COS 597C: Bayesian Nonparametrics

Lecturer: David Blei
Scribe: Chong Wang

Lecture # 6
October 22, 2007

Variational Inference for Dirichlet Process Mixtures

In this lecture, we begin with the basic concept of variational inference, with an emphasis on mean-field approach based on exponential families. Then Sethuraman's stick-breaking construction for Dirichlet process (DP) is briefly reviewed. Finally, we talk about applying variational inference for DP mixtures, based on Sethuraman's stick-breaking construction.

Variaitonal Inference

The basic idea of variational inference is to formulate the computation of a marginal or conditional probability in terms of an optimization problem. This (generally intractable) problem is then "relaxed", yielding a simplified optimization problem that depends on a number of free parameters, known as variational parameters. Solving for the variational parameters gives an approximation to the marginal or conditional probabilities of interest.

Notations:

$$\begin{aligned}&\text{data}: \ \mathbf{x} = \{x_1, x_2, \ldots, x_N\},\\ &\text{hidden variables}: \ \mathbf{w} = \{w_1, w_2, \ldots, w_M\},\\ &\text{model}: \ \theta.\end{aligned}$$

In Bayesian setting, we are usually interested in the posterior of $\mathbf{w}$ given $\mathbf{x}$ and θ: $p(\mathbf{w}|\mathbf{x},\theta)$. According to Bayes' theory, the computation of the posterior is

$$p(\mathbf{w}|\mathbf{x},\theta) = \frac{p(\mathbf{w},\mathbf{x}|\theta)}{p(\mathbf{x}|\theta)} = \frac{p(\mathbf{w},\mathbf{x}|\theta)}{\int p(\mathbf{w},\mathbf{x}|\theta)\mathrm{d}\mathbf{w}},$$

where the integral $p(\mathbf{x}|\theta) = \int p(\mathbf{w},\mathbf{x}|\theta)\mathrm{d}\mathbf{w}$ is usually computationally intractable. According to Jensen's inequality:

$$\begin{aligned}\log p(\mathbf{x}|\theta) &= \log \int p(\mathbf{w},\mathbf{x}|\theta)\mathrm{d}\mathbf{w}\\ &= \log \int q(\mathbf{w})\frac{p(\mathbf{w},\mathbf{x}|\theta)}{q(\mathbf{w})}\mathrm{d}\mathbf{w}\\ &= \log \mathbb{E}_q \frac{p(\mathbf{W},\mathbf{x}|\theta)}{q(\mathbf{W})}\\ &\geq \mathbb{E}_q \log \frac{p(\mathbf{W},\mathbf{x}|\theta)}{q(\mathbf{W})}\\ &= \mathbb{E}_q\left[\log p(\mathbf{W},\mathbf{x}|\theta)\right] - \mathbb{E}_q\left[\log q(\mathbf{W})\right]\\ &= \mathcal{L}(q). \qquad (1)\end{aligned}$$

In summary, we have $\log p(\mathbf{x}|\theta) \geq \mathcal{L}(q)$, where the equality holds only if $\frac{p(\mathbf{w},\mathbf{x}|\theta)}{q(\mathbf{w})} = p(\mathbf{x}|\theta)$, that is $q(\mathbf{w}) = p(\mathbf{w}|\mathbf{x},\theta)$, the posterior distribution of $\mathbf{w}$. In general, $\mathcal{L}(q)$ is a lower bound of $\log p(\mathbf{x}|\theta)$.

However, $q(\mathbf{w})$ without any constraints makes no simplification of the problem. In variational inference, we define an alternative family of distributions $q(\mathbf{w}|\boldsymbol{\nu})$, where $\boldsymbol{\nu}$ is called free variational parameters. Then, the optimization problem we want to solve is:

$$\arg\max_q \mathcal{L}(q) \Leftrightarrow \arg\min_q KL\left[q(\mathbf{w}|\boldsymbol{\nu}) \parallel p(\mathbf{w}|\mathbf{x},\theta)\right]. \tag{2}$$

There are no general rules to select q. The mean-field variational distribution is the simplest one with full factorization,

$$q(\mathbf{w}|\boldsymbol{\nu}) = \prod_{m=1}^{M} q_{\nu_m}(w_m). \tag{3}$$

Substitute (3) into (1), we rewrite (1) as

$$\begin{aligned} \log p(\mathbf{x}|\theta) &\geq \mathbb{E}_q\left[\log p(\mathbf{x}|\theta) + \log p(\mathbf{W}|\mathbf{x},\theta)\right] - \mathbb{E}_q\left[\log q(\mathbf{W})\right] \\ &= \mathbb{E}_q\left[\log p(\mathbf{x}|\theta) + \log p(\mathbf{W}|\mathbf{x},\theta) - \log q(\mathbf{W})\right] \\ &= \log p(\mathbf{x}|\theta) + \sum_{m=1}^{M} \mathbb{E}_q\left[\log p(W_m|W_1,\ldots,W_{m-1},\mathbf{x},\theta)\right] - \sum_{m=1}^{M} \mathbb{E}_q\left[\log q_{\nu_m}(W_m)\right] \end{aligned} \tag{4}$$

