COS 597C: Bayesian nonparametrics

Lecturer: David Blei | Lecture # 2
Scribes: Tao Yue, Eugene Brevdo | September 24, 2007

Markov Chain Monte Carlo Sampling for Dirichlet Process Mixture Models

In this lecture, we will begin by discussing Dirichlet processes, and progress onto Dirichlet process mixtures. Then we will review conjugacy and Markov Chain Monte Carlo (MCMC) methods. Finally, we will arrive at MCMC sampling for Dirichlet Process mixtures, which was presented in the assigned reading for this week's seminar.

Dirichlet Process (DP)

Let us make the connection explicit between the Dirichlet Process and the Chinese Restaurant Process presented last week. It is worth emphasizing that the Dirichlet Process is a distribution over distributions, which is parameterized by a scalar α and by the base distribution G_0. Note that the expression αG_0 is a measure which integrates to α.

Then, each draw from the Dirichlet Process would result in a distribution:

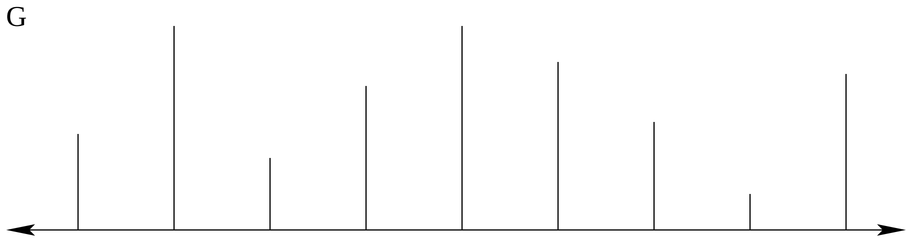
$$G \sim DP(\alpha G_0)$$

The space of G is the same space as G_0. If G_0 is a Gaussian, for example, then G would be a distribution over $\Re$. Similarly, the space of G would be $\Re^P$ for a multivariate Gaussian, the positive integers for a Poisson distribution, and $\Re^+$ for a Gamma distribution.

G is not present simply for the purposes of being integrated out. Sometimes, we do explicitly integrate G, although we will not do that for some time.

Some properties of a Dirichlet process:

1. Draws from $DP(\alpha G_0)$ are *discrete*. There is a positive probability of drawing certain numbers, unlike, say, a Gaussian where the probability of drawing any one number is always zero.



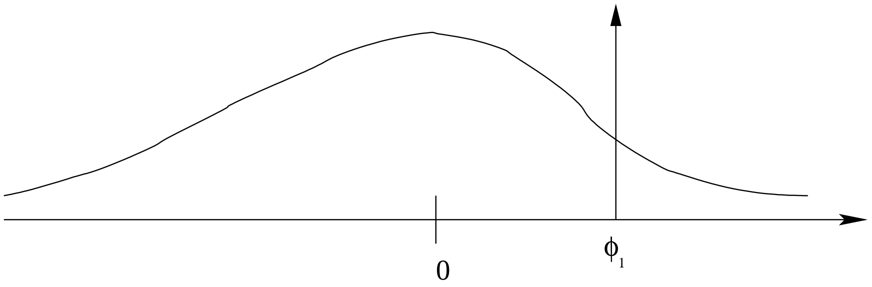
This is both interesting and limiting. The goal of a Dirichlet Process is to represent arbitrary distributions. This doesn't seem very arbitrary at first glance. But for now, simply consider it a property of the Dirichlet Process.

2. The posterior distribution of a Dirichlet Process is still a Dirichlet Process. This is a form of conjugacy, which we shortly will discuss in greater detail.

   Suppose we draw a value ϕ_n IID from the Dirichlet Process: (using Nealish notation rather than the notation we used last week):

$$G \sim DP(\alpha G_0)$$
$$\phi_n \overset{\text{iid}}{\sim} G$$

   Then the posterior distribution will be a Dirichlet process with a Dirac delta at ϕ_n. For example, after drawing the first value ϕ_1, the posterior distribution will look like this:



$$G|\{\phi_1, \alpha G_0\} \sim DP(\alpha G_0 + \delta_{\phi_1})$$

When we normalize the distribution to integrate to one, the presence of the Dirac delta δ_{ϕ_1} will give a greater probability to the value ϕ_1, and less to the initial distribution αG_0. Note that α functions as a scaling factor, where smaller α gives greater weight to ϕ_1 in the posterior distribution.

