Topic Model and Latent Dirichlet Allocation

Chunpai Wang

May 5, 2019

1 Topic Model and Latent Dirichlet Allocation

Here, we show an example of mean-field variational inference called Latent Dirichlet Allocation (LDA). In topic model, we assume each document is generated by a certain topic model, with the algorithm below:

Algorithm 1: Generating process of a Document; a word is represented as an one-hot vector w_n, and the total vocabulary size is V; a document is a sequence of N words denoted by $\mathbf{w} = (w_1, ..., w_N)$; a corpus is a collection of M document denoted by $\mathbf{D} = \{\mathbf{w}_1, ..., \mathbf{w}_M\}$

```
Draw N ~ Poisson(ξ); Decide on the number of words N the document will have.
Draw θ from the prior; Choose a topic mixture for the document, for example topic A has
  1/3 probability, topic B has 2/3 probability.
Draw β from the prior; Create K multinomial generators, and give each of them an ID from
  1 to K.
for each of the N words w_n do
    Choose a topic z_n from multinomial(θ); Pick a topic for current word, for example 1/3
      probability for topic A and 2/3 probability for topic B.
    Choose a word from p(w_n|z_n, β), a multinomial probability conditioned on the topic z_n;
      For example, if we choose the topic A, we might choose the word 1 with 1/4 probability
      in topic A, word 2 with 1/5 probability in topic A, etc.
end
```

Question: what is the β used for ? β_k is a vector, which represents the proportion of N total vocabularies in k^{th} topic. Thus, for each topic, the N vocabularies have different probability. We use θ_d to generate a topic k, then we choose the β_k which contains the proportion of each word to generate a word.

In details, we can build the generated model with some following assumptions

1. (Draw θ_d from the prior) we generate a topic distribution for a document. In this way, a document can have multiple topics. The prior θ is the distribution of the topic distribution. Note that, θ is a K-dimensional random variable. If the the prior is a Dirichlet distribution, the model is called Latent Dirichlet Allocation (LDA). The dimensionality K of the Dirichlet distribution (and thus the dimensionality of the topic variable z) is assumed known and fixed, which refers to the total number of possible topics.

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \cdots \theta_k^{\alpha_k - 1} \tag{1}$$

2. for each word, we select a topic from the topic distribution $z_{dn} \sim Multi(\theta_d)$, and then select a word for the document according to the probability of words given on the topic, $w_{dn} \sim Multi(\beta_{z_{dn}})$. β_k is the prior distribution of the word|topic distribution, which is a $k \times V$ matrix, where $\beta_{ij} = p(w^j = 1|z^i = 1)$ and V denotes by the total number of vocabularies. We assume the prior is a Dirichlet distribution as well, and η is the hyperparameter of Dirichlet distribution, which has V dimensionality as well.

3. Each document contains many topics, and we cannot directly see the topic distribution, thus, it is called "latent" Dirichlet allocation.

4. Another prior used in topic models is logistic normal.

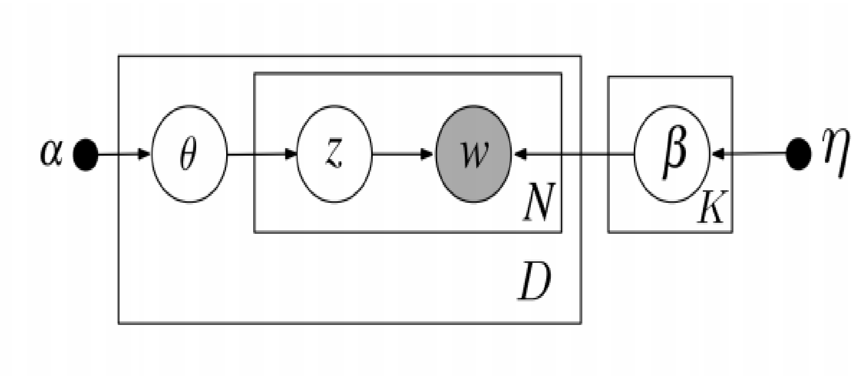

Figure 1: Graphical Model of LDA. The boxes are "plates" representing replicates. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document. D denotes the number of documents in a corpus; N denotes the number of words and corresponding topics in a document; There are three levels to the LDA representation. The parameter α, β are corpus-level parameters, assumed to be sampled once in the process of generating a corpus. The variables θ_d are document-level variables, sampled once per document. Finally, the variable z_{dn} and w_{dn} are word-level variables and are sampled once for each word in each document.

1.1 Some Probabilities

Given the paramters α and β, the joint distribution of a topic mixture θ, a set of N topics z, and a set of N words w is given by:

$$p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta) \tag{2}$$

where $p(z_n|\theta)$ is simply θ_i for the unique i such that $z_n^i = 1$. Integrating over θ and summing over z, we obtain the marginal distribution of a document:

$$p(\mathbf{w}|\alpha, \beta) = \int p(\theta|\alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n|\theta) p(w_n|z_n, \beta) \right) d\theta \tag{3}$$

Finally, taking the product of the marginal probabilities of single documents, we obtain the probability of a corpus:

$$p(\mathbf{D}|\alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn}, \beta) \right) d\theta_d \tag{4}$$

1.2 Gibbs Sampling for LDA

1.3 Variational Inference for LDA

We can use Gibbs-sampling to solve the LDA problem. However, we can also exploit the variational method. In LDA model, we care about the posterior distribution on hidden variables given the observed variable $p(\beta, \theta, \mathbf{z}|\mathbf{w}, \alpha, \eta)$,

$$p(\beta, \theta, \mathbf{z}|\mathbf{w}, \alpha, \eta) = \frac{p(\beta, \theta, \mathbf{z}, \mathbf{w}|\alpha, \eta)}{p(\mathbf{w}|\alpha, \eta)} \tag{5}$$

Directly calculating the probability is intractable, so we would instead solve $q(\beta, \theta, z)$, a close approximate estimation to the true posterior, which is achieved by approximate inference.

When we are doing the mean field approximation, we assume the variational approximation q over β, θ, z are independent. Thus we can use the fully factorized distribution:

$$q(\beta, \theta, z) = \prod_{k} q(\beta_k) \prod_{d} q(\theta_d) \prod_{n} q(z_{dn}) \tag{6}$$

where

- k denotes a topic
- d denotes a document
- n denotes a term in the document

Since the prior $\theta_d \sim Dirichlet(\alpha)$, $\beta_k \sim Dirichlet(\eta)$, then the approximated $q(\theta_d) = Dirichlet(\gamma)$ and $q(\beta_k) = Dirichlet(\lambda)$. Since $z_{dn} \sim Multi(\theta_d)$, then we assume $q(z_{dn} = k) = Multi(\phi_{dn}^k)$

References

[1] Eric Xing's Lecture https://www.cs.cmu.edu/~epxing/Class/10708-15/slides/lecture13-VI.pdf

[2] NIPS Tutorial on Variational Inference https://media.nips.cc/Conferences/2016/Slides/6199-Slides.pdf

[3] LDA Tutorial http://www.cnblogs.com/pinard/p/6831308.html
http://blog.echen.me/2011/08/22/introduction-to-latent-dirichlet-allocation/
https://bboalimoe.wordpress.com/2015/10/22/lda-%E6%80%BB%E7%BB%93-%E6%B5%85%E6%98%BE%E7%89%88%E6%9C%AC/