Variational Dirichlet Process Mixtures by Truncating Responsibilities

Michalis K. Titsias
Velestino 37500
GREECE
michalis_titsias@yahoo.gr

August 7, 2007

Abstract

In this technical note we present a new mean field variational method for Dirichlet Process mixture models. This method improves on the method presented in [1] by truncating only responsibilities and optimizing over the infinite set of variational parameter posteriors.

1 Variational Dirichlet Process mixtures

Assume two infinite sets $\boldsymbol{\phi} = \{\phi_1, \phi_2, \ldots\}$ and $\mathbf{v} = \{v_1, v_2, \ldots, \}$ of component parameter vectors and beta distributed random variables. Each ϕ_k is independently drawn from the prior $p_\phi(\phi_k|\lambda)$ with hyperparameter λ and each v_k is independently drawn from $B(v_k|1, \alpha)$ with hyperparameter α. Given these sets a Dirichlet Process (DP) mixture model is written as a mixture with infinite number of components

$$p(x|\mathbf{v}, \boldsymbol{\phi}) = \sum_{k=1}^{\infty} \pi_k(\mathbf{v}) p(x|\phi_k), \tag{1}$$

where the mixing coefficients $\boldsymbol{\pi}(\mathbf{v}) = \{\boldsymbol{\pi}_1(\mathbf{v}), \boldsymbol{\pi}_2(\mathbf{v}), \ldots\}$ are given through the stick breaking variables $\mathbf{v}$ so as $\pi_k(\mathbf{v}) = v_k \prod_{j=1}^{k-1}(1 - v_j)$. Note that the prior $p_\phi(\phi_k|\lambda)$ over component parameters is the base distribution of an underlying DP and α is the corresponding concentration parameter [3].

Let $X = \{\mathbf{x}^1, \ldots, \mathbf{x}^N\}$ be the data and $Z = \{\mathbf{z}^1, \ldots, \mathbf{z}^n\}$ the latent assignments represented by infinite indicator vectors. Each $\mathbf{z}^n$ is generated from a multinomial with parameters $\boldsymbol{\pi}(\mathbf{v})$ and then $\mathbf{x}^n$ is drawn from the component $p(\mathbf{x}^n|\phi_k)$ for which $z_k^n = 1$.

The posterior $p(\mathbf{v}, \boldsymbol{\phi}, Z|X)$ over the unknowns is intractable to be expressed analytically. Recently the variational mean field method has been applied to DP mixtures based on the above stick breaking representation [1, 2]. This technique is based on a factorized variational distribution of the form

$$q(\mathbf{v}, \boldsymbol{\phi}, Z) = \prod_{k=1}^{\infty} \left[q_v(v_k; \beta_k^v) q_\phi(\phi_k; \beta_k^\phi) \right] \prod_{n=1}^{N} q_z(\mathbf{z}^n; \gamma^n), \tag{2}$$

where $q_v(v_k; \beta_k^v)$ is a beta distribution with parameters $(\beta_{k,1}^v, \beta_{k,2}^v)$ and $q_\phi(\phi_k; \beta_k^\phi)$ is chosen to be conjugate to $p(x|\phi_k)$ and has parameters β_k^ϕ. Each $q_z(\mathbf{z}^n; \gamma^n)$ is a multinomial distribution with parameter vector

γ^n (such that $\sum_{k=1}^{\infty} \gamma_k^n = 1$). The variational lower bound of the marginal likelihood is written as follows

$$\begin{aligned} F &= \int_{\mathbf{v},\boldsymbol{\phi}} \left(\prod_{k=1}^{\infty} q_v(v_k) q_\phi(\phi_k) \right) \sum_{n=1}^{N} \sum_{k=1}^{\infty} \gamma_k^n \log \pi_k(\mathbf{v}) p(\mathbf{x}^n|\phi_k) d\mathbf{v} d\boldsymbol{\phi} \\ &+ \sum_{k=1}^{\infty} \int_{\phi_k} q_\phi(\phi_k) \log \frac{p_\phi(\phi_k|\lambda)}{q_\phi(\phi_k)} d\phi_k + \sum_{k=1}^{\infty} \int_{v_k} q_v(v_k) \log \frac{B(v_k|1,\alpha)}{q_v(v_k)} dv_k \\ &- \sum_{n=1}^{N} \sum_{k=1}^{\infty} \gamma_k^n \log \gamma_k^n, \end{aligned} \tag{3}$$

where we have expressed the summation over Z and we simplified our notation by writing $q_v(v_k)$ and $q_\phi(\phi_k)$ instead of $q_v(v_k; \beta_k^v)$ and $q_\phi(\phi_k; \beta_k^\phi)$. The above quantity involves infinite many factors and thus to practically maximize it we need to introduce truncation. In [1] an explicit truncation level T is used by setting $q_v(v_T = 1) = 1$ which results all the mixing coefficients for each component $k > T$ to be equal to zero. This yields also all the $q_z(z^n; \gamma^n)$ distributions to be truncated to give zero probability (responsibility) for all components $k > T$. In [2] a softer truncation method is presented where for each $k > T$, $q_v(v_k)$ and $q_\phi(\phi_k)$ are constrained to be equal to their priors $B(v_k|1, \alpha)$ and $p_\phi(\phi_k|\lambda)$, respectively.

Next we present a alternative approach. In our method we truncate only responsibilities (the $q_z(z^n; \gamma^n)$ distributions) and we optimize over all the infinite set of variational parameter posteriors.

