Latent Dirichlet Allocation

Paper Presentation

CSC 665 – Advanced Topics in Probabilistic Graphical Models

Marium Yousuf

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

Outline

- Background
- Latent Dirichlet Allocation
  - Notation and Terminology
  - General Model
- Variational Inference
- Application

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

Background

- Probabilistic Text Modeling
- Term Frequency-Inverse Document Frequency (TF-IDF)
  - “how important a word is in a document”
  - the **number of occurrences** of a particular word in a document set
  - Dimensionality (term-by-document matrix)
- Latent Semantic Indexing (LSI)
  - Significant compression
- Probabilistic LSI
  - Mixture model
- LSI and pLSI: dimensionality reduction methods

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

Background

- “bag-of-words”
  - **Exchangeability**
  - Order of words in a document is neglected
  - Same for documents
- De Finetti’s Theorem:
  - For any infinitely exchangeable sequence of RVs there exists some RV θ with density p(θ) s.t. the joint probability of any N observations has a mixture representation: $p(y_1, y_2, \ldots, y_N) = \int p(\theta) \prod_{i=1}^{N} p(y_i \mid \theta)\, d\theta$
  - Mixture models that capture the collection of exchangeable words and documents
- Latent Dirichlet Allocation

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

Latent Dirichlet Allocation (LDA)

- Three-level hierarchical Bayesian model
- Preserves statistical structure of each document from a corpus
- Each item in a collection of documents is modeled as a finite mixture over an underlying set of topics
- Note:
  - model for collections of discrete data
  - text modeling is just one instance

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

Notation and Terminology

- Word
  - Basic unit item from a vocabulary indexed {1, …, V}
  - Represented as unit-basis vectors:
    - the vth word represented by a V-vector w such that:

      $w^v = 1$ and $w^u = 0$ for $u \neq v$.
- Document
  - Sequence of N words: $\boldsymbol{w} = (w_1, w_2, \ldots, w_N)$
- Corpus
  - Collection of M documents: $D = \{\boldsymbol{w_1}, \boldsymbol{w_2}, \ldots, \boldsymbol{w_M}\}$

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

Generative Probabilistic Model

- LDA assumes the following for *each* document $\boldsymbol{w}$ in a corpus D:
  - $\theta \sim Dir(\alpha)$
  - For each of N words w_n:
    - Choose a topic $z_n \sim Multinomial(\theta)$
    - Choose a word w_n from $p(w_n | z_n, \beta)$, a multinomial probability conditioned on z_n.
- Assumptions:
  - Dimensionality of the Dirichlet distribution, k, is known and fixed.
  - β is a $k \times V$ matrix representing word probabilities:
    - $\beta_{ij} = p(w^j = 1, z^i = 1)$: probability that word j belongs to the i^{th} topic

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

Generative Probabilistic Model

- A Dirichlet R.V. θ
  - takes values in a $(k-1)$-simplex s.t. all points in the simplex $\theta_i \in [0,1]$ and $\sum_{i=1}^{k} \theta_i = 1$
  - has the probability density:

$$p(\theta|\alpha) = \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \cdots \theta_k^{\alpha_k - 1}, \qquad (1)$$

    - α is a k-vector with components $\alpha_i > 0$

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

Generative Probabilistic Model

- Given the parameters α and β, the joint probability distribution of a topic mixture θ, a set of *N* topics $\boldsymbol{z}$, and a set of *N* words $\boldsymbol{w}$ is given by:

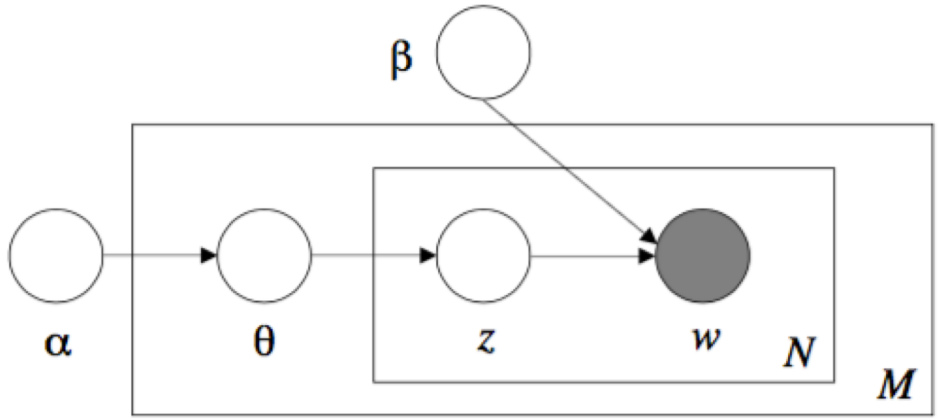
$$p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta) = p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta), \tag{2}$$

  - $p(Z_n \mid \theta)$ is θ_i for the unique i s.t. $z_n^i = 1$.
- Marginal Distribution:

$$p(\mathbf{w} \mid \alpha, \beta) = \int p(\theta \mid \alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n \mid \theta) p(w_n \mid z_n, \beta) \right) d\theta. \tag{3}$$

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

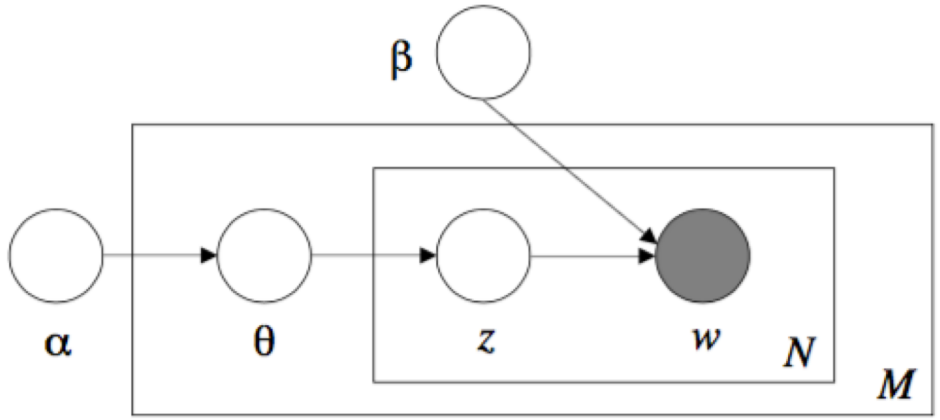
Graphical Model Representation



- Three-Levels:
  - α and β are the corpus level parameters
  - θ is the document-level variable
  - z and w are word-level variables

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

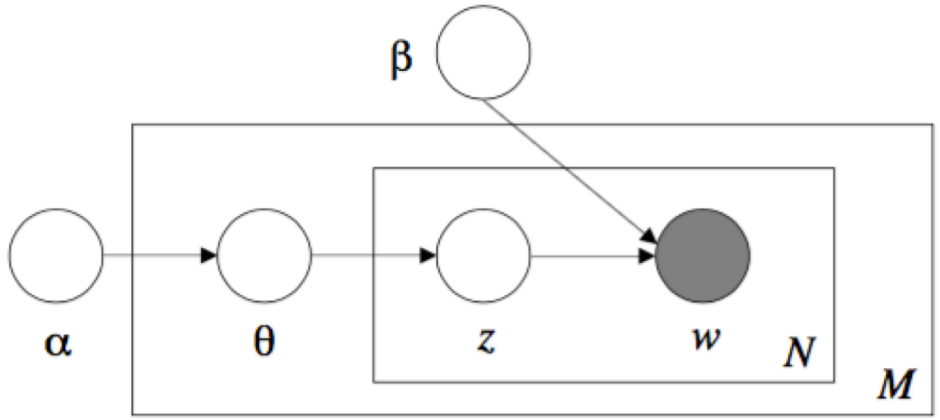
Graphical Model Representation



$$p(\mathbf{w}|\alpha,\beta) = \int p(\theta|\alpha)\left(\prod_{n=1}^{N}\sum_{z_n} p(z_n|\theta)p(w_n|z_n,\beta)\right)d\theta. \quad (3)$$

