Dirichlet Processes, Dependent Dirichlet Processes and Applications in Machine Learning

Zhang Zhihua

College of Computer Science & Technology
Zhejiang University
zhzhang@zju.edu.cn

May 2013
Outline

- Dirichlet Processes
- Dependent Dirichlet Processes
- Applications in Machine Learning
- References
Nonparametric Bayesian Methods

- Dirichlet processes (DPs) (Ferguson, 1973; Sethuraman, 1994) or DP mixture models (Lo, 1984) and Dependent DPs (MacEachern, 2000) are important nonparametric Bayesian modeling tools.
Nonparametric Bayesian Methods

- Dirichlet processes (DPs) (Ferguson, 1973; Sethuraman, 1994) or DP mixture models (Lo, 1984) and Dependent DPs (MacEachern, 2000) are important nonparametric Bayesian modeling tools.

- After Markov chain Monte Carlo (MCMC) algorithms (see, for example, Escobar and West, 1995; Bush and MacEachern, 1996; MacEachern and Muller, 1998; Neal, 2000) and Variational Bayes Algorithms (Blei and Jordan, 2005) were developed for DP mixture models, DP mixture models have been used very successfully in the literature.
Nonparametric Bayesian Methods

- Dirichlet processes (DPs) (Ferguson, 1973; Sethuraman, 1994) or DP mixture models (Lo, 1984) and Dependent DPs (MacEachern, 2000) are important nonparametric Bayesian modeling tools.

- After Markov chain Monte Carlo (MCMC) algorithms (see, for example, Escobar and West, 1995; Bush and MacEachern, 1996; MacEachern and Muller, 1998; Neal, 2000) and Variational Bayes Algorithms (Blei and Jordan, 2005) were developed for DP mixture models, DP mixture models have been used very successfully in the literature.

- Recent Developments in Statistics and Machine Learning: Duke University (David Dunson et al.), Berkeley (Mike Jordan and his students).
Dirichlet Process Mixture Models
Dirichlet Process Mixture Models

In a Dirichlet Process Mixture (DPM) model, the samples $x_i$ for $i = 1, \ldots, n$ are assumed to be drawn from a mixture component parameterized by $\theta_i \in \Theta$. The $\theta_i$s are in turn generated by the distribution $G$, which is assumed to follow a Dirichlet process prior. That is, the DPM is

$$x_0 \overset{ind}{\sim} F(\theta_i), \quad i = 1, \ldots, n,$$

$$[\theta_i | G] \overset{iid}{\sim} G, \quad i = 1, \ldots, n,$$

$$G \sim \text{DP}(\alpha G_0).$$
Dirichlet Process Priors

- If $G$ is drawn from the Dirichlet process $\text{DP}(\alpha G_0)$ with base probability measure $G_0$ and concentration parameter $\alpha > 0$ over $(\Theta, \mathcal{A})$ then for any finite partition $(A_1, \ldots, A_k)$ of $\mathcal{A}$,

$$(G(A_1), \ldots, G(A_k)) \sim \text{Dir}(\alpha G_0(A_1), \ldots, \alpha G_0(A_k)).$$

Here $\text{Dir}(\alpha_1, \ldots, \alpha_k)$ denotes the Dirichlet distribution with positive parameters $\alpha_1, \ldots, \alpha_k$. 
An Illustration for DPs
Dirichlet Processes

Two important representations
Dirichlet Processes

- Two important representations
  - The Pólya urn scheme (Blackwell and MacQuenn, 1973)
Dirichlet Processes

- Two important representations
  - The Pólya urn scheme (Blackwell and MacQuenn, 1973)
  - Stick-breaking priors (Sethuraman, 1994; Pitman and Yor, 1997)
The Pólya urn scheme
The Pólya urn scheme

