Gibbs Sampling on LDA

Jingbo Xia

Huazhong Agricultural University

xiajingbo.math@gmail.com

April 17, 2019

Table of contents I

Outline

Resources and Literature Reading

Literature that we mainly used:

1. LDA by Blei, Andrew Ng, and Michael Jordan, 2013 [1]. A work proposed LDA and solved it using VI.
2. Gibbs on LDA by Griffith [2]. One year after LDA. Solve LDA using Gibbs sampling.
3. FastLDA, KDD 09 paper [3]. We re-used the notations shown in this paper.
4. A more recent paper [4]. A little bit more details.

---

[1] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." Journal of machine Learning research 3, no. Jan (2003): 993-1022.

[2] Griffiths, Thomas L., and Mark Steyvers. "Finding scientific topics." Proceedings of the National academy of Sciences 101, no. suppl 1 (2004): 5228-5235.

[3] Porteous, Ian, David Newman, Alexander Ihler, Arthur Asuncion, Padhraic Smyth, and Max Welling. "Fast collapsed gibbs sampling for latent dirichlet allocation." In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 569-577. ACM, 2008.

[4] Darling, William M. "A theoretical and practical implementation tutorial on topic modeling and gibbs sampling." In Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies, pp. 642-647. 2011.

Outline

Gibbs Sampling

To sample X from the joint distribution $p(X) = p(X_1, \cdots, X_N)$, where there is no closed form solution for $p(X)$, but a representation for the conditional distributions is available, using Gibbs Sampling one would perform the following:

step 1: Randomly initialize X_i

step 2: For $t = 1, \cdots, T$

Step 2.1 $X_1^{t+1} \sim p\left(X_1 | X_2^{(t)}, X_3^{(t)}, \ldots, X_m^{(t)}\right)$

Step 2.2 $X_2^{t+1} \sim p\left(X_2 | X_1^{(t+1)}, X_3^{(t)}, \ldots, X_m^{(t)}\right)$

$\cdots$

Step 2.n $\cdots$

$\cdots$

Step 2.N $X_M^{t+1} \sim p\left(X_M | X_1^{(t+1)}, X_2^{(t+1)}, \ldots, X_{M-1}^{(t+1)}\right)$

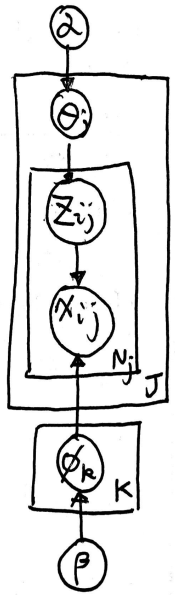
Outline

LDA

LDA: For each of N_j words in the j-th document

1. Sample $z_{i,j} \sim Multi(\theta_j)$. $(i = 1, 2, \cdots, N_j,\ j = 1, 2, \cdots, J)$. (Pleae note that $\theta_j \in \mathbb{R}^K$, and $\theta_{j,k} \in [0, 1]$, $\sum_{j=1}^{J} \theta_{j,k} = 1$.)
2. Sample $x_{i,j} \sim Multi(\phi_{z_{ij}})$. $(z_{ij} = 1, 2, \cdots, K)$

(Pleae note that $\phi_k \in \mathbb{R}^L$, and $\phi_{k,l} \in [0, 1]$, $\sum_{l=1}^{L} \phi_{k,l} = 1$.)



Observation: $X = \{x_{ij}\}$

Task: Latent Topic: $Z = \{z_{ij}\}$

Mixing Propotion: θ_j

Topic:ϕ_k $(k = 1, 2, \cdots, K)$

In addition, the vocabulary set $W = \{w_l\}$, $l = 1, 2, \cdots, L$.

Outline

Notations:

$N_{wkj} = \#\{i \mid x_{ij} = w, z_{ij} = k\}$

$N_{wk} = \sum_j N_{wkj}$

$N_{kj} = \sum_w N_{wkj}$

The core idea of Gibbs sampling is to sample $P(z_{ij} = k \mid Z^{\neg ij}, X, \alpha, \beta)$ which is re-written as $P(z_{ij}|Z^{\neg ij}, X, \alpha, \beta)$. We then have:

$$
\begin{aligned}
P(z_{ij}|Z^{\neg ij}, X, \alpha, \beta) &= \frac{P(z_{ij}, Z^{\neg ij}, X|\alpha, \beta)}{P(Z^{\neg ij}, X|\alpha, \beta)} \\
&\propto P(z_{ij}, Z^{\neg ij}, X|\alpha, \beta) \\
&= P(Z, X|\alpha, \beta) \\
\text{(Tricks again...)} \quad &= \int_0^1 \int_0^1 P(Z, X, \theta, \phi|\alpha, \beta) d\theta d\phi \\
&= \int_0^1 \int_0^1 P(Z|\theta) P(X|\phi_Z) P(\theta|\alpha) P(\phi|\beta) d\theta d\phi \\
&= \underbrace{\int_0^1 P(Z|\theta) P(\theta|\alpha) d\theta}_{(*)} \underbrace{\int_0^1 P(X|\phi_Z) P(\phi|\beta) d\phi}_{(**)}
\end{aligned}
\tag{1}
$$

Gibbs on LDA

We calculate (*) part of equation 1 first,

$$
\begin{aligned}
(*) \quad &= \int_{\mathbf{0}}^{\mathbf{1}} P(Z|\theta)P(\theta|\alpha)d\theta \\
&= \int_{\theta_1 \mathbf{0}}^{\mathbf{1}} \cdots \int_{\theta_J \mathbf{0}}^{\mathbf{1}} \prod_{i=1}^{N_j} \prod_{j=1}^{J} P(z_{ij}|\theta_j) \cdot \prod_{j=1}^{J} P(\theta_j|\alpha) d\theta_1 d\theta_2 \cdots d\theta_J \\
&= \prod_{j=1}^{J} [\int_{\mathbf{0}}^{\mathbf{1}} \prod_{i=1}^{N_j} P(z_{ij}|\theta_j) \cdot P(\theta_j|\alpha) d\theta_j] \qquad \color{red}{\text{Using equation}}\ (3, 4, 11) \\
&= \prod_{j=1}^{J} \frac{1}{B(\alpha)} \int_{\theta_{j,k} \in [0,1], \sum_k \theta_{j,k}=1} \prod_{k=1}^{K} \theta_{j,k}^{N_{kj}+\alpha_k-1} d\theta_j \\
&= \prod_{j=1}^{J} \frac{1}{B(\alpha)} B(N_{\cdot j} + \alpha).
\end{aligned} \tag{2}
$$

Here, $N_{\cdot j} + \alpha = (N_{1j}, \cdots, N_{Kj})$, and the multivariate Beta function $B(\alpha) = \int_{\mathbf{0}}^{\mathbf{1}} \prod_{k=1} x_k^{\alpha_k - 1} d\,\mathbf{x}$ also equals to $\frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{K} \alpha_k)}$. I put more details in the Appendix, please check them.

Gibbs on LDA

Note:

The main tricks used in the equation (2) are the counting of $\prod_{i=1}^{N_j} P(z_{ij}|\theta_j)$ and summary term of $B(N_{\cdot j} + \alpha)$.

Specifically, we have:

$$\prod_{i=1}^{N_j} P(z_{ij}|\theta_j) = \prod_{k=1}^{K} \theta_{j,k}^{N_{kj}} \tag{3}$$

$$B(N_{\cdot j} + \alpha) = \int_{\mathbf{0}}^{\mathbf{1}} \prod_{k=1}^{K} \theta_{j,k}^{N_{kj}+\alpha_k-1} d\theta_j \tag{4}$$

Gibbs on LDA

Similarly, we calculate (**) part of equation (1) subsequently, and we have

$$
\begin{aligned}
(**) \quad &= \int_{\mathbf{0}}^{\mathbf{1}} P(X|\phi_Z)P(\phi|\beta)d\phi \\
&= \int_{\phi_1 \mathbf{0}}^{\mathbf{1}} \cdots \int_{\phi_K \mathbf{0}}^{\mathbf{1}} \prod_{i=1}^{N_j} \prod_{j=1}^{J} P(x_{ij}|\phi_{z_{ij}}) \cdot \prod_{k=1}^{K} P(\phi_k|\beta) d\phi_1 d\phi_2 \cdots d\phi_K \\
&= \prod_{k=1}^{K} [\int_{\mathbf{0}}^{\mathbf{1}} \prod_{l=1}^{L} \phi_{k,l}^{N_{w_l k}} \cdot P(\phi_k|\beta) d\phi_k] \\
&= \prod_{k=1}^{K} \frac{1}{B(\beta)} \int_{\mathbf{0}}^{\mathbf{1}} \prod_{l=1}^{L} \phi_{k,l}^{N_{w_l k}+\beta_l-1} d\phi_k \\
&= \prod_{k=1}^{K} \frac{1}{B(\beta)} B(N_{\cdot k} + \beta),
\end{aligned} \tag{5}
$$

where $N_{\cdot k} + \beta = (N_{w_1 k} + \beta_1, \cdots, N_{w_L k} + \beta_1)$

Gibbs on LDA

To sum up equation (2) and equation (5), we have

$$P(Z, X|\alpha, \beta) = \prod_{j=1}^{J} \frac{B(N_{\cdot j} + \alpha)}{B(\alpha)} \cdot \prod_{k=1}^{K} \frac{B(N_{\cdot k} + \beta)}{B(\beta)}. \tag{6}$$

