Deep Topic Modeling by Multilayer Bootstrap Network and Lasso

Jianyu Wang, Xiao-Lei Zhang

Center for Intelligent Acoustics and Immersive Communications, School of Marine Science and Technology, Northwestern Polytechnical University

Abstract

This paper presents a deep topic model (DTM) for topic mining, which combines a deep dimensionality reduction algorithm, called multilayer bootstrap network and the linear regression together for improving the identifiability of topics. The motivations of this paper are:

- The model assumptions, such as the multinomial distribution, may not be always accurate.
- Traditional unsupervised topic modeling does not consider deep representation of documents.
- The optimization problem is NP-hard in the worst case due to the intractability of the posterior inference.

Introduction

Matrix factorization based topic modeling maps documents into a low-dimensional semantic space by decomposing the documents into a weighted combination of a set of topic distributions: $\mathbf{D} \approx \mathbf{CW}$ where $\mathbf{D}(:,d)$ represents the d-th document which is a column vector over a set of words with a vocabulary size of v, $\mathbf{C}(:,g)$ denotes the g-th topic which is a probability mass function over the vocabulary, and $W(g,d)$ denotes the probability of the g-th topic in the d-th document. The validness of NMF comes from the fact that the matrices $\mathbf{C}$ and $\mathbf{W}$ should be nonnegative. The objective function of NMF is generally as follows:

$$(\mathbf{C},\mathbf{W}) = \arg \min_{\mathbf{C}\geq 0;\mathbf{W}\geq 0} \parallel \mathbf{D} - \mathbf{CW} \parallel_F^2 \qquad (1)$$

An important weakness of this formulation is that there is no guarantee that the solutions of $\mathbf{C}$ and $\mathbf{W}$ are unique. Another problem of NMF is that it is formulated as a shallow learning method, which may not capture the nonlinearity of documents. Motivated by the above problems, this paper proposes a deep topic model (DTM), which learns a deep representation of the documents, i.e. $f(\mathbf{D})$, and the topic-word matrix $\mathbf{C}$ separately, under the assumption that if each of the components is good enough, then the overall performance can be boosted.

Object function

The objective of DTM is defined as

$$\min_{f(\cdot),\mathbf{C}} \frac{1}{2}\|\mathbf{C}f(\mathbf{D}) - \mathbf{D}\|_F^2 + \lambda\Omega(\mathbf{C}) \qquad (2)$$

where $f(\cdot)$ is an unsupervised deep model containing multiple layers of nonlinear transforms, $\Omega(\cdot)$ is a regularizer, and λ is a regularization hyperparameter. We optimize (2) in two steps. The first step learns $f(\mathbf{D})$ by MBN, which outputs the document-topic matrix $\mathbf{W}$. The second step learns the topic-word matrix $\mathbf{C}$ by Lasso, given $\mathbf{W} = f(\mathbf{D})$. The overall DTM algorithm is shown in Fig. 1.

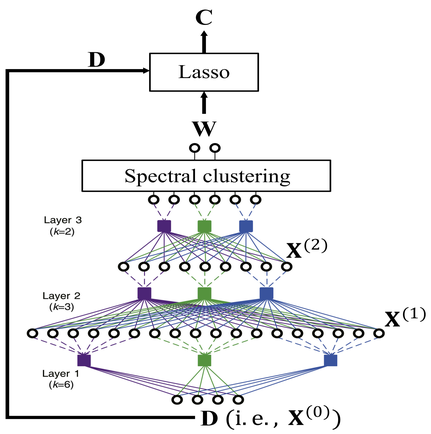

Figure 1: Deep topic model.

Structure of DTM

Multilayer bootstrap network (MBN) is trained layer-by-layer from bottom-up. To train the l-th layer, we simply need to focus on training each k_l-centroids clustering as follows:

- **Random sampling of input**. The first step randomly selects k_l data points from $\mathbf{X}^{(l-1)} = [\mathbf{x}_1^{(l-1)}, \ldots, \mathbf{x}_N^{(l-1)}]$ as the k_l centroids of the clustering, where N is the size of the corpus. If $l = 1$, then $\mathbf{X}^{(l-1)} = \mathbf{D}$.
- **One-nearest-neighbor learning**. The second step assigns any input $\mathbf{x}^{(l-1)}$ to one of the k_l clusters and outputs a k_l-dimensional indicator vector $\mathbf{h} = [h_1, \ldots, h_{k_l}]^T$, which is a one-hot sparse vector.

The output units of all k_l-centroids clusterings are concatenated as the input of their upper layer, i.e. $\mathbf{x}^{(l)} = [\mathbf{h}_1^T, \ldots, \mathbf{h}_V^T]^T$. From the above description, we can see that MBN does not make any model or data assumptions. Substituting the output of MBN, i.e. $\mathbf{W}$, to (2) derives:

$$\min_{\mathbf{C}} \frac{1}{2}\|\mathbf{CW} - \mathbf{D}\|_F^2 + \lambda\Omega(\mathbf{C}) \qquad (3)$$

Experiment results

Table 1: Comparison results on 20-newsgroups.

