

Latent Dirichlet Allocation and Beyond

Kangcheng Hou
kangchenghou@gmail.com

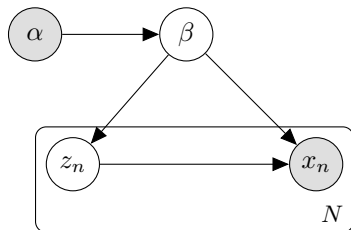
Zhejiang University

June 30, 2018

Agenda

- ▶ Latent variable model.
- ▶ Latent Dirichlet Allocation.
- ▶ Use Gibbs sampling.
- ▶ Use variational inference.
- ▶ Hierarchical dirichlet process.

Latent variable model



The joint distribution is

$$p(x, z, \beta | \alpha) = p(\beta | \alpha) \prod_{n=1}^N p(x_n, z_n | \beta)$$

The task is to infer the posterior distribution of local latent variables z and the global variable β given the data x .

$$p(z, \beta | x, \alpha)$$

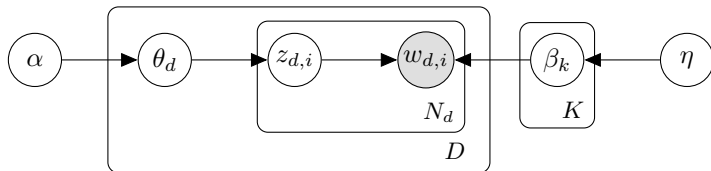
Two methods

To do inference, there are usually two methods:

1. MCMC, the basic idea is to construct a Markov chain such that if we sample points on this chain. We will get the posterior distribution.
2. Variational Inference

Introduction

Here is a PGM of LDA.



We want to sample the posterior distribution of the latent variable $z_{d,i}$ given $w_{d,i}$. One method is gibbs sampling, in every iteration, we sample z_i from distribution $p(z_i|z_{-i}, w)$. The joint distribution is

$$p(w, z|\alpha, \beta) = p(w|z, \beta)p(z|\alpha)$$

Posterior predictive of Dirichlet-Multinomial

Suppose we have dirichlet prior distribution,

$$p(\theta|\alpha) = \text{Dir}(\theta|\alpha) \propto \prod_{k=1}^K \theta_k^{\alpha_k-1}$$

Joint distribution I

The joint distribution is $p(w, z|\alpha, \beta) = p(w|z, \beta)p(z|\alpha)$. The first part is

$$p(w|z, \beta) = \int_{\phi_{1:K}} p(w|z, \phi_{1:K})p(\phi_{1:K}|\beta)d\phi_{1:K}$$

Let's see $p(w|z, \phi_{1:K})$ first,

$$p(w|z, \phi_{1:K}) = \prod_{i=1}^W \phi_{z_i, w_i}$$

Or we can rephrase it in another way, where we classify it by topic.

$$p(w|z, \phi_{1:K}) = \prod_{i=1}^W \phi_{z_i, w_i} = \prod_{k=1}^K \prod_{v=1}^V \phi_{k,v}^{n_k^{(v)}}$$

Joint distribution II

And

$$p(\phi_{1:K}|\beta) = \prod_{k=1}^K \prod_{v=1}^V \phi_{k,v}^{\beta_v-1}$$

Adding these two together, we have

$$p(w|z, \beta) = \prod_{k=1}^K \frac{B(n_k + \beta)}{B(\beta)}$$

n_k represents the word appearance frequencies in topic k . The topic distribution $p(z|\alpha)$ can be derived similarly,

$$p(z|\alpha) = \prod_{d=1}^D \frac{B(n_d + \alpha)}{B(\alpha)}$$

Joint distribution III

So the conditional distribution is

$$p(z, w | \alpha, \beta) = \prod_{k=1}^K \frac{B(n_k + \beta)}{B(\beta)} \prod_{d=1}^D \frac{B(n_d + \alpha)}{B(\alpha)}$$

$$p(z_i = k | z_{-i}, w, \alpha, \beta) = \frac{p(w, z) | \alpha, \beta}{p(w, z_{-i} | \alpha, \beta)} \propto p(w, z | \alpha, \beta)$$

