Bayesian Nonparametrics I

Instructor: David Blei; *Scribe*: Kui Tang

25 Mar 2016 (Revised 1 Apr 2016)

We begin by discussing the central problem of model selection, and quickly illustrate how Bayesian nonparametrics can help us with that problem and a lot more. We briefly introduce the notion of random measures, before reviewing the Chinese restaurant process (CRP) and infinite mixture models. We then formally define the Dirichlet process, demonstrate its properties as a random measure, and then derive the CRP from the definition of a Dirichlet process. We conclude with the stick-breaking process, another construction of the Dirichlet process, but did not have time to derive it.

1 Model Selection

Model selection is a whole field by itself; you've heard lots of acronyms such as Akaike information criterion (AIC), Bayesian information criterion (BIC), etc. Bayesian nonparametrics (BNP) is one way to do model selection, and a lot more.

In theory, there are an "infinite" number of components, but conditional on finite data, only a finite number of components will be used in the posterior.

Three things distinguish BNP from model selection, which can be advantageous when doing unsupervised learning with complicated latent variables:

1. Predictive distribution can grow with the number of components. [For instance, a test image could belong to a new cluster, one that never appeared in the training set. Traditional model selection methods can't accommodate this.]

2. BNP can express priors over combinatorial structures. [Think of learning latent data structures—trees, graphs, etc. It's (NP-)hard to cross-validate, for instance, over all possible trees.]

3. Through inference, BNP methods *search* and *estimate* at the same time. [This is "efficient": it will not waste time on structures that are not likely to explain the data. But cross-validation has to waste time estimating bad models.]

Of course, you still have to do model selection/validation with BNP. Issues of consistency (do you estimate the "true" number of clusters, tree structure, etc?) are subtle. See work by Miller and Harrison at Brown.[1]

2 Random Measures

The space of all probability measures is infinite-dimensional (while the space of *parametric* statistical models are, by definition, finite). Therefore, we begin by discussing probability distributions over probability measures (hence *Bayesian* nonparametrics).

2.1 Dirichlet Process

What is a Dirichlet distribution really? If we sample from an n-dimensional Dirichlet distribution, we get finite-dimensional vectors that sum to 1. So the n-dimensional Dirichlet distribution is really a random (probability) measure on the space $\Omega = \{1, \dots, n\}$.

What if we consider random measures on more general spaces? Let's start with something unexotic, $\Omega = \mathbb{R}$. One could imagine randomly drawing a density function $d\mu$ over Ω, which defines a random measure.

[1] http://papers.nips.cc/paper/4880-a-simple-example-of-dirichlet-process-mixture-inconsistency-for-the-number-of-components.pdf

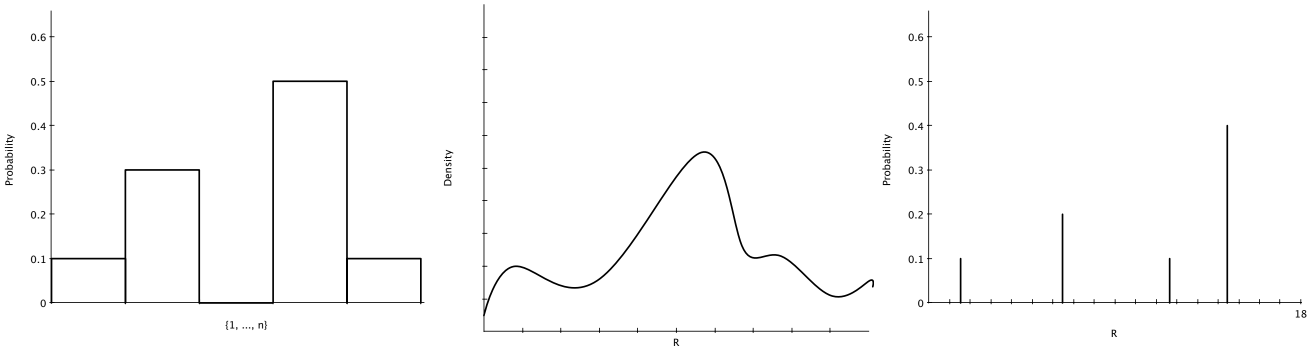

Figure 1: Illustration of a Dirichlet distribution (a finite-dimensional random (probability) measure), an arbitrary (absolutely continuous) random measure (represented as a density) over $\mathbb{R}$, and a Dirichlet process over $\mathbb{R}$. fig:random-measure

Instead of doing that, we will work with something more constrained (and thus having richer properties). We will consider the Dirichlet *process* over $\mathbb{R}$ as the infinite-dimensional analogue of the Dirichlet distribution. Whereas samples from the Dirichlet distribution assign probability mass to each of the n items in $\{1, \ldots, n\}$, samples from a Dirichlet process assign probability mass to a countably infinite subset of $\mathbb{R}$.

See Figure 1 for an illustration.

