Dirichlet Process

Yee Whye Teh, University College London

Related keywords: Bayesian nonparametrics, stochastic processes, clustering, infinite mixture model, Blackwell-MacQueen urn scheme, Chinese restaurant process.

Definition

The Dirichlet process is a stochastic proces used in Bayesian nonparametric models of data, particularly in Dirichlet process mixture models (also known as infinite mixture models). It is a distribution over distributions, i.e. each draw from a Dirichlet process is itself a distribution. It is called a Dirichlet process because it has Dirichlet distributed finite dimensional marginal distributions, just as the Gaussian process, another popular stochastic process used for Bayesian nonparametric regression, has Gaussian distributed finite dimensional marginal distributions. Distributions drawn from a Dirichlet process are discrete, but cannot be described using a finite number of parameters, thus the classification as a nonparametric model.

Motivation and Background

Probabilistic models are used throughout machine learning to model distributions over observed data. Traditional parametric models using a fixed and finite number of parameters can suffer from over- or under-fitting of data when there is a misfit between the complexity of the model (often expressed in terms of the number of parameters) and the amount of data available. As a result, model selection, or the choice of a model with the right complexity, is often an important issue in parametric modeling. Unfortunately, model selection is an operation that is fraught with difficulties, whether we use cross validation or marginal probabilities as the basis for selection. The Bayesian nonparametric approach is an alternative to parametric modeling and selection. By using a model with an unbounded complexity, underfitting is mitigated, while the Bayesian approach of computing or approximating the full posterior over parameters mitigates overfitting. A general overview of Bayesian nonparametric modeling can be found under its entry in the encyclopedia [**?**].

Nonparametric models are also motivated philosophically by Bayesian modeling. Typically we assume that we have an underlying and unknown distribu-

tion which we wish to infer given some observed data. Say we observe $x_1, \ldots, x_n$, with $x_i \sim F$ independent and identical draws from the unknown distribution F. A Bayesian would approach this problem by placing a prior over F then computing the posterior over F given data. Traditionally, this prior over distributions is given by a parametric family. But constraining distributions to lie within parametric families limits the scope and type of inferences that can be made. The nonparametric approach instead uses a prior over distributions with wide support, typically the support being the space of all distributions. Given such a large space over which we make our inferences, it is important that posterior computations are tractable.

The Dirichlet process is currently one of the most popular Bayesian nonparametric models. It was first formalized in [1][1] for general Bayesian statistical modeling, as a prior over distributions with wide support yet tractable posteriors. Unfortunately the Dirichlet process is limited by the fact that draws from it are discrete distributions, and generalizations to more general priors did not have tractable posterior inference until the development of MCMC techniques [3, 4]. Since then there has been significant developments in terms of inference algorithms, extensions, theory and applications. In the machine learning community work on Dirichlet processes date back to [5, 6].

Theory

The Dirichlet process (DP) is a stochastic process whose sample paths are probability measures with probability one. Stochastic processes are distributions over function spaces, with sample paths being random functions drawn from the distribution. In the case of the DP, it is a distribution over probability measures, which are functions with certain special properties which allow them to be interpreted as distributions over some probability space Θ. Thus draws from a DP can be interpreted as random distributions. For a distribution over probability measures to be a DP, its marginal distributions have to take on a specific form which we shall give below. We assume that the user is familiar with a modicum of measure theory and Dirichlet distributions.

Before we proceed to the formal definition, we will first give an intuitive explanation of the DP as an infinite dimensional generalization of Dirichlet distributions. Consider a Bayesian mixture model consisting of K components:

$$\begin{aligned} \pi|\alpha &\sim \mathrm{Dir}(\tfrac{\alpha}{K}, \ldots, \tfrac{\alpha}{K}) & \theta_k^*|H &\sim H \\ z_i|\pi &\sim \mathrm{Mult}(\pi) & x_i|z_i, \{\theta_k^*\} &\sim F(\theta_{z_i}^*) \end{aligned} \tag{1}$$

where π is the mixing proportion, α is the pseudocount hyperparameter of the Dirichlet prior, H is the prior distribution over component parameters θ_k^*, and $F(\theta)$ is the component distribution parametrized by θ. It can be shown that for large K, because of the particular way we parametrized the Dirichlet prior over π, the number of components typically used to model n data items

[1] Note however that related models in population genetics date back to [2].

becomes independent of K and is approximately $O(\alpha \log n)$. This implies that the mixture model stays well-defined as $K \to \infty$, leading to what is known as an infinite mixture model [5, 6]. This model was first proposed as a way to sidestep the difficult problem of determining the number of components in a mixture, and as a nonparametric alternative to finite mixtures whose size can grow naturally with the number of data items. The more modern definition of this model uses a DP and with the resulting model called a DP mixture model. The DP itself appears as the $K \to \infty$ limit of the random discrete probability measure $\sum_{k=1}^{K} \pi_k \delta_{\theta_k^*}$, where δ_θ is a point mass centred at θ. We will return to the DP mixture towards the end of this entry.