Thus we just sample the z_i according to the conditional distribution

$$p(z_i = k | z_{-i}, w, \alpha, \beta) \propto \prod_{k=1}^K \frac{B(n_k + \beta)}{B(\beta)} \prod_{d=1}^D \frac{B(n_d + \alpha)}{B(\alpha)}$$

We shall check out what is the same and what is different for different k where $z_i = k$. As a result, it can be shown that

$$p(z_i = k | z_{-i}, w, \alpha, \beta) \propto \Gamma(n_{k,w_i}^{-i}) \Gamma(n_{d,k}^{-i})$$

Joint distribution IV

The symbol here is quite messy, if you have question, please feel free to contact me. n_{k,w_i}^{-i} means except for this word, the frequency of word w_i in topic k . And $n_{d,k}^{-i}$ means that except for this word, the frequency of word in document d in topic k . This is collapsed Gibbs sampling. In implementing this, **what we need to do is keeps sampling** z , and that is enough. And after the sampling is done, the multinomial parameters can be derived as follows

$$p(\theta_d|w, z, \alpha) \sim \text{Dir}(\theta_m|n_m + \alpha)$$

$$p(\phi_d|w, z, \beta) \sim \text{Dir}(\phi_d|n_d + \beta)$$

Evidence Lower Bound(ELBO) I

$$\ln(p(X)) = \ln\left(\frac{p(X, Z)}{q(Z)}\right) - \ln\left(\frac{p(Z|X)}{q(Z)}\right)$$

Taking the expectation w.r.t $q(Z)$ on both sides

$$\begin{aligned}\ln(P(X)) &= \int q(Z) \ln\left(\frac{p(X, Z)}{q(Z)}\right) dZ - \int q(Z) \ln\left(\frac{p(Z|X)}{q(Z)}\right) dZ \\ &= \mathcal{L}(q) + KL(q||p)\end{aligned}$$

Evidence Lower Bound(ELBO) II

Alternative derivation using Jensen's inequality:

$$\begin{aligned}\ln(X) &= \ln \int_Z p(X, Z) dZ \\ &= \ln \int_Z p(X, Z) \frac{q(Z)}{q(Z)} dZ \\ &= \ln \mathbb{E}_{Z \sim q(Z)} \left[\frac{p(X, Z)}{q(Z)} \right] \\ &\geq \ln \mathbb{E}_{Z \sim q(Z)} \left[\frac{p(X, Z)}{q(Z)} \right] \\ &= \int q(Z) \ln \left(\frac{p(X, Z)}{q(Z)} \right) dZ\end{aligned}$$

Which is $\mathcal{L}(q)$. We are interested in finding the lower bound of $\ln(X)$. And when the selected $q(Z)$ is closest to $p(Z|X)$, the KL divergence is 0, therefore, minimum is achieved. There are two typical approaches of get $q(Z)$,

Evidence Lower Bound(ELBO) III

- ▶ Assume that $q(Z)$ has some factorization $q(Z) = \prod_{i=1}^M q_i(Z_i)$.
- ▶ Assume $q(Z)$ is in some families of distribution(e.g. exponential families).

Factorization form

Assume that $q(z)$ can be factorized as follows, $q(z) = \sum_{i=1}^M q_i(z_i)$

$$\begin{aligned}\mathcal{L}(q) &= \int q(z) \ln p(x, z) dz - \int q(z) \ln q(z) dz \\ &= \int \prod_{i=1}^M q_i(z_i) \ln p(x, z) dz - \int \prod_{i=1}^M q_i(z_i) \sum_{i=1}^M \ln q_i(z_i) dz\end{aligned}$$

If we just optimize $q_i(z_i)$ one at a time as follows, (Note that it is an iterative process)

$$\begin{aligned}\mathcal{L}(q) &= \int q_i(z_i) \mathbb{E}_{z_{-i}} [\ln p(x, z)] dz_i - \int q_i(z_i) \ln q_i(z_i) dz_i \\ &= \int q_i(z_i) \ln \tilde{p}(x, z_i) dz_i - \int q_i(z_i) \ln q_i(z_i) dz_i \\ &= \int q_i(z_i) \ln \frac{\tilde{p}(x, z_i)}{q_i(z_i)} = KL[\tilde{p}(x, z_i) || q_i(z_i)]\end{aligned}$$

