Collective Matrix Factorization for Co-clustering

Mrinmaya Sachan Shashank Srivastava
Carnegie Mellon University
{mrinmays, shashans}@cs.cmu.edu

ABSTRACT

We outline some matrix factorization approaches for co-clustering polyadic data (like publication data) using non-negative factorization (NMF). NMF approximates the data as a product of non-negative low-rank matrices, and can induce desirable clustering properties in the matrix factors through a flexible range of constraints. We show that simultaneous factorization of one or more matrices provides potent approaches for co-clustering.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data Mining

Keywords

Matrix Factorization, Co-clustering, Social Networks

1. INTRODUCTION

Real world data is often bimodal or dyadic. For example, *documents* comprise of *words*, *bloggers* generate *content* in social networks, *users* rate *movies* in recommendation systems, *customers* purchase *products* in retail data, etc. Such dyadic (polyadic) data can often be naturally represented as matrices, with rows and columns representing one of the entities. A relevant problem is to co-cluster both row and column entities, e.g., topic clusters for users and text, clusters of movies and users in recommendation systems, product-groups and consumer-groups in purchase records.

In this work, we focus on publication data. Publication data display a variety of linkages between entities. Documents often have multiple co-authors, and could be linked to other documents through citations. Documents or authors could also be seen as related to sets of abstract topics and these topics, in turn as mixtures over words. Accounting for these interactions to cluster on multiple dimensions presents challenges. The problem has been partially approached by topic models like LDA and Author Topic Model. However, these models omit some relationships between the data entities or make simplifying assumptions about the generative process of data to make inference tractable. We present an alternate perspective based on multiple matrix factorization that could avoid some of these problems.

2. METHOD

We employ the framework of matrix factorization that provides flexible formulations to co-cluster and induce desirable properties in resulting matrix factors. We consider broadly two sets of constraints to encode two different notions of clustering. First, we restrict the rows of factor matrices to lie on a probability simplex, i.e. the components are non-negative and normalized. This corresponds to a soft-clustering interpretation, where a point belongs to several clusters with different affinities. The alternate view is to treat the cluster assignments as hard partitions of data. This can be enforced by non-negativity and orthogonality constraints on factors.

Let A, D and W denote the numbers of authors, documents and words in our corpus respectively. Let $X_{(D\times W)}$, $Y_{(A\times D)}$ and $Z_{(D\times D)}$ represent the input matrices containing normalized counts/association strengths. Let T be a predetermied number of latent topics. Let $\theta_{(D\times T)}$ be a matrix that represents each document as a distribution over the topic space. Similarly, let matrices $\eta_{(A\times T)}$ and $\phi_{(T\times W)}$ represent author interests and topic distributions over the vocabulary space, respectively. The clustering problem is then to simultaneously factorize the input matrices X, Y and Z into θ, η, ϕ.

2.1 Clustering Documents and Words

We first describe a base case when we only have raw publication texts. In other words, we are given a word-document matrix X, that we wish to decompose to factors θ and ϕ. A reconstruction error such as a Squared Frobenius norm can be used to quantify the goodness of the approximation, $L_{SMF}(\theta,\phi) = \|X-\theta\phi\|_F^2$. The objective is non-convex overall, but is individually convex in each of the two factors θ and ϕ. This allows for the application of block gradient descent to solve the unconstrained optimization.

We could introduce orthogonality constraints. Adding these constraints to our previous formulation would restrict θ and ϕ to be cluster indicator matrices and lead to the following biconvex optimization:

$$\text{Min}\|X-\theta\phi^T\|_F^2 \quad \text{s.t.} \quad \theta\geq 0, \phi\geq 0, \theta^T\theta = I, \phi^T\phi = I$$

However, this would be an over-constrained problem as orthogonal factor approximations for X may not exist. For example, it is easy to see that the constraints restrict the elements in both θ and ϕ to be less than unity, which would give a very poor approximation if elements in X are extremely large. This is resolved in [1] by adding a $T\times T$ factor matrix Σ to account for the difference in scales of X, and provide additional degrees of freedom to get a close low-rank approximation for X. The optimization is:

$$\text{Min } \|X-\theta\Sigma\phi^T\|_F^2 \text{ s.t.} \theta\geq 0, \phi\geq 0, \Sigma\geq 0, \theta^T\theta = I, \phi^T\phi = I$$

