Nonnegative matrix factorization for interactive topic modeling and document clustering

Sarah Lu, Jody Shu, Jashaina Thomas

February 11, 2020

Nonnegative matrix factorization (NMF) approximates a nonnegative matrix by the product of two low-rank nonnegative matrices. Since it gives a semantically meaningful result that is easily interpretable in clustering applications, NMF has been widely used as a clustering method especially for document data, and as a topic modeling method. These notes focus on the broad spectrum of NMF in the context of clustering and topic modeling.

1 Introduction to Nonnegative Matrix Factorization

Nonnegative matrix factorization (NMF) is a dimension reduction method and factor analysis method related to low-rank approximations of matrices in which the low-rank factor matrices are constrained to have only nonnegative elements.
Suppose a nonnegative matrix $A \in \mathbb{R}^{mxn}$ is given. When the desired lower dimension is k, the goal of NMF is to find the two matrices $W \in \mathbb{R}^{mxk}$ k and $H \in \mathbb{R}^{kxn}$ n having only nonnegative entries such that

$$A \approx WH.$$

Each data point, which is represented as the column of A, can be approximated by an additive combination of the nonnegative basis vectors, which are represented as the columns of W. k is assumed to satisfy $k < min\{m, n\}$. The matrices W and H are found by solving an optimization problem defined with the Frobenius norm (a distance measure between two given matrices).

$$min\{W \geq 0, H \geq 0\} f(W, H) = ||A - WH||_F^2$$

All entries of W and H are nonnegative. Because of this, the result of NMF can be viewed as document clustering and topic modeling results directly.

2 Nonnegative Matrix Factorization for Clustering

Consider the low-rank approximation where $A \in \mathbb{R}_{+}^{mxn}$, $W \in \mathbb{R}_{+}^{mxk}$,$H \in \mathbb{R}_{+}^{kxn}$, and $k << min(m, n)$ is the pre-specified lower rank. The columns of A represent n data points in an m dimensional space. Each column of H is the k-dimensional representation of a data point. If we can use H to derive an assignment of the n data points into k groups, clustering can be viewed as a special type of dimension reduction.
An example of dimension reduction is NMF:

$$min\{W \geq 0, H \geq 0\}f(W, H) = ||A - WH||_F^2$$

The columns of W provide the basis of a latent k-dimensional space, and the columns of the second factor H provide the representation of $a_1, ..., a_n$ in the latent space. The columns of W are interpreted as k cluster representatives, and the i-th column of H contains the soft clustering membership of the i-th data point for the k clusters. In document clustering and topic modeling, the basis vectors in W represent k topics, and the coefficients in the i-th column of H indicate the topic proportions for a_i, the i-th document.To obtain a hard clustering result, we simply choose the topic with the largest weight, i.e., the largest element in each column of H.
If $k < rank(A)$, the columns of W are linearly independent due to $rank(A) \leq nonnegative-rank(A)$. Therefore, NMF performs well when different clusters correspond to linearly independent vectors.
The success of NMF as a clustering method depends on the underlying data set, and its greatest success has been in the area of document clustering. Recently, NMF has been applied to topic modeling, a task related to document clustering, and achieved satisfactory results. In a document data set, data points are often represented as unit-length vectors and embedded in a linear subspace. For a term-document matrix A, a basis vector w_j is interpreted as the keyword-wise distribution of a single topic. When these distributions of k topics are linearly independent, which is usually the case, NMF can properly extract the ground-truth clusters determined by the true cluster labels.
The output types include: a keyword-wise topic representation (the columns of W), and a topic-wise document representation (the columns of H).

3 Optimization Framework for Nonnegative Matrix Factorization (NMF)

In section 3, authors discuss block coordinate descent (BCD) and multiplicate updating (MU) methods to approximates a nonnegative matrix by the product of two low-rank nonegative matrices, i.e. A $\approx$ WH . . . (1). MU is more inferior solutions than BCD due to its slow convergence which in the later section, the authors show the better resulting clustering by BCD algorithm.

Basically, BCD is an algorithm takes turns solving W and H matrices and A is given as an input. The follwings are the subproblems for the constraints of the BCD and are called nonnegatiity constrained least squares (NLS);therefore; the BCD has also been referred to the alternating nonnegatie least squares (ANLS).

$$\min_{W\geq 0}\|H_T W_T - A_T\|_F^2 \ldots (8a)$$
$$\min_{H\geq 0}\|WH - A\|_F^2 \ldots (8b)$$

One needs to initialize H or W first to solve (8a) and (8b) iteratively as stated in Algorithm 1 (Please see the original paper.). The initialization of H or W matrices can be randomly and pick the solution with the smallest cost value, and other proposed method is to run the K-means with the cosine similarity as the distance function, and use the final outputs centroids as the initialization of W or H. In addition, there are other methods to solve the subproblems.

The objective function of NMF is nonconvex, therefore, it's no gurarantee to find a local minimum, but can only find the stationarity of a limit point as stated in the following theorem 1.