Exponential family distribution I

Exponential family distribution is a rich family of distribution which computational tractable which is suitable for variational inference.

$$p(x) = h(x)e^{\eta^\top T(x) - A(\eta)}$$

Probability distribution in exponential family has several advantages. We can see that it is very easy to get the maximum likelihood estimation.

$$\ln \prod_{i=1}^N p(x_i|\eta) = \sum_{i=1}^N \ln h(x_i) + \eta^\top \left(\sum_{i=1}^N T(x_i) \right) - NA(\eta)$$

So if we want to get the maximum likelihood estimation,

$$\frac{dA(\eta)}{d\eta} = \frac{\sum_{i=1}^N T(x_i)}{N}$$

Exponential family distribution II

Let's see an example of one-dimensional Gaussian distribution,

$$\mathcal{N}(x|\mu, \sigma^2) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

We transform it into the exponential family distribution,

$$p(x|\eta) = h(x)e^{\eta^\top T(x) - A(\eta)} = \exp\left(\begin{bmatrix} x \\ x^2 \end{bmatrix}^\top \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} - \left(\frac{-\eta_1^2}{4\eta_2} + \frac{1}{2} \ln \eta_2\right) - \frac{1}{2} \ln 2\pi\right)$$

$$\begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} = \begin{bmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{bmatrix}$$

and

$$A(\eta) = \frac{-\eta_1^2}{4\eta_2} + \frac{1}{2} \ln \eta_2 \quad T(x) = \begin{bmatrix} x \\ x^2 \end{bmatrix} \quad \theta = \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix} = \begin{bmatrix} \frac{-\eta_1}{2\eta_2} \\ \frac{-1}{2\eta_2} \end{bmatrix}$$

Exponential family distribution III

Now we use the previous conclusion to get the maximum likelihood estimation.

$$\frac{dA(\eta)}{d\eta} = \frac{\sum_{i=1}^N T(x_i)}{N}$$

$$A(\eta) = \frac{-\eta_1^2}{4\eta_2} + \frac{1}{2} \ln \eta_2$$

$$\begin{bmatrix} -\frac{\eta_1}{2\eta_2} \\ \frac{\eta_1^2}{4\eta_2^2} + \frac{1}{2\eta_2} \end{bmatrix} = \begin{bmatrix} \frac{\sum_{i=1}^N x_i}{N} \\ \frac{\sum_{i=1}^N x_i^2}{N} \end{bmatrix}$$

With

$$\theta = \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix} = \begin{bmatrix} \frac{-\eta_1}{2\eta_2} \\ \frac{-1}{2\eta_2} \end{bmatrix}$$

it is easy to solve μ and σ^2 . Moreover, conjugate distribution in exponential family distribution are easy to compute.

Exponential family distribution IV

We now show one important property of exponential family distribution that, for

$$p(x|\lambda) = h(x) \exp(T(x)^\top \lambda - A_g(\lambda))$$

We have

$$\nabla_\lambda A_g(\lambda) = \mathbb{E}_{p(x|\lambda)}[T(x)]$$

From

$$\int_x p(x|\lambda) dx = \int_x h(x) \exp(T(x)^\top \lambda - A_g(\lambda)) dx = 1$$

Exponential family distribution V

$$\nabla_{\lambda} \int_x h(x) \exp(T(x)^{\top} \lambda - A_g(\lambda)) dx = 0$$

$$\int_x \nabla_{\lambda} h(x) \exp(T(x)^{\top} \lambda - A_g(\lambda)) dx = 0$$

$$\int_x \nabla_{\lambda} h(x) \exp(T(x)^{\top} \lambda - A_g(\lambda)) dx = 0$$

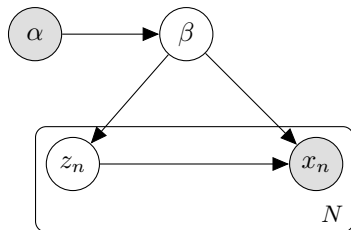
$$\int_x p(x|\lambda) (T(x) - \nabla A_g(\lambda)) dx = 0$$

$$\nabla_{\lambda} A_g(\lambda) = \mathbb{E}_{p(x|\lambda)}[T(x)]$$

Thus the equation holds.

