals (the matrix rows), clustering methods approximate the data by partitioning its rows in a small number of partitions and by replacing the rows in each partition by a single row vector. Similarly, in feature reduction methods, the columns are partitioned, and the elements of each partition are replaced by a single column vector. Both techniques are heavily used in data analysis, since instead of analyzing the differences from one individual or variable to another, one can simply analyze the differences between the few row or column vectors representing groups of individuals or variables.

To simplify analysis even further, one can consider a co-clustering approach [1], where data are simultaneously partitioned into individuals/features groups. Co-clustering can be cast as a matrix tri-factorization problem [2], where data is approximated by a small real-valued matrix multiplied on its left and right by two binary-constrained selection matrices. In this project, an algorithm for calculating this approximate tri-factorization based on unconstrained gradient descent will be studied. The main difficulties of the project will lie on how to relax the binary constraints, reparametrize the approximation problem and add smooth penalization terms so that the initially constrained problem becomes unconstrained.

An extension of the algorithm to tackle the problem of higher-order co-clustering will also be considered. In higher-order co-clustering [3], a tensor [4] of data is available and the objective is to partition simultaneously the rows, columns and tubes of the tensor. Such co-clustering problem may appear in applications where data naturally have a third diversity, for example multivariate time series.

The algorithms will be coded python and tested both artificially generated datasets and real datasets (e.g.: on the dataset considered in [5]).

Mots-clés : Clustering, co-clustering, matrix factorizations, tensors, gradient descent.

Références bibliographiques :

[1] Brault, V., & Lomet, A. (2015). Methods for co-clustering: a review. Journal de la Société Française de Statistique, 156(3), 27-51.
[2] Ding, C., Li, T., Peng, W., & Park, H. (2006). Orthogonal nonnegative matrix t-factorizations for clustering. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 126-135).
[3] Jegelka, S., Sra, S., & Banerjee, A. (2009). Approximation algorithms for tensor clustering. In International Conference on Algorithmic Learning Theory (pp. 368-383). Springer, Berlin, Heidelberg.
[4] Kolda, T. G., & Bader, B. W. (2009). Tensor decompositions and applications. SIAM review, 51(3), 455-500.
[5] Battaglia, E., & Pensa, R. G. (2019). Parameter-Less Tensor Co-clustering. In International Conference on Discovery Science (pp. 205-219). Springer, Cham.

Nom et affiliation de l'encadrant : Rodrigo CABRAL FARIAS, I3S, Polytech Nice MAM