Theorem 1 If a minimum of each subproblem in (8) is attained at each step, every limit point of the sequence $\{(W, H)^{(i)}\}$ generated by the **BCD** framework is a stationary point of $\min_{H\geq 0, W\geq 0}\|A - WH\|_F^2 \ldots (2)$

One should note that the minimum of each subproblem is not unique and this solution (stationary point), the Karush-Kuh-Tucher(KKT) condition is satisfied the follows:

$W \geq 0, H \geq 0, \ldots$ (9a)
$\nabla f_W = 2WHH^T - 2AH^T \geq 0,\ \nabla f_H = 2W_T WH - 2W_T A \geq 0 \ldots (9b),$
$W. * \nabla f_W = 0, H. * \nabla f_H = 0 \ldots (9c).$

As for the MU algorithm, it does not have the convergence property as **Theorem 1**, however, it is simple and easy to implement and it is more popular than BCD and the solutions for the subproblems (8) are not optimal. The update rule of MU is similar to gradient descent algorithm with certain chosen step lengths for making sure the result is nonnegative. Furthermore, the solutions of MU are denser than BCD framework, therefore, it's harder to explain the clustering results.

With the numerical analysis to find the solutions for the subproblems (8), a stop criteria between each objective function is set to be less than a pre-defined threshold ε as follow:
$|f(W^{(i-1)}, H^{(i-1)}) - f(W^{(i)}, H^{(i)})| \leq \varepsilon$

However, this is not a good method since the above-mentioned criteria may be satisfied before finding a stationary point.

Instead, the projected gradient matrices at the i-th iteration by $\nabla^p f_W^{(\mathrm{i})}$ and $\nabla^p f_H^{(\mathrm{i})}$ is defined as follows:

$\nabla(i) = \sqrt{\|\nabla^p f_W^{(\mathrm{i})}\|_F^2 + \|\nabla^p f_H^{(\mathrm{i})}\|_F^2} \ldots (14)$
Where conditions (9) can be rephrased as $\nabla^P f_W = 0, \nabla^P f_H = 0 \ldots (13)$, Using equation (14), the stopping criterion becomes $\frac{\nabla(i)}{\nabla(1)} \leq \varepsilon \ldots (15)$ where $\nabla(1)$ is the first iteration of (W, H).
When there are hard clustering result, the authors impose extra constraints into the NMF formula (2) by adding two regularization terms to it. The following are the two terms to be added to promote sparsity.
$\phi(W) = \alpha\|W\|_F^2, and\ \Psi(H) = \beta \sum_{i=1}^n \|H(:,i)\|_1^2$ where H(:,i) represents the i-th column of H.
Moreover, the authors introduce Weakly-Supervised NMF ($WSNMF$) which is semi-supervised algorithm which enhances the visual analytics environment for the users. The formulation introduces $W_r\ and\ H_r$ which are the referenced matrices of W and H matrices respectively, and these referenced matrices are similar to W and H matrices. In addition, there are $M_W\ and\ M_H$ for the diagonal mask/weight matrices. The formula is as follows:
$f(W, H, D_H) = \min\limits_{W,H,D_H} \|A - WH\|_F^2 l + \|(W - W_r)M_W\|_F^2 + \|(H - H_r D_H)M_H\|_F^2 \ldots (19)$
where D_H is a diagonal matrix.
To optimize formula (19), one has to use iterative method to update W, H(,i) and D_H as follows:
W$\leftarrow \underset{W \geq 0}{\text{argmin}} \| [\begin{matrix} H_T \\ M_W \end{matrix}] \| W_T - [\begin{matrix} A_T \\ M_W W_r^{\mathrm{T}} \end{matrix}] \|_F^2 \ldots (20)$
H(:,i)$\leftarrow \underset{H(:,i) \geq 0}{\text{argmin}} \| [\begin{matrix} W \\ M_H(i) I_k \end{matrix}] \| H(:,i) - [\begin{matrix} A(:,i) \\ M_H(i) D_H(i) H_r(:,i) \end{matrix}] \|_F^2 \ldots (21)$
$D_H(i) = \leftarrow \begin{cases} \frac{H_r(:,i)^T \cdot H(:,i)}{\|H_r(:,i)\|_2^2}. & \text{if } M_H(i) \neq 0 \\ 0. & \text{otherwise} \ldots (21) \end{cases}$
where (:,i) indicates the i-th column of a matrix.

