Latent Dirichlet Allocation

Subtitle

Cynthia Cook
cmc496@psu.edu

Advisor: David Hunter
dhunter@stat.psu.edu

Department of Statistics
Pennsylvania State University
November 3, 2015

Latent Dirichlet Allocation

Subtitle

Cynthia Cook

Abstract

Will include once I have a clear research plan...

1. Introduction

For the time being, I am studying the statistical properties of the Latent Dirichlet Allocation topic model.

2. Background on LDA

With an increase in the amount of digitized text, topic modeling has become an important area of research. One of the most popular methods for topic modeling is Latent Dirichlet Allocation (LDA) first described in [2]. LDA is a probabilistic hierarchical generative model, which uses the Bayesian framework to discover hidden structure, or *topics*, within a text Corpus. For a more thorough background in topic modeling, refer to [4].

2.1. Generative Process

Given a set of K topics $k = 1, ...K$ and a set of D documents $d = 1, ..., D$, also referred to as a corpus, the vocabulary is the set of unique words contained in the corpus such that $V = w_{1,1}, ..., w_{d,n_d}, ..., w_{D,N_D}$, where N_d is number of unique words in document d. Then the number of words in V is $N = \sum_{d=1}^{D} N_d$. Let $c_{d,k}$ be vector of length K containing the count of times each topic was assigned to any word in document d, and let $C_{k,n}$ be the vector of length N containing the count of times each word in V is assigned to topic k. The goal of this algorithm as stated by Blei, "to find a probabilistic model of a corpus that not only assigns high probability to members of the corpus, but also assigns high probability to other 'similar' documents" ([2], pp.996). The generative process for a corpus is as follows:

- For each topic,
  1. Draw $\beta_k \sim Dir(\eta)$, a distribution over the vocabulary. Each parameter is of dimension N.
- For each document,
  2. Draw $\theta_d \sim Dir(\alpha)$, a distribution over the topics. Each parameter is of dimension K.
  - For each word in document d,
    3. Draw $z_{d,n} \sim Mult(\theta_d)$, a topic assignment (i.e. $z_{d,n} \in \mathbb{Z} \in \{1, ..., K\}$).
    4. Draw $w_{d,n} \sim Mult(\beta_{z_{d,n_d}})$, the nth word in document d.

This process requires that we know the number of topics K a priori. Blei also assumed that the hyperparameters $\eta \in (0, 1)$ and $\alpha \in (0, 1)$ are fixed and estimated through the given data. Note that $\sum_{n=1}^{N} \eta_n = \sum_{k=1}^{K} \alpha_k = 1$. Assuming this model, we can define the joint probability distribution of a corpus as:

$$p(\beta, \theta, z_{d,n}, w_{d,n}|\alpha, \eta) = \prod_{k=1}^{K} p(\beta_k|\eta) \prod_{d=1}^{D} \left(p(\theta_d|\alpha) \prod_{n=1}^{N_d} p(z_{d,n}|\theta_d) p(w_{d,n}|\beta_1, ..., \beta_K, z_{d,n}) \right), \tag{1}$$

Let $B(\cdot)$ be the Beta function, then

$$
\begin{aligned}
p(\beta|\eta) &= \prod_{k=1}^{K} \frac{1}{B(\eta)} \prod_{n=1}^{N} \beta_{k,n}^{\eta_n - 1} \\
p(\theta|\alpha) &= \prod_{d=1}^{D} \frac{1}{B(\alpha)} \prod_{k=1}^{K} \theta_{d,k}^{\alpha_k - 1} \\
p(z|\theta) &= \prod_{d=1}^{D} \prod_{k=1}^{K} \theta_{d,k}^{c_{d,k}} \\
p(w|\beta, z) &= \prod_{d=1}^{D} \prod_{n=1}^{N} \beta_{k,n}^{C_{k,n}}
\end{aligned}
\tag{2}
$$

Rewriting 1, we obtain the following joint pdf, which also happens to be our posterior distribution for a corpus.

