A supplment material to understand Latent Dirichlet Allocation

Saki Kuzushima

2019-05-07

The Latent Dirichlet Allocation model is very popular among the political scientists, and many people have summarized the model and the inference. However, understanding LDA requires some basis. This material aims to provide essential definitions and theorems to understand LDA and their variants. First section sketches the model and the infernece of LDA, and the second section describes some definitions and theorems essential to understand the model and the vaious inferences.

Part I

1. Latent Dirichlet Allocation model(s)

1.1 The original model

We firstly define the key terms used in the model. - Word: an element of a vocabulary. - Vocabulary: a set of V unique words appear in the entire corpus. - Document: a sequence of N words, denoted by $d = (w_1, w_2, , ..w_N)$ - Corpus: a set of M documents, denoted by $D = (d_1, d_2...d_M)$ - Topic: a distribution of words. We assume that the number of topics is K.

LDA is the following generating probabilistic model.

- For each document, $d_j = (w_{j,1}, w_{j,2}, , ..w_{j,N}), j \in [1...M]$,
  - $\theta_j \sim Dirichlet(\alpha)$
  - For each word, $w_n, n \in [1...N]$
    - $z_{j,n} \sim Mutlnimoial(\theta_j)$
    - $w_{j,n} \sim Multinomial(\phi_{z_{j,n}})$

For instance, we have three topics on a news article: sports, politics, and technology ($K = 3$). θ_j, which is the distribution of these topics over the document d_j was drawn from the Dirichlet distribution. We assume that the sampled $\theta_j = [\frac{1}{3}, \frac{1}{4}, \frac{5}{12}]$, where the order of topics are [sports, politics, technology]. Substantively, this means that the document d_j is a mixture of 1/3 "sports", 1/4 "poliitcs", and 5/12 "technology".

Next, we draw the topic of each word $z_{j,n}$ from $Multinomial(\theta)$. We assume this happens to be "politics." Then, we sample each word $w_{j,n}$ from $Multinomial(\phi_{zj,n})$.

Assuming that the documents and the words are independent, $d_j \perp\!\!\!\perp d_{j'}$ and $w_{j,n} \perp\!\!\!\perp w_{j,n'}$, we can write the joint distribution of W, Z, Θ given α, Φ in the following way.

$$p(W, Z, \Theta|\alpha, \Phi) = \prod_{j=1}^{M} p(\theta_j|\alpha) \prod_{j=1}^{K} p(z_{j,n}|\theta_j) p(w_{j,n}|\phi_{z_{j,n}}) \tag{1}$$

1.2 Smoothed LDA

Alternatively, we can assign a prior to ϕ. This model is called smoothed LDA.

- For each topic, $\phi_i \sim Dirichlet(\beta), i \in [1, ..., K]$
- For each document, $d_j = (w_{j,1}, w_{j,2}, , ..w_{j,N}), j \in [1...M]$,
  - $\theta_j \sim Dirichlet(\alpha)$
  - For each word, $w_n, n \in [1...N]$
    - $z_{j,n} \sim Mutlnimoial(\theta_j)$
    - $w_{j,n} \sim Multinomial(\phi_{z_{j,n}})$

In this model, we can write the joint distribution of W, Z, Θ given α, β.

$$p(W, Z, \Theta|\alpha, \beta) = \prod_{i=1}^{K} p(\phi_j|\beta) \prod_{j=1}^{M} p(\theta_j|\alpha) \prod_{j=1}^{K} p(z_{j,n}|\theta_j) p(w_{j,n}|\phi_{z_{j,n}}) \tag{2}$$

2. Inference

2.1 Variational Inference (original method by Blei et al.)

The original inference method uses variational inference. Variational inference essentially approximate the intractable posterior with another distribution that varies across some parameters, and tune the parameters to minimize the difference between the posterior and the approximate distribution. The difference of the posterior and the approximate distributions are measured by the Kullback-Leibler divergence.

For the two pdf p and q, which are defined on the same probability space, the Kullback-Leibler divergence between P and Q is

$$KL(g||p) = -\mathbb{E}_g\Big(\log(\frac{p(\theta|y)}{g(\theta)})\Big) = -\int \log(\frac{p(\theta|y)}{g(\theta)}) g(\theta) d\theta$$

for more information about KL divergence, this blog is helpful

Part II

1. The Dirichlet is a conjugate prior of the multinomial

The Dirichelt distribution and the multinomial distribution are the key distributions for the latent Dirichlet allocation. First we review the key features of the two distributions, and confirm that the Dirichlet distribution is a conjugate prior to the multinomial distribution.

