Stat260: Bayesian Modeling and Inference **Lecture Date: April 28, 2010**

Lecture 24: Dirichlet distribution and Dirichlet Process

Lecturer: Michael I. Jordan *Scribe: Vivek Ramamurthy*

1 Introduction

In the last couple of lectures, in our study of Bayesian nonparametric approaches, we considered the Chinese Restaurant Process, Bayesian mixture models, stick breaking, and the Dirichlet process. Today, we will try to gain some insight into the connection between the Dirichlet process and the Dirichlet distribution.

2 The Dirichlet distribution and Pólya urn

First, we note an important relation between the Dirichlet distribution and the Gamma distribution, which is used to generate random vectors which are Dirichlet distributed. If, for $i \in \{1, 2, \cdots, K\}$,

$$Z_i \sim \text{ Gamma}(\alpha_i, \beta) \text{ independently,}$$

then

$$S = \sum_{i=1}^{K} Z_i \sim \text{ Gamma} \left(\sum_{i=1}^{K} \alpha_i, \beta \right)$$

and

$$V = (V_1, \cdots, V_K) = (Z_1/S, \cdots, Z_K/S) \sim \text{ Dir}(\alpha_1, \cdots, \alpha_K)$$

Now, consider the following Pólya urn model. Suppose that

- X_i - color of the ith draw
- $\mathcal{X}$ - space of colors (discrete)
- $\alpha(k)$ - number of balls of color k initially in urn.

We then have that

$$p(X_i = k | X_1, \cdots, X_{i-1}) = \frac{\alpha(k) + \sum_{j<i} \delta_{X_j}(k)}{\alpha(\mathcal{X}) + i - 1}$$

where $\delta_{X_j}(k) = 1$ if $X_j = k$ and 0 otherwise, and $\alpha(\mathcal{X}) = \sum_k \alpha(k)$. It may then be shown that

$$p(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n) = \frac{\alpha(x_1)}{\alpha(\mathcal{X})} \prod_{i=2}^{n} \frac{\alpha(x_i) + \sum_{j<i} \delta_{X_j}(x_i)}{\alpha(\mathcal{X}) + i - 1}$$

$$= \frac{\alpha(1)[\alpha(1)+1]\cdots[\alpha(1)+m_1-1]\alpha(2)[\alpha(2)+1]\cdots[\alpha(2)+m_2-1]\cdots\alpha(C)[\alpha(C)+1]\cdots[\alpha(C)+m_C-1]}{\alpha(\mathcal{X})[\alpha(\mathcal{X})+1]\cdots[\alpha(\mathcal{X})+n-1]}$$

where $1, 2, \cdots, C$ are the distinct colors that appear in $x_1, \cdots, x_n$ and $m_k = \sum_{i=1}^{n} 1\{X_i = k\}$.

3 The Pitman-Yor process

This section is a small aside on the Pitman-Yor process, a process related to the Dirichlet Process.

Recall that, in the stick-breaking construction for the Dirichlet Process, we define an infinite sequence of Beta random variables as follows:

$$\beta_i \sim \text{ Beta}(1, \alpha_0) \qquad\qquad i = 1, 2, \cdots$$

Then, we define an infinite sequence of mixing proportions as follows:

$$\begin{aligned} \pi_1 &= \beta_1 \\ \pi_k &= \beta_k \prod_{j<k} (1 - \beta_j) \qquad\qquad k = 2, 3, \cdots \end{aligned}$$

The Pitman-Yor process $PY(d, \alpha, G_0)$ is a related probability distribution over distributions. The parameters of this process are $0 \leq d < 1$ a discount parameter, a strength parameter $\theta > -d$ and a base distribution over G_0. In the special case of $d = 0$, the Pitman-Yor process is equivalent to the Dirichlet process.

Under the Pitman-Yor process, the infinite sequence of Beta random variables is defined as

$$\beta_i \sim \text{ Beta}(1 - d, \alpha + kd) \qquad\qquad i = 1, 2, \cdots$$

As in the Dirichlet Process, we complete the description of the Pitman-Yor process via

$$\begin{aligned} G &= \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k} \\ \theta_i | G &\sim G \end{aligned}$$

Hence, due to the way β_i are drawn, the Pitman-Yor process has a longer tail than the Dirichlet Process, that is, more sticks have non-trivial amounts of mass. The Pitman-Yor process may be seen as a Chinese restaurant process with a rule that favors starting a new table. In the homework, you will be required to show that the expectation of the total number of occupied tables in the Chinese restaurant scales as $O(\alpha n^d)$ under the Pitman-Yor process, $PY(d, \alpha, G_0)$. This is known as a power-law and is in contrast to the logarithmic growth for the Dirichlet Process that we discuss in the next section. Many natural phenomena follow power-law distributions, and in these cases, the Pitman-Yor process may be a better choice for a prior than the Dirichlet Process.

