Topic Modeling: Latent Dirichlet Allocation

Mohamed Kleit [1] **Fatima-Zahra Banani** [1] **Robert Chamoun** [1]

Abstract

We will explore Latent Dirichlet Allocation (LDA) (Blei et al., 2003) in the context of topic modeling. Topic modeling aims at discovering hidden themes in a collection of documents and annotating them according to these topics. LDA is a generative probabilistic model as it assumes documents where generated according to a generative process. LDA tries to capture the intuition that documents generally exhibit multiple topics. Indeed, in this three-level Bayesian model, each document is modeled as a finite mixture over an underlying set of topics and each topic can be modeled as an infinite mixture over an underlying set of topic probabilities. As per topic modeling, this infinite mixture provides an explicit representation of each document. We will verify through our own implementation, that the variational methods and EM algorithm presented in (Blei et al., 2003) provide an efficient way of conducting inference and parameter estimation. In order to do this, we will try to reproduce the results from the dataset (Blei) associated to the paper and will apply our implementation to the collection of lecture notes from the class.

1. Motivation

Topic modeling brings a solution to the difficulties faced when searching and exploring large collections of documents that were neither annotated nor labeled prior to their utilisation. Its goal is to discover the main themes running through large and unstructured collections of documents so that they can be organized according to the inferred themes. Different topic models have been used in the past to accomplish this task, which we describe below in order to contextualise the situation. They all share three fundamental assumptions:

- The documents have a latent semantic structure, which are the topics in our case.
- We can infer topics from the multiple occurrences of words across different documents.
- Every word is associated to a topic and all documents exhibit a mixture of topics.

In other words, topic models are built around the idea that the structure of the documents is actually a result of some latent variables - the topics.

1.1. Latent Semantic Indexing (LSI)

One of the foundation technique of topic modeling, LSI (Landauer et al., 1998) aims to find a low-dimension representation of the documents and words. Given documents d_m where $m \in (1, ..., M)$ and vocabulary words w_n where $n \in (1, ..., N_d)$, we construct a document-term matrix $X \in R^{m \times n}$ where $x_{i,j}$ describes the occurrence of word w_j in document d_i. However, raw-counts do not account for the significance of each word in a document. We therefore replace the raw counts in X with a term frequency-inverse document frequency score which assigns a weight for the j^{th} term in the i^{th} document. Intuitively, a term will have a large weight if it occurs frequently in a document but not so frequently across a collection of documents. We then reduce the dimensionality of our matrix X using truncated Singular Value Decomposition (SVD) as follows:

$$X \simeq U_t \Sigma_t V_t^T \tag{1}$$

where U relates our document-topic matrix, V relates a term-topic matrix and Σ is a diagonal matrix of singular values where we keep only the t most significant dimensions/topics. We can then use our decomposition to find similarities between documents or between words. However, LSI presents some pitfalls. Indeed, its representation is less efficient than the two models we discuss below and we need really large set of documents and vocabulary to get accurate results.

1.2. Probabilistic Latent Semantic Indexing (pLSI)

Rather than using SVD, pLSI (Hofmann, 1999) uses a probabilistic graphical model where we treat documents as the

[1] Department of Computer Science and Operations Research, University of Montreal, Montreal, Canada. Correspondence to: Mohamed Kleit 1061121 <mohamed.kleit@umontreal.ca>, Fatima-Zahra Banani 20181291 <fatima-zahra.banani@umontreal.ca>, Robert Chamoun 987719 <robert.chamoun@umontreal.ca>.

IFT 6269 project report - Université de Montréal

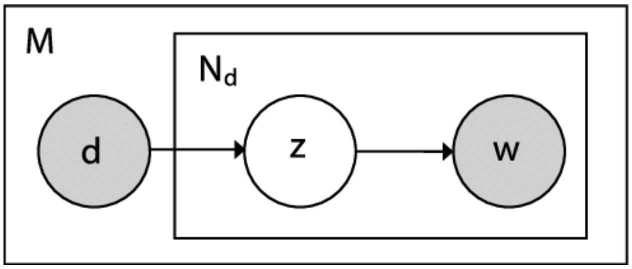

Figure 1. pLSI graphical model

result of a generative probabilistic process that includes latent topics. The graphical model is depicted below: where d is the m_{th} document in our collection of documents, z is the topic associated to the n^{th} word w in document d. The joint probability of observing word d and w is thus given by:

