Matrix Factorization for Clustering: NMF and Beyond

Da Kuang

Advisor: Haesun Park

1 Introduction

Clustering is a fundamental problem in unsupervised and semi-supervised machine learning. Besides classical approaches like K-means, recent approaches are developed through the formulation into nonnegative matrix factorization (NMF). Optimized by a multiplicative update rule algorithm or alternating least squares algorithm, NMF can achieve much higher accuracy of clustering than K-means in the high-dimensional case (hundreds to millions) [3], especially when nonnegativity is imposed in each iteration of the alternating least squares process (ANLS) [2].

Generally speaking, there are two ways of formulating clustering problem into NMF framework. First, let J be the objective function using Euclidean distance. We can express J in a straightforward way [3]:

$$J = \|A - CH^T\|_F^2 \tag{1}$$

where $A = [a_1, \cdots, a_n] \in \mathbb{R}^{m\times n}$ is n nonnegative data points in m dimensions, $C = [c_1, \cdots, c_k] \in \mathbb{R}^{m\times k}$ is the centroid points of k clusters, and $H \in \mathbb{R}^{n\times k}$ denotes clustering assignment matrix, i.e. $H_{ij} = 1$ if data point a_i belongs to cluster j, and $H_{ij} = 0$ otherwise.

However, the above formulation is inflexible to include extensions, such as kernel trick in kernel K-means clustering [4], or must-link and cannot-link constraints in semi-supervised clustering [5]. Another formulation, introduced by Ding et al. [1], is adaptable to these extensions:

$$J = \|W - HH^T\|_F^2 \tag{2}$$

where $W = A^TA$ is the pairwise similarity matrix. W can also be a general kernel matrix, or include constraints in semi-supervised case. Under the assumption of orthogonality $H^TH = I$, (2) is an acceptable relaxation of (1) in terms of clustering assignment: each row of H contains only one positive entry, indicating assignment of the corresponding data point.

The formulation (2) is called symmetric NMF. We are seeking to solve it by ANLS algorithm due to its superior performance, but a whole new set of optimization procedures have to be applied. Also, beyond NMF, there are more general matrix factorization methods worth considering, and this leads to some fundamental aspects of clustering problem.

2 Project

(1) Computational issues in solving symmetric NMF using ANLS algorithm

We will develop related optimization theory, such as iterating framework, stopping criterion, and proof of convergence. And there is opportunity to reduce time and space complexity, considering the symmetric property of the matrix W. Currently, we propose the following iterating procedures:

$$H_2 \leftarrow \min_{H\geq 0} \|H_0H^T - W\|_F^2 \tag{3}$$

$$H \leftarrow (1-\alpha)H_0 + \alpha H_2 \tag{4}$$

We fix the left factor H as H_0, and solve for the right factor H^T as a least squares problem, where H^T is restricted to be nonnegative; then, compute a weighted sum of previous H_0 and new result H_2. Proof of convergence exists for a similar algorithm using multiplicative update rule. Here, we propose to prove convergence in the ANLS case based on the convergence theory developed in two-block coordinate descent framework [2], as well as derive a stopping criterion using KKT condition.

(2) Sparsity vs. Orthogonality

In the above iteration (3)(4), the optimization for H does not include any constraint except nonnegativity. Ideally, for the purpose of clustering, we wish to restrict H to be exactly an assignment matrix. However, due to computational reasons, this constraint is often relaxed to sparsity or orthogonality on H (soft clustering), which may also have a reasonable interpretation. One version of the modified formulations can be: (sparsity on (1), orthogonality on (2))

$$J = \|A - CH^T\|_F^2 + \beta \sum_{j=1}^{n} \|h_j\|_1^2 \tag{5}$$

$$J = \|W - HH^T\|_F^2 + \beta \|H^T H - I\|_F^2 \tag{6}$$

where h_j is the j-th row of H. In the nonsymmetric form, sparsity with L_1 norm can be used. But the crucial point is that the symmetric form (2) results from a derivation which is based upon the assumption of orthogonality $H^T H = I$. Therefore, simply imposing sparsity in the symmetric form seems to make little sense. Since it is not so obvious how to combine the penalty term in (6) into ANLS algorithm, we will investigate how much these two constraints can differ in terms of clustering performance.

(3) Kernel clustering using symmetric NMF; Matrix factorization beyond NMF

We have discovered that NMF is less likely to be superior than K-means in the low-dimensional case; that is to say, the formulation into NMF limits the ability to exploit the structure of data within some subspace. However, the data points are not necessarily restricted to where they are supposed to be, but can be mapped into high-dimensional space using kernels. Literature on kernel clustering is numerous, but hardly contains any approaches using symmetric NMF.

If the centroid vectors of each cluster are made nearly orthogonal through certain kernel, we can expect H would naturally result in the form of assignment matrix, even if nonnegativity or sparsity is not imposed. In this sense, we can further explore general matrix factorization and nonlinear programming techniques for clustering, both classical and novel. For example, by viewing clustering as a dimension reduction method, and K-means as a lower rank approximation of the original data matrix, we propose to design new implementations of K-means by matrix factorization. Basically, K-means seeks a pointwise way to solve the formulation of clustering as a matrix factorization (1) or (2). Efforts in a matrix-factorization way will shed light on a different look of K-means and clustering problem.

References

[1] C. Ding, X. He, and H. D. Simon. On the equivalence of nonnegative matrix factorization and spectral clustering. In *Proceedings of the SIAM International Conference on Data Mining*, pages 606–610, 2005.

[2] H. Kim and H. Park. Non-negative matrix factorization based on alternating non-negativity constrained least squares and active set method. *SIAM Journal in Matrix Analysis and Applications*, 30(2):713–730, 2008.

[3] J. Kim and H. Park. Sparse Nonnegative Matrix Factorization for Clustering. *Technical Report GT-CSE-08-01*, 2008.

[4] B. Kulis, S. Basu, I. Dhillon, and R. Mooney. Semi-supervised graph clustering: a kernel approach. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 457–464, 2005.

[5] T. Li, C. Ding, and M.I. Jordan. Solving consensus and semi-supervised clustering problems using nonnegative matrix factorization. In *Proceedings of the IEEE International Conference on Data Mining*, pages 577–582, 2007.