4 Choosing the Number of Clusters

The authors used the idea of a consensus matrix. For a data set with n samples, the (i,j)-th entry of a consensus matrix $\tilde{C} \in \mathbb{R}^{n \times n}$ is the co-clustered frequency of the i-th and j-th samples over multiple runs of NMF. Basically, there are T subsets generated by random sampling, each with sampling rate r, and run NMF algorithm on each subset with the same number of clusters k. Define the elements in matrices $C^{(t)}\ and\ S^{(t)}$ as follows:

$$c_{ij}^{(\mathrm{t})} = \begin{cases} 1, & \text{if } the\ i^{th}\ and\ the\ j^{th}\ documents\ belong\ to\ the\ same. \\ & cluster\ using\ A_t; \\ 0, & \text{otherwise.} \end{cases} \tag{23}$$

$$s_{ij}^{(\mathrm{t})} = \begin{cases} 1, & \text{if } \textit{boththe } i^{th} \textit{ and the } j^{th} \textit{ documents appear in } A_t. \\ 0, & \text{otherwise.} \end{cases} \quad (24)$$

Obviously, if $c_{ij}^{(\mathrm{t})} = 1 \Rightarrow s_{ij}^{(\mathrm{t})} = 1$

The element of consensus matrix $\tilde{c}_{ij} = \frac{\sum_{t=1}^{T} c_{ij}^{(\mathrm{t})}}{\sum_{t=1}^{T} s_{ij}^{(\mathrm{t})}} \ldots (25)$. And the dispersion coefficient ρ becomes:

$$\rho = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} 4(\tilde{c}_{ij} - 0.5)^2 \ldots (26)$$

where the $0 \leq \rho < 1$. The number of clusters is chosen to be the maximum of ρ.

5 Experimental Results

In section 5, authors present the empirical evidences that support NMF as a successful document clustering and topic modeling method. They compare the clustering quality between K-means and NMF; Within the NMF algorithms, they compare the multiplicative updating (MU) algorithm and the alternating nonnegative least squares (ANLS) algorithm in terms of their clustering quality and convergence behavior, as well as sparseness and consistency in the solution.
The data are described as follows:

Data set	Terms	Documents	Ground-truth clusters
TDT2	26,618	8,741	20
Reuters	12,998	8,095	20
20 Newsgroups	36,568	18,221	20
RCV1	20,338	15,168	40
NIPS14-16	17,583	420	9

They process each term-document matrix A in two steps. First, they normalize each column of A to have a unit $L2$-norm. Conceptually, this makes all the documents have equal lengths. Next, they compute the normalized-cut weighted version of A:

$$D = diag(A^T A 1_n), A \leftarrow A D^{-1/2}$$

For K-means clustering, they used the standard K-means with Euclidean distances, through a batch-update phase and a more time-consuming online-update phase in Matlab.
For the ANLS algorithm for NMF, they used the block principal pivoting algorithm to solve the NLS subproblems and the stopping criterion was $\varepsilon = 10^{-4}$. For the MU algorithm for NMF, they used another stopping criterion:

$$\|H^{i-1} - H^{i}\|_F / \|H^{i}\|_F \leq \varepsilon$$

Clustering quality: Normalized Mutual information (NMI) is calculated by:

$$\text{NMI} = \frac{I(C_{ground_truth}, C_{computed})}{[H(C_{ground_truth}) + H(C_{computed})]/2} = \frac{\sum_{h,l} n_{h,l} \log \frac{n_{h,l}}{n_h n_l}}{(\sum_h n_h \log \frac{n_h}{n} + \sum_l n_l \log \frac{n_l}{n})/2}$$

6 UTOPIAN: User-driven Topic Modeling via Interactive NMF

In this section, the authors present a visual analytics system called UTOPIAN (User-driven Topic Modeling Based on Interactive NMF). UTOPIAN provides a visual overview of the NMF topic modeling result. Beyond the visual exploration of the topic modeling result in a passive manner, UTOPIAN provides various interaction capabilities that can actively incorporate user inputs to topic modeling processes.

Topic keyword refinement: This interaction allows users to change the weights corresponding to keywords so that the meaning of the topic can be refined.

Topic merging: This interaction merges multiple topics into one.

Topic splitting: It splits a particular topic into the two topics. To guide this splitting process, users can assign the reference information for the two topics.

Document-induced topic creation: This interaction creates a new topic by using user-selected documents as seed documents.

Keyword-induced topic creation. It creates a new topic via user-selected keywords. For instance, given the summary of topics as their representative keywords, users might want to explore more detailed (sub-)topics about particular keywords.

7 Conclusions and Future Directions

In this paper, the authors have presented nonnegative matrix factorization (NMF) for document clustering and topic modeling. They have first introduced the NMF formulation and its applications to clustering. Next, they have presented the flexible algorithmic framework based on block coordinate descent (BCD) as well as its convergence property and stopping criterion. Based on the BCD framework, they discussed two important extensions for clustering, the sparse and the weakly-supervised NMF,and their method to determine the number of clusters. Experimental results on various real-world document data sets show the advantage of their NMF algorithm in terms of clustering quality, convergence behavior, sparseness, and consistency. Finally, they presented a visual analytics system called UTOPIAN for interactive visual clustering and topic modeling and demonstrated its interaction capabilities such as topic splitting/merging as well as keyword-/document-induced topic creation.