Latent Dirichlet Analysis for Document Topic Modelling

Xiang Li
xngli@umich.edu

Zheng Luo
zlou@umich.edu

Yan Chen
yanchenm@umich.edu

Sajan Patel
sajanptl@umich.edu

Abstract

Abstract goes here. Talk about how LDA can be used for topic modeling. Our method and implementation is built on the Blei et. al. JMLR paper. We recreated the experiment for document modeling from the paper. We also applied LDA to the Yelp Dataset. We present our implementation of the method, experiment, and our analysis of the results in this paper.

1 Introduction

Here, describe the problem statement: Given a text document, model the topics of the document. (expand on that more).

1.1 Related Works

Based on Blei, briefly talk about unigram, mixture of unigram, and plsi.

2 Notation

We used similar notation as those denoted in the paper:
N = number of words in total.
z_j = the j-th topic
d_i = the i-th document
w_i = the i-th word in the document

3 LDA

The latent Dirichlet allocation (LDA) model explains the generation of text documents, which can be viewed as samples from a mixture of multinomial distributions over a vocabulary of words. Each multinomial mixture component is called a topic. The general process of the model to write a document is the following:

- The number of words N in document $\sim$ Poisson (ζ)
- The topic mixture θ for the document $\sim$ Dir (α) (with a fixed set of K topics)
- Then for each word w_i in the document:
  1. Choose a topic t_i based on multinomial distribution with parameter θ from step (2)

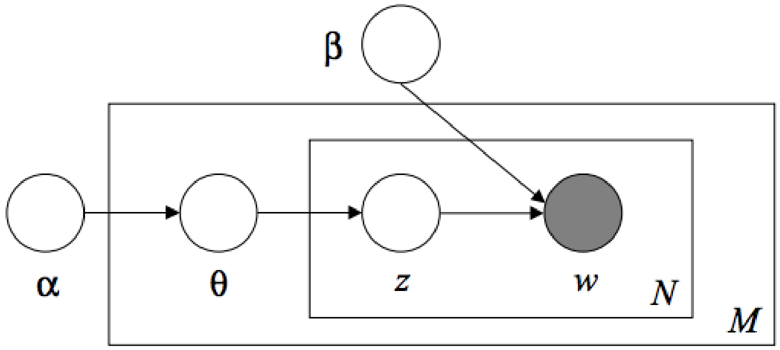

Figure 1: Graphic model of LDA. The boxes are "plates" representing replicates. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document.

2. Use this topic z_i to generate word itself by using the existing probability for each word in this topic (i.e. $p(w_i|z_i, \beta)$

This is the generative model for a collection of documents. LDA then tries to backtrack from the (training) documents to find a set of topics that are likely to have generated the collection. Now the question is how does LDA backtrack to find the parameters in this model. Suppose we have a set of documents D, and we set the number of topics to be K. What we want is to use LDA to learn two things 1) the topic representation of each document and 2) the words associated to each topic.

There are two methods to learn these two things: collapsed Gibbs sampling [1] and variational inference [2, 3]. We first discuss Gibbs sampling method here:

3.1 Gibbs Sampling Method

- For each document, randomly assign each word in the document to one of the K topics.
  1. this step gives topic representation of all the documents
  2. this step gives word distributions of all the topics
  3. since randomly assign topics to each word is very native, so we need to improve it
- For each word w_k in document d_i
  1. For each topic t_j that this word belongs to, compute:
     (a) $p(z_j|d_i) = \frac{numberofwordsassignedtoz_jind_i}{totalnumberofwordsind_i}$
     (b) $p(w_k|z_j) = \frac{numberofwordsassignedtoz_jind_i}{numberofwordsassignedtoz_jforalldocs}$
- we compute the product of i) and ii) above which gives the new topics to assign to this word.
- repeating step 2 over and over until it reaches a steady state where the assignments make good sense.
- Use this model to estimate the topic mixtures of each document and words associated to each topic, which are the two things we want to learn.

3.2 Inference Method

Show derivation and algorithm

Algorithm 1 Estimation

procedure
Initializatize: $\phi_{ni}^{0} = 1/K$ for all i and n
Initializatize: $\gamma_i^0 = \alpha_i + N/K forall i$
repeat
    for n=1 to N
    $i \leftarrow$ *patlen*
top:
    if $i >$ *stringlen* **then return** false
    $j \leftarrow$ *patlen*
loop:
    if *string*$(i) =$ *path*(j) **then**
        $j \leftarrow j - 1.$
        $i \leftarrow i - 1.$
        goto *loop*.
        close;
    $i \leftarrow i + \max(\mathit{delta}_1(\mathit{string}(i)), \mathit{delta}_2(j)).$
    goto *top*.

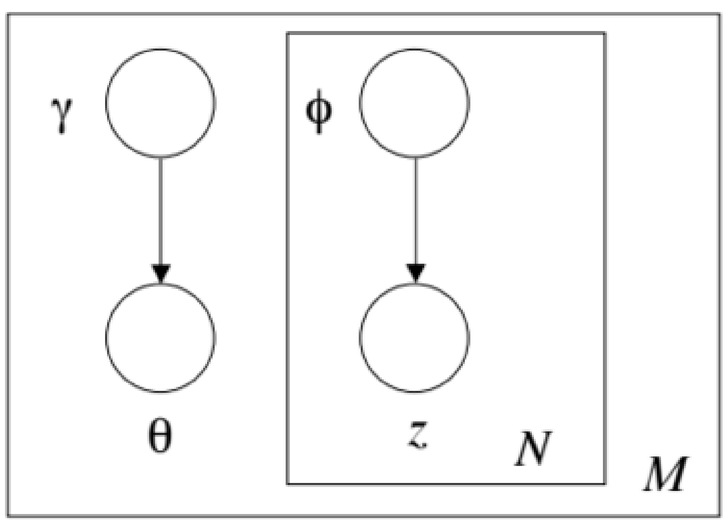

Figure 2: Graphical model representation of the variational distribution used to approximate the posterior in LDA.

4 Our Implementation

5 Experimental Evaluation

5.1 Setup

Describe recreation of Document Modeling and Yelp experiment from Blei et. al. here.

5.2 Metrics

Describe perplexity here.

5.3 Results

Show graphs of the perplexity and histogram of topics.

Discuss results in table if also needed.

Also compare gibbs vs inference in each.

6 Conclusion

Discuss the subsequent conclusions we gained from this reimplementation of LDA. Summarize advantages and disadvantages as well.

Acknowledgements

Describe efforts and work division here.

References

[1]. Griffiths, Tom, and Mark Steyvers. ”A probabilistic approach to semantic representation.” Proceedings of the 24th annual conference of the cognitive science society. 2002.

[2]. Jordan, Michael I., et al. ”An introduction to variational methods for graphical models.” Machine learning 37.2 (1999): 183-233.

[3]. Teh, Yee Whye, et al. ”Hierarchical dirichlet processes.” Journal of the american statistical association 101.476 (2006).

[4]. Blei, David M., Andrew Y. Ng, and Michael I. Jordan. ”Latent dirichlet allocation.” the Journal of machine Learning research 3 (2003): 993-1022.