Variational Inference I

Now let's see how to implement do variational inference for the following model.



The joint distribution is

$$p(x, z, \beta | \alpha) = p(\beta | \alpha) \prod_{n=1}^N p(x_n, z_n | \beta)$$

Variational Inference II

We make the following assumptions, the conditional distribution belongs to exponential family distribution,

$$p(\beta|x, z, \alpha) = h(\beta) \exp(\eta_g(x, z, \alpha)^\top T(\beta) - A_g(\eta_g(x, z, \alpha)))$$

and

$$p(z_{nj}|x_n, z_{n,-j}, \beta) = h(z_{nj}) \exp(\eta_l(x_n, z_{n,-j}, \beta)^\top T(z_{nj}) - A_l(\eta_l(x_n, z_{n,-j}, \beta)))$$

The evidence lower bound is

$$\mathcal{L}(q) = \mathbb{E}_q[\log p(x, z, \beta)] - \mathbb{E}_q[\log q(z, \beta)]$$

We select the distribution q such that it is both expressive and easier to optimize, (Note that we only restrict the function form for p , but have not specifies the factorization form of p .)

$$q(z, \beta) = q(\beta|\lambda) \prod_{n=1}^N \prod_{j=1}^J q(z_{nj}|\phi_{nj})$$

Variational Inference III

Other than the factorization of $q(z, \beta)$, we restrict $q(z, \beta)$ and $q(z_{nj}|\phi_{nj})$.

$$q(\beta|\lambda) = h(\beta) \exp(\lambda^\top T(\beta) - A_g(\lambda))$$

$$q(z_{nj}|\phi_{nj}) = h(z_{nj}) \exp(\phi_{nj}^\top T(z_{nj}) - A_l(\phi_{nj}))$$

Now let's alter the parameter λ, ϕ to maximize the ELBO. Rather than $\mathcal{L}(q)$, we can write it as $\mathcal{L}(\lambda, \phi)$. It turns out we can optimize λ and ϕ in turn. We first fix ϕ and optimize λ .

Variational Inference IV

$$\begin{aligned}\mathcal{L}(q) &= \mathbb{E}_q[\log p(x, z, \beta)] - \mathbb{E}_q[\log q(z, \beta)] \\&= \mathbb{E}_q[\log p(\beta|x, z)] + \underbrace{\mathbb{E}_q[\log p(x, z)]}_{\text{not related to } \beta} - \mathbb{E}_q[\log q(z, \beta)] \\&= \mathbb{E}_q[\log h(\beta)] + \mathbb{E}_q[\eta_g(x, z)]^\top \mathbb{E}_q[T(\beta)] - \mathbb{E}_q[A_g(\eta_g(x, z))] \\&\quad - \mathbb{E}_q[\log h(\beta)] - \mathbb{E}_q[T(\beta)^\top \lambda] + \mathbb{E}[A_g(\lambda)] \\&= \mathbb{E}_q[\eta_g(x, z)]^\top \mathbb{E}_q[T(\beta)] - \mathbb{E}_q[A_g(\eta_g(x, z))] - \mathbb{E}_q[T(\beta)^\top \lambda] \\&\quad + \mathbb{E}[A_g(\lambda)] \\&= \mathbb{E}_q[\eta_g(x, z)]^\top \mathbb{E}_q[T(\beta)] - \mathbb{E}_q[T(\beta)]^\top \lambda + A_g(\lambda) \\&= \nabla_\lambda A_g(\lambda)^\top (\mathbb{E}_{q(z|\phi)}[\eta_g(x, z)] - \lambda) + A_g(\lambda)\end{aligned}$$

Maximize w.r.t $\mathcal{L}(\lambda)$ a.k.a $\mathcal{L}(q)$, $\frac{d\mathcal{L}(\lambda)}{d\lambda} = 0$, we have

$$\nabla_\lambda^2 A_g(\lambda)^\top (\mathbb{E}_{q(z|\phi)}[\eta_g(x, z)] - \lambda) = 0$$

Variational Inference V

Assume that $\nabla_{\lambda}^2 A_g(\lambda) \neq 0$, so we have

$$\mathbb{E}_{q(z|\phi)}[\eta_g(x, z)] - \lambda = 0$$

That's cool! Right? What have we done so far?