The multinomial distribution has the following pmf. If

$$X \sim Multinomial(\theta, n)$$

where $X = (X_1, ..., X_n)$, $\theta = (\theta_1, ...\theta_k)$. Then,

$$p(x|\theta, n) = \frac{n!}{\prod_{i=1}^{K} x_i!} \prod_{i=1}^{K} \theta_i^{x_i} \tag{3}$$

where $x = (x_1, ..., x_n)$, and $\sum_{i=1}^{K} x_i = n$.

The Dirichlet distribution has the following pdf. If

$$\theta \sim Dirichlet(\alpha)$$

Then,

$$p(\theta|\alpha) = \frac{1}{B(\alpha)} \prod_{i=1}^{K} \theta_i^{\alpha_i - 1} \tag{4}$$

where $\sum_{i=1}^{K} \theta_i = 1$ and $\theta_i \geq 0, \forall i$. (i.e. θ belongs to the $K-1$ simplex.)

We can intuitively think how these distributions are related. We can think of the multinomial distribution as a result of n consecutive rolls of a die, where $X = (X_1, ..., X_6)$ is the number of each face showing up during the n rolls. Then, the Dirichlet distribution is the distribution over the probability of each face showing up in each roll. (If a die is fair, then $\theta = (1/6, ...1/6)$).

We can mathmatically show that the Dirichlet distribution is a conjugate prior of the multinomial distribution. The posterior of θ is

$$p(\theta|X, \alpha) \propto p(X|\theta, \alpha)p(\theta|\alpha)$$

Then, the RHS is

$$\begin{aligned} & p(X|\theta, \alpha)p(\theta|\alpha) \\ &= \frac{n!}{\prod_{i=1}^{K} x_i!} \prod_{i=1}^{K} \theta_i^{x_i} \frac{1}{B(\alpha)} \prod_{i=1}^{K} \theta_i^{\alpha_i - 1} \\ &\propto \prod_{i=1}^{K} \theta_i^{x_i + \alpha_i - 1} \end{aligned}$$

This is the kernel of the Dirichlet pdf. So we find that the posterior is also the Dirichlet distribution. By adjusting the constant term, we find that

$$p(\theta|X) = \frac{1}{B(\alpha + x)} \prod_{i=1}^{K} \theta_i^{x_i + \alpha_i - 1} \tag{5}$$

Therefore, the posterior distribution of θ is

$$\theta^{post} \sim Dirichlet(\alpha + x)$$

1.1 The relationship of the kernel of Dirichlet and Beta and Gamma functions

It is also helpful to see the integral of the kernel of Dirichlet is a Beta function, and therefore a ratio of gamma functions. This equation is useful in the derivation of the collapsed gibbs sampler.

$$
\begin{aligned}
& p(\theta|\alpha) = \frac{1}{B(\alpha)} \prod_{i=1}^{K} \theta_i^{\alpha_i - 1} \\
\iff & 1 = \int_\theta p(\theta|\alpha) d\theta = \int_\theta \frac{1}{B(\alpha)} \prod_{i=1}^{K} \theta_i^{\alpha_i - 1} d\theta \\
\iff & B(\alpha) = \int_\theta \prod_{i=1}^{K} \theta_i^{\alpha_i - 1} d\theta \\
\iff & \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{K} \alpha_i)} = \int_\theta \prod_{i=1}^{K} \theta_i^{\alpha_i - 1} d\theta
\end{aligned}
$$

2. The Dirichlet distribution belongs to the exponential family

The fact that the Dirichlet distribution belongs to the exponential family is useful. While the expression of the Dirichlet in terms of the exponential family is widely available, the derivation is not. The following sketches the conversion from the usual expression of the pdf of Dirichlet distribution to the form in terms of the exponential family.

The pdf of the Dirichlet distribution is

$$
p(\theta|\alpha) = \frac{1}{B(\alpha)} \prod_{i=1}^{K} \theta_i^{\alpha_i - 1} \tag{6}
$$

where $\theta = [\theta_1 .... \theta_K]^\intercal$, and $\sum_{i=1}^{K} \theta_i = 1$, $\theta_i \geq 0, \forall i \in \{1, 2, ..., K\}$.

The exponential family has the following general pdf

$$
p(\theta|\eta) = h(\theta) \exp(\eta^\intercal t(\theta) - A(\eta)) \tag{7}
$$

where $t(\theta)$ is the sufficient statistic, η is called the natural parameter, $A(\eta)$ is the log normalization factor, and $h(x)$ is the base measure.