2 Truncating responsibilities

Let introduce a truncation level T so as all mixture components $k > T$ obtain zero responsibility, i.e. $\gamma_k^n = 0$, for $k > T$ and $n = 1, \ldots N$. This is the only truncation assumption we make. By using the truncation over the responsibilities the lower bound in Equation (3) is written as

$$\begin{aligned} F &= \int_{\mathbf{v},\boldsymbol{\phi}} \left(\prod_{k=1}^{T} q_v(v_k) q_\phi(\phi_k) \right) \sum_{n=1}^{N} \sum_{k=1}^{T} \gamma_k^n \log \pi_k(\mathbf{v}) p(\mathbf{x}^n|\phi_k) d\mathbf{v} d\boldsymbol{\phi} \\ &+ \sum_{k=1}^{\infty} \int_{\phi_k} q_\phi(\phi_k) \log \frac{p_\phi(\phi_k|\lambda)}{q_\phi(\phi_k)} d\phi_k + \sum_{k=1}^{\infty} \int_{v_k} q_v(v_k) \log \frac{B(v_k|1,\alpha)}{q_v(v_k)} dv_k \\ &- \sum_{n=1}^{N} \sum_{k=1}^{T} \gamma_k^n \log \gamma_k^n, \end{aligned} \tag{4}$$

where we used the fact that the distributions $q_v(v_k)$ and $q_\phi(\phi_k)$ for $k > T$ integrate to one in the first row of Equation (3) (to derive this note that $\pi_k(\mathbf{v})$ depends only on the first k stick breaking weights). The above bound can be globally maximized over the variational distributions $q_v(v_k)$ and $q_\phi(\phi_k)$ for $k > T$. Clearly this involves minimizing separate KL divergences of these distributions with the respective priors, thus the optimal solution is

$$q_\phi(\phi_k) = p_\phi(\phi_k|\lambda), \quad q_v(v_k) = B(v_k|1, \alpha), \quad k > T. \tag{5}$$

This result is very intuitive. A condition over a variational parameter posterior distribution is an approximate Bayes rule (likelihood is not expressed exactly) and a sufficient condition for Bayes to set a parameter posterior equal to the prior is the data to provide zero information about the parameters. The variational posterior over the parameters of a mixture component (at local maxima conditions) is equal to the prior if the responsibilities of this component are zero for all data[1]. Substituting equations (5)

[1]In our case this is also a necessary condition as it can be shown by examining the posterior beta distributions for the stick breaking weights.

back to (4) we obtain

$$
\begin{aligned}
F &= \int_{\mathbf{v},\boldsymbol{\phi}} \left(\prod_{k=1}^{T} q_v(v_k) q_\phi(\phi_k)\right) \sum_{n=1}^{N}\sum_{k=1}^{T} \gamma_k^n \log \pi_k(\mathbf{v}) p(\mathbf{x}^n|\phi_k) d\mathbf{v} d\boldsymbol{\phi} \\
&+ \sum_{k=1}^{T} \int_{\phi_k} q_\phi(\phi_k) \log \frac{p_\phi(\phi_k|\lambda)}{q_\phi(\phi_k)} d\phi_k + \sum_{k=1}^{T} \int_{v_k} q_v(v_k) \log \frac{B(v_k|1,\alpha)}{q_v(v_k)} dv_k \\
&- \sum_{n=1}^{N}\sum_{k=1}^{T} \gamma_k^n \log \gamma_k^n. \qquad (6)
\end{aligned}
$$

This bound now involves finite number of factors and can be maximized iteratively over $q_v(v_k)$, $q_\phi(\phi_k)$ for $k \leq T$ and over γ_k^ns with the constrain that $\sum_{k=1}^{T} \gamma_k^n = 1$ for each n.

There are two main differences with the method in [1]: i) we freely optimize over $q_v(v_T)$ while in [1] $q_v(v_T = 1) = 1$ and ii) the mixing coefficients are not truncated and thus it holds $\sum_{k=1}^{T} E_{q_v}[\pi_k(\mathbf{v})] < 1$ where E_q denotes expectation with respect to the distribution q. Since our method is less constrained than the approach in [1], it can achieve a better lower bound. Note also that the predictive mixture distribution remains an infinite mixture of the following form

$$
\begin{aligned}
p(x|X,\lambda,\alpha) &= \sum_{k=1}^{\infty} E_{q_v}[\pi_k(\mathbf{v})] E_{q_{\phi_k}}[p(x|\phi_k)] \\
&= \sum_{k=1}^{T} E_{q_v}[\pi_k(\mathbf{v})] E_{q_{\phi_k}}[p(x|\phi_k)] + \left(1 - \sum_{k=1}^{T} E_{q_v}[\pi_k(\mathbf{v})]\right) E_{p_\phi}[p(x|\phi)], \qquad (7)
\end{aligned}
$$

Clearly this distribution is a $T+1$-component mixture with the $T+1$th component associated with all infinite number of components ($k > T$) for which the parameter posteriors are equal to the priors.

The presented method is complementary to the method in [2]. In both methods the variational parameter posteriors for $k > T$ are equal to the priors, however this is achieved via different ways. In [2] all these posteriors are constrained to be equal to the priors and the bound is not optimized over them. Also since they use responsibilities with infinite support (γ_k^ns are not truncated) the former posteriors do not satisfy optimality variational conditions. In contrast we truncate only responsibilities and we maximize over all (infinite) set of variational parameter posteriors.

References

[1] D. M. Blei and M. I. Jordan. Variational inference for Dirichlet Process mixtures. *Journal of Bayesian Analysis*, 1(1):121–144, 2005.

[2] K. Kurihara, M. Welling, and N. Vlassis. Accelerated Variational Dirichlet Process mixtures. In *NIPS 19*, 2007.

[3] J. Sethuraman. A constructing definition of Dirichlet priors. *Statist. Sinica*, 4:639–650, 1994.