COS 597C: Bayesian Nonparametrics

Lecturer: David Blei Lecture #3
Scribes: Jordan Boyd-Graber and Francisco Pereira October 1, 2007

1 Gibbs Sampling with a DP

First, let's recapitulate the model that we're using. We assume that each table has an associated parameter ϕ which comes from some base measure G_0 for $k = 1 \ldots$

$$\phi_k^* \sim G_0. \tag{1}$$

Each data point comes from some table given by

$$z_n \sim \mathrm{CRP}(\alpha, z_{1:n-1}). \tag{2}$$

And then the data itself comes from the parameter associated with that table:

$$x_n \sim p(\cdot\,|\phi_{z_n}^*) \tag{3}$$

Last time, we left off with the predictive distribution

$$p(x|x_{1:N}) = \sum_{z_{1:N}} p(z_{1:N}|x_{1:N}) p(x|z_{1:N}, x_{1:N}), \tag{4}$$

where the sum ranges over all possible seating arrangements $z_{1:N}$. The quantity on the left can also be viewed as the expectation of $p(x|z_{1:N}, x_{1:N})$ under the distribution $p(z_{1:N}|x_{1:N})$, i.e. $\mathbb{E}_{p(z_{1:N}|x_{1:N})}[p(x|z_{1:N},, x_{1:N})]$.

1.1 Sampling Equations

Last time, we showed that $p(x|z_{1:N}, x_{1:N})$ can be computed exactly when $p(\cdot\,|\phi_{z_n}^*)$ is conjugate to G_0, and the other half is just the well-defined Chinese Restaurant Process. The problem, however, is that it is intractable to sum over all of the seating possibilities. Thus, we need approximate inference to handle the sum over z. What we can do is take the average of S samples from the true distribution to estimate the expectation

$$\mathbb{E}_{p(z_{1:N}|x_{1:N})}[p(x|z_{1:N},, x_{1:N})] \approx \frac{1}{S} \sum_{s=1}^{S} p(x|z_{1:N}^s, x_{1:N}) \tag{5}$$

using $z_{1:N}^s$ to denote the assignment to tables for sample s. To get these samples from the posterior, we use Gibbs sampling. We fix all but one z (denoted by z_{-i}) and compute the table probabilities for that z_i conditioned on all of the others, which is given by

$$\begin{aligned} p(z_n|z_{-n}, x_{1:N}) &= \frac{p(z_n, x_n|z_{-n}, x_{-n})}{p(x_n|z_{-n}, x_{-n})} && (6)\\ &\propto p(z_n, x_n|z_{-n}, x_{-n}) && (7)\\ &= p(z_n|z_{-n}, x_{-n}) p(x_n|z_n, z_{-n}, x_{-n}) && (8) \end{aligned}$$

after applying the chain rule. At this point, we can observe that $p(z_n|z_{-n}, x_{-n}) = p(z_n|z_{-n})$, as z_n is independent of x_{-n} given their respective table assignments z_{-n}, and is just given by the Chinese Restaurant Process. Note that z_n can take $K_{-n} + 1$ values, where K_{-n} is the number of tables occupied when we consider the variables z_{-n}.

The other term, however, requires us to marginalize over ϕ_{z_n}, which gives us

$$p(x_n|z_n, z_{-n}, x_{-n}) = \int_{\phi^*_{z_n}} p(x_n|\phi^*_{z_n})p(\phi^*_{z_n}|z_{-n}, x_n) \tag{9}$$

This can be viewed as the expected value of $p(x_n|\phi^*_{z_n})$ under the posterior distribution $p(\phi^*_{z_n}|z_{-n}, x_{-n})$. This depends on the distribution for generating the data, but if it's in the same family as G_0, then it's almost always possible to compute this exactly. For instance, if $G_0 \sim N(0, \sigma_0^2)$ and $P(X|\phi^*)$ is $N(\phi^*, \sigma_x^2)$, then this posterior is also normally distributed.

For Gibbs sampling we go through all of the seating assignments for a single iteration; after some number of iterations called the burn-in period, we begin taking samples from the distribution with a periodicity called the lag.

1.2 Implementation Details

- Autocorrelation is usually used to determine the lag and burn-in, as this is a measure of independence. However, in practice, these values are usually reported without justification.

- The indices of assignment aren't necessarily - and usually aren't - consistent across iterations of the Gibbs sampler. The trickiest part of the implementation is representing the table assignments.

- While you might expect α to have the biggest effect on the number of clusters (choosing a new table with probability proportional to α), the more relevant factor is the variance of the base measure. For example, if $G_0 : N(0, \sigma_0^2)$, a really small σ_0^2 will cause you to have a lot of clusters.

- Hyperparameters can also have a prior distribution put on them, but there is a "moment of truth" at some point when a parameter will have to be set, either by choice or resorting to cross-validation.