We use coordinate ascent to optimize the lower bound. To optimize with respect to ν_i, reorder $\mathbf{w}$ such that w_i is last in the list. Isolate the terms containing ν_i,

$$l_i = \mathbb{E}_q\left[\log p(W_i|\mathbf{W}_{-i},\mathbf{x},\theta)\right] - \mathbb{E}_q\left[\log q_{\nu_i}(W_i)\right] \tag{5}$$

Suppose that $p(w_i|\mathbf{w}_{-i},\mathbf{x},\theta)$ is in the exponential family:

$$p(w_i|\mathbf{w}_{-i},\mathbf{x},\theta) = h(w_i)\exp\left\{g(\mathbf{w}_{-i},\mathbf{x},\theta)^T w_i - a(g(\mathbf{w}_{-i},\mathbf{x},\theta))\right\}, \tag{6}$$

and $q_{\nu_i}(w_i)$ is in the same family, which means $q_{\nu_i}(w_i) = h(w_i)\exp\left\{{\nu_i}^T w_i - a(\nu_i)\right\}$. The maximum of (5) is achieved by setting $\frac{\partial l_i}{\partial \nu_i} = 0$, which gives:

$$\nu_i = \mathbb{E}_q\left[g(\mathbf{w}_{-i},\mathbf{x},\theta)\right]. \tag{7}$$

DP Mixtures based on the Stick-Breaking Construction

Consider two infinite collections of independent random variables, $V_i \sim \text{Beta}(1,\alpha)$ and $\eta_i^* \sim G_0$, $i = 1, 2, \ldots$, Sethuraman's stick-breaking construction of DP is:

$$\begin{aligned} \pi(\mathbf{v}_i) &= v_i \prod_{j=1}^{i-1}(1-v_j) \\ G &= \sum_{i=1}^{\infty} \pi(\mathbf{v}_i)\delta_{\eta_i^*}. \end{aligned} \tag{8}$$

And we have $G \sim \text{DP}(\alpha G_0)$. Based on Sethuraman's stick-breaking construction, DP mixtures is stated as follows:

1. Draw $V_i|\alpha \sim \text{Beta}(1,\alpha), i = 1, 2, \ldots$
2. Draw $\eta_i^*|\lambda \sim G_0(\cdot\,|\lambda), i = 1, 2, \ldots$

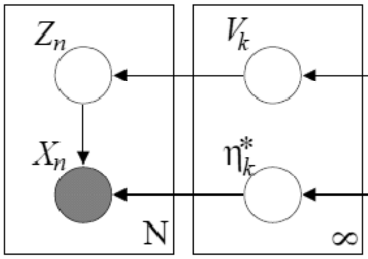

Figure 1: Graphical model representation of an exponential family DP mixture based on Sethuraman's stick-breaking construction.

3. For the nth data point:
   (a) Draw $Z_n \sim \mathrm{Mult}(\pi(\mathbf{v}))$
   (b) Draw $X_n | z_n \sim F(\cdot\, | \eta_{z_n} *)$

The graphical model of DP mixtures is shown as Figure 1. We restrict ourselves to DP mixtures for which the observable data are drawn from an exponential family distribution, and where the base distribution for the DP is the corresponding conjugate prior.

The distribution of X_n conditioned on Z_n and $\{\eta_1^*, \eta_2^*, \ldots\}$ is:

$$p(x_n | z_n, \eta_1^*, \eta_2^*, \ldots) \prod_{n=1}^{\infty} \left(h(x_n) \exp\left\{ \eta_i^* x_n - a(\eta_i^*) \right\} \right)^{\mathbf{1}[z_n = i]} . \tag{9}$$

And the base distribution $G_0(\cdot\, | \lambda)$ is the conjugate prior to $F(\cdot\, | \eta)$:

$$p(\eta^* | \lambda) = h(\eta^*) \exp\{\lambda_1^T \eta^* + \lambda_2(-a(\eta^*)) - a(\lambda)\}. \tag{10}$$

The hidden variables here are $\mathbf{W} = \{\mathbf{Z}_{1:N}, \boldsymbol{\eta}^*_{1:\infty}, \mathbf{V}_{1:\infty}\}$.

