Latent Dirichlet Analysis for Document Topic Modelling

Xiang Li
xngli@umich.edu

Zheng Luo
zluo@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)

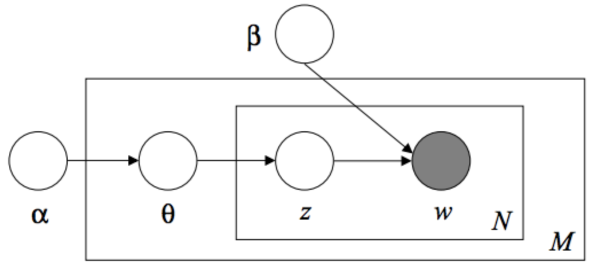

Figure 1: Graphic model of LDA. The boxes are "plates" representing replicates. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document.

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) $p(z_j|d_i) = \frac{number of words assigned to z_j in d_i}{total number of words in d_i}$
     (b) $p(w_k|z_j) = \frac{number of words assigned to z_j in d_i}{number of words assigned to z_j for all docs}$
- 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

4 Our Implementation

5 Evaluations and Empirical Results

5.1 Dataset

To evaluate our algorithm, we replicated one of the experiments implemented by Blei et al. on document modeling. We used AP dataset which contains 2346 (verify this number) news articles

Algorithm 1 Estimation

procedure
Initializatize: $\phi_{ni}^0 = 1/K$ for all i and n
Initializatize: $\gamma_i^0 = \alpha_i + N/K$ for all i
repeat
    for $n = 1$ to N **do**
        for $i = 1$ to K **do**
            $\phi_{ni}^{t+1} := \beta_{iw_n} \exp\{\varphi(\gamma_i^t) - \varphi(\sum_{j=1}^K \gamma_j^t)\}$
        end
    end
    $\gamma^{t+1} := \alpha + \sum_{n=1}^N \phi_n^{t+1}$
    $t \leftarrow t + 1$
until convergence

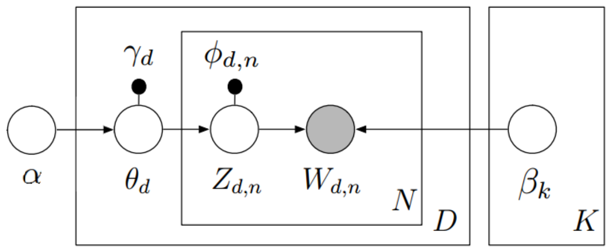

Figure 2: Graphical model representation of the variational distribution used to approximate the posterior in LDA.

from the associated press [5]. The dataset was partitioned in to 80 percent training set and 20 percent test data.

5.2 Metrics and Results

In AP dataset, each document is unlabeled. Therefore, here we are doing unsupervised learning with a purpose to estimate the likelihood of the test dataset. In natural language processing, the likelihood of a document is usually called perplexity, which is the inverse of the average per-word log likelihood. The perplexity of on the text corpus is defined by

Describe perplexity here (need help with Yan).

While implementing our algorithms, we first monitors the convergence of the algorithm. This was done by monitoring the perplexity (defined below) of training data, as shown Figure 3. We can see that in both cases the algorithm converges after enough number of iterations, although the numbers of iterations required to reach convergence are very different.

6 Conclusion

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

Author Contributions

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.

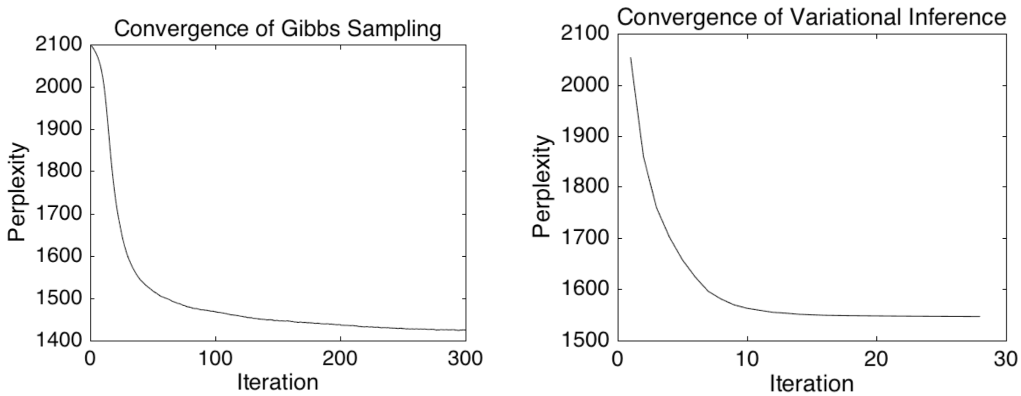

Figure 3: Convergence of LDA using Gibbs Sampling (*left*) and Variational Inference (*right*)

“Law”	**“War”**	**“Trade”**	**“Politics”**	**“Company”**
court	united	percent	government	million
state	states	market	soviet	year
federal	military	prices	party	billion
case	war	stock	union	company
department	president	dollar	south	new
law	american	trade	gorbachev	workers
attorney	iraq	year	political	based
judge	officials	late	west	last
office	aid	oil	country	corp
former	israel	higher	president	co

Figure 4: List of topics generated by LDA.

[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.

[5] `http://www.cs.columbia.edu/~blei/lda-c/`. It should be noted that the size of the dataset here is about one eighth of the dataset used by Blei et al. So there might be some discrepancy between our result and those presented in Blei et al.