$$P(D, W) = P(D) \sum_{Z} P(Z|D)P(W|Z) \tag{2}$$

In our case, $P(D)$, $P(Z|D)$ and $P(W|Z)$ are the parameters of the model. Now, $P(Z|D)$ and $P(W|Z)$ are modeled as multinomial distributions and can be estimated using expectation-maximization (EM). However, some issues arise regarding pLSI such as the lack of parameters for $P(D)$ which makes it difficult to assign probabilities to new documents. It is also prone to overfitting as the number of parameters for $P(Z|D)$ grows linearly with the number of documents.

2. Latent Dirichlet Allocation (LDA)

We thus generalize the pLSI model with LDA which is a three-level Bayesian graphical model. As mentioned previously, LDA is a probabilistic model and it has a corresponding generative process by which it assumes all documents arose.

2.1. Notation and terminology

Let us define a few terms before diving into the inherent characteristics of LDA (Blei et al., 2003).

- A word w as an unit element from a vector of size V, which corresponds to the fixed vocabulary defined across all documents in the corpus. We define the word such that the v^{th} word in the vocabulary is represented by a vector of size V such that $w^v = 1$ and $w^u = 0$ for $v \neq u$.

- A document is defined as a sequence of N words denoted by a vector **w** $= (w_1, w_2, ..., w_N)$ where w_n is the n^{th} word in the document d.

- We define a corpus as a collection of documents of size M where a document is defined as described above.

2.2. Graphical model

The advantage that LDA provides with respect to pLSI is the fact that it can generalize better to unseen documents. The graphical model is depicted as below:

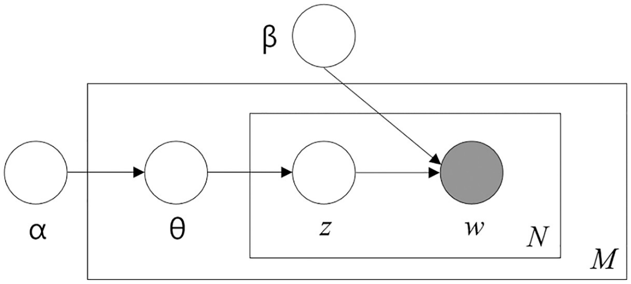

Figure 2. LDA graphical model

We assume here that α is the parameter for a Dirichlet distribution which acts as a prior on the latent variable θ. θ represents the topic proportions for document d in our corpus D. We assume here α is a fixed vector of positive real values of dimension K. K is a hyper parameter representing the number of topics we assume to be running through the entire corpus. α thus represents the topic proportions across the corpus. β is a $k \times V$ matrix which parameterizes the word probabilities. We have $\beta_{i,j} = p(w^j = 1|z^i = 1)$. As for α, we assume β to be fixed - we will try to estimate. A more intuitive way to define β would be describe it as a distribution over the vocabulary V for topic k. We define z as the topic assignments for each word n in document d. Finally, w is the only observed variable in our model and represents the observed word n in document d. Given the parameters α and β that we want to estimate, we can factorize the joint distribution of a topic mixture θ, a set of N topics **z** and a set of N words **w** as follows:

$$p(\theta_d, Z_d, W_d|\alpha, \beta) = p(\theta_d|\alpha) \prod_{n=1}^{N} p(z_n|\theta_d)p(w_n|z_n, \beta) \tag{3}$$

2.3. A generative process

LDA assumes that each document **w** was generated by a simple process described below. The intuitive idea is that each document is represented as a random mixture over latent topics and each topic is characterized by a distribution over words.

- Draw a distribution over topics: $\theta_d \sim Dirichlet(\alpha)$

- For each word n:
  - Draw a topic from the distribution over topics: $z_{d,n} \sim Mult(\theta_d)$
  - Draw a word from the corresponding topic: $w_{d,n} \sim Mult(\beta_{z_{d,n}})$

As mentioned earlier, LDA is a three-level graphical model. Indeed, we have corpus-level parameters in α and β which we assume are sampled once in the generative process. Then, θ_d is a document-level variable, sampled once per document. Finally, $z_{d,n}$ and $w_{d,n}$ are sampled once for each word in each document: they are word-level variables.