Rather than n-dimensional probability vectors, in a Dirichlet process, we get a countably infinite probability vector. However, we must also keep track of where we put the probability mass (the atoms). So we can write a Dirichlet process as

$$G(\cdot) = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}(\cdot) \tag{1}$$ {eq:dirich

where π_k is the probability mass assigned to the k'th atom, θ_k is the location of the k'th atom, and δ is a Dirac delta function. Note that $\sum_{k=1}^{\infty} \pi_k = 1$.

We will revisit the representation (1) constantly.

3 Chinese Restaurant Process

[Why Chinese restaurant? This term was coined by Jim Pitman and Lester Dubins, probabilistic by Berkeley, who were inspired by the “seemingly endless” seating capacity of local restaurants.]

Consider patrons walking into a restaurant with infinitely many tables; at each table can sit infinitely many patrons. Table k has n_k patrons. When a new customer i walks in (the first customer is 1, other customer numbers increase monotonically), she chooses to sit at table with probability

$$P(\text{customer } i \text{ sits at occupied table } k) = \frac{n_k}{i - 1 + \alpha} \tag{2}$$ {eq:crp-ex

$$P(\text{customer } i \text{ sits at unoccupied table}) = \frac{\alpha}{i - 1 + \alpha} \tag{3}$$ {eq:crp-ne

The parameter α is the *concentration.* It determines the likelihood for customers to sit at new tables.

This process induces a *partition* of the customers—a clustering of the customers according to which table they sit at. For example, suppose the following sequence of events occurs:

1. Customer 1 sits at table 1 (she must).
2. Customer 2 sits at table 1.
3. Customer 3 sits at table 2.
4. Customer 4 sits at table 1.
5. Customer 5 sits at table 3.

Then the partition is $\{\{1, 2, 4\}, \{1\}, \{1\}\}$. Note that we disregard the order of the tables, as well as the order in which patrons sit down at each table.

Some important properties of the partition induced by the CRP:

1. The process is *exchangeable* in the sense that if customers arrive in a different order, the probability of a given *partition* is the same. [However, the identities of the tables certainly are not.]

2. A new cluster can expand the partition.

Exchangeability implies easy *Gibbs sampling.* In Gibbs sampling, we need the conditional distribution for where customer i sits, conditioned on where all the other customers sits. Exchangeability means we can pretend that customer i is now the last customer to come in. Then the conditional distributions are just given by (2)—(3), where the n_k are obtained by counting how many other customers sit at each table.

3.1 Mixture Models

Suppose each table has its own dish. That is, a finite-dimensional parameter θ_k^* at table k. Now, in addition to sitting down at the table according to (2)—(3), each customer draws a random variable from the distribution $x_i | z_i = k \sim F(\theta_k^*)$, where F is a parametric distribution.

[Somebody asked about concentration parameter. Dave then discussed the example of Gaussian mixture models, where it is apparent that the *variances* of components has influences the number of clusters in the posterior. Small variances will induce many clusters, as each cluster can capture few points, while large variances will induce fewer clusters. Dave says that this has much greater impact on the number of posteriors components than α.

However, for small data, small α yields what is called "size bias": the posterior will concentrate on just one or two clusters.

Finally, a single mixture models seems not to be very sensitive to α (so long as we have lots of data). But concentration matters much more for hierarchical models, because each component model in the hierarchy is conditioned on less data.]

[Q: If E[number of clusters] $\propto \log \alpha$, then do we scale α according to the number of data points?
A: Not really—the likelihood term dominates.]

4 Dirichlet Process

A Dirichlet process over Ω has two parameters—a base measure G_0 on Ω and a concentration parameter α. (In practice, we will only use probability measures as our base measure G_0.) We write it as

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

Two facts:

1. $E[G] = G_0$. Note that G is a stochastic process indexed by partitions of Ω while G_0 is a probability measure on Ω. We can prove this using the stochastic process representation (below).

2. Draws from the Dirichlet process are discrete (even if G_0 is continuous). [This is not obvious, and requires proof!] Moreover, we can represent the measure G as

$$G(\cdot) = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_i}(\cdot)$$

where δ is the Dirac delta and $\sum_{k=1}^{\infty} \pi_k = 1$.

Let us now proceed more formally.

Consider a probability space Ω and a finite partition $\mathcal{A}$ of Ω. [That is, $\mathcal{A}$ is a collection of subsets $A_i \subset \Omega$ such that $\cup_{i=1}^{k} A_i = \Omega$ and the A_i are disjoint.] Now consider the vector $v(A) := [G(A_1), \ldots, G(A_k)]$, where each component is just the measure G applied to each subset A_i. Since G is a random probability measure, $v(A)$ is a random vector on the $k-1$ simplex (it is a random vector, and it sums to 1).

Now G is a Dirichlet process if and only if for any finite partition $\mathcal{A}$, we have

$$v(A) \sim \mathrm{Dir}\left(\alpha G_0(A_1), \ldots, \alpha G_0(A_k)\right)$$

i.e. the random vector $v(\mathcal{A})$ induced by the partition $\mathcal{A}$ has a Dirichlet distribution.