- The posterior will be multimodal. This is not a problem for our purposes, both because we want to find a single mode and also because modes will correspond to different yet equivalent assignments (e.g. points 1,4,3 sit at table A and 4 and 6 at table B is equivalent to reversing the table choices).

- If G_0 is not conjugate, you have to use Metropolis-Hastings, usually Algorithm 8 in Neal's paper.

1.3 Score

The score is proportional to the posterior (for a particular sample s) and is given by

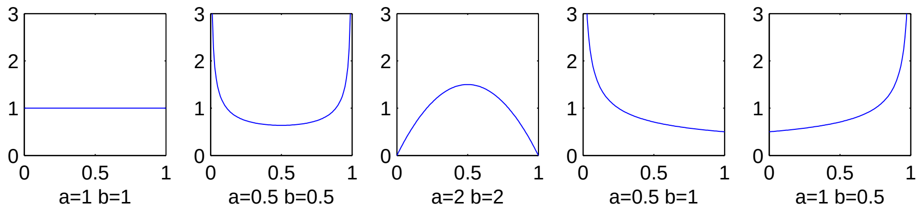

Figure 1: The probability density function of the Beta distribution for different settings of the parameters a and b.

$$\log(p(z^s_{1:N})p(x_{1:N}|z^s_{1:N})). \tag{10}$$

The goal is to get samples from a mode of the posterior; an increase in score followed by a plateau as the number of iterations increases is used as an indication that the samples are coming from a high probability region. Computing this is a good idea because it allows you to assess convergence. This still remains a tricky problem, however, as discussed in the Neal article, and tt's usually not addressed thoroughly in papers.

2 Dirichlet Distribution

The Dirichlet distribution is a distribution over vectors with k elements such that $G_i > 0$ and $\sum G_i = 1$, which means that the Dirichlet distribution is a distribution over the $k-1$ simplex. When $k = 2$, this is the β distribution, the p.d.f. of which is illustrated in Figure 1 for various values of the a and b parameters.

This distribution takes parameter α, a k-dimensional vector, and its probability density function is

$$p(G|\alpha) = \frac{\Gamma(\sum \alpha_i)}{\prod \Gamma(\alpha_i)} G_1^{\alpha_i - 1} \dots G_k^{\alpha_k - 1}. \tag{11}$$

We will now discuss properties of the Dirichlet distribution that will also apply to the Dirichlet process.

2.1 Influence of the parameter

In order to have an idea of how the setting of the parameters α_i influence the samples drawn from a Dirichlet, Figure 2 shows several draws at different settings (letting all α_i be the same value). If $\alpha_i < 1$ we have sparse distributions that concentrate the mass on a few or even a single value. If $\alpha_i >> 1$, they are centered around a point on the simplex (a uniform distribution). If $\alpha_i = 1$, the same probability is given to all the points on the simplex.

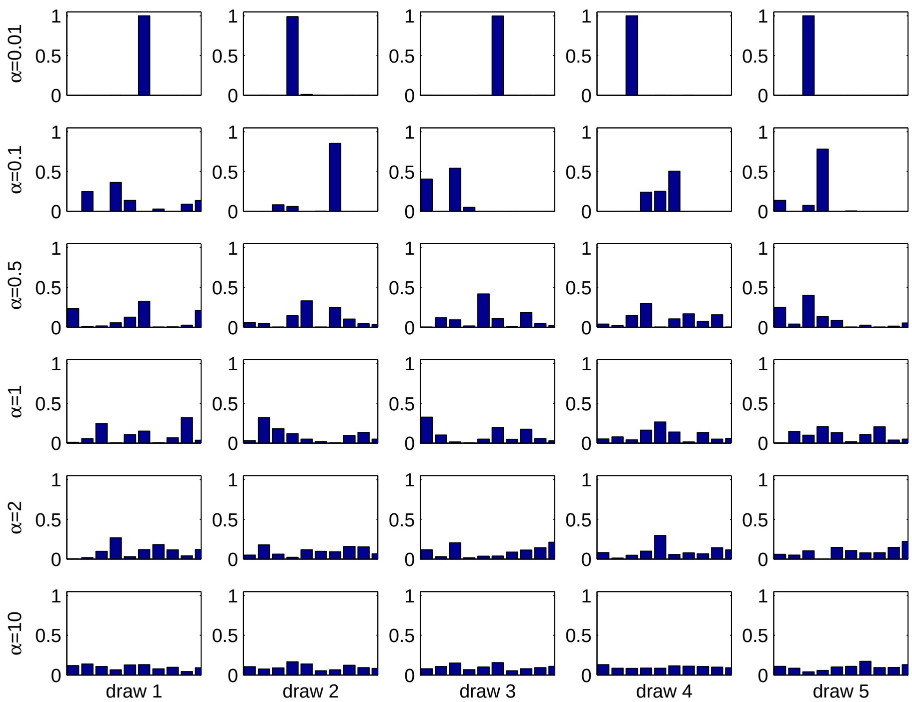

Figure 2: Each row shows five draws from a Dirichlet distributions with a given value of α_i.