3. Parameter Estimation

In order to use LDA we should compute the posterior distribution of the hidden variables Z_d (representing the topics) given a document W_d.

$$p(\theta_d, Z_d | W_d, \alpha, \beta) = \frac{p(\theta_d, Z_d, W_d | \alpha, \beta)}{p(W_d | \alpha, \beta)} \quad (4)$$

But this distribution is intractable to compute in general. So we're going to use variational inference instead to approximate the posterior.
The family of variational distributions on the latent variables that is used in this work is given by the graphical model in Figure 3 .

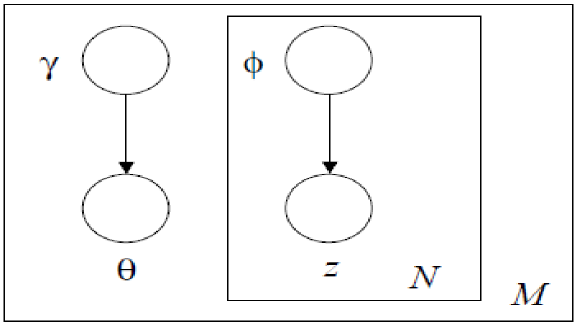

Figure 3. The Graphical Model of the variational distribution q.

This family is characterized by the following distribution:

- For a document d:

$$q(\theta_d, Z_d | \gamma_d, \phi_d) = q(\theta_d | \gamma_d) \prod_{n=1}^{N} q(Z_d^n | \phi_d^n) \quad (5)$$

- For each document d:
  - γ_d is the variational Dirichlet parameter.
  - $(\phi_d^1, \dots, \phi_d^N)$ are the variational multinomial parameters for the latent variables.

We have to choose α^* and β^* that maximizes the log likelihood of the corpus D.

$$\begin{aligned} \alpha^*, \beta^* &= argmax \log p(D | \alpha, \beta) \\ &= argmax \sum_{d=1}^{M} \log p(W_d | \alpha, \beta) \end{aligned} \quad (6)$$

Using Jensen's inequality:

$$\begin{aligned} \log p(W_d | \alpha, \beta) &\geq \int \sum_{Z} q(\theta, Z | \gamma_d, \phi_d) \log \frac{p(\theta, Z, W_d | \alpha, \beta)}{q(\theta, Z | \gamma_d, \phi_d)} d\theta \\ &= L(\gamma_d, \phi_d; \alpha, \beta) \end{aligned} \quad (7)$$

And we can easily prove that:

$$log(p(W_d | \alpha, \beta)) = \\ L(\gamma_d, \phi_d; \alpha, \beta) + D(q(\theta_d, Z_d | \gamma_d, \phi_d || p(\theta_d, Z_d | W_d, \alpha, \beta))$$

From the last equality we can conclude that maximizing the lower bound $L(\gamma_d, \phi_d; \alpha, \beta)$ with respect to γ_d and ϕ_d is equivalent to minimizing the KL divergence between the true and variational posterior probabilities.

3.1. E-step updates

For each document d, we are going to maximize the lower bound $L(\gamma_d, \phi_d; \alpha^t, \beta^t)$ with respect to γ_d and ϕ_d.

- For the multinomial parameters $(\phi_d^1, \dots, \phi_d^N)$ the maximization is constrained since for each latent variable $Z_d^n : \sum_{i=1}^{k} \phi_{d\,i}^n = 1$. Using the Lagrange Multiplier approach we can find the following update:

$$\phi_{d\,i}^n \propto \beta_{iw_n} exp(\psi(\gamma_{di}) - \psi(\textstyle\sum_{j=1}^{k} \gamma_{dj}))$$

- Taking the derivative of $L(\gamma_d, \phi_d; \alpha^t, \beta^t)$ with respect to γ_d and setting it to zero, we can conclude that:

$$\gamma_{di} = \alpha_i^t + \textstyle\sum_{n=1}^{N} \phi_{d\,i}^n$$

We alternate between these two updates until convergence to finally find the values of γ_d^{t+1} and ϕ_d^{t+1}.