- Integrating over $G$ results in a Pólya urn scheme for the $\theta_i$:

$$\theta_1 \sim G_0(\theta_1),$$

$$[\theta_i|\theta_1, \ldots, \theta_{i-1}] \sim \frac{\alpha G_0(\theta_i) + \sum_{l=1}^{i-1} \delta(\theta_i|\theta_l)}{\alpha + i - 1},$$

where $\delta(\theta_i|\theta_l)$ is a point mass at $\theta_l$. 
The Pólya urn scheme

- Integrating over $G$ results in a Pólya urn scheme for the $\theta_i$:

$$\theta_1 \sim G_0(\theta_1),$$

$$[\theta_i|\theta_1, \ldots, \theta_{i-1}] \sim \frac{\alpha G_0(\theta_i) + \sum_{l=1}^{i-1} \delta(\theta_i|\theta_l)}{\alpha + i - 1},$$

where $\delta(\theta_i|\theta_l)$ is a point mass at $\theta_l$.

- We see that as $\alpha \to 0$, all the $\theta_i$ are identical to $\theta_1$, which in turn follows $G_0$. When $\alpha \to \infty$, the $\theta_i$ become iid $G_0$. 
The Pólya urn scheme

- Integrating over \( G \) results in a Pólya urn scheme for the \( \theta_i \):

\[
\theta_1 \sim G_0(\theta_1),
\]

\[
[\theta_i|\theta_1, \ldots, \theta_{i-1}] \sim \frac{\alpha G_0(\theta_i) + \sum_{l=1}^{i-1} \delta(\theta_i|\theta_l)}{\alpha + i - 1},
\]

where \( \delta(\theta_i|\theta_l) \) is a point mass at \( \theta_l \).

- We see that as \( \alpha \to 0 \), all the \( \theta_i \) are identical to \( \theta_1 \), which in turn follows \( G_0 \). When \( \alpha \to \infty \), the \( \theta_i \) become iid \( G_0 \).

- Since the \( \theta_i \) are exchangeable, the Pólya urn scheme can be written as

\[
[\theta_i|\theta_{-i}] \sim \frac{\alpha G_0(\theta_i) + \sum_{l \neq i} \delta(\theta_i|\theta_l)}{\alpha + n - 1},
\]

were \( \theta_{-i} \) represents \( \{\theta_l : l \neq i\} \).
The clustering property of the DP prior

- The discreteness of the random distribution $G$ leads to the well-known clustering property of the DP, which plays a central role in nonparametric Bayesian inference and computation.
The clustering property of the DP prior

- The discreteness of the random distribution $G$ leads to the well-known clustering property of the DP, which plays a central role in nonparametric Bayesian inference and computation.
- The clustering property allows DPs to formalize the notion of “borrowing strength” across related studies.
The clustering property of the DP prior

Assume that there are \( c \) distinct values among the \( \theta_i \) as \( \Phi = \{\phi_1, \ldots, \phi_c\} \), and that there are \( n_k \) occurrences of \( \phi_k \) such that \( \sum_{k=1}^{c} n_k = n \). The vector of configuration indicators, \( \mathbf{w} = (w_1, \ldots, w_n) \), is defined by \( w_i = k \) if and only if \( \theta_i = \phi_k \) for \( i = 1, \ldots, n \). Thus \( (\Phi, \mathbf{w}) \) is an equivalent representation of \( \Theta \), and hence the DP is also defined as

\[
[\theta_i|\theta_{-i}] \sim \frac{\alpha G_0(\cdot) + \sum_{k=1}^{c} n_k(-i) \delta(\theta_i|\phi_k)}{\alpha + n - 1},
\]

where \( n_k(-i) \) refers to the cardinality of cluster \( k \), with \( \theta_i \) removed, and

\[
\phi_k \overset{iid}{\sim} G_0(\cdot), \ k = 1, \ldots, c.
\]
An Alternative View: Chinese Restaurant Processes
Stick-Breaking Priors