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

Graphical Model Representation



$$p(\mathbf{w}|\alpha,\beta) = \int p(\theta|\alpha)\left(\prod_{n=1}^{N}\sum_{z_n} p(z_n|\theta)p(w_n|z_n,\beta)\right)d\theta. \quad (3)$$

- Probability of a corpus: $$p(D|\alpha,\beta) = \prod_{d=1}^{M}\int p(\theta_d|\alpha)\left(\prod_{n=1}^{N_d}\sum_{z_{dn}} p(z_{dn}|\theta_d)p(w_{dn}|z_{dn},\beta)\right)d\theta_d.$$

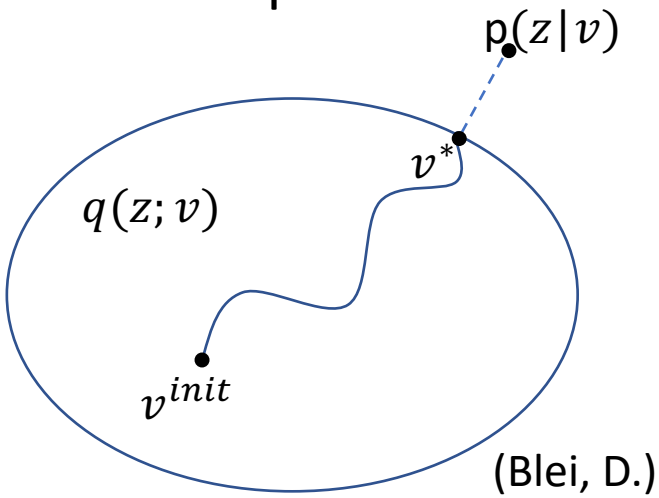
Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

Variational Inference

- To use LDA for text corpora model:
  - Compute posterior probability of latent variables given a document

$$p(\theta, \mathbf{z} | \mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, \mathbf{w} | \alpha, \beta)}{p(\mathbf{w} | \alpha, \beta)}.$$

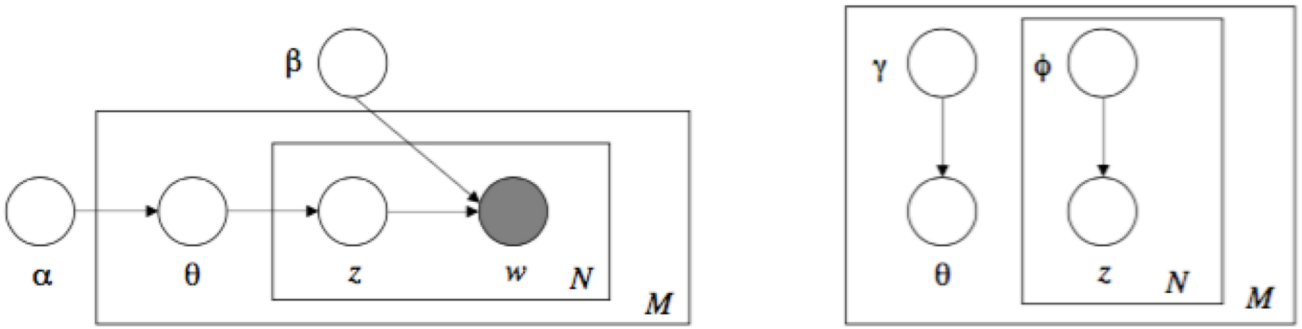
  - $p(\boldsymbol{w} | \alpha, \beta)$, marginal distribution of a document, is intractable to compute
- Variational Inference
  - Approximate inference technique
    - Turns inference into optimization



(Blei, D.)

