Topic Modeling for Scientific Documents

Jethro Kuan
WING-NUS

Abstract

Topic models are statistical models, used to discover abstract topics that occur in a collection of documents. In this paper, we focus on the technique of topic modeling, particularly in its application to the domain of scientific documents. We discuss different approaches to topic modeling, and weigh in on their strengths and shortcomings.

1 Introduction

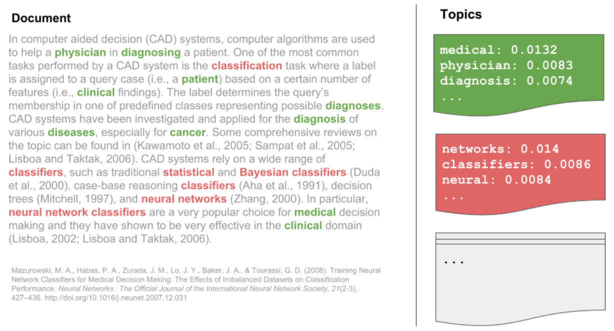
In this information age, the availability of knowledge is insufficiently met with the tools to navigate it. Archives such as Arxiv see an exponential growth in the number of documents being hosted. Developing new tools for browsing, and searching these documents is a technological challenge that calls for research in statistical modeling.

Rather than relying on keyword matching, tools like Semantic Scholar use machine learning to process scientific documents and discover meaningful structure, empowering researchers to discover papers more relevant to their work. One such statistical modeling technique is topic modeling.

Topic modeling attempts to discover abstract topics within documents. Topic models capture the intuition that if a document is of a particular topic, then words from the topic should appear more frequently.

A topic model trained on a corpora of scientific documents could learn topics such as "Neural Networks", "Biology" and "Medicine". Topics attribute a high probability to words that relate to the topic. For example, a "Neural Networks" topic would give attribute a high probability to "classifiers", and a low probability to "bacteria". The quotations around the topics are to make explicit that the labels are human interpretations of what these topics may be. Topic modeling is an unsupervised problem, and is often trained on unlabeled data. Automatic labeling of these topics is an active area of research [9], [8].

Topic modeling may discover that document in Figure 1 relates to "Medicine" and "Neural Networks", because these topics give rise to the colored words in the document.


Document

In computer aided decision (CAD) systems, computer algorithms are used to help a **physician** in **diagnosing** a patient. One of the most common tasks performed by a CAD system is the **classification** task where a label is assigned to a query case (i.e., a **patient**) based on a certain number of features (i.e., **clinical** findings). The label determines the query's membership in one of predefined classes representing possible **diagnoses**. CAD systems have been investigated and applied for the **diagnosis** of various **diseases**, especially for **cancer**. Some comprehensive reviews on the topic can be found in (Kawamoto et al., 2005; Sampat et al., 2005; Lisboa and Taktak, 2006). CAD systems rely on a wide range of **classifiers**, such as traditional **statistical** and **Bayesian classifiers** (Duda et al., 2000), case-base reasoning **classifiers** (Aha et al., 1991), decision trees (Mitchell, 1997), and **neural networks** (Zhang, 2000). In particular, **neural network classifiers** are a very popular choice for **medical** decision making and they have shown to be very effective in the **clinical** domain (Lisboa, 2002; Lisboa and Taktak, 2006).

Mazurowski, M. A., Habas, P. A., Zurada, J. M., Lo, J. Y., Baker, J. A., & Tourassi, G. D. (2008). Training Neural Network Classifiers for Medical Decision Making: The Effects of Imbalanced Datasets on Classification Performance. *Neural Networks : The Official Journal of the International Neural Network Society, 21*(2-3), 427–436. http://doi.org/10.1016/j.neunet.2007.12.031

Figure 1: According to the trained topic model, topics like "Medicine" and "Neural Networks" generate the colored words of the document

In addition, we can view these abstract topics as a form of clustering. Researchers are better able to navigate the large corpora of scientific knowledge, by exploring documents that have similar topic proportions.

The earliest known technique for topic modeling is Latent Semantic Indexing (LSI) [5]. The first probabilistic approach, Probabilistic Latent Semantic Analysis (pLSA), was proposed in 1999 [7]. It was only in 2003 when Blei et al. introduced Latent Dirchlet Allocation (LDA), which remains to this date the predominant technique for topic modeling.

Most of the state-of-the-art topic models are generative models, assuming that latent variables govern the generative process of a document. According to these models, a document is produced from a distribution of topics, and the topics themselves are distributions over the vocabulary of the corpora.

In this paper, we briefly discuss the predominant model: Latent Dirichlet Allocation (LDA) and its variants. In addition, we explore other models that are not derivatives of LDA, such as the Replicated Softmax.

