Latent Dirichlet Allocation

Subtitle

Cynthia Cook
cmc496@psu.edu

Advisor: David Hunter
dhunter@stat.psu.edu

Department of Statistics
Pennsylvania State University

November 11, 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 in the Corpus say $w_{1,1}, ..., w_{d,n_d}, ..., w_{D,n_D}$, where n_d is the number of uniques words in document d and $V = \sum_{d=1}^{D} n_d$. 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 for each topic over the vocabulary.

- For each document,
  2. Draw $\theta^d \sim Dir(\alpha)$, a distribution for each document over the topics.

    – 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 $w_{d,n} \sim Mult(\beta^{z_{d,n}})$, 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 can be estimated through the given data. Assuming this model, we can define the joint probability distribution of a corpus as:

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

We will use the following dot notation:

$$c_{k,d,v} = \sum_{n=1}^{n_d} I(z_{d,n} = k \& w_{d,n} = v); \qquad \text{\# of times word } v \text{ is assigned to topic } k \text{ in document } d$$

$$c_{k,.,v} = \sum_{d=1}^{D} c_{k,d,v}; \qquad \text{\# of words in document } d \text{ assigned to topic } k$$

$$c_{k,d,.} = \sum_{v=1}^{V} c_{k,d,v}; \qquad \text{\# of times word } v \text{ is assigned to topic } k \text{ in any document}$$

$$c_{k,.,.} = \sum_{d=1}^{D} \sum_{v=1}^{V} c_{k,d,v}; \qquad \text{\# of words in corpus assigned to topic } k$$

Let $\Gamma(\cdot)$ be the Gamma function, then:

$$p(\beta^k|\eta) = \frac{\Gamma(\sum_{v=1}^{V} \eta_v)}{\prod_{v=1}^{V} \Gamma(\eta_v)} \prod_{v=1}^{V} (\beta_v^k)^{\eta_v - 1}$$

$$p(\theta^d|\alpha) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} (\theta_k^d)^{\alpha_k - 1}$$

$$p(z_{d,n}|\theta) = \frac{1}{\prod_{k=1}^{K} c_{d,k,.}!} \prod_{k=1}^{K} (\theta_k^d)^{c_{d,k,.}}$$

$$p(w_{d,n}|\beta, z) = \frac{1}{\prod_{v=1}^{V} c_{k,.,v}!} \prod_{v=1}^{V} (\beta_v^{z_{d,n}})^{c_{k,.,v}}$$

2.2. Posterior Inferance

In practice, we do not know either β or θ. All we are given is a set of documents providing the $w_{d,n}$ parameters. Most often, we are interested in the latent variables β and θ. Rewriting 1

$$p(w, z, \theta, \beta|\alpha, \eta) = \left(\prod_{k=1}^{K} \frac{\Gamma(\sum \eta_v)}{\prod_{v=1}^{V} (\Gamma(\eta_v))} \prod_{v=1}^{V} (\beta_v^k)^{(\eta_v - 1 + c_{k,.,v})} \right) * \left(\prod_{d=1}^{D} \frac{\Gamma(\sum \alpha_k)}{\prod_{k=1}^{K} (\Gamma(\alpha_k))} \prod_{k=1}^{K} (\theta_k^d)^{(\alpha_k - 1 + c_{k,d,.})} \right) \tag{2}$$

Integrating out the latent variables, we obtain the following posterior distribution of a corpus given the hyperparameters

$$p(w|\theta, \beta, z) = \int_\theta \int_\beta \sum_z \left(\prod_{k=1}^{K} \frac{\Gamma(\sum \eta_v)}{\prod_{v=1}^{V} (\Gamma(\eta_v))} \prod_{v=1}^{V} (\beta_v^k)^{\eta_v - 1 + c_{k,.,v}} \right) * \left(\prod_{d=1}^{D} \frac{\Gamma(\sum \alpha_k)}{\prod_{k=1}^{K} (\Gamma(\alpha_k))} \prod_{k=1}^{K} (\theta_k^d)^{\alpha_k - 1 + c_{k,d,.}} \right) d\beta d\theta \tag{3}$$

This distribution can be used for maximum likelihood estimation and to find the distribution of the latent variables. However, it has been proven to be intractable [2]. 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]. We will explore both Gibbs sampling and variational methods in the following two subsections.

2.2.1. Gibbs

Explain Gibbs Sampling some...Note that the conditional distribution of β is independent of all variables except for η. Its update equation at step t:

$$p(\beta^{(t)}|\beta^{(t-1)}, \theta, z, w, \alpha, \eta) \propto p(\beta(t)|\beta^{(t-1)}, \eta) \sim Dir(\eta + c_{k,.,v})$$

Simerarily, we eaily find the update at step t for θ:

$$p(\theta^{(t)}|\theta^{(t-1)}, \beta, z, w, \alpha, \eta) \propto p(\theta(t)|\theta^{(t-1)}, \alpha) \sim Dir(\alpha + c_{k,d,.})$$

The update equation at step t for the topic assignments z do not follow such a simple form. Within each document d, the update equation for word n is as follows:

$$p(z^t_{d,n}|z^{(t-1)}_{-(d,n)}, \theta, \beta, w, \alpha, \eta) = \frac{p(z, w|\theta, \beta)}{p(z^t_{-(d,n)}, w|\theta, \beta)} \propto p(z, w|\theta, \beta) = p(z^{(t-1)}_{d,n}|\theta)p(w_{d,n}|\beta, z^{(t-1)}_{d,n})$$

Psuedo Code...

This update although simple, cheap, and accurate, converges slowly. The collapsed Gibbs sampler was created after noticing that the algorithm could be speeded up by integrating out θ from the Markov Chain's state space. The collapsed Gibbs sampler allows us to sample on a topic-by-topic basis. After some careful calculations [11], one can show that for word n in document d, the probability of word $w_{d,n}$ being assigned topic $z_{d,n} = k$ is

$$p(z^t_{d,n} = k|z^{(t-1)}_{-(d,n)}, \theta, \beta, w) = c^{-1} * \frac{(c^{-(d,n)}_{z_{d,n},d,.} + \alpha_{z_{d,n}}) * (c^{-(d,n)}_{z_{d,n},.,w_{d,n}} + \beta_{w_{d,n}})}{c^{-(d,n)}_{z_{d,n},.,.} + \sum_{v=1}^{V} \beta_v}$$

where c is the normalizing constant equal to

$$\sum_{k=1}^{K} \frac{(c^{-(d,n)}_{k,d,.} + \alpha_k) * (c^{-(d,n)}_{k,.,w_{d,n}} + \beta_{w_{d,n}})}{c^{-(d,n)}_{k,.,.} + \sum_{v=1}^{V} \beta_v}$$

Psuedo Code Here...

2.2.2. Variational Methods

2.2.3. Other Methods Here

Maybe...

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.

[11] Carpenter, Bob. (2010), “Integrating out Multinomial Parameters in Latent Dirichlet Allocation and Naive Bayes for Collapsed Gibbs Sampling.” *LingPipe*: carp@lingpipe.com.