In fact, this definition satisfies the hypotheses for Kolmogorov's extension theorem. [We merely asserted this and did not go into details during class.] Therefore, we have shown that these exists a stochastic process G, defining a collection of random variables $v(A)$ indexed by A.

These properties actually suffice to derive Gibbs samplers for Dirichlet processes, even though they contain no hints on any construction or algorithm. This is because we can derive the CRP from these properties as we show below, and the CRP naturally lends itself to a Gibbs sampler.

4.1 The Dirichlet Process is Self-Conjugate

Consider

$$\begin{aligned} G &\sim DP(\alpha G_0) \\ \theta_1 &\sim G \end{aligned}$$

What is the distribution of $G|\theta_1$? Let's consider the induced Dirichlet distribution on the partition $\mathcal{A} := A_{1:k}$. Then we look at $G(A_{1:k})|\theta_1$, which is just a just a Dirichlet distribution

$$G(A_{1:k})|\theta_1 \sim \text{Dir}\left(\alpha G_0(A_1) + \delta_{\theta_1}(A_1), \ldots, \alpha G_0(A_k) + \delta_{\theta_1}(A_k)\right) \tag{4}$$

{eq:partit

Why is this the case? Recall that the Dirichlet distribution is conjugate to the categorical distribution. If we observe a point belonging to class k, then we increment the Dirichlet parameter in its k'th dimension. In a Dirichlet process, we do not observe categorical vectors, but instead $\theta_1 \in \Omega$. However, θ_1 belongs to exactly one subset A_k. Observing θ_1 thus means incrementing the parameter for the k'th part, hence (4).

Since this is true for any partition $\mathcal{A}$, we have that $G|\theta_1$ is also a Dirichlet process with parameter $DP(\alpha G_0 + \delta_{\theta_1})$. [Note that the posterior cannot really be decomposed into α multiplied by another measure, which is why Dave does not like to write $DP(a, G_0)$ like others do.]

Essentially, our posterior base measure is equal to our old base measure, plus a spike at θ_1.

4.2 The Dirichlet and Chinese Restaurant Processes are the Same

Now let us consider the posterior predictive distribution

$$\begin{aligned} p(\theta|\theta_{1:n}) &= \frac{\int G(\theta) p(G|\theta_{1:n}) dG}{\int \left(\int G(\theta) p(G|\theta_{1:n}) dG\right) d\theta} \\ &= \frac{E[G|\theta_{1:n}](\theta)}{\int E[G|\theta_{1:n}](\theta) d\theta} \\ &= \frac{\alpha G_0 + \sum_i \delta_{\theta_i}}{\int \left(\alpha G_0 + \sum_i \delta_{\theta_i}\right) d\theta}(\theta) \\ &= \frac{\alpha G_0 + \sum_i \delta_{\theta_i}}{\alpha + n}(\theta) \\ &= \frac{\alpha}{\alpha + n} G_0(\theta) + \sum_i \frac{1}{\alpha + n} \delta_{\theta_i}(\theta) \end{aligned}$$

Note that the integral $\int G(\theta) p(G|\theta_{1:n}) dG$ marginalizes the stochastic process and yields a measure over Ω. We desire a probability measure, so we must explicitly normalize this quantity. Of course we must assume the necessary conditions for these integrals to exist.

The third line applies the property $E[G] = G_0$ to the posterior Dirichlet process.

To simplify the denominator, we note that $\int G_0 d\theta = 1$ because G_0 is a probability measure, and $\int \left(\sum_i \delta_{\theta_i}\right) d\theta = n$ because each existing value of θ_i is hit exactly once by the integral.

The last line is exactly the CRP conditional probabilities (2)—(3)!

The Dirichlet process mixture is really a corollary. Instead of sampling just θ_i, we treat θ_i as a finite-dimensional parameter sample sample data $x_i \sim \theta_i$. You can think of this as placing a little bump around θ_i. [But not really, because θ_i and x_i live in different spaces. But it is a good visualization.]

5 The Stick Breaking Construction

We've discussed the Chinese restaurant process, a concrete construction of the DP. As a reminder, we can write the random measure G as

$$G = \sum_k \pi_k \delta_{\theta_i} \tag{5}$$

{eq:G-rand

$$\theta_i \sim G_0 \qquad \forall i \tag{6}$$

{eq:theta-

The stick breaking construction is a constructive method to arrive at (5)—(6). Consider an infinite collection of Beta random variables v_k with

$$\begin{aligned} v_k &\sim \text{Beta}(1, \alpha) \\ \pi_k &:= v_k \prod_i (1 - v_i) \end{aligned}$$

Consider this process iteratively. For $k = 1$, we simply get $\pi_1 = v_1$, which is a proportion of a stick of length 1. Now we are left with a stick of length $(1 - v_1)$. For $k = 2$, break off v_k fraction of this stick. Then we get a stick of length $v_2(1 - v_1)$, and are left with a stick of length $(1 - v_2)(1 - v_1)$, which we continue breaking off.