Similarly,

$$P(Z^{\neg \mathrm{ij}}, X|\alpha, \beta) = \prod_{j=1}^{J} \frac{B(N_{\cdot j}^{\neg \mathrm{ij}} + \alpha)}{B(\alpha)} \cdot \prod_{k=1}^{K} \frac{B(N_{\cdot k}^{\neg \mathrm{ij}} + \beta)}{B(\beta)}, \tag{7}$$

where $Z^{\neg \mathrm{ij}} = Z - \{z_{\mathrm{ij}}\}$, $N_{\cdot k}^{\neg \mathrm{ij}} + \beta = (N_{w_1 k}^{\neg \mathrm{ij}} + \beta_1, \cdots, N_{w_L k}^{\neg \mathrm{ij}} + \beta_1)$.

Gibbs on LDA

$$
\begin{aligned}
&P(z_{\mathrm{ij}} = \mathrm{k} | Z^{\neg \mathrm{ij}}, X, \alpha, \beta) = \frac{P(Z,X|\alpha,\beta)}{P(Z^{\neg \mathrm{ij}},X|\alpha,\beta)} \\
&= \prod_{j=1}^{J} \frac{B(N_{\cdot j}+\alpha)}{B(N_{\cdot j}^{\neg \mathrm{ij}}+\alpha)} \cdot \prod_{k=1}^{K} \frac{B(N_{\cdot k}+\beta)}{B(N_{\cdot k}^{\neg \mathrm{ij}}+\beta)} \\
&= \prod_{j=1}^{J} \frac{\prod_{k=1}^{K} \Gamma(N_{kj}+\alpha_k)}{\prod_{k=1}^{K} \Gamma(N_{kj}^{\neg \mathrm{ij}}+\alpha_k)} \frac{\Gamma(\sum_{k=1}^{K} N_{kj}^{\neg \mathrm{ij}}+\alpha_k)}{\Gamma(\sum_{k=1}^{K} N_{kj}+\alpha_k)} \cdot \prod_{k=1}^{K} \frac{\prod_{l=1}^{L} \Gamma(N_{w_l k}+\beta_l)}{\prod_{l=1}^{L} \Gamma(N_{w_l k}^{\neg \mathrm{ij}}+\beta_l)} \frac{\Gamma(\sum_{l=1}^{L} N_{w_l k}^{\neg \mathrm{ij}}+\beta_l)}{\Gamma(\sum_{l=1}^{L} N_{w_l k}+\beta_l)} \\
&= \prod_{j=1}^{J} \frac{\prod_{k=1}^{K} \Gamma(N_{kj}+\alpha_k)}{\prod_{k=1}^{K} \Gamma(N_{kj}^{\neg \mathrm{ij}}+\alpha_k)} \cdot \prod_{j=1}^{J} \frac{\Gamma(\sum_{k=1}^{K} N_{kj}^{\neg \mathrm{ij}}+\alpha_k)}{\Gamma(\sum_{k=1}^{K} N_{kj}+\alpha_k)} \cdot \prod_{k=1}^{K} \frac{\prod_{l=1}^{L} \Gamma(N_{w_l k}+\beta_l)}{\prod_{l=1}^{L} \Gamma(N_{w_l k}^{\neg \mathrm{ij}}+\beta_l)} \cdot \prod_{k=1}^{K} \frac{\Gamma(\sum_{l=1}^{L} N_{w_l k}^{\neg \mathrm{ij}} \cdots}{\Gamma(\sum_{l=1}^{L} N_{w_l k} \cdots} \\
&= \frac{\Gamma(N_{\mathrm{kj}}+\alpha_{\mathrm{k}})}{\Gamma(N_{\mathrm{kj}}^{\neg \mathrm{ij}}+\alpha_{\mathrm{k}})} \cdot \frac{\Gamma(N_{\mathrm{j}}^{\neg \mathrm{ij}}+\sum_{k=1}^{K} \alpha_k)}{\Gamma(N_{\mathrm{j}}+\sum_{k=1}^{K} \alpha_k)} \cdot \frac{\prod_{l=1}^{L} \Gamma(N_{w_l \mathrm{k}}+\beta_l)}{\prod_{l=1}^{L} \Gamma(N_{w_l \mathrm{k}}^{\neg \mathrm{ij}}+\beta_l)} \cdot \frac{\Gamma(\sum_{l=1}^{L} N_{w_l \mathrm{k}}^{\neg \mathrm{ij}}+\beta_l)}{\Gamma(\sum_{l=1}^{L} N_{w_l \mathrm{k}}+\beta_l)} \\
&= \Gamma(N_{\mathrm{kj}}^{\neg \mathrm{ij}} + \alpha_{\mathrm{k}}) \cdot Constant \cdot \frac{\Gamma(N_{w_{\mathrm{l}} \mathrm{k}}+\beta_l)}{\Gamma(N_{w_{\mathrm{l}} \mathrm{k}}^{\neg \mathrm{ij}}+\beta_l)} \cdot \frac{\Gamma(\sum_{l=1}^{L} N_{w_l \mathrm{k}}^{\neg \mathrm{ij}}+\beta_l)}{\Gamma(\sum_{l=1}^{L} N_{w_l \mathrm{k}}+\beta_l)} \quad (w_{\mathrm{l}} = x_{\mathrm{ij}}) \\
&= \Gamma(N_{\mathrm{kj}}^{\neg \mathrm{ij}} + \alpha_{\mathrm{k}}) \cdot Constant \cdot \Gamma(N_{w_{\mathrm{l}} \mathrm{k}}^{\neg \mathrm{ij}} + \beta_l) \cdot \frac{1}{\Gamma(\sum_{l=1}^{L} N_{w_l \mathrm{k}}^{\neg \mathrm{ij}}+\beta_l)} \\
&\propto \Gamma(N_{\mathrm{kj}}^{\neg \mathrm{ij}} + \alpha_{\mathrm{k}}) \cdot \frac{\Gamma(N_{w_{\mathrm{l}} \mathrm{k}}^{\neg \mathrm{ij}}+\beta_l)}{\Gamma(\sum_{l=1}^{L} N_{w_l \mathrm{k}}^{\neg \mathrm{ij}}+\beta_l)} \\
&:= a_{\mathrm{kj}} \cdot b_{w_{\mathrm{l}} \mathrm{k}}
\end{aligned} \tag{8}
$$