The Stick-Breaking Prior is defined by

\[
P(\cdot) = \sum_{k=1}^{K} w_k \delta(\cdot | \phi_k), \quad \phi_k \overset{iid}{\sim} G_0
\]

\[
w_1 = q_1 \text{ and } w_k = q_k \prod_{l=1}^{k-1} (1 - q_l), \quad q_k \sim \text{Beta}(a_k, b_k)
\]

where the number of atoms \( K \) can be finite or infinite.
Stick-Breaking Priors

- The Stick-Breaking Prior is defined by

\[
P(\cdot) = \sum_{k=1}^{K} w_k \delta(\cdot | \phi_k), \quad \phi_k \overset{iid}{\sim} G_0
\]

\[
w_1 = q_1 \text{ and } w_k = q_k \prod_{l=1}^{k-1} (1 - q_l), \quad q_k \sim \text{Beta}(a_k, b_k)
\]

where the number of atoms \( K \) can be finite or infinite.

- In the finite case, setting \( q_K = 1 \) guarantees that \( \sum_{k=1}^{K} w_k = 1 \) with probability 1.
In the infinite case,

\[ \sum_{k=1}^{\infty} w_k = 1 \text{ a.s. iff } \sum_{k=1}^{\infty} \mathbb{E}(\log(1 - q_k)) = -\infty. \]

Alternatively, it is sufficient to check that

\[ \sum_{k=1}^{\infty} \log(1 + a_k/b_k) = +\infty \] (Ishwaran and James, 2001).
Stick-Breaking Priors

- In the infinite case,

\[
\sum_{k=1}^{\infty} w_k = 1 \text{ a.s. iff } \sum_{k=1}^{\infty} \mathbb{E}(\log(1 - q_k)) = -\infty.
\]

Alternatively, it is sufficient to check that

\[
\sum_{k=1}^{\infty} \log(1 + a_k/b_k) = +\infty \quad \text{(Ishwaran and James, 2001)}.
\]

- For example, \( a_k = 1 - a \) and \( b_k = b + ka \) for \( 0 \leq a \leq 1 \) and \( b > -a \) leads to the two-parameter Poisson-Dirichlet process, as known as the Pitman-Yor process.
Stick-Breaking Priors

- In the infinite case,

\[
\sum_{k=1}^{\infty} w_k = 1 \text{ a.s. iff } \sum_{k=1}^{\infty} E(\log(1 - q_k)) = -\infty.
\]

Alternatively, it is sufficient to check that
\[
\sum_{k=1}^{\infty} \log(1 + a_k/b_k) = +\infty \text{ (Ishwaran and James, 2001)}.
\]

- For example, \(a_k = 1 - a\) and \(b_k = b + ka\) for \(0 \leq a \leq 1\) and \(b > -a\) leads to the two-parameter Poisson-Dirichlet process, as known as the Pitman-Yor process.

- Especial cases: \(q_k \overset{iid}{\sim} \text{Beta}(1, \alpha)\) where \(\alpha > 0\), that is, \(a = 0\) and \(b = \alpha\) (Sethuraman, 1994); \(q_k \overset{iid}{\sim} \text{Beta}(1-a, ka)\) (Pitman and Yor, 1997).
An Illustration for Stick-Breaking Priors

stick-breaking weights

\( \alpha = 5 \)

stick indices

\( \alpha = 5 \)
The Stick-Breaking Representation and Pólya urn scheme

- The Pólya urn scheme makes inference and computation more feasible. Moreover, clustering property of the DP is very useful in practical applications. For example, it can be used for automatical choice of the number of classes in clustering analysis.
The Stick-Breaking Representation and Pólya urn scheme

- The Pólya urn scheme makes inference and computation more feasible. Moreover, clustering property of the DP is very useful in practical applications. For example, it can be used for automatical choice of the number of classes in clustering analysis.