3. Because the Dirichlet Process is a distribution of distributions, the expectation is a distribution. Let the numerator be the base measure, and normalize to get:

$$\begin{aligned}\mathsf{E}[G|\alpha G_0] &= \frac{\alpha G_0}{\int \alpha G_0(\phi)d\phi} \\ &= \frac{\alpha G_0}{\alpha \int G_0(\phi)d\phi} \\ &= \frac{\alpha G_0}{\alpha} \\ &= G_0\end{aligned}$$

Notice that G is discrete, but the expectation is not discrete. It is interesting that the expectation does not possess the same properties as each individual draw from the distribution.

The DP and the Chinese Restaurant Process

After N draws from a Dirichlet Process, the posterior distribution is:

$$G|\phi_{1:N}, \alpha G_0 \sim DP\left(\alpha G_0 + \sum_{n=1}^{N} \delta_{\phi_n}\right)$$

Then the probability of the next draw from the Dirichlet Process is:

$$\begin{aligned}P(|\phi_{1:N}) &= \int G P(G|\phi_{1:N})dG \\ &= \mathsf{E}[G|\phi_{1:N}] \\ &= \frac{\alpha G_0 + \sum_{n=1}^{N} \delta_{\phi_n}}{\alpha + N}\end{aligned}$$

In other words, with probability proportional to α, the next draw will be a Gaussian. With probability proportional to 1, we will get a value that has already been drawn.

To put it more formally:

$$P(\phi|\phi_{1:N}) = \begin{cases} \text{draw from } G_0 & \text{with probability} \frac{\alpha}{\alpha+N} \\ \phi_N & \text{with probability} \frac{1}{\alpha+N} \end{cases}$$

The denominator is simply a normalizing factor, as follows:

$$\int \alpha G_0 + \sum_{n=1}^{N} \delta_{\phi_n} = \alpha \int G_0(\phi) d\phi + \sum_{n=1}^{N} \int \delta_{\phi_n} d\phi \qquad = \alpha + N$$

Notice that the denominator $\alpha + N$ is the same as in the Chinese Restaurant Process. So in order to get to the Chinese Restaurant Process from the Dirichlet Process, we need to argue that the ϕ_n's have a clustering structure.

$$P(\phi_1, ..., \phi_N) = P(\phi_1)P(\phi_2|\phi_1)...P(\phi_N|\phi_{1:(N-1)})$$

Suppose $\phi_{1:N}$ takes on the unique values $\phi^*_{1:K}$. Then with probability $\frac{\alpha}{\alpha+N}$, we will obtain a Gaussian. Because Gaussians do not repeat, this corresponds to a new value. Similarly, with probability $\frac{\text{number of occurrences}}{\alpha+N}$, we will obtain an existing value. Arbitrarily map the discrete real values to table numbers, and we see that the partition structure of the Dirichlet Process is exactly that of the Chinese Restaurant Process.

Just to recap, then:

$$\begin{aligned} G &\sim DP(\alpha G_0) \\ \phi_n &\overset{\text{iid}}{\sim} G \\ P(\phi_1 ... \phi_N) &= \int \prod_{n=1}^{N} P(\phi_n|G)P(G)dG \end{aligned}$$

As an aside, note that this is a different model from other distributions such as:

$$\mu \sim N(0,1)\phi_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$$

The Dirichlet Process gives our Gs very special properties.

Summary

$$G|\phi_1 \sim DP(\alpha G_0 + \delta_{\phi_1})$$

$$\begin{aligned} P(G|\phi_1) &\propto P(G, \phi_1) \\ &= P(G)P(\phi_1|G) \\ &= \mathrm{DP} \cdot G(\phi_1) \end{aligned}$$

$$\begin{aligned} P(G|\phi_1, \phi_2) &\propto P(G, \phi_2|\phi_1) \\ &= P(G|\phi_1)P(\phi_2|G) \\ &= \mathrm{DP} \cdot G(\phi_2) \\ &\sim \mathrm{DP}(\alpha G_0 + \delta_{\phi_1} + \delta_{\phi_2}) \end{aligned}$$

Here, we have defined the Dirichlet Process implicitly, but understanding its properties led to the development of a model in which it is identical to the Chinese Restaurant Process. Later on, we shall define the Dirichlet Process more constructively.

