Topic Modeling via Nonnegative Matrix Factorization on Probability Simplex

Anonymous Author(s)
Affiliation
Address
email

Abstract

One important goal of document modeling is to extract a set of informative topics from a text corpus and produce a reduced representation of each document. In this paper, we propose a novel algorithm for this task based on nonnegative matrix factorization on a probability simplex. We further extend our algorithm by removing global and generic information to produce more diverse and specific topics. In contrast to other matrix factorization methods, such as latent semantic indexing by singular value decomposition, our model has a solid statistical foundation and is based on a generative model for text corpus. In contrast to purely probabilistic approach, such as probabilistic latent semantic indexing (pLSI) solved by Expectation-Maximization, our method is based on efficient block coordinate decent optimization. Experiments demonstrate that the new method generates more meaningful and diverse topics compared with pLSI and LDA with faster convergence behavior.

1 Introduction

There has been increasing interests in topic modeling to analyze a large corpus of documents and distill a set of meaningful topics. Current methods can be roughly divided into two categories, 1) matrix decomposition methods and 2) probabilistic methods. The most prominent examples in the first category is latent semantic indexing (LSI) [4] and nonnegative matrix factorization (NMF) based methods [9, 1]. However, they often ignore probabilistic constraints imposed by the data generating process and lack solid statistical foundations.

The most well-known probabilistic methods are probabilistic latent semantic indexing (pLSI) [6, 7] and latent Dirichlet allocation (LDA) [3]. The parameters can be estimated with variational inference or sampling algorithms [3, 5, 2], which suffers from run time efficiency problem.

In this paper, we propose a fast nonnegative matrix factorization algorithm that combines the best of both worlds. It is efficient and respects the constraints imposed by the probabilistic generative models as in pLSI. In addition, we extend our algorithm to remove global and generic information in a principled way and generate more diverse topics in cases of small number of topics.

2 NMF Formulation with Probability Constraints

Given N documents with a M-sized vocabulary, we create the empirical conditional word distribution matrix $\boldsymbol{A} = \hat{P}(w|d)$, with each of its column summing to one. We approximate the nonnegative matrix $\mathbf{A} \in \mathbb{R}^{M \times N}$ by a product of two lower-rank nonnegative matrices, $\mathbf{W} \in \mathbb{R}^{M \times K}$ and $\mathbf{H} \in \mathbb{R}^{K \times N}$, where K is the number of topics. We interpret $\mathbf{W}$ as topic distributions and $\mathbf{H}$ as per document mixture proportions. Thus each column is required to have unit L_1 norm, making it a probability distribution.

Due to difficulties in optimizing with constraints, we transform the objective function into a regularized form

$$\min_{\boldsymbol{W}\geq 0, \boldsymbol{H}\geq 0} \|\boldsymbol{A} - \boldsymbol{W}\boldsymbol{H}\|_F^2 + \alpha\|\mathbf{1}_M^\top \mathbf{W} - \mathbf{1}_K^\top\|_F^2 + \beta\|\mathbf{1}_K^\top \mathbf{H} - \mathbf{1}_N^\top\|_F^2, \tag{1}$$

where $\mathbf{1}_K$ is an all-ones vector of size K, and α and β are regularization parameters.

Our formulation is closely related to pLSI. To see this, recall that the columns of $\boldsymbol{W}$ are the topic vectors $P(w|z)$, and the columns of $\boldsymbol{H}$ are the document specific mixture proportions $P(z|d)$. The matrix product $\boldsymbol{W}\boldsymbol{H}$ is the conditional distribution of words per document $P(w|d)$. By minimizing the Frobenius norm $\|\boldsymbol{A} - \boldsymbol{W}\boldsymbol{H}\|_F^2$, we obtain factors $\boldsymbol{W}$ and $\boldsymbol{H}$ that approximate the empirical conditional word distribution $\boldsymbol{A}$. In contrast, pLSI maximizes the likelihood of data under the model, which is equivalent to minimizing the Kullback-Leibler (KL) divergence of the empirical joint distribution $\hat{P}(w, d)$ and the model $P(w, d)$.

Compared with pLSI, our method minimizes L_2 distance instead of KL divergence for ease of optimization. Also, we approximate the conditional word distribution instead of the joint distribution. In fact, the extra factor $P(d)$ in pLSI is not useful in discovering topics [3].