4 Expected number of occupied tables in the Chinese Restaurant Process

Going back to the Dirichlet Process, the stick-breaking construction/GEM distribution (named so by Ewens (1990) after Griffiths, Engen and McCloskey), of the β_i is given by $Beta(1, \alpha)$. In the resulting Chinese Restaurant Process, we have

$$P(\text{new table on } i\text{th draw}) = \frac{\alpha}{\alpha + i - 1}$$

We define

$$W_i = \begin{cases} 1 & \text{if } i\text{th draw is a new table} \\ 0 & \text{otherwise} \end{cases}$$

The expected number of occupied tables is then given by

$$E\left(\sum_{i=1}^{n} W_i\right) = \sum_{i=1}^{n} \frac{\alpha}{\alpha+i-1} \sim \alpha \log\left(\frac{\alpha+n}{\alpha}\right)$$

Moreover, we can also easily compute the probability of any sequence of draws. Consider, for instance, the sequence of draws given by $W = (1,1,0,0,1)$. The probability of this sequence is given by

$$\begin{aligned} P(W = (1,1,0,0,1)) &= \frac{\alpha}{\alpha} \cdot \frac{\alpha}{\alpha+1} \cdot \frac{2}{\alpha+2} \cdot \frac{3}{\alpha+3} \cdot \frac{\alpha}{\alpha+4} \\ &= \frac{\alpha^3 \Gamma(\alpha)}{\Gamma(\alpha+n)}(1 \cdot 1 \cdot 2 \cdot 3 \cdot 1) \end{aligned}$$

In general, it may be shown that the probability of drawing k tables in n draws is given by

$$P(k \text{ tables}|\alpha, n) = S(n,k)\frac{\alpha^k \Gamma(\alpha)}{\Gamma(\alpha+n)}$$

where $S(n,k)$ is a Stirling number of the first kind. The following identity may be shown to hold for Stirling numbers of the first kind:

$$S(n+1, m+1) = S(n,m) + nS(n, m+1)$$

5 Gibbs Sampling α for Dirichlet Process mixtures

Our goal now, is to sample α in a Gibbs sampler for Dirichlet Process mixtures. The sampling distributions used are

$$\begin{aligned} G &\sim DP(\alpha, G_0) \\ \theta_i|G &\sim G \\ X_{ij}|\theta_i &\sim F_{\theta_i} \end{aligned}$$

In addition, we also need to determine a prior on α. Toward this end, let $k =$ the number of occupied tables in a Chinese restaurant process. By the last computation, we have that

$$p(k|\alpha) \propto \frac{\alpha^k \Gamma(\alpha)}{\Gamma(\alpha+n)}$$

We also have that the Beta function is

$$B(\alpha_1, \alpha_2) = \int_0^1 x^{\alpha_1 - 1}(1-x)^{\alpha_2 - 1}dx = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(\alpha_1+\alpha_2)}$$

$$\Longrightarrow \frac{\Gamma(\alpha)}{\Gamma(\alpha+n)} = \frac{(\alpha+n)B(\alpha+1,n)}{(\alpha+n)\Gamma(\alpha+n)}$$

The above identity may be verified by checking that

$$B(\alpha+1, n) = \frac{\Gamma(\alpha+1)\Gamma(n)}{\Gamma(\alpha+1+n)} = \frac{\alpha\Gamma(\alpha)\Gamma(n)}{(\alpha+n)\Gamma(\alpha+n)}$$

Hence, it then follows that

$$p(\alpha|k) \propto p(\alpha)\alpha^{k-1}(\alpha+n)\int_0^1 x^{\alpha}(1-x)^{n-1}dx$$

Introducing $X \sim \text{Beta}(\alpha+1, n)$, we observe that $p(\alpha)$ must be chosen to be conjugate to the Gamma distribution. Since the Gamma distribution is conjugate to itself, it follows that $p(\alpha)$ must be chosen to be a Gamma distribution.

6 The Dirichlet Process

In this section, we discuss why the Dirichlet Process is named the Dirichlet Process.

