25th International Conference on Pattern Recognition (ICPR 2020)

Deep Topic Modeling by Multilayer Bootstrap Network and Lasso

Jianyu Wang and Xiao-Lei Zhang

Center of Intelligent Acoustics and Immersive Communications
Northwestern Polytechnical University

alexwang96@mail.nwpu.edu.cn

xiaolei.zhang@nwpu.edu.cn

Contents

1. Motivation

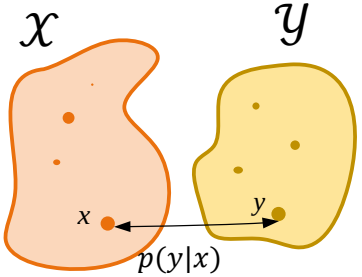
2. Methods

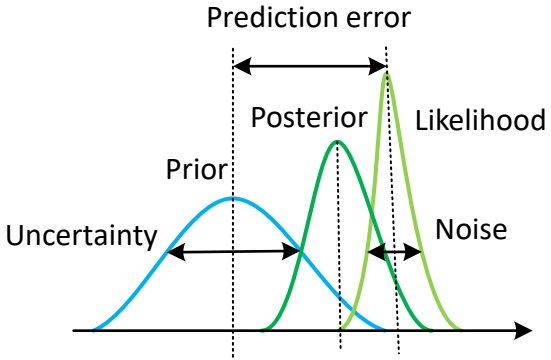
3. Experiments

4. Conclusions

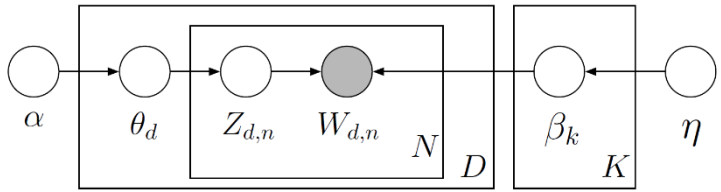
Motivation

Model & Data assumptions





Shallow model



Latent Dirichlet analysis

Difficult Optimization

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})}$$

Variational inference

MCMC

$$p(\mathcal{D}) = \int p(\mathcal{D}, \theta)\, d\theta$$

Methods

Model: word-document matrix $\boldsymbol{D = CW}$.

$\boldsymbol{C}$: whose columns are topics.

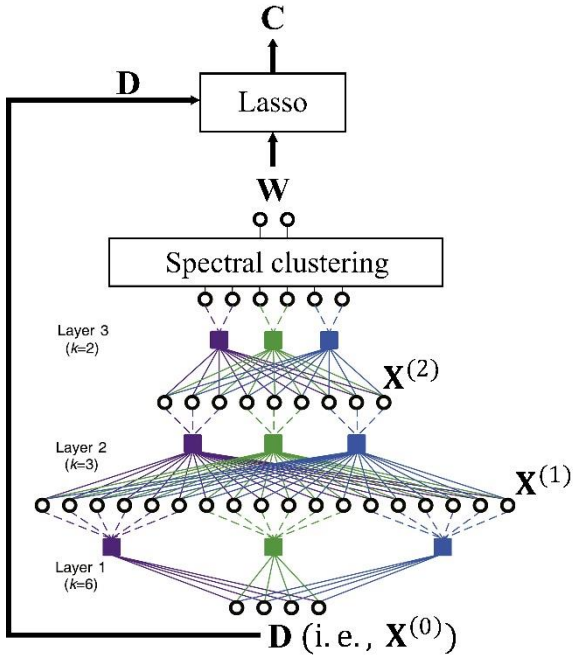
$\boldsymbol{W}$: weights of the topics in the documents.

Deep topic modeling:

Avoid inaccurate model assumptions.

Capture the deep latent representations of documents.

No parameter tuning.



Experiments

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+Lasso	-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+Lasso	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 discovery

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

Conclusions

1. Extending the linear matrix factorization problem to its nonlinear case.
2. Estimating the topic-document matrix and word-topic matrix separately by MBN and Lasso independently.
3. Achieving the state-of-the-art performance.

THE END!

THANK YOU FOR YOUR WATCHING!