Topic-weak-correlated Latent Dirichlet Allocation

Yimin TAN, Zhijian OU

ozj@tsinghua.edu.cn

Department of Electronic Engineering, Tsinghua University, Beijing

Propose: TWC-LDA for topic modeling, which constrains different topics to be weak-correlated. This is technically achieved by placing a special prior over the topic-word distributions.

Superiority: in semantically meaningful topic discovery and document classification.

Motivation to propose TWC-LDA

In the basic LDA, both priors are assumed to be dirichlet.

The prior over the topic proportion, $p(\theta_d)$	The prior over the topic-word distribution, $p(\beta)$

------ exploring new priors ------

Use the logistic normal prior [2][3] or the Dirichlet tree prior [4] to develop correlated topic models.

few works

the main issue addressed in this paper

Why we care about the priors over the topic-word distributions, $p(\beta)$

- Not merely for smoothing in estimating the topic-word probabilities.
  Have practical effects, e.g. [5] using nested CRP, [6] using Gaussian Markov random fields.
- The *topic* term in the LDA is more a metaphor.
  Topics are expected to be distinct in order to convey information.
- Reduce the overlapping between the topic-word distributions.

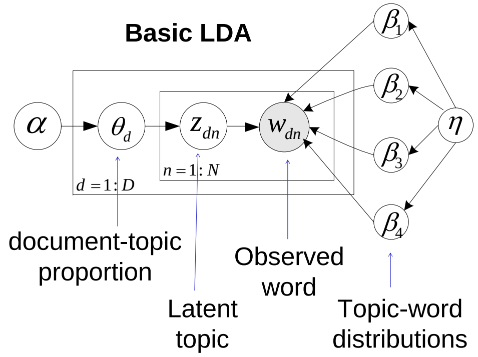
[2] Blei, Lafferty. A correlated topic model of Science. Annals of Applied Statistics, 2007.
[3] Mimno, et al. Gibbs Sampling for Logistic Normal Topic Models with Graph-Based Priors. NIPS 2008.
[4] Tam, Schultz. Correlated latent semantic model for unsupervised LM adaptation. ICASSP 2007.
[5] Blei, et al. Hierarchical topic models and the nested Chinese restaurant process. NIPS 2003.
[6] Wang, Thiesson, Meek, Blei. Markov topic models. AISTATS 2009.
[7] Wallach, et al. Rethinking lda: Why priors matter. NIPS 2009.

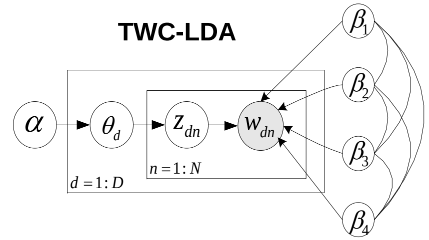
Compare TWC-LDA with some related LDA researches

Correlated topic model [2]	TWC-LDA
aims at capturing the correlation between the occurrences of latent topics	focuses on incorporating the weak correlation between the topics themselves

LDA using asymmetric dirichlet prior over document-topic distributions [7] employs computational-intensive Gibbs sampling.	The seeming consequence of [7] and TWC-LDA is similar - being robustness to stop-words, their modeling motivation are different. uses efficient variational inference.

Topic-weak-correlated LDA (TWC-LDA)





TWC-LDA: placing a special prior over β

$$p(\beta) = \frac{1}{Z}\exp\left\{-\rho\sum_{m\neq n}\beta_m\beta_n^T\right\}$$

This prior incorporates the interaction of different topics and forces them to have weak correlations.

Variational Inference

Basic Idea: minimize the Kullback-Leibler distance KL($q\|p$)

$$p(\theta, z, \beta \mid d) \approx q(\theta \mid \gamma_d)\, q(z_{1:N} \mid \phi_{d,1:N})\, q(\beta)$$

Experiment Results

(1) Synthetic dataset

- 400 words equally divided into 4 topics
- hyperparameter α_1=5, α_2=α_3=α_4=0.5
- 6000 documents (30 words per document)

Four topics by LDA				Four topics by TWC-LDA			
5	61	78	10	2	342	261	184
40	7	79	17	99	385	284	175
78	2	61	95	78	368	297	155
23	82	83	47	43	361	202	117
98	98	82	26	95	390	247	187
99	11	236	67	47	313	213	178
119	46	37	99	44	321	286	112
37	19	64	344	10	380	209	163
12	79	8	83	11	302	295	185
70	95	20	59	46	354	208	103

(2) TREC AP corpus (stop-words removed)

Basic LDA

topic 1	topic 2	topic 3	topic 4
i	court	soviet	government
years	case	gorbachev	president
new	attorney	new	people
first	trial	i	national
two	judge	air	new
like	charge	people	communist
just	prison	two	congress
people	sentence	africa	years
last	federal	flight	last

TWC-LDA

topic 1	topic 2	topic 3	topic 4
i	court	soviet	bill
new	case	united	senate
years	drug	government	committee
people	judge	military	budget
two	attorney	states	congress
state	trial	president	tax
last	charges	war	rep
time	prison	foreign	sen
first	investigation	official	house

(3) Year 1994 China daily corpus (raw)

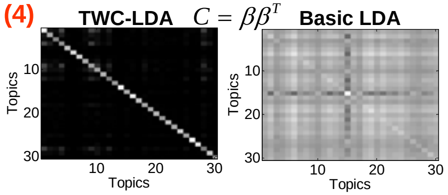
Basic LDA

topic 1	topic 2	topic 3	topic 4
的	的	的	的
是	人	体育	是
了	到	了	在
产品	和	和	和
在	有	比赛	艺术
和	他	训练	观众
企业	来	有	音乐
市场	是	到	了

TWC-LDA

topic 1	topic 2	topic 3	topic 4
犯罪	文化	的	十
机关	出版	在	二
案件	历史	和	三
治安	读者	上	八
公安	时代	中	百
打击	传统	有	九
法院	读者	对	七
法律	书	为	千

(4)

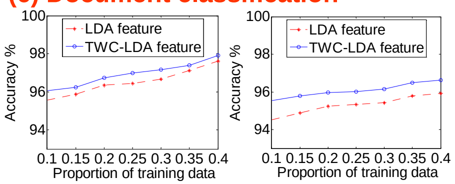


(5)

$$W = \sum_{m\neq n}\beta_m\beta_n^T$$

Corpus	*W* of LDA	*W* of TWC-LDA
TREC-AP	0.0416	0.0078
China Daily	3.2922	0.0113

(6) Document classification



Reuters-21578 dataset - "EARN", "GRAIN"