Consider a set Φ and a partition $A_1, A_2, \cdots$ of Φ such that $\cup_k A_k = \Phi$. We would like to construct a random probability measure on Φ, i.e., a random probability measure G such that for all i, $G(A_i)$ is a random variable. For this, we need to specify a joint distribution for $(G(A_1), G(A_2), \cdots, G(A_k))$ for any k and any partition.

Ferguson (1973) showed that G is a Dirichlet process with parameters α_0 and G_0, i.e. $G \sim DP(\alpha_0, G_0)$ if for any partition $A_1, \cdots, A_k$, we have that

$$(G(A_1), \cdots, G(A_k)) \sim \mathrm{Dir}[\alpha_0 G_0(A_1), \cdots, \alpha_0 G_0(A_k)]$$

Sethuraman (1994) showed that the Dirichlet Process is an infinite sum of the form $G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$ that obeys the definition of the stick-breaking process.

We wish to give an overview of why these definitions are equivalent. First, we present three facts, two of which you need to prove in the homework and one which we simply state as a fact.

1. (Homework) Suppose that U and V are independent k-vectors. Let $U \sim \mathrm{Dir}(\alpha)$ and $V \sim \mathrm{Dir}(\gamma)$. Let $W \sim (\sum_i \alpha_i, \sum_i \gamma_i)$, independently of U and V. It may then be shown that

$$WU + (1-W)V \sim \ \mathrm{Dir}(\alpha + \gamma)$$

2. (Homework) Let e_j denote a unit basis vector. Let $\beta_j = \gamma_j / \sum_i \gamma_i$. In this case, it may be shown that

$$\sum_j \beta_j \mathrm{Dir}(\gamma + e_j) = \ \mathrm{Dir}(\gamma)$$

3. Let W, U be a pair of random variables where W take values in $[1, 1]$ and U takes values in a linear space. Suppose V is a random variable taking values in the same linear space as U and which is independent of (W, U) and satisfies the distributional equation

$$V \stackrel{st}{=} U + WV$$

where the notation $\stackrel{st}{=}$ stands for "has same distribution". If $P(|W| = 1) \neq 1$, then there is a unique distribution for V that satisfies this equation.

By the stick-breaking construction,

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k} = \pi_1 \delta_{\phi_1} + (1 - \pi_1) \sum_{k=2}^{\infty} \pi_k \delta_{\phi_k}$$

which gives us the distributional equation

$$G \stackrel{st}{=} \pi_1 \delta_{\phi_1} + (1 - \pi_1) G$$

Evaluating the LHS and RHS on a partition $(A_1, \cdots, A_k)$ gives us

$$V \stackrel{st}{=} \pi_1 X + (1 - \pi_1) V \tag{1}$$

where $\pi_1 \sim \text{Beta}(1, \alpha_0)$, X is $k-$vector that takes on the value e_j with probability $G_0(A_k)$, and V is independent of X and π_1.

We show that the $k-$dimensional Dirichlet distribution $V \sim \text{Dir}(G_0(A_1), \ldots, G_0(A_k))$ satisfies Equation (1) and therefore, by fact 3, V is the unique distribution to satisfy this. Therefore, V constructed via the stick-breaking construction and V as defined in the Dirichlet Process are equivalent.

Now to show that the $k-$dimensional Dirichlet distribution satisfies Equation (1). Let V on the RHS of Equation (1) be the $k-$dimensional Dirichlet distribution $\text{Dir}(G_0(A_1), \ldots, G_0(A_k))$. By definition, the $k-$dimensional Dirichlet distribution $\text{Dir}(e_j)$ assigns probability 1 to partition A_j . Now conditioning on $X = e_j$, the distribution of $\pi_1 X + (1 - \pi_1)V$ is $\pi_1 \text{Dir}(e_j) + (1 - \pi_1)\text{Dir}(G_0(A_1), \ldots, G_0(A_k))$. By fact 1, this is distributed $\text{Dir}((G_0(A_1), \ldots, G_0(A_k)) + e_j)$. Now integrating over the distribution of X where

$$P\{X = e_j\} = G_0(A_j) = \frac{\alpha_j}{\sum_i \alpha_i}$$

and using fact 2, we see that the RHS is distributed $\text{Dir}(G_0(A_1), \ldots, G_0(A_k))$.

Therefore, the stick-breaking construction and the mathematical definition of the Dirichlet Process are equivalent.

References

[1] Ferguson, T. S. 1973. A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics* 1 (2): 209-230.

[2] Sethuraman, J. 1994. A Constructive Definition of Dirichlet Priors. *Statistica Sinica* 4: 639-650.