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

Variational Inference



- Graphical model representation of the variational distribution used to approximate the posterior in LDA ($p(\theta, \mathbf{z} | \mathbf{w}, \alpha, \beta)$)

$$q(\theta, \mathbf{z} | \gamma, \phi) = q(\theta | \gamma) \prod_{n=1}^{N} q(z_n | \phi_n), \qquad (4)$$

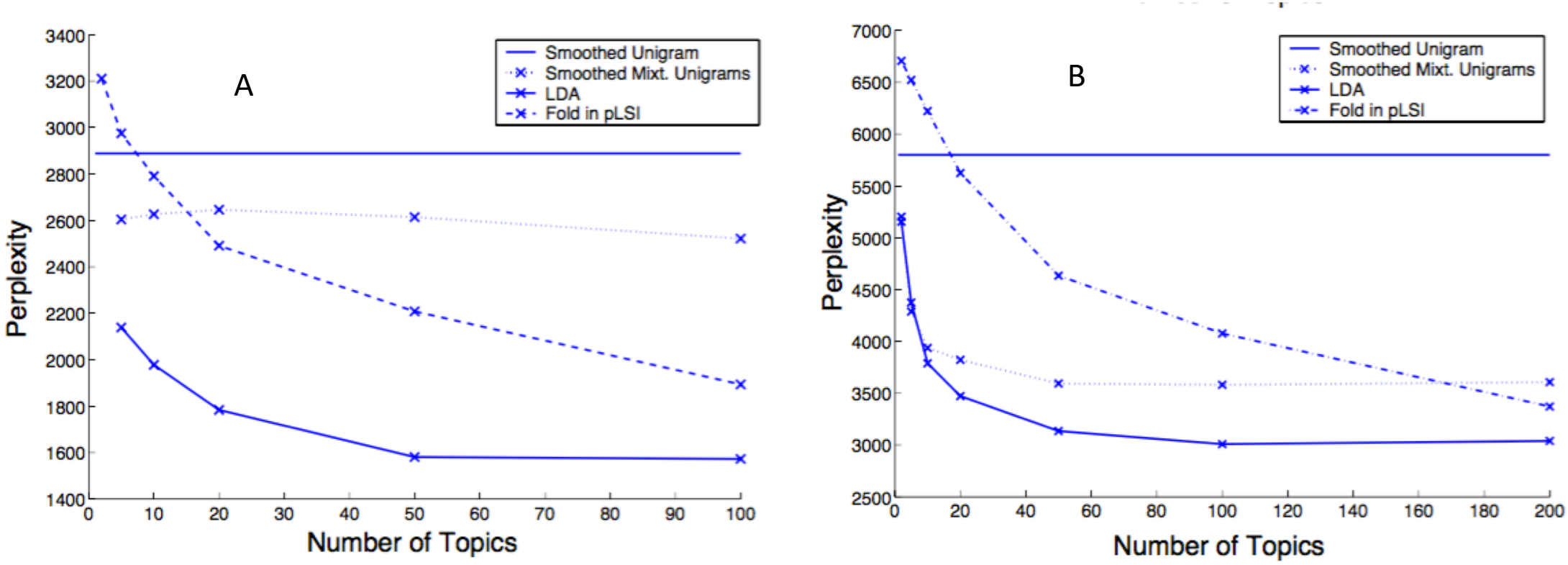
Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

Document Modeling - Application

- Comparison of generalization performance by pLSI, LDA (as well as Unigram and Mixture Unigram)
- Two Corpora:
  - A: 5,225 abstracts with 28,414 unique terms
  - B: 16,333 newswire articles with 23,075 unique terms
- 90% of the dataset used for model training and 10% for testing
- Generalization performance measured by *perplexity*
  - monotonically decreasing in the likelihood of the test data
  - Goal: to achieve high likelihood (lower perplexity score)

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

Document Modeling - Application



Compared to other latent variable models, LDA performs significantly better.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.