Dirichlet Process Mixtures

$$P(\phi_1, ..., \phi_N) = P(\phi_1)P(\phi_2|\phi_1)...P(\phi_N|\phi_{1:(N-1)})$$

This is equivalent to the CRP mixture:

$$\begin{aligned} \phi_i^* &\sim G_0 \\ z_n &\sim CRP(\alpha; z_1, ..., z_{n-1}) \\ x_n &\sim P(x|\phi_{z_n}^*) \end{aligned}$$

From now on, we can equivalent use the CRP form or the DP form:

$$\begin{aligned} G &\sim DP(\alpha G_0) \\ \phi_n &\overset{\text{iid}}{\sim} G \\ x_n &\sim P(x|\phi_n) \end{aligned}$$

Computing Posteriors

Conjugacy

A basic definition of conjugacy: the posterior is in the same family as the prior. The best reference for conjugacy is Gelman; it's crystal-clear there.

We saw one example of this already:

$$\begin{aligned} G &\text{ is a DP} \\ \phi &\sim G \\ G|\phi &\text{ is still a DP} \end{aligned}$$

More examples of conjugacy:

Prior	$P_0(\theta)$	Dirichlet	Beta	Gaussian
Likelihood	$P(X\|\theta)$	Discrete	Binomial	Gaussian
Posterior	$P(\theta\|X)$	Dirichlet	Beta	Gaussian

A more formal definition: Let

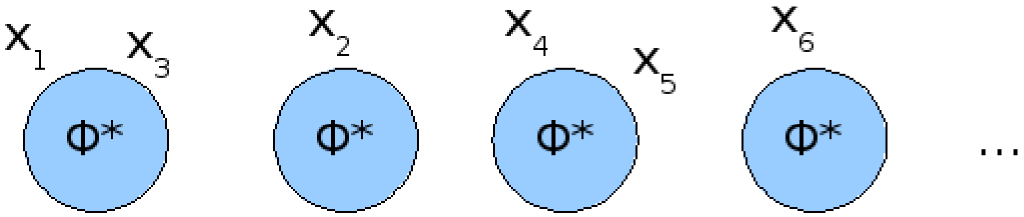
$$\begin{aligned} \phi &\sim G_0 \\ X &\overset{\text{iid}}{\sim} F(\phi) \\ \text{Call } \phi|X &\sim \hat{G}_{0,X} \end{aligned}$$

Then if $\hat{G}_{0,X}$ is in the same family as G_0, then G_0 is a conjugate prior.

In general, *every* exponential family has a conjugate prior.

We use conjugate priors to make Bayesian computation easier. Aside: a reference prior is the case where you let the prior be totally uninformative, (the KL divergence between the prior and the posterior when the number of observations is asymptotically large is a maximum), i.e. we let the data speak for itself. Then a conjugate prior is the exact opposite of the reference prior, in that we are maximizing rather than minimizing the effect of the prior.

Let's assume that G_0 is the conjugate prior to $P(X|\phi)$



Here, the X_i's are chosen from a distribution according to the parameter at the table.

A Basic Review of MCMC

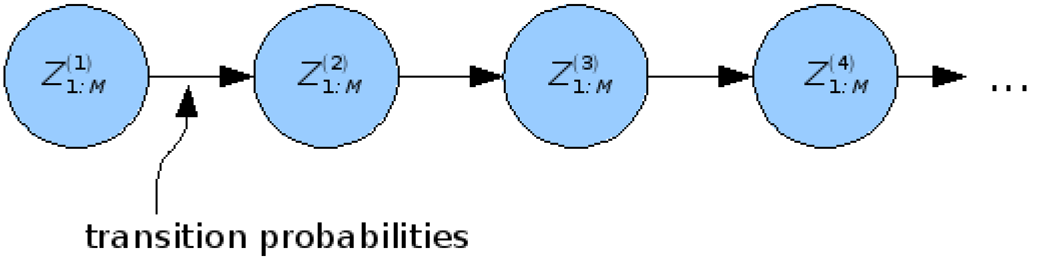
A more thorough exposition may be found in (Neal 93). Neal is a very thorough author who posts source code, compiler source code, and PRNG seed values, so you can always reproduce his results exactly. Gelman gives a practical approach, which is suitable for our purposes.