2 LDA

Latent Dirichlet Allocation is widely considered to be the simplest topic model. LDA models each document as a mixture of topics, where a topic β_k is a probability distribution over a fixed vocabulary of terms. Training LDA requires fixing the number of topics, K. We can draw the graphical model for LDA as in Figure 2.

At risk of being pedantic, we explain the graphical model in detail. η is the topic hyperparameter, which produces a topic distribution β_k of the Dirichlet family. There are a total of K topics. Similarly, α is a hyperparameter that produces the per-document topic proportions θ_d. These Dirichlet distributions are of dimension $K-1$, because there are a total of K topics. There are D such topic proportions, where D is the total number of documents. $Z_{d,n}$ is the per-word topic assignment, drawn from the particular θ_d. Finally, $W_{d,n}$ is the nth word in the dth document, an observed variable. It is simple to see that $P(W_{d,n}|Z_{d,n},\beta_k) = \beta_{z_{d,n},w_{d,n}}$.

A high α value encodes the belief that documents contain a mixture of many topics, rather than being largely represented by a few topics. Similarly, a high η value encodes the belief that topics has high probability for a large number of words in the vocabulary.

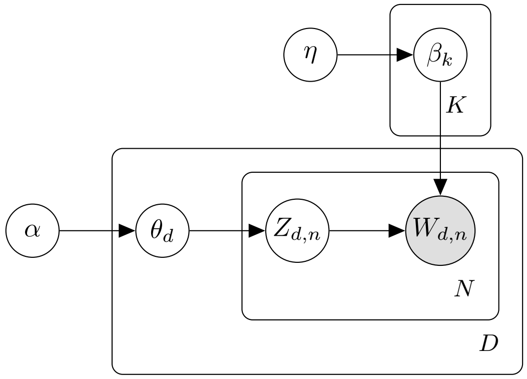


Figure 2: Plate notation for LDA, a directed graphical model.

We can fully specify the model by looking at the joint distribution $\mathcal{J}$ of all the latent and observed variables.

$$\mathcal{J} = \left(\sum_{k=1}^{K} P(\beta_k|\eta)\right)\left(\sum_{d=1}^{D} P(\theta_d|\alpha)\right) \left(\sum_{n=1}^{N} \theta_{d,Z_{d,n}}\beta_{Z_{d,n},W_{d,n}}\right) \quad (1)$$

The generative process of a document is as follows:

Algorithm 1 Generative Process of LDA

```
for each document $d$ in $D$ do
    Draw topic distribution $\theta_d \sim Dir(\alpha)$
    for each word $n_d$ in $N_d$ do
        Sample topic $z_{d,n} \sim Multinomial(\theta)$
        Sample word $w_{d,n} \sim Multinomial(\beta_{z_{d,n}})$
```

If we fix the topic distributions $\beta_{1:K}$, we can compute the per-document posterior θ given the document.

$$P(\theta|w_{1:n},\alpha,\beta_{1:K}) = \frac{P(\theta|\alpha)\prod_{n=1}^{N} P(z_n|\theta)P(w_n|z_n,\beta_{1:K})}{\int_\theta P(\theta|\alpha)\prod_{n=1}^{N}\sum_{z=1}^{K} P(z_n|\theta)P(w_n|z_n,\beta_{1:K})} \quad (2)$$

The denominator is intractable to compute, due to the coupling between θ and β under the multinomial assumption. [4]. Hence, we rely on techniques for approximate inference of the posterior. We discuss these techniques in section 6.

Why does LDA work? The Dirichlet distribution encourages sparsity, encoding the belief that the document-topic distribution has few topics per document, and the topic-word distribution has few words per topic. These two beliefs work against each other, and LDA discovers this sparsity balance, which gives rise to the structure of the textual data.

2.1 Statistical Assumptions

As Mackay quips, we cannot make inference without statistical assumptions. LDA makes several assumptions, some rendering it less suited for application to the domain of scientific documents.

LDA models documents as "bag-of-words": words within the document are interchangeable. [4] I think this is a reasonable assumption to make, given the task is to discover themes within the document.

LDA also assumes that the order of documents do not matter. Exchangeability of both words and documents allows LDA to model the joint distribution as a mixture model. I believe this assumption to be invalid in the domain of scientific documents. The meaning of keyphrases used in scientific literature change over time. For example, the landscape of research neural networks is vastly different now, as compared to the 1990s, and LDA will fail to capture these differences.

The use of the Dirichlet distribution also encodes statistical assumptions about the correlation between topics. Under the Dirichlet, components of θ_d are nearly independent. This leads to the modeling assumption that the presence of one topic is not correlated with the presence of another [3]. As explained by Blei and Lafferty, this assumption is strong and unrealistic in the domain of scientific documents. An article about genetics is also highly likely to be about health and disease, and unlikely to be about astronomy [3].