Metric	Model	T=5	T=10	T=15	T=20	rank
ACC	LDA	0.7013	0.5915	0.5187	0.4900	5.5
	LTM	0.8184	0.7109	0.6412	0.5996	2
	SNPA	0.4078	0.3079	0.2538	0.1744	7
	AnchorFree	0.7595	0.6423	0.5207	0.4485	5.25
	SC	0.8124	0.6960	0.5849	0.4773	3.5
	DTM	**0.8747**	**0.7323**	**0.6471**	**0.6538**	1
	DPFM	0.7730	0.6439	0.5785	0.5328	3.75
Coh.	LDA	**-509.76**	-574.40	-617.87	-759.13	2
	LTM	-893.72	-901.88	-896.35	-855.95	6
	SNPA	-813.89	-843.59	-786.52	-760.83	4.5
	AnchorFree	-565.95	**-572.25**	**-571.92**	**-596.10**	1.5
	SC	-674.59	-762.26	-836.73	-890.21	5.25
	DTM	-653.63	-728.21	-818.98	-862.67	4.25
	DPFM	-1234.22	-959.51	-696.45	-517.01	4.5
SimC.	LDA	22.34	66.38	116.2	196	3.75
	LTM	28	30.62	31.42	**26**	2.75
	SNPA	31.9	157.02	413.48	549	5
	AnchorFree	32.7	195.52	600.14	1235	6.25
	SC	**3.44**	**11.68**	**24.32**	52	1.25
	DTM	3.54	13.48	28.04	86.02	2.25
	DPFM	118.10	296.22	712	890.14	6.75

Table 2: Comparison results on TDT2.

Metric	Model	T=5	T=10	T=15	T=20	rank
ACC	LDA	0.7013	0.6413	0.5941	0.6093	5.75
	LTM	0.9443	0.7705	0.6861	0.6458	3
	SNPA	0.6986	0.5612	0.4694	0.4610	7
	AnchorFree	0.9383	0.7756	0.7420	0.7352	2.25
	SC	0.7943	0.6739	0.6266	0.5819	5.25
	DTM	**0.9778**	**0.9148**	**0.8170**	**0.7842**	1
	DPFM	0.8037	0.7305	0.6849	0.6776	3.75
Coh.	LDA	-509.76	-574.40	-617.87	-642.48	4.5
	LTM	-634.29	-597.61	-579.34	-616.12	4.25
	SNPA	-610.96	-668.08	-660.27	-679.49	6
	AnchorFree	-407.25	-466.23	**-494.75**	**-531.64**	1.5
	SC	-441.52	-517.57	-542.88	-629.02	3.25
	DTM	**-373.89**	**-451.45**	-526.38	-648.51	2.5
	DPFM	-803.90	-715.69	-676.80	-627.00	6
SimC.	LDA	8.02	30.48	65.08	104.82	4
	LTM	24.74	23.34	23.26	20.76	3.5
	SNPA	29.36	74.78	189.44	271.5	6
	AnchorFree	6.18	30.42	84.18	150.04	3.25
	SC	1.06	10	19.02	35.68	2.5
	DTM	**0.3**	**1.98**	**5.6**	**12.32**	1
	DPFM	112.22	287.76	690.20	1056.20	7

Topics discovered by AnchorFree and DTM

AnchorFree

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
netanyahu	**asian**	bowl	tornadoes	**economic**
israeli	**asia**	super	florida	indonesia
israel	**economic**	broncos	central	**asian**
palestinian	**financial**	denver	storms	**financial**
peace	**percent**	packers	ripped	imf
arafat	**economy**	bay	victims	**economy**
palestinians	market	green	tornado	**crisis**
albright	stock	football	homes	**asia**
benjamin	**crisis**	game	killed	monetary
west	markets	san	people	currency

DTM

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
netanyahu	asian	bowl	florida	nigeria
israeli	percent	super	tornadoes	abacha
israel	indonesia	broncos	tornado	military
palestinian	asia	denver	storms	police
peace	economy	packers	killed	nigerian
albright	financial	green	victims	opposition
arafat	market	game	damage	nigerias
palestinians	stock	bay	homes	anti
talks	economic	football	ripped	elections
west	billion	elway	nino	arrested

Summary

In this paper, we have proposed a deep topic model based on MBN and Lasso. The novelty of DTM lies in the following three respects. First, we extended the linear matrix factorization problem to its nonlinear case. Second, we estimated the topic-document matrix and word-topic matrix separately by MBN and Lasso independently, which simplifies the optimization problem of (2). At last, we applied MBN and Lasso to the unsupervised topic modeling for the first time. Particularly, MBN, as an unsupervised deep model, overcomes the weaknesses of the model assumptions, anchor word assumption, and shallow learning, which accounts for the advantage of DTM over the 5 representative comparison methods. Experimental results on 20-newsgroups, TDT2 and Reuters-21578 have demonstrated the effectiveness of the proposed method.