One way to solve (1) is to use the block coordinate descent framework [8, 10], alternating between solving for $\mathbf{H}$ (with fixed $\mathbf{W}$) and solving for $\mathbf{W}$ (with fixed $\mathbf{H}$). However, due to difficulty in solving for $\boldsymbol{W}$ we introduce an auxiliary variable $\boldsymbol{Z}$, which serves as a proxy of $\boldsymbol{W}$ and hence decouples the non-negativity and the unit L_1 norm constraints. Finally, we solve the following optimization problem

$$\min_{\mathbf{W}\geq 0, \mathbf{H}\geq 0} \|\mathbf{A} - \mathbf{W}\mathbf{H}\|_F^2 + \gamma\|\mathbf{W} - \mathbf{Z}\|_F^2 + \alpha\|\mathbf{1}_M^\top \mathbf{Z} - \mathbf{1}_K^\top\|_F^2 + \beta\|\mathbf{1}_K^\top \mathbf{H} - \mathbf{1}_N^\top\|_F^2, \tag{2}$$

where γ is a regularization parameter.

3 Three-Block Coordinate Descent Algorithm

The optimization problem (2) can be solved by alternating updates for $\boldsymbol{H}$, $\boldsymbol{W}$ and $\boldsymbol{Z}$ through the following three subproblems:

$$\boldsymbol{H} \leftarrow \operatorname*{argmin}_{\boldsymbol{H}\geq 0} \left\| \begin{pmatrix} \boldsymbol{A} \\ \sqrt{\beta}\mathbf{1}_N^\top \end{pmatrix} - \begin{pmatrix} \boldsymbol{W} \\ \sqrt{\beta}\mathbf{1}_K^\top \end{pmatrix} \boldsymbol{H} \right\|_F^2 \tag{3}$$

$$\boldsymbol{W} \leftarrow \operatorname*{argmin}_{\boldsymbol{W}\geq 0} \left\| \begin{pmatrix} \boldsymbol{A}^\top \\ \sqrt{\gamma}\boldsymbol{Z}^\top \end{pmatrix} - \begin{pmatrix} \boldsymbol{H}^\top \\ \sqrt{\gamma}\boldsymbol{I}_K^\top \end{pmatrix} \boldsymbol{W}^\top \right\|_F^2 \tag{4}$$

$$\boldsymbol{Z} \leftarrow \operatorname{argmin} \left\| \begin{pmatrix} \sqrt{\gamma}\boldsymbol{W} \\ \sqrt{\alpha}\mathbf{1}_K^\top \end{pmatrix} - \begin{pmatrix} \sqrt{\gamma}\boldsymbol{I}_M \\ \sqrt{\alpha}\mathbf{1}_M^\top \end{pmatrix} \boldsymbol{Z} \right\|_F^2 \tag{5}$$

We use ANLS/BPP [8] to solve the subproblems (3) and (4).

The least squares problem (5) has a special structure, and we can apply the Sherman-Morrison formula to obtain a direct solution: $\boldsymbol{Z} \leftarrow \mathbf{W} - \frac{\alpha}{\gamma+M\alpha}\mathbf{1}\mathbf{1}^\top\mathbf{W} + \left(\frac{\alpha}{\gamma} - \frac{M\alpha^2}{\gamma(\gamma+M\alpha)}\right)\mathbf{1}\mathbf{1}^\top$, where the main cost of computing $\mathbf{1}^\top\mathbf{W}$ can be carried out very efficiently.

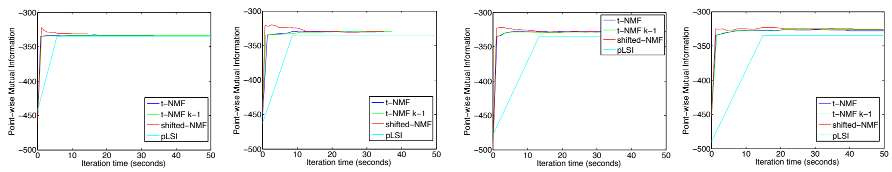
Theoretically, we need to use large regularization parameters so that the the unit L_1 constraints on the columns of $\boldsymbol{W}$ and $\boldsymbol{H}$ are satisfied. However, large regularization parameters put too much emphasis on the constraints during the early stage of iterations. Therefore, we start with reasonably small values and then adaptively increase them as iteration progresses [11]. This approach is reminiscent of simulated annealing where one starts with high temperature parameters and gradually decreases the temperature to zero. The overall algorithm is summarized in Algorithm 1, and we name the algorithm t-NMF for topic NMF.