It is known that the Dirichlet distribution belongs to the exponential family, and the parameters are

- The natural parameter: $\eta = \alpha = [\alpha_1 ... \alpha_K]^\intercal$
- The sufficient statistic: $t(\theta) = \log \theta = \log[\theta_1, ... \theta_K]^\intercal$.
- The log normalization factor: $A(\eta) = \sum_{i=1}^{K} \log \Gamma(\eta_i) - \log \Gamma(\sum_{i=1}^{K} \eta_i)$
- The base measure: $h(x) = \frac{1}{\prod_{i=1}^{K}} \theta_i$

The following sketches why the parameters are expressed in these ways.

First, recall that $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)})$. Then the pdf of the Dirichlet distribution is

$$
\begin{aligned}
p(\theta|\alpha) &= \frac{1}{B(\alpha)} \prod_{i=1}^{K} \theta_i^{\alpha_i - 1} \\
&= \frac{\Gamma(\sum_{i=1}^{K} \alpha_i)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \cdot \frac{\prod_{i=1}^{K} \theta_i^{\alpha_i}}{\prod_{i=1}^{K} \theta_i}
\end{aligned}
$$

Taking the exponential and log,

$$\begin{aligned}
p(\theta|\alpha) &= \frac{\Gamma(\sum_{i=1}^K \alpha_i)}{\prod_{i=1}^K \Gamma(\alpha_i)} \cdot \frac{\prod_{i=1}^K \theta_i^{\alpha_i}}{\prod_{i=1}^K \theta_i} \\
&= \frac{1}{\prod_{i=1}^K \theta_i} \exp\Big[\log \frac{\Gamma(\sum_{i=1}^K \alpha_i)}{\prod_{i=1}^K \Gamma(\alpha_i)} \prod_{i=1}^K \theta_i^{\alpha_i} \Big] \\
&= \frac{1}{\prod_{i=1}^K \theta_i} \exp\Big[\log \Gamma(\sum_{i=1}^K \alpha_i) - \log \prod_{i=1}^K \Gamma(\alpha_i) + \log \prod_{i=1}^K \theta_i^{\alpha_i} \Big] \\
&= \frac{1}{\prod_{i=1}^K \theta_i} \exp\Big[\log \Gamma(\sum_{i=1}^K \alpha_i) - \sum_{i=1}^K \log \Gamma(\alpha_i) + \sum_{i=1}^K \alpha_i \log \theta_i \Big] \\
&= \frac{1}{\prod_{i=1}^K \theta_i} \exp\Big[\alpha^\intercal \log \theta - \big\{ \sum_{i=1}^K \log \Gamma(\alpha_i) - \log \Gamma(\sum_{i=1}^K \alpha_i) \big\} \Big]
\end{aligned}$$

Comparison of this expression with the general pdf of the exponential family yields the parameters listed above.

3. The expectation of the sufficient statistic of the exponential family is the derivative of the log normalizing factor

Note that this is useful in the derivation of the original variational inference for LDA (in Blei et al 2003).

Theorem:

$$\mathbb{E}_\theta[t(\theta)] = \frac{\partial}{\partial \eta} A(\eta) \tag{8}$$

Proof:

Using the fact that the pdf must integrate to one, we first express $a(\eta)$ by the other parameters.

$$\begin{aligned}
1 = \int_\theta p(\theta|\eta) d\theta &= \int_\theta h(\theta) \exp(\eta\ t(\theta) - A(\eta)) d\theta \\
\iff 1 &= \frac{1}{\exp A(\eta)} \int_\theta h(\theta) \exp(\eta\ t(\theta)) d\theta \\
\iff \exp A(\eta) &= \int_\theta h(\theta) \exp(\eta\ t(\theta)) d\theta \\
\iff A(\eta) &= \log \int_\theta h(\theta) \exp(\eta\ t(\theta)) d\theta
\end{aligned}$$

Let $g(\eta) = \frac{1}{\exp A(\eta)}$. Then this is also equivalent to

$$1 = g(\eta) \int_\theta h(\theta) \exp(\eta\ t(\theta)) d\theta \tag{9}$$

Taking the derivative of Eq (3) with respect to η,

$$\begin{aligned}
0 &= \frac{\partial}{\partial \eta}\Big[g(\eta)\int_\theta h(\theta)\exp(\eta\ t(\theta))d\theta\Big] \\
&= g'(\eta)\int_\theta h(\theta)\exp(\eta\ t(\theta))d\theta + g(\eta)\frac{\partial}{\partial\eta}\int_\theta h(\theta)\exp(\eta\ t(\theta))d\theta \\
&= \frac{g'(\eta)}{g(\eta)} + g(\eta)\int_\theta h(\theta)\exp(\eta\ t(\theta))t(\theta)d\theta \\
&= \frac{g'(\eta)}{g(\eta)} + \int_\theta g(\eta)h(\theta)\exp(\eta\ t(\theta))t(\theta)d\theta \\
&= \frac{g'(\eta)}{g(\eta)} + \mathbb{E}_\theta[t(\theta)]
\end{aligned}$$

i.e.

$$\mathbb{E}_\theta[t(\theta)] = -\frac{g'(\eta)}{g(\eta)} = -\frac{\partial}{\partial\eta}log(g(\eta)) \tag{10}$$

because $\frac{\partial}{\partial\eta}log(g(\eta)) = \frac{1}{g(\eta)}\cdot\frac{\partial}{\partial\eta}g(\eta)$.