2.2 Gamma Distribution

In order to generate draws from a Dirichlet we can take advantage of the following property

$$z_i \sim \text{Gamma}(\alpha_i), \tag{12}$$

then independent of $\sum z_i$,

$$\langle \frac{z_j}{\sum z_i}, \ldots, \frac{z_k}{\sum z_i} \rangle \sim \text{Dir}(\alpha) \tag{13}$$

by drawing one value from each of the variables z_i and combining them into a vector that will be a draw from a Dirichlet.

2.3 Partitions

A draw from a Dirichlet can be viewed as a distribution over X (e.g. the 20 loadings on a D&D 20-sided die). If we have a partition $A_1, \ldots, A_m$ of X (e.g. roll $1-8$ severe maiming, roll $9-16$ flesh wound, roll $17-20$ victory would be three partitions) then, taking $G(A_i) = \sum_{x \in A_i} \alpha_x$, we have that, if $G \sim \text{Dir}(\alpha)$, then for **any** partition $A_1, \ldots, A_m$ of X:

$$\langle G(A_1), \ldots, G(A_m) \rangle \sim \text{Dir}(\alpha'). \tag{14}$$

for $\alpha' = (\sum_{j \in A_1} \alpha_j, \ldots, \sum_{j \in A_m} \alpha_j)$.

In particular, this tell us that the marginal of G_i when we divide into just two partitions is

$$\langle A_1, A_2 \rangle \sim \text{Dir}(\alpha_i, \sum_{j \neq i} \alpha_j), \tag{15}$$

a Beta distribution. We can consider α as a measure on our discrete space X.

If we have $A_1, \ldots, A_m$ partitioning X, then

$$G(j|A_i) = \frac{G(j)}{G(A_i)} \tag{16}$$

for $j \in A_i$ we have two properties.

First, $\langle G(A_1), \ldots, G(A_m) \rangle, G(\cdot|A_1), \ldots, G(\cdot|A_m)$ are independent of each other. Moreover, $G(\cdot|A_m) \sim \text{Dir}(\alpha_{A_i})$ (in other words, the α restricted to A_i). This means that if you know the partition, then it tells you nothing about what's inside.

Secondly, this partition is neutral to the right. Thus, if we have a hierarchy of subsets

$$B_1 \supset B_2 \supset B_3 \supset \cdots \supset B_m, \tag{17}$$

then $G(B_1) \perp G(B_2|B_1) \perp \cdots \perp G(B_m|B_{m-1})$.

3 Expectation

The expectation of $G \sim \text{Dir}(\alpha)$ is given by

$$\mathbb{E}[G] = \frac{\alpha}{\sum_i \alpha_i} = \frac{\alpha}{\alpha(X)} = \bar{\alpha}, \tag{18}$$

where $\alpha(X)$ makes it sum to one and thus be a distribution on the k-simplex.

4 Posterior

Suppose that we have $G \sim \mathrm{Dir}(\alpha)$ and we have x_n i.i.d. from G. Then we have

$$\begin{aligned}
p(G|x_1 \dots x_n) &\propto p(G) \prod_{i=1}^{n} p(x_i|G) &(19)\\
&= p(G) \prod_{i=1}^{n} G_{x_i} &(20)\\
&= \prod_{i=1}^{k} G^{\alpha_i - 1 + n_i}, &(21)\\
& &(22)
\end{aligned}$$

In other words, we still have a Dirichlet, but with the number of x_i observed for each component added to our parameter α_i. This is analogous to the setting where we have an urn with multicolored balls, and for each ball that we remove from the urn we place another ball of the same color back in (i.e. adding one to the numerator for each x_i).

For a particular x_{n+1}, we then have

$$p(x_{n+1}|x_1, \dots, x_n) = \int G(x_{n+1}) p(G|x_1, \dots, x_n) = \frac{\alpha + \sum \delta_{x_i}}{\alpha(X) + n}. \tag{23}$$

We can thus interpret α as an unnormalized guess at G and, as we condition on observed data, the posterior becomes a convex combination of our prior and our empirical observation estimates. We can also view this as a new measure $\alpha' = \alpha + \sum_{i=1}^{n} \delta_{x_i}$

The predictive distribution can also be viewed as a convex combination of the distribution mean $\bar{\alpha}$ and the empirical distribution where the probability of X having outcome i is given by $\frac{n_i}{n}$ (i.e. the fraction of the n observations where the outcome was seen). More formally

$$p(x_{n+1} = i|x_1, \dots, x_n) = \frac{\alpha(X)}{\alpha(X) + n} \bar{\alpha} + \frac{n}{\alpha(X) + n} \frac{n_i}{n} \tag{24}$$

and it's clear that the magnitude of α determines how many observations are necessary before the empirical distribution has more influence than $\bar{\alpha}$ on the prediction.