Let
$Z_{1:M}$: all of the hidden variables
$X_{1:N}$: the observations
θ: the fixed values (hyperparameters)

We want the posterior

$$P(Z_{1:M}|X_{1:N}, \theta)$$

How do we do this? The basic idea of MCMC is to build a Markov Chain on $Z_{1:M}^{(t)}$, t being the iteration index, such that the stationary distribution $Z_{1:M}^{(\infty)}$ of this MC is the posterior.



For large T, $Z_{1:M}^{T}$ should approximate a draw from π, the stationary distribution.

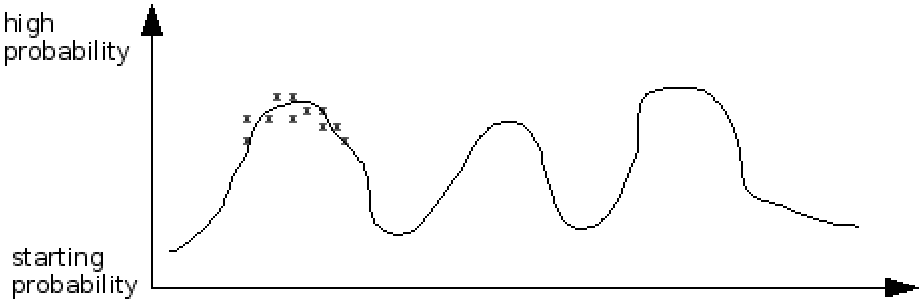
Once we have collected B samples from π, we can estimate the posterior

$$P(Z_{1:M}|X_{1:N},\theta) \approx \frac{1}{B}\sum_{i=1}^{B}\delta_{Z_{1:M}}(i)$$

However, to do this we do not take *all* the iterations of samples, because this may not lead us to a good estimate. Some basic terminology and concepts in MCMC sampling:

- Burn-in: The number of iterations between the first sample and the first sample from the approximate stationary distribution.
- Lag: The number of iterations between actual samples taken after burn-in.

To estimate sufficient independence of samples taken after burn-in, one common metric is the autocorrelation. Typically we care only about one of the modes, so this is a reasonable heuristic to use.



Gibbs Sampling

Gibbs Sampling is the simplest of all MCMC algorithms. It is a special case of Metropolis-Hastings.

Let $Z_{-i} = Z \backslash Z_i = \{Z_1, \cdots, Z_{i-1}, Z_{i+1}, \cdots, Z_M\}$. We sample

$$Z_i \sim P(Z_i | Z_{-i}, X_{1:N})$$

The proof of convergence of this sampling to π is given in (Neal 93).

When G_0 is conjugate, we only need $P(Z_{1:N}|X_{1:N}, G_0, \alpha)$, where $Z_{1:N}$ are the table assignments. That is, we do not need the ϕ's; we've marginalized out the DP parameters. Quick example: $G_0 \sim \mathcal{N}(0, 10), \alpha = 1.0$.

Sometimes we look for the predictive distribution $P(X|X_{1:N})$. This can be calculated by marginalizing out a number of parameters, and using conjugate priors. The remainder is a derivation of how to estimate this distribution.

$$P(X|X_{1:N}) = \sum_{Z_{1:N}} P(Z_{1:N}|X_{1:N}) P(X|Z_{1:N}, X_{1:N}) \tag{1}$$

$$P(X|Z_{1:N}, X_{1:N}) = \sum_{i=1}^{1+K_n} P(Z = i|Z_{1:N}) P(X|Z = i, Z_{1:N}, X_{1:N}) \tag{2}$$

$$P(X|Z = i, Z_{1:N}, X_{1:N}) = \int P(\phi_i^*|Z = i, Z_{1:N}, X_{1:N}) P(X|\phi_i^*) d\phi_i^* \tag{3}$$

Where in (**??**) we marginalize over the partition structure and use the chain rule, in (**??**) we marginalize over all possible current and future tables.

In (**??**), the first term in the integral is a conjugate prior, so we can calculate the integral analytically, and represent it as $b_i(X; Z_{1:N}, X_{1:N})$. Note that we sample to approximate the predictive distribution:

$$P(X|X_{1:N}) \approx \frac{1}{B} \sum_{b=1}^{B} P(X|Z_{1:N}^{(b)}, X_{1:N})$$