4 Shifted Non-negative Matrix Factorization

Algorithm 1 can also be extended for hierarchical topic modeling, where at each level, we only find a small number of topics for a partition of the corpus, and recursively partition the documents and

Algorithm 1 tNMF: Nonnegative Matrix Factorization on Probability Simplex

1: **input**: Empirical conditional distribution $\mathbf{A} \in \mathbb{R}_+^{M \times N}$, the number of topcis K, regularization parameter $\alpha > 0, \beta > 0, \gamma > 0$.
2: **output**: Topics $\mathbf{W} \in \mathbb{R}_+^{M \times K}$ and mixing proportions $\mathbf{H} \in \mathbb{R}_+^{K \times N}$.
3: **repeat**
4: **if** Difference in objective from one iteration to the next is below some threshold and (α, β, γ) has not been increased for a certain number of iterations **then**
5: Increase the values for α, β, and γ.
6: **end if**
7: $\boldsymbol{H} \leftarrow \operatorname{argmin}_{\boldsymbol{H} \geq 0} \left\| \begin{pmatrix} \boldsymbol{A} \\ \sqrt{\beta}\mathbf{1}_N^\top \end{pmatrix} - \begin{pmatrix} \boldsymbol{W} \\ \sqrt{\beta}\mathbf{1}_K^\top \end{pmatrix} \boldsymbol{H} \right\|_F^2$
8: $\boldsymbol{W} \leftarrow \operatorname{argmin}_{\boldsymbol{W} \geq 0} \left\| \begin{pmatrix} \boldsymbol{A}^\top \\ \sqrt{\gamma}\boldsymbol{Z}^\top \end{pmatrix} - \begin{pmatrix} \boldsymbol{H}^\top \\ \sqrt{\gamma}\boldsymbol{I}_K^\top \end{pmatrix} \boldsymbol{W}^\top \right\|_F^2$
9: $\mathbf{Z} \leftarrow \mathbf{W} - \frac{\alpha}{\gamma + M\alpha}\mathbf{1}\mathbf{1}^\top \mathbf{W} + \left(\frac{\alpha}{\gamma} - \frac{M\alpha^2}{\gamma(\gamma + M\alpha)} \right) \mathbf{1}\mathbf{1}^\top$
10: **until** stopping criterion is reached



(a) PMI $k = 3$ (b) PMI $k = 5$ (c) PMI $k = 8$ (d) PMI $k = 10$

Figure 1: Convergence speed comparison using Point-wise Mutual Information (PMI).

apply NMF. The problem, however, is that the obtained topics are similar to each other when K is small, often consisting of repeated terms that are global to the corpus. In light of this situation, we propose shifted NMF by explicitly picking out a global topic in hope for more diverse topics.

Specifically, we fix the first topic vector as the average of $\boldsymbol{A}$, *i.e.*, $\boldsymbol{W}_{:,1} = \frac{1}{N}\sum_i \boldsymbol{A}_{:,i}$. The corresponding mixture proportion, *i.e.*, the first row of $\boldsymbol{H}$, is allowed to change so that it optimizes the objective function. The intuition is that by fixing a global topic, the remaining topics are tilted to explain more specific contents in the corpus, thus arriving at more diverse topics. In addition, the proportion of the global topic is not the same for every document, and it is optimized together with other mixture proportions. The algorithm is similar to t-NMF except that $\boldsymbol{W}$ and $\boldsymbol{Z}$ are now both one dimension smaller than their counterparts in t-NMF.

5 Experiments

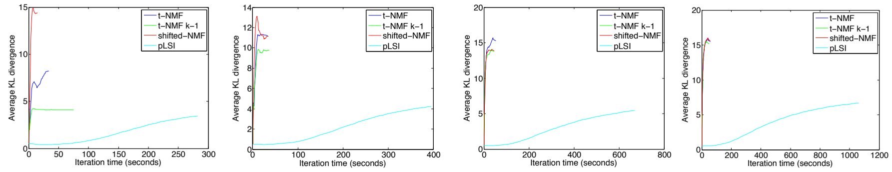
We did experiments on the NIPS dataset, which contains 1739 documents from 2000 to 2012 proceedings, with vocabulary size 13648. We used Point-wise Mutual Information (PMI) [1] to evaluate topic quality: $PMI = \frac{1}{K}\sum_i \sum_{s,t \in \mathcal{T}_i, s<t} \log \frac{D_{st}+\epsilon}{D_s D_t}$, where D_{st} is the number of documents in which keywords s and t co-occur, D_s is the number of documents keyword s occurs, and ϵ is a small constant for smoothing. We also computed Average KL divergence (AKL) to measure the distinctiveness of the topics: $AKL = \frac{2}{K(K-1)}\sum_{i<j}\left(\sum_t W_{ti} \log \frac{W_{ti}}{W_{tj}}\right)$.