Variants of LDA relax these statistical assumptions, or make other assumptions in place. We discuss Dynamic Topic Modeling (DTM) in section 4 and Correlated Topic Modeling(CTM), and briefly mention the rest in the appendix.

3 Correlated Topic Modeling

The Correlated Topic Model (CTM) addresses the model assumption that topic proportions are not correlated. In scientific documents, this assumption is unlikely to hold true.

Instead of drawing from a Dirichlet distribution, the CTM uses a logistic normal distribution. CTM draws a real valued random vector from a multivariate Gaussian, and maps it to the simplex to obtain a multinomial parameter [3]. The $K \times K$ covariance matrix Σ models dependencies between the topics. This tweak is evident in the generative process shown in Algorithm 2, contrasting it with Algorithm 1.

Algorithm 2 Generative Process of CTM

```
for each document $d$ in $D$ do
    Draw topic distribution $\beta_d \sim \mathcal{N}(\mu, \Sigma)$
    for each word $n_d$ in $N_d$ do
        Sample topic $z_{d,n} \sim Multinomial(\pi(\beta_d))$
        Sample word $w_{d,n} \sim Multinomial(\beta_{z_{d,n}})$
```

π maps the multinomial parameters to the mean parameters, $\pi(x) = \frac{e^x}{\sum_i e^{x_i}}$

We can also visualize CTM with a graphical model.

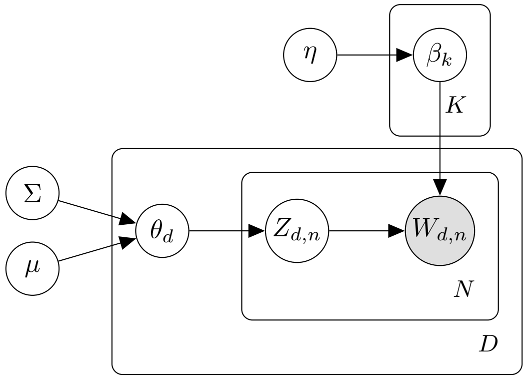


Figure 3: Plate notation for CTM.

The Multinomial and Guassian distributions are not conjugate distributions. This causes inference via Gibbs sampling is difficult. Hence, variational inference is used instead.

4 Dynamic Topic Modeling

Dynamic Topic Modeling (DTM) was proposed to remove the assumption that documents are *exchangeable.* [2]

The order of documents are important for scientific documents, since both the content, and the meaning of words evolve over time.

In DTM, data is divided by discrete time slices. The topics associated with time slice t evolve from the time slice $t - 1$. Because the Dirichlet distribution is not amenable to sequential modeling, we use the Gaussian distribution to model the sequence of random variables.

The generative process for time slice t is as follows:

Algorithm 3 Generative Process of DTM

```
Draw topic distribution $\beta_t|\beta_{t-1} \sim N(\beta_{t-1}, \sigma^2 I)$
Draw $\alpha_t|\alpha_{t-1} \sim N\left(\alpha_{t-1}, \delta^2 I\right)$
for each document $w$ do
    Draw $\eta_{w,t} \sim N\left(\alpha_t, a^2 I\right)$
    for each word at position $n$ do
        Sample topic $z_{t,n} \sim Multinomial(\pi(\eta_{w,t}))$
        Sample word $w_{t,d,n} \sim Multinomial(\beta_{t,z,n})$
```

Similar to CTM, the use of the logistic normal distribution causes variational inference to be the preferred approximate inference technique.

Further extensions of this approach include the continuous Dynamic Topic Models (cDTM), which removes the discretization of the time slices. [11] This model has been used to predict the timestamp of documents.

5 Critique of Models

One important component we have yet to discuss is the choice of K, the number of topics used to model the corpora. Notice that the value of K affects model complexity. In both LDA and DTM, documents are modeled as mixtures of K distributions. The choice of K is therefore also an important parameter that requires tuning. This is a pitfall of parametric models. A misfit between the complexity of the model and the amount and quality of data available can lead to severe underfitting or overfitting [10].

6 Inference Methods

Approximating intractable probability densities is a well-studied problem in modern statistics. This problem arises often in Bayesian statistics, where computing posterior probablity densities in requires inference over latent variables. Many learning algorithms have been developed, including collapsed Gibbs Sampling, Variational Inference, Collapsed Variational Inference, and MAP estimation. Each of these approximation techniques have their own strength and shortcomings.

The two inference methods are briefly discussed below, and a comparison between them relegated to the appendix in subsection 6.3.

6.1 MCMC Sampling

Historically, Markov Chain Monte Carlo (MCMC) sampling has been the dominant technique for approximating posterior densities. In MCMC, we construct an ergodic Markov chain on the latent variable z, whose stationary distribution is the posterior $P(z|x)$. Samples are drawn from the stationary distribution, and used to approximate the posterior empirically.

In Gibbs sampling, the space of the Markov Chain is the space of the configurations of the hidden variables. In Gibbs sampling, the next state is reached by sequentially sampling all variables from the distribution, conditioned on all the current sampled values.

After a “burn-in” period, the samples would be drawn from the posterior distribution.

Algorithm 4 Gibbs Sampling

```
$x^0 \leftarrow q(x)$
for $i = 1, 2, 3, \ldots$ do
    for $d = 1, 2, 3, \ldots, D$ do
        $x_d^i \sim P(X_1 = x_1 | X_k = x_k^{i-1}$ for $k = \{1..n \setminus d\})$
```

A full treatment of Gibbs Sampling applied to LDA can be found in [6].

6.2 Variational Inference

In variational inference, the posterior distribution over a set of unobserved variables $p(Z|H)$ is approximated by a distribution $q(Z)$, selected to be in a family that can approximately model the true posterior. Inference is performed by minimizing the distance between $p(Z|H)$ and $q(Z)$. One common metric used is the KL-divergence. Here, we discuss variational inference for LDA.

Mean field variational inference (MFVI) breaks the coupling between θ and z by introducing free variational parameters γ over θ and ϕ over z and dropping the edges between them. This results in an approximate posterior $q(\theta, z|\gamma, \phi) = q_\gamma(\theta) \prod_n q_\phi(z_n)$. This is illustrated in the graphical model in Figure 4.

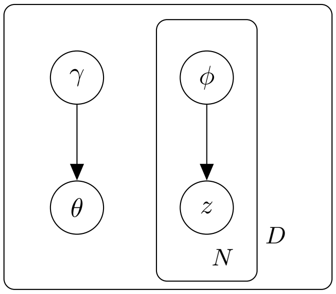


Figure 4: Graphical model for the approximate posterior for variational inference.

To best approximate the true posterior, we frame it as an optimization problem, minimizing L where:

$$L(\gamma, \phi|\alpha, \beta) = D_{KL}\left[q(\theta, z|\gamma, \phi) || p(\theta, z|\alpha, \beta)\right] - \log p(w|\alpha, \beta) \tag{3}$$

This optimization has closed form coordinate descent equations for LDA, because the Dirichlet is conjugate to the Multinomial distribution. This computational convenience comes at the expense of robustness, making it difficult to apply to other more complicated topic models [4].

6.3 Choosing an Inference Method

How do we know which technique to use to approximate the posterior density? MCMC methods are computationally more intensive, but provide samples that are approximately exact from the target posterior density. In contrast, VI methods view the problem as an optimization problem, which allows it to utilize efficient learning algorithms such as stochastic optimization. This is much quicker to compute, and is suited for larger datasets.

MCMC methods, however, cover a large family of sampling methods. Gibbs sampling requires that the prior and posterior are conjugate distributions. When this is not possible, such as in DTM, VI methods can perform better than other methods in the MCMC family.

A closer look at the different inference approximation algorithms, however, shows that the performance differences can be explained away by setting certain smoothing hyperparameters [1].

7 Appendix

7.1 Alternatives to LDA

References

[1] Arthur Asuncion, Max Welling, Padhraic Smyth, and Yee Whye Teh. On smoothing and inference for topic models. *CoRR*, 2012. URL `http://arxiv.org/abs/1205.2662v1`.

[2] David M Blei and John D Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120. ACM, 2006.

[3] David M Blei and John D Lafferty. A correlated topic model of science. *The Annals of Applied Statistics*, pages 17–35, 2007.

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

[5] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41 (6):391, 1990.

[6] Tom Griffiths. Gibbs sampling in the generative model of latent dirichlet allocation. *Stanford University*, 2002.

[7] Thomas Hofmann. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 289–296. Morgan Kaufmann Publishers Inc., 1999.

[8] Jey Han Lau, Karl Grieser, David Newman, and Timothy Baldwin. Automatic labelling of topic models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1536–1545. Association for Computational Linguistics, 2011.

[9] Qiaozhu Mei, Xuehua Shen, and ChengXiang Zhai. Automatic labeling of multinomial topic models. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 490–499. ACM, 2007.

[10] Yee Whye Teh. Dirichlet process. In *Encyclopedia of machine learning*, pages 280–287. Springer, 2011.

[11] Chong Wang, David Blei, and David Heckerman. Continuous time dynamic topic models. *CoRR*, 2012. URL `http://arxiv.org/abs/1206.3298v2`.