Because $g(\eta) = \frac{1}{\exp(A\eta))} = -\exp(A(\eta))$, Eq (4) is equivalent to

$$\mathbb{E}[t(\theta)] = -\frac{\partial}{\partial\eta}log(g(\eta)) = \frac{\partial}{\partial\eta}A(\eta) \tag{11}$$

4. Derivation steps for the collapsed Gibbs Sampler for the smoothed LDA

This section describes the steps for the derivation of the collapsed Gibbs Sampler for the smoothed LDA. It aims to fill the gaps of the lecture slides by Shiraito for POLSCI798. The model is specified in Part I, smoothed LDA. (Notation has changed. Need to fix.) Because the following equations involve many products, let us assume that the product term,$\prod$, ends at the $\times$.

Joint posterior distribution

$$p(Z,\theta,\eta|W,\alpha,\beta) \tag{12}$$

$$\propto \underbrace{p(\eta|\beta)p(\theta|\alpha)}_{\text{prior}} \times \underbrace{p(Z|\theta)p(W|Z,\eta)}_{\text{likelihood}} \tag{13}$$

$$= \prod_{k=1}^{K} p(\eta_k|\beta) \times \prod_{j=1}^{J} p(\theta_j|\alpha) \times \prod_{j=1}^{J}\prod_{i=1}^{N_j} p(Z_{ij}|\theta_j) \times \prod_{j=1}^{J}\prod_{i=1}^{N_j} p(W_{ij}|Z_{ij},\eta_j) \tag{14}$$

$$= \prod_{k=1}^{K} p(\eta_k|\beta) \times \prod_{j=1}^{J} p(\theta_j|\alpha) \prod_{i=1}^{N_j} p(Z_{ij}|\theta_j)p(W_{ij}|Z_{ij},\eta_j) \tag{15}$$

$$\tag{16}$$

Joint posterior distribution after collapsing θ and η

$$p(Z|W) \propto \int_{\theta}\int_{\eta} p(Z,\theta,\eta|W)\ d\eta d\theta \tag{17}$$

$$= \int_{\theta}\int_{\eta} \prod_{k=1}^{K} p(\eta_k|\beta) \times \prod_{j=1}^{J} p(\theta_j|\alpha) \prod_{i=1}^{N_j} p(Z_{ij}|\theta_j)\ d\eta d\theta \tag{18}$$

$$= \int_{\theta} \prod_{j=1}^{J} p(\theta_j|\alpha) \prod_{i=1}^{N_j} p(Z_{ij}|\theta_j)\ d\theta\ \times \int_{\eta} \prod_{k=1}^{K} p(\eta_k|\beta) \times \prod_{j=1}^{J}\prod_{i=1}^{N_j} p(Z_j|\theta_j)p(W_{ij}|Z_j,\eta)\ d\eta \tag{19}$$

Conditional posterior

$$p(Z_{i^*j^*} = k|Z_{-i^*j^*}, W)$$
$$\propto \int_{\theta_{j^*}} p(\theta_{j^*}|\alpha) \prod_{i=1}^{N_j} p(Z_{i^*j^*} = k, \theta_{j^*}) \prod_{i\neq i^*} p(Z_{ij}|\theta_{j^*})\ d\theta_{j^*}$$
$$\times \int_{\eta_k} p(\eta_k|\beta) \times p(W_{i^*j^*}|Z_{i^*j^*} = k, \eta_k) \times \prod_{ij\neq i^*j^*, Z_{ij}=k} p(W_i j|Z_{ij} = k, \eta_k)\ d\eta_k$$

NOTE: diagram of data generating process is helpful - insert here

Reference

Blei, D.M., Ng, A.Y. and Jordan, M.I., 2003. Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), pp.993-1022.

Geigle, C., 2016. Inference Methods for Latent Dirichlet Allocation.

Gelman, A., Stern, H.S., Carlin, J.B., Dunson, D.B., Vehtari, A. and Rubin, D.B., 2013. Bayesian data analysis. Chapman and Hall/CRC.

Will Kurt. A blog post about KL divergence

Lecture Slides by Yuki Shiraito, POLSCI798, 2019.