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

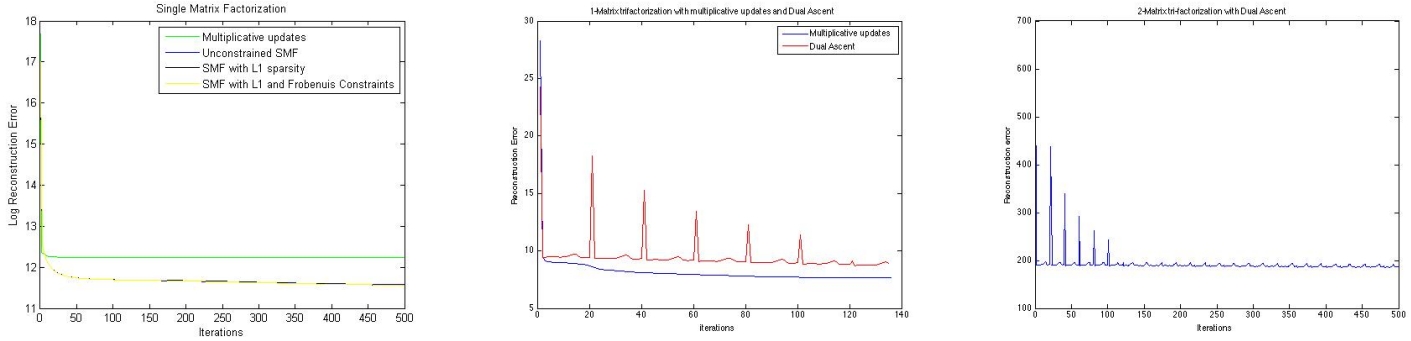

Figure 1: Reconstruction Error vs Number of Iterations plots: (a) SMF: Unconstrained, Constrained and Multiplicative Updates and (b) SMF Multiplicative Updates vs Dual Ascent, and (c) BMF Dual Ascent.

Refer to [1] for multiplicative gradient descent updates.

2.2 Clustering Authors, Documents and Words

We extend the previous formulation when we also have author information. We seek non-negative factors θ, η and ϕ for matrices $X_{(D\times W)}$ and $Y_{(A\times D)}$ and choose weighted sum of the reconstruction losses for X and Y as the objective.

$$L_{BMF}(\theta,\eta,\phi) = \alpha\|X - \theta\Sigma_1\phi\|_F^2 + (1-\alpha)\|Y - \eta\Sigma_2\phi\|_F^2$$

First, we look at normalization constraints, which constrain row sums of the factors to be unity, and view each row of θ, η and ϕ as a distribution. We enforce L_1 regularization on factor matrices to enforce sparsity. Also, we want our matrix factors $\theta\eta^T$ and $\eta\phi$ to be low rank. This regularization can be approximated by penalizing the Frobenius norms of θ, η and ϕ. We optimize the objective using projected gradient descent. PGD makes an update for the unconstrained optimization and then projects the modified factors to the constrained space at each iteration.

For a hard clustering assignments, orthogonality constraints present a natural choice. Under these constraints, however, PGD becomes intractable as there are no efficient ways to project a matrix to the space of orthogonal matrices. Here we resort to a Dual Ascent approach. In each iteration of the algorithm, we exactly solve for the primal variables θ, η, ϕ that minimize the augmented lagrangian using gradient updates, and then make partial updates on the dual variables in the direction of the positive gradients. To ensure non-negativity, we also project the factor components to the positive quadrant after each iteration.

2.3 Clustering Authors, Documents and Words with Citation information

Finally, we also add citation information. Let Z be the $D \times D$ citation matrix. Let D_{ij} be the number of times i cites j or j cites i. Thus, maintaining the matrix factorization objectives in our earlier formulation, we also want to decompose Z into $\eta\eta^T$.

$$L_{TMF}(\theta,\eta,\phi) = \alpha\|X - \theta\eta^T\|_F^2 + \beta\|Y - \eta\phi\|_F^2 + \gamma\|Z - \eta\eta^T\|_F^2$$

Here, the matrix differentials cannot be computed in closed form due to the appearance of the quadratic term $\eta\eta^T$. Hence, we resort to an equivalent relaxation, where we introduce and additional matrix ψ to approximate η^T and enforce a constraint that ψ and η^T are close to each other. We choose the weights α, β and γ by hold-out cross validation.

3. EXPERIMENTS

We evaluate our approach on the Cora (2480 authors, 2410 documents and 2961 words) and NIPS datasets (2037 authors, 1740 documents and 13649 words). Figure 1 plots the decrease in reconstruction error of various formulations with iterations. We note that variations with different penalty functions converge at a similar rate. On the other hand, for 1-matrix tri-factorization, multiplicative updates converge much quicker than Dual Ascent. The Dual Ascent algorithm shows a characteristic decline within each epoch (which corresponds to minimizing w.r.t. the primal variables for a particular value of dual variables), and an upward spike that corresponds to a partial update of the dual variables in the optimization problem. Overall, the objective decreases, but at a much slower rate than multiplicative updates. For 2-matrix tri-factorization, Dual Ascent retains its characteristic shape, but convergence is slower and needs many more iterations. In general, orthogonal constraints provide more interpretable results (in terms of a manual judgment of topics) than simplex constraints. To illustrate the topic modeling ability of the models, we give a glimpse of the topics extracted by the 3-Factor orthogonal NMF on factorization of the document word and author word matrices respectively in Table 1 and 2.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Report	Learning	Neural	Results	University
Revised	Machine	Networks	Problem	Department
Research	Reinforcement	Network	Method	Statistics
Internal	Knowledge	Recurrent	Approach	Science
Journal	Learn	Training	Model	Computer

Table 1: 3 Factor Orthogonal NMF topics on $X_{D\times W}$

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Paper	Learning	Research	Algorithm	Model
Algorithm	Machine	Office	Model	Paper
Results	Concept	Supported	System	Data
Model	Theory	Partially	Show	Models
Algorithms	Task	Naval	Performance	Class

Table 2: 3 Factor Orthogonal NMF topics on $X_{A\times W}$

In conclusion, collective NMF with constraints seem to provide a flexible framework to co-cluster data to induce desirable properties in resulting matrix factors.

4. REFERENCES

[1] C. Ding, T. Li, W. Peng, and H. Park. Orthogonal nonnegative matrix t-factorizations for clustering. KDD '06, pages 126–135, New York, NY, USA, 2006.