Gibbs on LDA

Use tcolorbox and columns to better my slides

Geez! It is done, finally, like this way...
But notice that $P(Z_{\mathrm{ij}} = \mathrm{k} | Z^{\neg \mathrm{ij}}, X, \alpha, \beta)$ is proportional to $a_{\mathrm{kj}} \cdot b_{w_{\mathrm{l}}\mathrm{k}}$, we need to scale it to $0 \sim 1$.

Let $\Delta = \sum_{k}^{K} a_{k\mathrm{j}} b_{w_{\mathrm{l}} k}$ is the normalization constant , then we have

$$P(z_{\mathrm{ij}} = \mathrm{k} | Z^{\neg \mathrm{ij}}, X, \alpha, \beta) = (a_{\mathrm{kj}} \cdot b_{w_{\mathrm{l}}\mathrm{k}}) / \Delta.$$

For simplicity, finally we can remove the mark in red:

$$P(z_{ij} = k | Z^{\neg ij}, X, \alpha, \beta) = (a_{kj} \cdot b_{w_l k}) / \Delta, \text{where } x_{ij} = w_l. \tag{9}$$

Gibbs on LDA

Equation (9) provides the prob for each k ($k = 1, 2, \cdots, K$) that z_{ij} that might be. Then by following the distribution, z_{ij} is sampled for a fixed i, j pair in the two-layer iteration of $i : 1 \to N_j$, $j : 1 \to J$.

In each round, for a given z_{ij} (when the observation $x_{ij} = w_l$), N_{kj}, N_k, and $N_{w_l k}$ are updated, and the parameter $\phi_{w_l,k}$ and $\theta_{k,j}$ are uploaded by the following rule:

$$\hat{\phi}_{k,l} = \frac{N_{w_l k} + \beta_l}{\sum_{l'=1}^{L} N_{w_{l'} k} + \beta_{l'}} = \frac{N_{w_l k} + \beta_l}{N_{w_k} + |\beta|},$$

$$\hat{\theta}_{j,k} = \frac{N_{kj} + \alpha_k}{\sum_{j'=1}^{J} N_{kj'} + \alpha_{k'}} = \frac{N_{kj} + \alpha_k}{N_k + |\alpha|},$$

where $|\cdot|$ is to sum up the vector terms.