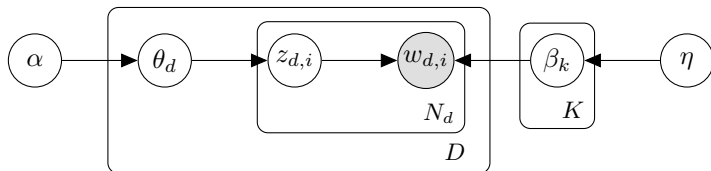
- ▶ We first specifies a graphical model, and specifies the latent variables e.g. z, β .
- ▶ We define the distribution so that the conditional distribution for latent varaibels given other variables is always exponential family distribution.
- ▶ For joint distribution of latent variable e.g. $q(z, \beta)$, we assume that $q(z, \beta)$ factorizes into $q(z|\lambda)q(\beta|\phi)$.

Variational Inference VI

- ▶ Then, we have a very nice conclusion from our previous lengthy derivation that if we optimize the parameters e.g. λ, ϕ one by one, we just do it in this way. Repeat between $\lambda = \mathbb{E}_{q(z|\phi)}[\eta_g(x, z)]$ and $\phi = \mathbb{E}_{q(z|\lambda)}[\eta_l(x, \beta)]$ which is **set the parameter to be the expectation of natural parameter conditioning on all other latent variables.**

Variational Inference for LDA I

We will now do variational inference for the following model,



Here is a pipeline for doing variational inference.

- ▶ The graphical model encodes the conditional independence information of the model.
- ▶ We assign the exponential family distribution to the distribution such that we can get exponential family conditional distribution for future ease.
- ▶ Write the q distribution for each full conditional distribution and do optimization.

Variational Inference for LDA II

For LDA, The document-topic distribution is

$$\theta_d \sim \text{Dir}(\alpha)$$

The topic-word distribution is

$$\phi_k \sim \text{Dir}(\beta)$$

The topic assignment of each word is

$$z_{d,i} \sim \text{Mult}(\theta_d)$$

And for each word,

$$w_{d,i} \sim \text{Mult}(\theta_{z_{d,i}})$$

Variational Inference for LDA III

We have three kinds of latent variables, θ_d , $z_{d,i}$ and ϕ_k respectively. We now derive the conditional distribution of these given all other variables. First, we derive the conditional distribution of $z_{d,n}$, it can be shown that the conditional distribution of $z_{d,n}$ is

$$\begin{aligned} & p(z_{d,n} = k | \theta_d, w_{d,n}, \phi_k) \\ \propto & p(w_{d,n} | z_{d,n} = k, \phi_k) p(z_{d,n} = k | \theta_d) \\ = & \phi_{k, w_{d,n}} \theta_{d,k} \\ = & \exp(\ln[\phi_{k, w_{d,n}} \theta_{d,k}]) \end{aligned}$$

Variational Inference for LDA IV

For conditional distribution for the topic distribution of each document θ_d , it can be shown that

$$p(\theta_d | \text{all other variables}) = p(\theta_d | z_d)$$

Variational Inference for LDA V

$$\begin{aligned} & p(\theta_d | z_d) \\ \propto & p(\theta_d | \alpha) \prod_{n=1}^{N_d} p(z_{d,n} | \theta_d) \\ = & \prod_{k=1}^K (\theta_{d,k}^{\alpha_k - 1} \prod_{n=1}^{N_d} \theta_{d,k}^{\delta(z_{d,n}, k)}) \\ = & \prod_{k=1}^K (\theta_{d,k}^{\alpha_k - 1 + n_k}) \quad \text{where } n_k = \sum_{n=1}^{N_d} \delta(z_{d,n}, k) \\ = & \exp\left(\sum_{k=1}^K (\alpha_k - 1 + n_k) \log \theta_{d,k}\right) \end{aligned}$$