- The Stick-Breaking Representation makes modeling more powerful. But the corresponding computation usually requires a truncated technique in the infinite case.
Dependent Dirichlet Processes

Dependent Dirichlet Processes (DDPs) provide a general framework to describe dependency among a collection of stochastic processes. A principled approach to this direction is to treat the weights in the stick-breaking representation as stochastic functions (DeIorio et al, 2004).
Dependent Dirichlet Processes

- Dependent DPs (DDPs) provide a general framework to describe dependency among a collection of stochastic processes.
Dependent Dirichlet Processes

- Dependent DPs (DDPs) provide a general framework to describe dependency among a collection of stochastic processes.
- A principled approach to this direction is to treat the weights in the stick-breaking representation as stochastic functions (DeIorio et al, 2004).
Dependent Dirichlet Processes

Other ways of achieving dependence among random measures include the hierarchical DP model (Teh et al, 2006), the use of linear combinations of realizations of independent DPs (Muller et al, 2004) and kernel-weighted mixture of DPs (Dunson et al, 2007). These are specialized approaches that can make use of generalized Pólya urn schemes for posterior inference and prediction.

It is also desirable to address this issue under the "single-p" DPP (MacEachern, 2000), because inference and computation for the resulting model can proceed via a relatively straightforward application of the well-established MCMC techniques.
Other ways of achieving dependence among random measures include the hierarchical DP model (Teh et al, 2006), the use of linear combinations of realizations of independent DPs (Muller et al, 2004) and kernel-weighted mixture of DPs (Dunson et al, 2007). These are specialized approaches that can make use of generalized Pólya urn schemes for posterior inference and prediction.
Dependent Dirichlet Processes

- Other ways of achieving dependence among random measures include the hierarchical DP model (Teh et al, 2006), the use of linear combinations of realizations of independent DPs (Muller et al, 2004) and kernel-weighted mixture of DPs (Dunson et al, 2007). These are specialized approaches that can make use of generalized Pólya urn schemes for posterior inference and prediction.

- It is also desirable to address this issue under the “single-$p$” DPP (MacEachern, 2000), because inference and computation for the resulting model can proceed via a relatively straightforward application of the well-established MCMC techniques.
Assume there are a collection of random probability measures $G_j$ on $(\Phi, B)$. In general, there are two extreme constructions for the $G_j$. For the first construction, $G_j$ are treated as independent DPs given hyperparameters $\theta$, so the model is equivalent to the separate submodels. For the second one, the model is treated as a single conventional DP, i.e., $G_1 = \cdots = G_q$. Clearly, the first case allows too little sharing of strength in many applications, while the second case enforces too much sharing.
Dependent Dirichlet Processes

- Assume there is a collection of a collection of random probability measures $G_j$ on $(\Phi, \mathcal{B})$.
- In general, there are two extreme constructions for the $G_j$. For the first construction, $G_j$ are treated as independent DPs given hyperparameters $\theta$, so the model is equivalent to the $m$ separate submodels. For the second one, the model is treated as a single conventional DP, i.e., $G_1 = \cdots = G_q$. Clearly, the first case allows too little sharing of strength in many applications, while the second case enforces too much sharing.
The Kernel Weighted Mixture of DPs

Assume there a set of samples \( \{(x_i, y_i), i = 1, \ldots, n\} \) where \( x_i \in \mathbb{R}^p \) is an input vector and \( y_i \) is the corresponding output.

The kernel weighted mixture of DPs (Dunson et al, 2007) is specified as

\[
G_x = \sum_{l=1}^{n} b_l(x) G_l^*, \quad G_l^* \overset{iid}{\sim} \text{DP}(\alpha G_0), \text{ for } l = 1, \ldots, n,
\]

where \( b_l(x) \) is a kernel-based weight.
Capture relationships among multiple studies

