Brief Introduction to LDA

Xiaobao Wu
Tsinghua University
xiaobao.wu1996@gmail.com

1 Introduction

Latent Dirichlet Allocation(LDA) is the most widely used topic model. It was first proposed by(Blei et al., 2003). In the LDA model, one document covers several topics. Each topic is a distribution of words. Those words are relative to that topic. Like topic "Sports" may contain "football", "basketball", etc. From my personal point of view, the mathematics part of LDA is charming and elegant. In the next section, we will introduce the generative process of LDA. Then we will talk about the inference method of LDA.

2 Generative Process

1. For each topic, draw a word distribution, $\vec{\phi}_k \sim \mathrm{Dir}(\vec{\beta})$
2. For each document,
   (a) Draw a topic distribution $\vec{\theta}_m \sim \mathrm{Dir}(\vec{\alpha})$
   (b) For each word
      i. draw a topic $z_{m,n} \sim \mathrm{Mult}(\vec{\theta}_m)$
      ii. draw a word $w_{m,n} \sim \mathrm{Mult}(\vec{\phi}_k)$ where $k = z_{m.n}$

2.1 Conjugate Prior

Dirichlet distribution is a conjugate prior for multinomial distribution. Here we present a complete proof. Through Bayes' formula, we get

$$p(\vec{\theta}_m|\vec{z}_m) = \frac{p(\vec{z}_m|\vec{\theta}_m)p(\vec{\theta}_m|\vec{\alpha})}{p(\vec{z}_m|\vec{\alpha})}$$

where $\vec{z}_m$ is the topic propotion of all words in m-th document.

Now, we first consider the numerator $p(\vec{z}_m|\vec{\theta}_m)p(\vec{\theta}_m|\vec{\alpha})$:

$$p(\vec{z}_m|\vec{\theta}_m)p(\vec{\theta}_m|\vec{\alpha}) = \prod_{k=1}^{K} \theta_k^{n_{m,k}} \frac{1}{\Delta(\vec{\alpha})} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \quad (1)$$

$$= \frac{1}{\Delta(\vec{\alpha})} \prod_{k=1}^{K} \theta_k^{n_{m,k} + a_k - 1} \quad (2)$$

where $n_{m,k}$ means the number of words assigned to topic k in m-th document.

With the result above, we can easily get the normalization factor:

$$\begin{align} p(\vec{z}_m|\vec{\alpha}) &= \int \frac{1}{\Delta(\vec{\alpha})} \prod_{k=1}^{K} \theta_k^{n_{m,k}+a_k-1} d\vec{\theta}_m \tag{3} \\ &= \frac{1}{\Delta(\vec{\alpha})} \int \prod_{k=1}^{K} \theta_k^{n_{m,k}+a_k-1} d\vec{\theta}_m \tag{4} \\ &= \frac{\Delta(\vec{n}_m + \vec{\alpha})}{\Delta(\vec{\alpha})} \tag{5} \end{align}$$

Finally, it is proved:

$$\begin{align} p(\vec{\theta}_m|\vec{z}_m) &= \frac{p(\vec{z}_m|\vec{\theta}_m)p(\vec{\theta}_m|\vec{\alpha})}{p(\vec{z}_m|\vec{\alpha})} \tag{6} \\ &= \frac{1}{\Delta(\vec{n}_m + \vec{\alpha})} \prod_{k=1}^{K} \theta_k^{n_{m,k}+a_k-1} \tag{7} \\ &= \mathrm{Dir}(\vec{n}_m + \vec{\alpha}) \tag{8} \\ & \tag{9} \end{align}$$

For another conjugate, we consider the process to generate all words in corpus instead of words in a document. Similariy, we get

$$p(\vec{\phi}_k|\vec{w}) = \mathrm{Dir}(\vec{n}_k + \vec{\beta})$$

where $\vec{n}$ means the number of words assigned to topic k. where $\vec{n}_k = (n_k^{(1)}, n_k^{(2)}, \cdots, n_k^{(V)})$ means the number of words assigned to topic k.

3 Inference Process

Then we can use the collapsed Gibbs Sampling to inference the parameters in LDA. Gibbs sampling is a kind of Markov Chain Monte Carol (MCMC) method.

$$p(\vec{w}, \vec{z}|\vec{\alpha}, \vec{\beta}) = p(\vec{w}|\vec{z}, \vec{\beta})p(\vec{z}|\vec{\alpha}) \tag{10}$$

$$\begin{align} p(\vec{w}|\vec{z}, \vec{\beta}) &= \int p(\vec{w}|\vec{z}, \Phi)p(\Phi|\vec{\beta})d\Phi \tag{11} \\ &= \int \prod_{z=1}^{K} \frac{1}{\Delta(\vec{\beta})} \prod_{t=1}^{V} \phi_{z,t}^{n_z^{(t)}+\beta_t-1} d\vec{\phi}_z \tag{12} \\ &= \prod_{z=1}^{K} \frac{\Delta(\vec{n}_z + \vec{\beta})}{\Delta(\vec{\beta})} \tag{13} \end{align}$$