$$
p(\beta, \theta, z_{d,n}, w_{d,n}|\alpha, \eta) = \left(\prod_{k=1}^{K} \frac{1}{B(\eta)} \prod_{n=1}^{N} \beta_{k,n}^{\eta_{k,n} - 1 + C_{k,n}} \right) * \left(\prod_{d=1}^{D} \frac{1}{B(\alpha)} \prod_{k=1}^{K} \theta_{d,k}^{\alpha_{d,k} - 1 + c_{d,k}} \right) \tag{3}
$$

When applying this model to real text data, we do not know either β or θ. They are latent or hidden, and what we are most interested in finding the posterior distribution of the latent variables given the corpus and hyperparameters. Given K, z, α, η to find θ, β we marginalize 3 with respect to the latent variables to obtain the following likelihood equation:

$$
p(w|\alpha, \eta) = \int_{\theta} \int_{\beta} \sum_{z} \left(\prod_{k=1}^{K} \frac{1}{B(\eta)} \prod_{n=1}^{N} \beta_{k,n}^{\eta_{k,n} - 1 + C_{k,n}} \right) * \left(\prod_{d=1}^{D} \frac{1}{B(\alpha)} \prod_{k=1}^{K} \theta_{d,k}^{\alpha_{d,k} - 1 + c_{d,k}} \right) d\beta d\theta \tag{4}
$$

This distribution has been shown to be intractable for exact calculation [2]. However several options are available to approximate this posterior.

2.2. Posterior Inferance

Given a document the distribution of latent topics is found through estimation of the posterior distribution. However, this is very challenging and (computationally demanding) since the posterior distribution of the latent variables given a document is intractable. The development of several fast and accurate approximation methods applied to LDA has provided a user with several options including collapsed Gibbs sampling [7], collapsed variational Bayesian inference [10], maximum likelihood estimation [9], and maximum a posteriori estimation [6].

2.2.1. Gibbs

2.2.2. Variational Methods

2.2.3. Other Methods Here

3. Implementation/Data Example

In [2], variational methods are applied to solve the problem. These methods were coded in C by Blei (refer to [3]) and have since been utilized in the *topicmodels* package in R [8]. Another package, *lda*, in R has been developed by Chang, which uses his own C code with collapsed Gibbs sampling for posterior approximation [5]. On the website Cross Validated, there is a discussion of which package is better for users, with no definitive answer [1]. Refer to my 540 project and what I find there...

3.1. Data

What and how I got it...

3.2. lda results

4. Conclusions

[1] Anonymous. (2012), "Two R Packages for Topic Modeling, lda and topicmodels?" *Cross Validated*: http://stats.stackexchange.com/questions/24441/two-r-packages-for-topic-modeling-lda-and-topicmodels.

[2] Blei, D., Ng, A., and Jordan, M. (2003), "Latent Dirichlet Allocation." *Journal of Machine Learning Research*: 3 993-1022.

[3] Blei, D. (2004),"LDA-C."

[4] Blei, D. (2012), "Survaying a Suit of Algorithms that offer a solution to managing large document archives: Probalistic Topic Models." *Communications fo the ACM*: vol.55, no.4, 77-84.

[5] Chang, J. (2015), "Package 'lda'." *CRAN*.

[6] Chien, J., and Wu, M. (2008), "Adaptive Bayesian Semantic Analysis." *Audio, Speech, and Language Processing*, IEEE Transactions on: 16(1), 198-207.

[7] Griffiths, L. and Steyvers, M. (2004), "Finding Scientific Topics." *PNAS*: 1(Suppl 1), 5228-5235.

[8] Grun, B. and Hornik, K. (2015), "Package 'topicmodels'." *CRAN*.

[9] Hofmann, T. (2001), "Unsupervised Learning by probabilistic Latent Semantic Analysis." *Machine Learning*: 42(1), 177-196.

[10] Teh, Y. W., Newman, D., and Welling, M. (2007), "A Collapsed Variational Bayesian Inference Algorithm for Latent Dirichlet Allocation." *NIPS*: 3, 1353-1360.