- In order to model relationships among multiple studies, we consider a nonparametric hierarchical model. Let $y_i = (y_{i1}, \ldots, y_{iq})^T$, $y_{j} = (y_{1j}, \ldots, y_{nj})^T$ denote the response vector in study $j$. The model is

$$[y_{ij} | b_{ij}] \overset{ind}{\sim} F(y_{ij} | b_{ij}), \; j = 1, \ldots, q \text{ and } i = 1, \ldots, n_j;$$

$$[b_{ij} | G_j] \overset{iid}{\sim} G_j, \; i = 1, \ldots, n_j \text{ for each } j.$$
The Nested Dirichlet Process

The Nested Dirichlet Process (Rodriguez et al., 2008) is defined by

\[ G_j(\cdot) \sim Q \triangleq \sum_{k=1}^{\infty} \pi^*_k \delta(\cdot | G^*_k), \]

\[ G^*_k(\cdot) \triangleq \sum_{l=1}^{\infty} w_{lk} \delta(\cdot | \phi_{lk}), \]

\[ \phi_{lk} \sim G_0, \]

with \( w_{lk} = u_{lk} \prod_{s=1}^{l-1} (1-u_{sk}), \quad \pi^*_k = v_k \prod_{s=1}^{k-1} (1-v_s), \)
\[ v_k \sim \text{Beta}(1, \alpha), \quad \text{and} \quad u_{lk} \sim \text{Beta}(1, \beta). \]
The Conditional Autoregressive DP

- We model the $G_j$ in autoregressive form of

$$G_j = \omega_{jj} G_j^* + \sum_{l \neq j} \omega_{jl} G_l, \quad j = 1, \ldots, q,$$  \hspace{1cm} (1)

where $0 \leq \omega_{jl} < 1$ and $\sum_{l=1}^q \omega_{jl} = 1$.

- From (1), we get the following conditional autoregressive model

$$E(G_j(A)|G_l(A), l \neq j) = \omega_{jj} G_0(A) + \sum_{l \neq j} \omega_{jl} G_l(A)$$

for any Borel set $A \in \mathcal{A}$. We thus say the $G_j$ defined by (1) follow conditional autoregressive DPs.
Graphical Representations

Figure: Graphical Representations: (a) the conditional autoregressive DP model, (b) $q$ independent DP mixture models, and (c) DP mixture model.
The Spatial DP Model

- The spatial DP (sDP) model (Gelfand et al., 2005) is

\[
\begin{align*}
[y_j | s_j] & \overset{ind}{\sim} F(\cdot | s_j), \quad j = 1, \ldots, q, \\
[s_j | G] & \overset{iid}{\sim} G, \quad j = 1, \ldots, q, \\
[G | \alpha, G_0] & \sim \text{DP}(\alpha G_0).
\end{align*}
\]

Furthermore, the base distribution is defined as a Gaussian process (GP). Specifically, this model describes the dependence among the response variates via DP, and the dependence among the instances via GP.
The Matrix-Variate DP Model

- The Matrix-Variate DP mixture (Zhang et al., 2010) is

\[
[y_i|\Theta_i] \overset{ind}{\sim} F(\Theta_i), \quad i = 1, \ldots, n,
\]

\[
[\Theta_i|G] \overset{iid}{\sim} G, \quad i = 1, \ldots, n,
\]

\[
G \sim \text{DP}(\alpha G_0).
\]

Here \(\Theta_i\)'s are a collection of matrices of the same size (e.g., \(p \times q\)), and we assume that the base probability measure \(G_0\) follows a matrix-variate distribution. We thus refer to the resulting DP as a matrix-variate DP (MATDP).