Dirichlet Process

For a random distribution G to be distributed according to a DP, its marginal distributions have to be Dirichlet distributed [1]. Specifically, let H be a distribution over Θ and α be a positive real number. Then for any finite measurable partition $A_1, \ldots, A_r$ of Θ the vector $(G(A_1), \ldots, G(A_r))$ is random since G is random. We say G is Dirichlet process distributed with base distribution H and concentration parameter α, written $G \sim \mathrm{DP}(\alpha, H)$, if

$$(G(A_1), \ldots, G(A_r)) \sim \mathrm{Dir}(\alpha H(A_1), \ldots, \alpha H(A_r)) \tag{2}$$

for every finite measurable partition $A_1, \ldots, A_r$ of Θ.

The parameters H and α play intuitive roles in the definition of the DP. The base distribution is basically the mean of the DP: for any measurable set $A \subset \Theta$, we have $E[G(A)] = H(A)$. On the other hand, the concentration parameter can be understood as an inverse variance: $V[G(A)] = H(A)(1 - H(A))/(\alpha + 1)$. The larger α is, the smaller the variance, and the DP will concentrate more of its mass around the mean. The concentration parameter is also called the strength parameter, referring to the strength of the prior when using the DP as a nonparametric prior over distributions in a Bayesian nonparametric model, and the mass parameter, as this prior strength can be measured in units of sample size (or mass) of observations. Also, notice that α and H only appear as their product in the definition (2) of the DP. Some authors thus treat $\widetilde{H} = \alpha H$, as the single (positive measure) parameter of the DP, writing $\mathrm{DP}(\widetilde{H})$ instead of $\mathrm{DP}(\alpha, H)$. This parametrization can be notationally convenient, but loses the distinct roles α and H play in describing the DP.

Since α describes the concentration of mass around the mean of the DP, as $\alpha \to \infty$, we will have $G(A) \to H(A)$ for any measurable A, that is $G \to H$ weakly or pointwise. However this not equivalent to saying that $G \to H$. As we shall see later, draws from a DP will be discrete distributions with probability one, even if H is smooth. Thus G and H need not even be absolutely continuous with respect to each other. This has not stopped some authors from using the DP as a nonparametric relaxation of a parametric model given by H. However, if smoothness is a concern, it is possible to extend the DP by convolving G with kernels so that the resulting random distribution has a density.

A related issue to the above is the coverage of the DP within the class of all distributions over Θ. We already noted that samples from the DP are discrete, thus the set of distributions with positive probability under the DP is small. However it turns out that this set is also large in a different sense: if the topological support of H (the smallest closed set S in Θ with $H(S) = 1$) is all of Θ, then any distribution over Θ can be approximated arbitrarily accurately in the weak or pointwise sense by a sequence of draws from $\mathrm{DP}(\alpha, H)$. This property has consequence in the consistency of DPs discussed later.

For all but the simplest probability spaces, the number of measurable partitions in the definition (2) of the DP can be uncountably large. The natural question to ask here is whether objects satisfying such a large number of conditions as (2) can exist. There are a number of approaches to establish existence. [1] noted that the conditions (2) are consistent with each other, and made use of Kolmogorov's consistency theorem to show that a distribution over functions from the measurable subsets of Θ to $[0, 1]$ exists satisfying (2) for all finite measurable partitions of Θ. However it turns out that this construction does not necessarily guarantee a distribution over probability measures. [1] also provided a construction of the DP by normalizing a gamma process. In a later section we will see that the predictive distributions of the DP are related to the Blackwell-MacQueen urn scheme. [7] made use of this, along with de Finetti's theorem on exchangeable sequences, to prove existence of the DP. All the above methods made use of powerful and general mathematical machinery to establish existence, and often require regularity assumptions on H and Θ to apply these machinery. In a later section, we describe a stick-breaking construction of the DP due to [8], which is a direct and elegant construction of the DP which need not impose such regularity assumptions.

Posterior Distribution

Let $G \sim \mathrm{DP}(\alpha, H)$. Since G is a (random) distribution, we can in turn draw samples from G itself. Let $\theta_1, \ldots, \theta_n$ be a sequence of independent draws from G. Note that the θ_i's take values in Θ since G is a distribution over Θ. We are interested in the posterior distribution of G given observed values of $\theta_1, \ldots, \theta_n$. