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. Tools like Semantic Scholar use machine learning to process scientific documents and extract meaningful structure, empowering researchers to discover papers more relevant to their work. One such technique is topic modeling.

2 Topic Modeling

Discovering underlying structure within scientific documents is an unsupervised learning problem. Generative models, such as topic modeling, uncover the underlying semantic structure of a document collection based on a hierarchical Bayesian analysis of the texts. 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.

3 LDA

Latent Dirichlet analysis 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.

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.

The generative process is described as follows:

```
for each document $w$ do
  Draw topic distribution $\theta \sim Dir(\alpha)$;
  for each word at position $n$ do
    Sample topic $z_n \sim Multinomial(\theta)$
    Sample word $w_n \sim Multinomial(\beta_{z_n})$
```

We can compute the probability of a word, given the LDA parameters:

$$p(w|\alpha, \beta) = \int_\theta \left(\sum_{n=1}^{N} \sum_{z_n=1}^{k} p(w_n|z_n, \beta) p(z_n|\theta) \right) p(\theta|\alpha) d\theta \tag{1}$$

Posterior inference over the hidden variables θ and z is intractable due to the coupling between θ and β under the multinomial assumption. (Blei et al., 2003). Hence, we rely on techniques for approximate inference of the posterior. These techniques are discussed in section 4.

3.1 Statistical Assumptions

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

1. The order of documents do not matter.

The meaning of keyphrases used in scientific literature change over time. For example, neural networks meant a different thing two decades ago.

2. Bag of Words

Variants of LDA relax these statistical assumptions, or make other assumptions in place. We discuss Dynamic Topic Modeling (DTM) in section 5, and briefly mention the rest in the appendix.

4 Inference Methods

Approximating intractable probability densities is a well-studied problem in modern statistics. This problem arises often in Bayesian statistics, where inference over latent variables are required to compute the posterior probability densities. 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.

A comprehensive look at the different inference approximation algorithms, shows that the performance differences can be explained away by setting certain smoothing hyperparameters (Asuncion et al., 2012). Nevertheless, we will take a closer look at them.

4.1 Gibb's Sampling

4.2 Variational Inference

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 $q(\theta, z|\gamma, \phi) = q_\gamma(\theta) \prod_n q_\phi(z_n)$.

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

$$L(\gamma, \phi|\alpha, \beta) = D_{KL}\left[q(\theta, z|\gamma, \phi)||p(\theta, z|\alpha, \beta)\right] - \log p(w|\alpha, \beta) \tag{2}$$

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.

5 Dynamic Topic Modeling

Dynamic Topic Modeling (DTM) was proposed to remove the assumption that documents are *exchangeable*. (Blei and Lafferty, 2006)

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

In DTM, data is assumed to be 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, the Gaussian distribution is used instead to model the sequence of random variables.

The generative process for time slice t is as follows:

```
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 $w_{t,d,n} \sim Multinomial(\beta_{t,z,n})$
```

π maps the multinomial parameters to the mean parameters, $\pi\left(\beta_{k,t}\right)_w = \frac{exp(\beta_{k,t,w})}{\sum_w exp(\beta_{k,t,w})}$

The Multinomial and Guassian distributions are not conjugates, inference via Gibb's sampling is difficult. Hence, posterior inference is accomplished via variational inference instead.

Further extensions of this approach include the continuous Dynamic Topic Models (cDTM), which removes the discretization of the time slices. (Wang et al., 2012) This model has been used to predict the timestamp of documents.n

6 Appendix

6.1 Extension to Inference Methods

6.2 Alternatives to LDA

References

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.

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.

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

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