Inference and Representation: Latent Dirichlet Allocation

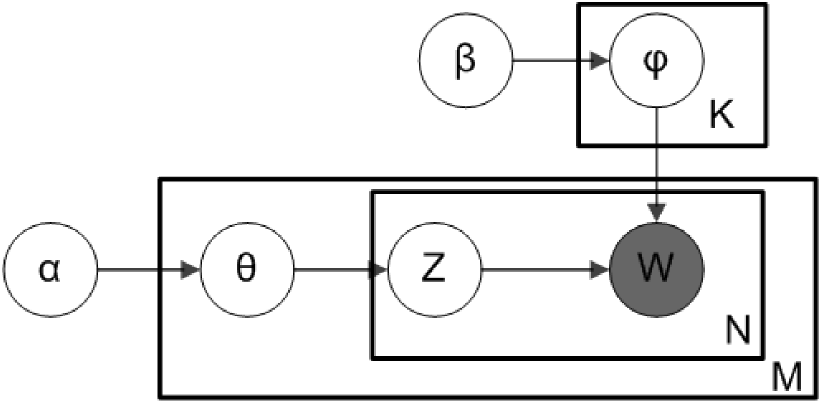
Rahul G. Krishnan

New York University

Lab 5, October 4, 2016

Outline

LDA (Blei et. al)



Generative Model

- Given α, β as parameters for a Dirichlet distribution
- For each topic k, $\beta_k \sim \text{Dir}(\alpha)$ where $k \in \{1, \dots, K\}$
- β_k is a vector that sums to 1 representing the word probabilities for topic k
- For each document d, $\theta_d \sim \text{Dir}(\alpha)$ where $d \in \{1, \dots, M\}$
- θ_k is a vector that sums to 1 representing the topic proportions for document d
- For every word n in document d
    - $z_{n,d} \sim \text{Mult}(\theta_i)$ is a categorical random variable (with cardinality K) whose assignment is the topic for the current word
    - $w_{n,d} \sim \text{Mult}(\beta_{z_{n,d}})$ is a categorical random variable with cardinality $|V|$ (vocabulary size)

Choosing K

- No one right choice.
- Different choices lead to different results
- Choice of K also interacts with the choice of α and encapsulates prior knowledge
- eg. small K and $\alpha < 1$ means you believe that there exist few disjoint topics within your corpora

Author-topic model (Rosen-Zvi et al., UAI '04)

- Goal: topic models that take into consideration author *interests*
- Training data: corpora with label for who wrote each document
  - Papers from NIPS conference from 1987 to 1999
  - Twitter posts from US politicians
- Why do this?
- How to do this?

Author-topic model (Rosen-Zvi et al., UAI '04)

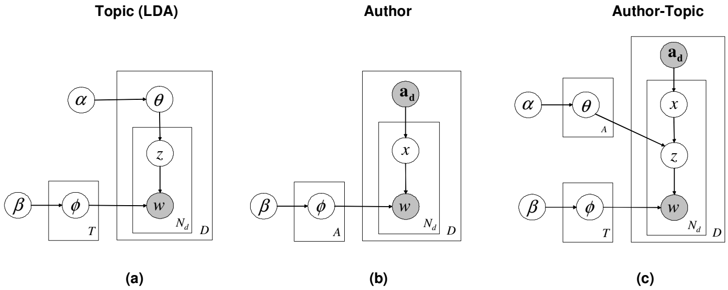


Figure 1: Generative models for documents. (a) Latent Dirichlet Allocation (LDA; Blei et al., 2003), a topic model. (b) An author model. (c) The author-topic model.

Most likely author for a topic

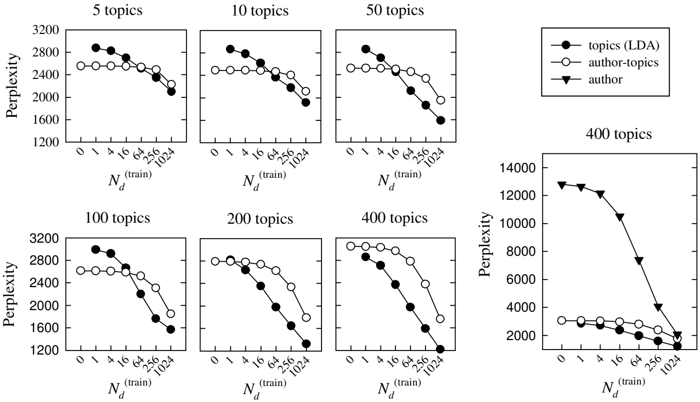
TOPIC 31	
WORD	**PROB.**
SPEECH	0.0823
RECOGNITION	0.0497
HMM	0.0234
SPEAKER	0.0226
CONTEXT	0.0224
WORD	0.0166
SYSTEM	0.0151
ACOUSTIC	0.0134
PHONEME	0.0131
CONTINUOUS	0.0129
AUTHOR	**PROB.**
Waibel_A	0.0936
Makhoul_J	0.0238
De-Mori_R	0.0225
Bourlard_H	0.0216
Cole_R	0.0200
Rigoll_G	0.0191
Hochberg_M	0.0176
Franco_H	0.0163
Abrash_V	0.0157
Movellan_J	0.0149

TOPIC 61	
WORD	**PROB.**
BAYESIAN	0.0450
GAUSSIAN	0.0364
POSTERIOR	0.0355
PRIOR	0.0345
DISTRIBUTION	0.0259
PARAMETERS	0.0199
EVIDENCE	0.0127
SAMPLING	0.0117
COVARIANCE	0.0117
LOG	0.0112
AUTHOR	**PROB.**
Bishop_C	0.0563
Williams_C	0.0497
Barber_D	0.0368
MacKay_D	0.0323
Tipping_M	0.0216
Rasmussen_C	0.0215
Opper_M	0.0204
Attias_H	0.0155
Sollich_P	0.0143
Schottky_B	0.0128

TOPIC 71	
WORD	**PROB.**
MODEL	0.4963
MODELS	0.1445
MODELING	0.0218
PARAMETERS	0.0205
BASED	0.0116
PROPOSED	0.0103
OBSERVED	0.0100
SIMILAR	0.0083
ACCOUNT	0.0069
PARAMETER	0.0068
AUTHOR	**PROB.**
Omohundro_S	0.0088
Zemel_R	0.0084
Ghahramani_Z	0.0076
Jordan_M	0.0075
Sejnowski_T	0.0071
Atkeson_C	0.0070
Bower_J	0.0066
Bengio_Y	0.0062
Revow_M	0.0059
Williams_C	0.0054

TOPIC 100	
WORD	**PROB.**
HINTON	0.0329
VISIBLE	0.0124
PROCEDURE	0.0120
DAYAN	0.0114
UNIVERSITY	0.0114
SINGLE	0.0111
GENERATIVE	0.0109
COST	0.0106
WEIGHTS	0.0105
PARAMETERS	0.0096
AUTHOR	**PROB.**
Hinton_G	0.2202
Zemel_R	0.0545
Dayan_P	0.0340
Becker_S	0.0266
Jordan_M	0.0190
Mozer_M	0.0150
Williams_C	0.0099
de-Sa_V	0.0087
Schraudolph_N	0.0078
Schmidhuber_J	0.0056

Perplexity as a function of number of observed words



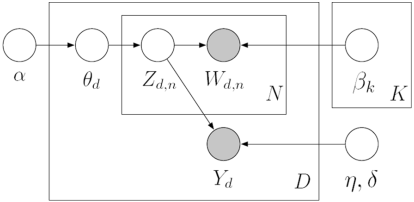
$$\text{perplexity}(\mathbf{w}_{test,d} \mid \mathbf{a}_d) = \exp\left[-\frac{\sum_{d=1}^{M} \ln p(\mathbf{w}_{test,d}|\mathbf{a}_d)}{\sum_{d=1}^{M} N_{test,d}}\right]$$

Adding supervision to LDA

- What if, in addition to words, you had labels for a document?
- Possible labels:
  - Sentiment: Is the document generally positive or negative?
  - Content: Dollar value of the item that the document describes.
- Your topics might be useful as *latent representations* for the words in the document.

Supervised Topic Models

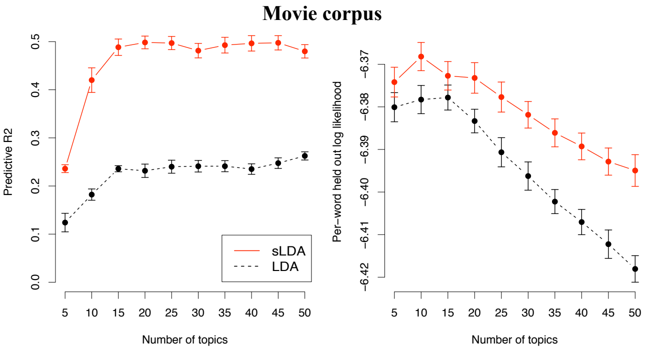
- Supervised LDA:



- The inferred θ or **z** can be used as features in many prediction tasks.
- Performance can be improved by jointly training the representation and the predictor.

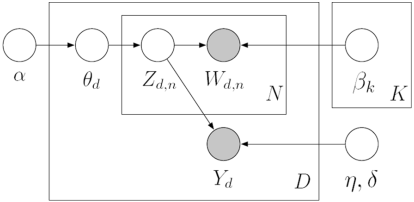
Evaluation

- Supervised LDA vs LDA (where a separate classifier is trained on the documents' topics)
- Dataset: Predicting movie ratings from reviews



Design Question

- Bayesian Network Design Question: Why not condition Y_d on θ_d rather than $Z_{d,n}$?



Group Excercise

- Grab a worksheet!
- Form groups of $3-4$ with people sitting around you
- Write all your names on the top left corner!
- Read the instructions
- Please write legibly
- You will have 20 minutes