Outline

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x}\, dx.$$

$$\begin{aligned}
\Gamma(z+1) &= \int_0^\infty x^{z} e^{-x}\, dx \\
&= \Big[-x^z e^{-x} \Big]_0^\infty + \int_0^\infty z x^{z-1} e^{-x}\, dx \\
&= \lim_{x\to\infty} (-x^z e^{-x}) - (0 e^{-0}) + z \int_0^\infty x^{z-1} e^{-x}\, dx
\end{aligned}$$

Recognizing that $-x^z e^{-x} \to 0$ as $x \to \infty$,

$$\Gamma(z+1) = z \int_0^\infty x^{z-1} e^{-x}\, dx = z\Gamma(z) \quad .$$

Note: $\Gamma(n) = 1 \cdot 2 \cdot 3 \cdots (n-1) = (n-1)!$

[5] https://en.wikipedia.org/wiki/Gamma_function

Beta function: $\mathrm{B}(x,y)=\int_0^1 t^{x-1}(1-t)^{y-1}\,dt$, for Re $x>0$, Re $y>0$.

Multivariate Beta function: $\mathrm{B}(\alpha_1,\alpha_2,\ldots\alpha_n)=\dfrac{\Gamma(\alpha_1)\,\Gamma(\alpha_2)\cdots\Gamma(\alpha_n)}{\Gamma(\alpha_1+\alpha_2+\cdots+\alpha_n)}$.

The general definition of multivariate Beta function comes from a property of the Beta function, $\mathrm{B}(x,y)=\dfrac{\Gamma(x)\,\Gamma(y)}{\Gamma(x+y)}$.

Relationship between gamma function and beta function

$\Gamma(x)\Gamma(y)=\int_{u=0}^{\infty}e^{-u}u^{x-1}du\cdot\int_{v=0}^{\infty}e^{-v}v^{y-1}dv$
$=\int_{v=0}^{\infty}\int_{u=0}^{\infty}\;e^{-u-v}u^{x-1}v^{y-1}\,du\,dv$

$$\begin{aligned}\Gamma(x)\Gamma(y)\quad &=\int_{z=0}^{\infty}\int_{t=0}^{1}e^{-z}(zt)^{x-1}(z(1-t))^{y-1}\big|J(z,t)\big|dtdz\\ &=\int_{z=0}^{\infty}\int_{t=0}^{1}e^{-z}(zt)^{x-1}(z(1-t))^{y-1}zdtdz\\ &=\int_{z=0}^{\infty}e^{-z}z^{x+y-1}dz\cdot\int_{t=0}^{1}t^{x-1}(1-t)^{y-1}dt\\ &=\Gamma(x+y)\mathrm{B}(x,y).\end{aligned}\tag{10}$$

Dirichlet distribution [7]

$X \sim Dirichlet(\boldsymbol{\alpha})$:

$$f(x_1, \ldots, x_K | \alpha_1, \ldots, \alpha_K) = \frac{1}{\mathrm{B}(\boldsymbol{\alpha})} \prod_{i=1}^{K} x_i^{\alpha_i - 1},$$

where $\sum_{i=1}^{K} x_i = 1$, $x_i \geq 0$ for all $i \in [1, K]$,
$\mathrm{B}(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}, \qquad \boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K).$

[7] https://en.wikipedia.org/wiki/Dirichlet_distribution

Multinomial distribution: $X \sim \text{Multi}(P)$

$$P(X_1 = m_1, X_2 = m_2, \cdots, X_N = m_N) = \frac{(\sum_{n=1}^{N} m_n)!}{\prod_{n=1}^{N} m_n!} p_1^{m_1} p_2^{m_2} \cdots p_N^{m_N}.$$

Categorical distribution: $X \sim \text{Multi}(P)$
By default, we also denote it in the same way, but actually some restriction is set here: ($m_n = 0$ or 1, and $\sum_{n=1}^{N} m_n = 1$).
$P(X_k = 1, X_{\neg k} = 0) = p_k$.

Note: In LDA, $Z_{ij} \sim \text{Multi}(\theta_j)$, and we have:

$$P(z_{ij} = k | \theta_j) = P(z_{ij,k} = 1, z_{ij,\neg k=0}) = \theta_{j,k} \tag{11}$$

Acknowledgement

Thank students who attended the seminar for the related discussion.

April 17, 2019