$$\begin{align} p(\vec{z}|\vec{\alpha}) &= \int p(\vec{z}|\Theta)p(\Theta|\vec{\alpha})d\Theta \tag{14} \\ &= \int \prod_{m=1}^{M} \frac{1}{\Delta(\alpha)} \prod_{k=1}^{K} \theta_{m,k}^{n_m^{(k)}+\alpha_k-1} d\vec{\theta}_m \tag{15} \\ &= \prod_{m=1}^{M} \frac{\Delta(\vec{n}_m + \vec{\alpha})}{\Delta(\vec{\alpha})} \tag{16} \end{align}$$

$$
\begin{aligned}
p(\vec{z},\vec{w}|\vec{\alpha},\vec{\beta}) &= p(\vec{w}|\vec{z},\vec{\beta})p(\vec{z}|\vec{\alpha}) \qquad (17)\\
&= \prod_{z=1}^{K} \frac{\Delta(\vec{n}_z+\vec{\beta})}{\Delta(\vec{\beta})} \cdot \prod_{m=1}^{M} \frac{\Delta(\vec{n}_m+\vec{\alpha})}{\Delta(\vec{\alpha})} \qquad (18)
\end{aligned}
$$

Here we condider the word $i=(m,n)$, the n-th word in the m-th document. Note that $\vec{w}=\{w_i=t,\vec{w}_{\neg i}\}$ and $\vec{z}=\{z_i=k,\vec{z}_{\neg i}\}$. Hyperparameters $\vec{\alpha}$ and $\vec{\beta}$ are omitted.

$$
\begin{aligned}
p(z_i=k|\vec{z}_{\neg i},\vec{w}) &= \frac{p(\vec{w},\vec{z})}{p(\vec{w},\vec{z}_{\neg i})} \qquad (19)\\
&= \frac{p(\vec{w}|\vec{z})}{p(\vec{w}_{\neg i}|\vec{z}_{\neg i})p(w_i)} \cdot \frac{p(\vec{z})}{p(\vec{z}_{\neg i})} \qquad (20)\\
&\propto \frac{p(\vec{w}|\vec{z})}{p(\vec{w}_{\neg i}|\vec{z}_{\neg i})} \cdot \frac{p(\vec{z})}{p(\vec{z}_{\neg i})} \qquad (21)\\
&= \frac{\Gamma(n_k^{(t)}+\beta_t)\,\Gamma(\sum_{t=1}^{V} n_{k,\neg i}^{(t)}+\beta_t)}{\Gamma(n_{k,\neg i}^{(t)}+\beta_t)\,\Gamma(\sum_{t=1}^{V} n_k^{(t)}+\beta_t)} \cdot \frac{\Gamma(n_m^k+\alpha_k)\,\Gamma(\sum_{k=1}^{K} n_{m,\neg i}^k+\alpha_k)}{\Gamma(n_{m,\neg i}^k+\alpha_k)\,\Gamma(\sum_{k=1}^{K} n_m^k+\alpha_k)} \qquad (22)\\
&= \frac{\Gamma(n_{k,\neg i}^{(t)}+\beta_t+1)\,\Gamma(\sum_{t=1}^{V} n_{k,\neg i}^{(t)}+\beta_t)}{\Gamma(n_{k,\neg i}^{(t)}+\beta_t)\,\Gamma([\sum_{t=1}^{V} n_{k,\neg i}^{(t)}+\beta_t]+1)} \cdot \frac{\Gamma(n_{m,\neg i}^k+\alpha_k+1)\,\Gamma(\sum_{k=1}^{K} n_{m,\neg i}^k+\alpha_k)}{\Gamma(n_{m,\neg i}^k+\alpha_k)\,\Gamma([\sum_{k=1}^{K} n_m^k+\alpha_k]-1+1)} \qquad (23)\\
&= \frac{n_{k,\neg i}^{(t)}+\beta_t}{\sum_{t=1}^{V} n_{k,\neg i}^{(t)}+\beta_t} \cdot \frac{n_{m,\neg i}^{(t)}+\alpha_k}{[\sum_{k=1}^{K} n_m^k+\alpha_k]-1} \qquad (24)\\
&\propto \frac{n_{k,\neg i}^{(t)}+\beta_t}{\sum_{t=1}^{V} n_{k,\neg i}^{(t)}+\beta_t}(n_{m,\neg i}^k+\alpha_k) \qquad (25)
\end{aligned}
$$

$$
p(\theta_m|\vec{z}_m,\vec{\alpha}) = Dir(\vec{\theta}_m|\vec{n}_m+\vec{\alpha}) \qquad (26)
$$

$$
p(\phi_k|\vec{z},\vec{w},\vec{\beta}) = Dir(\vec{\phi}_k|\vec{n}_k+\vec{\beta}) \qquad (27)
$$

$$
\begin{aligned}
\vec{\theta}_{m,k} &= \frac{n_m^{(k)}+\alpha_k}{\sum_{k=1}^{K} n_m^{(k)}+\alpha_k} \qquad (28)\\
\vec{\phi}_{k,t} &= \frac{n_k^{(t)}+\beta_t}{\sum_{t=1}^{V} n_k^{(t)}+\beta_t} \qquad (29)
\end{aligned}
$$

where we use $\Gamma(x)=(x-1)\Gamma(x-1)$ and $\frac{\Gamma(x+m)}{\Gamma(x)}=\prod_{i=1}^{m}(x+i-1)$.

References

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