- As a concrete example, let \(G_0\) follow a matrix-variate normal distribution of the form

\[
G_0(\cdot|\Sigma, \Lambda) = N_{p,q}(\cdot|M, A \otimes B).
\]
Graphical Representations

Figure: Graphical representations under regression setting: (a) MATDP and (b) sDP.
Let $\theta_i = (\mu_i, \Sigma_i)$ be the parameter of Gaussians for $i = 1, \ldots, N$;

- Let $x_i \sim N(\cdot | \theta_i)$, $\theta_i \sim G$ and $G \sim \text{DP}(\alpha G_0)$;
- $G_0$ is defined as a Normal-Inverse Wishart $\text{NIW}$. 
Dirichlet Process Mixture of Gaussian Density

- Different values of $\alpha$ yield different number of components.
Dirichlet Process Mixture of Gaussian Density

- The Two Moon Data: a more illustrative example for automatically determining the number of components.
We consider a regression problem on a training dataset \( \{(x_i, y_i)\}_{i=1}^{n} \), where \( x_i \subset \mathbb{R}^p \) is an input or covariate vector and \( y_i \in \mathbb{R}^q \) is a \( q \)-dimensional continuous vector of responses, and we also treat multivariate classification problems, specifically the multi-class classification problem and the multi-label prediction problem. In the latter problem, the label associated with an input vector \( x \) is \( y \in \{0, 1\}^q \). Unlike the conventional multi-class classification problem in which \( x \) belongs to one and only one class, in the multi-label problem \( x \) is allowed to belong simultaneously to several classes.
Nonlinear Supervised Learning Models

We are currently interested in jointly modeling two types of relationships: the dependency among the data instances and the dependency among the response (or input) variates. This is a challenging and interesting issue, because it provides leverage on problems where the data are not iid (independent and identically distributed) while the (co)variates are not independent.
Dirichlet Process Multinomial Logit (dpMNL) models

The specification of dpMNL (Shahbaba and Neal, 2009) is

\[
\begin{align*}
(x_i, y_{ij}) | b_{ij}, \sigma^2, \mu_i, \Sigma_i & \quad \overset{\text{ind}}{\sim} \quad \text{MNL}(y_{ij} | x_i^T b_{ij}) \times \\
N_n(x_i | \mu_i, \Sigma_i), & \quad i = 1, \ldots, n; \\
(b_{ij}, \mu_i, \Sigma_i) | G_j & \quad \overset{\text{iid}}{\sim} \quad G_j, \quad i = 1, \ldots, n; \\
G_j | \nu, G_0 & \quad \overset{\text{iid}}{\sim} \quad \text{DP}(\alpha G_0).
\end{align*}
\]
Dirichlet Process Latent Factor Models (DP-LFM)

The specification of DP-LFM (Zhang et al., 2013) is

\[
\begin{align*}
x_i & \sim N(\cdot | A_i \eta_i + \mu_i, \Sigma_i), \\
y_i & \sim F(\cdot | B_i \eta_i + \nu_i, \Lambda_i), \\
\eta_i & \sim N(\cdot | 0, I_r) \\
\theta_i | G & \sim G(\cdot), \\
G & \sim DP(\alpha G_0),
\end{align*}
\]

where \( \Sigma_i = \text{diag}(\sigma_{i1}^2, \ldots, \sigma_{ip}^2) \), \( \Lambda_i = \text{diag}(\lambda_{i1}^2, \ldots, \lambda_{iq}^2) \), and the \( \theta_i = \{A_i, B_i, \mu_i, \nu_i, \Sigma_i, \Lambda_i\} \) are the parameters following a joint distribution generated from the DP prior \( DP(\alpha G_0) \).
Dirichlet Process Mixture of Gaussian Density

(a) dpMNL

(b) DP-LFM
Summary

- Introduce two definitions of DP priors: Pólyn urn scheme and stick-braking representation.
- Introduce modeling approaches to DDPs.
- Some Applications in Machine Learning.
Several Active Issues

▶ Extensions: Beta Processes (Indian Buffet Processes) and Pólyn Tree Processes
Several Active Issues

- Extensions: Beta Processes (Indian Buffet Processes) and Pólyn Tree Processes
- Efficient Computation: Sequential Monte Carlo and Online Variational EM algorithms.
References: DP Models


References: Computations


References: Computations


References: DDPs


References: DDPs


References: DDPs


References: Applications


Thanks

Questions & Comments!