We compared with a Matlab implementation of pLSI in terms of convergence speed and present the comparison in Figures 1 and 2. Our NMF-based algorithms converge quickly in terms of both PMI and average KL divergence. The shifted-variant achieves the fastest convergence when k is small.

We summarize the topic quality in Table 1 and list the top 10 keywords from each topic in Table 2. Shifted NMF is able to produce more diverse topics by allowing a global topic to explain the generic content in the corpus. t-NMF and LDA discover topics with more repetition of global keywords such as “data” and “model”. In contrast, pLSI performs poorly by selecting non-descriptive tokens, and it is not robust to noise.

References

[1] S. Arora, R. Ge, and A. Moitra. Learning topic models–going beyond svd. In *Foundations of Computer Science (FOCS), 2012 IEEE 53rd Annual Symposium on*, pages 1–10. IEEE, 2012.



(a) KL div $k = 3$ (b) KL div $k = 5$ (c) KL div $k = 8$ (d) KL div $k = 10$

Figure 2: Convergence speed comparison using average KL divergence.

Table 1: Performance comparison of shifted NMF, tNMF, LDA and pLSI on NIPS dataset. AKL is Average KL divergence between pair-wise topic vectors, and PMI is Point-wise Mutual Information.

Method	K	AKL	PMI	K	AKL	PMI
shifted NMF	3	**15.32**	**-330.99**	8	14.54	-328.82
t-NMF	3	7.99	-332.07	8	**14.61**	**-327.14**
LDA	3	4.26	-332.36	8	5.72	-329.65
pLSI	3	4.01	-333.32	8	5.87	-330.62
shifted NMF	5	11.30	**-329.62**	10	15.18	-328.02
tt-NMF	5	**11.33**	-331.49	10	**16.23**	**-327.78**
LDA	5	4.93	-331.03	10	5.90	-327.85
pLSI	5	4.88	-332.71	10	6.72	-328.56

[2] D. M. Blei and M. I. Jordan. Variational inference for dirichlet process mixtures. *Bayesian analysis*, 2006.

[3] D. M. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *JMLR*, 3:993–1022, 2003.

[4] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *JASIS*, 1990.

[5] T. L. Griffiths and M. Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[6] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57. ACM, 1999.

[7] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Mach. Learn.*, 2001.

[8] J. Kim and H. Park. Fast nonnegative matrix factorization: An active-set-like method and comparisons. *SIAM J. Sci. Comput.*, 33(6):3261–3281, Nov. 2011.

[9] D. Kuang, C. Ding, and H. Park. Symmetric nonnegative matrix factorization for graph clustering. In *SIAM International Conference on Data Mining (SDM '12)*, pages 106–117, 2012.

[10] L. Li, G. Lebanon, and H. Park. Fast bregman divergence nmf using taylor expansion and coordinate descent. In *KDD 2012*, page 307315. ACM, 2012.

[11] J. Nocedal and S. Wright. *Numerical Optimization*. Springer, 1999.

Table 2: Top key words discovered by shifted NMF, tNMF, LDA and pLSI for different K on NIPS dataset.

Method	K	Topics
shifted NMF	3	network,learning,model,neural,input,function,figure,data,time,networks data,model,algorithm,set,function,learning,models,training,distribution,number model,neurons,figure,input,time,cells,neuron,cell,neural,visual
t-NMF	3	model,figure,time,system,data,neurons,models,cells,input,visual learning,function,algorithm,data,set,error,training,state,problem,number network,neural,networks,input,output,units,training,hidden,layer,weights
LDA	3	model,figure,time,neurons,input,neuron,system,neural,visual,cells network,training,neural,input,networks,units,output,set,learning,hidden learning,function,model,data,algorithm,state,set,error,linear,problem
pLSI	3	kwon,unreliable,cart,shades,finite,minimising,awi,stark,proliferation,thalamus narrowing,martin,cropped,rts,englewood,stochasticity,aij,automatically,instructive,rgb closeness,concentric,adds,sooner,pairwise,dispersed,measurable,medicine,ile,sea