Variational Inference for LDA VI

For conditional distribution for the word distribution of each topic ϕ_k ,

$$\begin{aligned} & p(\phi_k | Z, W) \\ \propto & p(\phi_{k,v} | \beta) \prod_{d=1}^D \prod_{n=1}^{N_d} p(w_{d,n} | \phi_k)^{\delta(z_{d,n}, k)} \\ = & \prod_{v=1}^V \phi_{k,v}^{\beta_v - 1} \prod_{d=1}^D \prod_{n=1}^{N_d} p(w_{d,n} | \phi_k)^{\delta(z_{d,n}, k)} \\ = & \prod_{v=1}^V \phi_{k,v}^{\beta_v - 1} \prod_{d=1}^D \prod_{n=1}^{N_d} \phi_{k, w_{d,n}}^{\delta(z_{d,n}, k)} \\ = & \exp\left(\sum_{v=1}^V (\beta_v - 1 + n_v) \log \phi_{k,v}\right) \end{aligned}$$

Variational Inference for LDA VII

where n_v is the number of occurrence of word v which belongs to topic k .

So now we have the three conditional distribution of latent parameters. Now we can perform variational inference for our graphical model.

$$p(z_{d,n} = k | \theta_d, w_{d,n}, \phi_k) = \exp(\ln[\beta_{k,w_{d,n}} \theta_{d,k}])$$

$$p(\theta_d | z_d) = \exp\left(\sum_{k=1}^K (\alpha_k - 1 + n_k) \log \theta_{d,k}\right)$$

$$p(\beta_k | Z, W) = \exp\left(\sum_{v=1}^V (\eta_v - 1 + n_v) \log \beta_{k,v}\right)$$

Variational Inference for LDA VIII

We assume that the variational function is in the form of

$$q(z_{d,n}) = \text{Mult}(\phi_{d,n})$$

$$q(\theta_d) = \text{Dir}(\gamma_d)$$

$$q(\beta_k) = \text{Dir}(\lambda_k)$$

According to the property of exponential family, we have the following conclusion

$$\eta(\phi_{d,n}) = \mathbb{E}_{q(\theta_d)q(\beta_k)}(\ln[\beta_{k,w_{d,n}}\theta_{d,k}])$$

$$\eta(\gamma_d) = \mathbb{E}_{q(z_{d,n})q(\beta_k)}([\alpha_1 - 1 + n_1 \cdots \alpha_K - 1 + n_K]^\top)$$

$$\eta(\beta_k) = \mathbb{E}_{q(z_{d,n})q(\theta_d)}([\eta_1 - 1 + n_1 \cdots \eta_V - 1 + n_V]^\top)$$

Variational Inference for LDA IX

$$\begin{aligned}\eta(\phi_{d,n}) &= \mathbb{E}_{q(\theta_d)q(\beta_k)}(\ln[\beta_{k,w_{d,n}}\theta_{d,k}]) \\ &= \mathbb{E}_{q(\theta_d)}[\ln \theta_{d,k}] + \mathbb{E}_{q(\beta_k)}[\ln \beta_{k,w_{d,n}}] \\ &= \Psi(\gamma_{d,k}) - \Psi\left(\sum_{k=1}^K \gamma_{d,k}\right) + \Psi(\lambda_{k,w_{d,n}}) - \Psi\left(\sum_{v=1}^V \lambda_{k,v}\right)\end{aligned}$$

$$\eta(\gamma_d) = [\alpha_1 - 1 + \phi_{d,n}^1 \cdots \alpha_K - 1 + \phi_{d,n}^K]^\top$$

$$\eta(\beta_k) = \eta - 1 + \sum_{d=1}^D \sum_{n=1}^N w_{d,n} \phi_{d,n}^k$$

Huh! That's our updating rules for the variational inference for LDA.

Perplexity