Variational Inference for DP Mixtures

First, we have

$$\begin{aligned} \log p(\mathbf{x}_{1:N} | \alpha, \lambda) &\geq \mathbb{E}_q \left[\log p(\mathbf{V}_{1:\infty} | \alpha) \right] + \mathbb{E}_q \left[\log p(\boldsymbol{\eta}^*_{1:\infty} | \lambda) \right] \\ &+ \sum_{n=1}^{N} \left(\mathbb{E}_q [\log p(Z_n | \mathbf{V}_{1:\infty})] + \mathbb{E}_q [\log p(x_n | Z_n)] \right) \\ &- \mathbb{E}_q [\log q(\mathbf{Z}_{1:N}, \boldsymbol{\eta}^*_{1:\infty}, \mathbf{V}_{1:\infty})]. \end{aligned} \tag{11}$$

For the variational distribution, we consider the truncated stick-breaking representations. We fix a value T and let $q(v^T = 1) = 1$ (as shown in Figure 2); this implies that the mixture proportions $\pi(\mathbf{v})$ are equal to zero for $t > T$ according to (8). Thus, the proposed variational distribution (fully factorized) is:

$$q(\mathbf{v}, \boldsymbol{\eta} *, \mathbf{z}) = \prod_{t=1}^{T-1} q_{\gamma_t}(v_t) \prod_{t=1}^{T} q_{\tau_t}(\eta_t^*) \prod_{n=1}^{N} q_{\phi_n}(z_n), \tag{12}$$

where $q_{\gamma_t}(v_t)$ are beta distributions, $q_{\tau_t}(\eta_t^*)$ are exponential family distributions with natural parameters τ_t (within the same family of G_0), and $q_{\phi_n}(z_n)$ are discrete multinomial distributions.

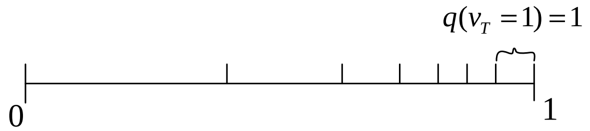

Figure 2: Truncated stick-breaking representations.

All of the terms in the lower bound of Equation (11) involve standard computations in the exponential family, except for the third term $\mathbb{E}_q[\log p(Z_n|\mathbf{V}_{1:\infty})]$. We rewrite it as follows:

$$\begin{aligned}\mathbb{E}_q[\log p(Z_n|\mathbf{V}_{1:\infty})] &= \mathbb{E}_q\left[\log\left(\prod_{i=1}^{\infty}(1-V_i)^{\mathbf{1}[Z_n>i]}V_i^{\mathbf{1}[Z_n=i]}\right)\right] \\ &= \sum_{i=1}^{\infty}\left(\mathbb{E}_q[\mathbf{1}[Z_n>i]\log(1-V_i)]+\mathbb{E}_q[\mathbf{1}[Z_n=i]\log V_i]\right) \\ &= \sum_{i=1}^{T}\left(q(z_n>i)\mathbb{E}_q[\log(1-V_i)]+q(z_n=i)\mathbb{E}_q[\log V_i]\right), \qquad (13)\end{aligned}$$

where

$$\begin{aligned}q(z_n=i) &= \phi_{n,i} \\ q(z_n>i) &= \sum_{j=i+1}^{T}\phi_{n,j} \\ \mathbb{E}_q[\log V_i] &= \Psi(\gamma_{i,1})-\Psi(\gamma_{i,1}+\gamma_{i,2}) \\ \mathbb{E}_q[\log(1-V_i)] &= \Psi(\gamma_{i,2})-\Psi(\gamma_{i,1}+\gamma_{i,2}). \qquad (14)\end{aligned}$$

Ψ is the digamma function. Using Equation (7), we have the coordinate ascent algorithm as follows:

$$\begin{aligned}\gamma_{t,1} &= 1+\sum_n \phi_{n,t} \\ \gamma_{t,2} &= \alpha+\sum_n\sum_{j=i+1}^{T}\phi_{n,j} \\ \tau_{t,1} &= \lambda_1+\sum_n \phi_{n,t}x_n \\ \tau_{t,2} &= \lambda_2+\sum_n \phi_{n,t} \\ \phi_{n,t} &\propto \exp(S_t), \qquad (15)\end{aligned}$$

where,

$$S_t = \mathbb{E}_q[\log V_t]+\sum_{i=1}^{t-1}\mathbb{E}_q[\log(1-V_i)]+\mathbb{E}_q[\eta_t^*]^T X_n-\mathbb{E}_q[a(\eta_t^*)]. \qquad (16)$$