3.2. M-step updates

In this step we are going the maximize

$$L(\gamma^{t+1}, \phi^{t+1}; \alpha, \beta) = \sum_{d=1}^{M} L(\gamma_d^{t+1}, \phi_d^{t+1}; \alpha, \beta) \quad (8)$$

with respect to α and β.

- To find β^{t+1} we are going to use the Lagrange Multiplier approach again since : $\sum_{j=1}^{V} \beta_{ij} = 1$ for each $i \in [1, k]$:

  $$\beta_{i,j} = \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi_{d\,i}^{n} w_{d\,j}^{n} \tag{9}$$

  w_d^n is the hot encoding of the nth word in document d.
- Taking the derivative of $L(\gamma^{t+1}, \phi^{t+1}; \alpha, \beta)$ with respect to α_i gives:

  $$\frac{\delta L}{\delta \alpha_i} = M(\psi(\textstyle\sum_{j=1}^{k} \alpha_j) - \psi(\alpha_i)) + \sum_{d=1}^{M}(\psi(\gamma_{di}) - \psi(\sum_{j=1}^{k} \gamma_{dj})$$

  Since this derivative depends on α_j where $j \neq i$, we must use an iterative method to find the maximal α. The iterative method chosen in this work is Newton-Raphson algorithm to find the root of $\frac{\delta L}{\delta \alpha}$

4. Implementation

In this section we present our implementation of LDA in *Python* and discuss our results on real data.

4.1. Library Used

- *scikit learn* for pre-processing.
- *scipy* for probability distributions.

4.2. Dataset Description

The dataset used is the same one used in the authors implementation. it can be found in this link.

It consist of 2246 small documents. We used 2000 documents for training our model and kept the rest for testing (classification).

4.3. Pre-processing

Tokenization of the documents: LDA uses the bag of word assumption (the order of words apparitions in a document is ignored). This allows us to use a compact representation of documents in a matrix of M x V. Each line in the matrix represents a document and the values are integer that represent the number of occurrences of a word in the vocabulary. This is similar to the one hot encoding trick that we used in the class.

In addition to tokenization, we included the following filters:

- Removed non-alphabetical characters.
- Removed words with only 1 occurrence across the corpus.
- Removed words that occur in 95 % or more of the documents.

Finally, Due to the limited memory of our hardware and for runtime reasons, we limited the vocabulary to only the 3000 most frequent words of the corpus.

4.4. Hyperparameter Selection

We chose our topic vector α to be of size 100 to match the original implementation of the author.

4.5. Parameter Initialization

As stated in section 3, LDA with with variational EM has 4 sets of parameters to estimate. The parameters initialized to the following value:

Table 1. Parameter initialization. K represent the number of topics, V the size of the vocabulary , M the number of document and N[d] the length of each document.

Parameter	Dimension	Initialization
α	[K]	np.random.gamma(shape=np.ones((K)), scale=1/K)
β	[KxV]	np.random.dirichlet(np.ones(V), K)
γ	[MxK]	alpha + (N[d] / K)
ϕ	[MxN[d]xK]	np.ones((N[d], K))/ K

4.6. Stopping Criteria

In the paper, the author doesn't state very clearly the stopping criterion used in its implementation. So we chose to stop the model once the parameter L2 norm of γ from one iteration to the next is is below a given threshold.

5. Empirical Results

5.1. Topic Extraction

The canonical use of topic modeling is to find a list of topics across a corpus of text. We can attempt to understand the meaning of a topic using the words with the highest probability for the given topic (table 2).

5.2. Document Classification

LDA can also be used to classify previously unseen document to one of the K-topics. First, we need to perform variational inference on the unseen document and use the trained gamma parameter to assign the document to the most likely topic. We scored the holdout documents and made a manual comparison of document vs topics. (See table 3 in appendix A for a sample of documents.)

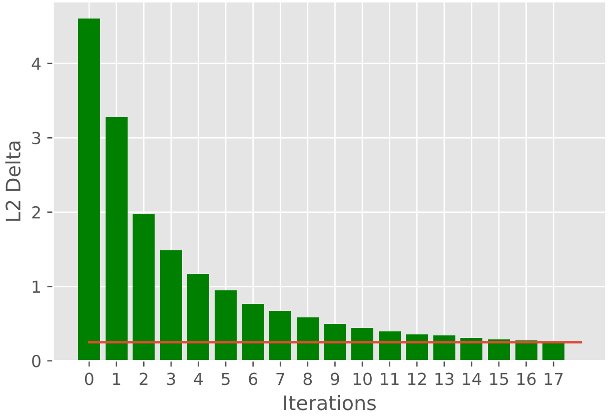

Figure 4. Change in γ parameter across EM iterations. The green bars represent the L2 change of γ parameters. The red line represent the stopping criterion used (0.25 in our case).

Table 2. Topic Sample from LDA model: each column represents the top 10 words from the topic. We can see that the topics are easy to interpret. For full list of topic, see the joined file: *TopicsList.csv*

Topic Sample				
Topic 1: Politics	Topic 2: Finance	Topic 3: Trial	Topic 4: Crime	Topic 5: Korea
trade	york	years	authorities	roh
committee	shares	guilty	city	hostages
senate	board	case	injured	guard
new	million	trial	shot	officials
time	exchange	prison	hospital	coast
abortion	trading	said	night	korean
bush	stocks	judge	killed	korea
president	index	court	people	south
souter	market	attorney	police	north
said	stock	charges	said	said

5.3. Discussion

We have presented latent Dirichlet allocation, a generative probabilistic model for discrete data. We derived the parameters of the model using variational inference algorithms and presented the results of out implementation on text data. The LDA model can be extended in a number of ways, one possible avenue is to use Gibbs sampling instead of variational method, another would be to organize our topics into a hierarchy (Blei et al., 2004).

References

Blei, D. M. Latent dirichlet allocation. URL http://www.cs.columbia.edu/~blei/lda-c/.

Blei, D. M. Probabilistic topic models. *Commun. ACM*, 55 (4):77–84, April 2012. ISSN 0001-0782. doi: 10.1145/2133806.2133826. URL http://doi.acm.org/10.1145/2133806.2133826.

Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003. ISSN 1532-4435. doi: http://dx.doi.org/10.1162/jmlr.2003.3.4-5.993. URL http://www.cs.columbia.edu/~blei/papers/BleiNgJordan2003.pdf.

Blei, D. M., Griffiths, T. L., Jordan, M. I., and Tenenbaum, J. B. Hierarchical topic models and the nested Chinese restaurant process. In *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.

Hofmann, T. Probabilistic latent semantic indexing. In Gey, F., Hearst, M., and Tong, R. (eds.), *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '99), August 15-19, 1999, Berkeley, CA, USA*, pp. 50–57. ACM Press, New York, NY, USA, 1999.

Landauer, T. K., Foltz, P. W., and Laham, D. An introduction to latent semantic analysis. *Discourse Processes*, 25(2-3):259–284, 1998. doi: 10.1080/01638539809545028. URL https://doi.org/10.1080/01638539809545028.

Appendix A: Sample of Document Classification

Table 3. Sample of the first lines of documents from the test data with the predicted topic. Full results on the test set can be found in the following file: *ClassificationHoldout.csv*

Documents	**Predicted Topic**
A federal judge approved a settlement Thursday that keeps the state from restricting abortions to women less than weeks pregnant. U.S. District Judge John Nordberg agreed to the settlement between Illinois Attorney ...	federal, panel, department, appeals, decision, city, environmental, state, court, said
The stock market showed a small loss today as traders studied the latest data on economic growth. The Dow Jones average of industrials dropped . to ,. by noontime on Wall Street ...	york, shares,board, million exchange, trading, stocks, index, market, stock
A former programmer has been convicted of planting a computer "virus" in his employer's system that wiped out , records and was activated like a time bomb, doing its damage two days after he was fired ...	years guilty, case, trial, prison, said, judge, court, attorney, charges
Bloodstains on a pillowcase and exercise bar found in Joel Steinberg's apartment came from his former lover, an FBI expert testified at Steinberg's trial on charges he beat his illegally adopted -year-old daughter to death ...	parents, wife, children, police, meese, hospital, family, year, mrs, said
Prime Minister Ingvar Carlsson, head of Sweden's caretaker Cabinet, is ready to lead a new government to replace the one which resigned last week, the national news agency TT reported Thursday. Earlier Thursday, the Social Democratic leader informed Parliament ...	today, cabinet, democratic, new, political opposition, minister, government,said,party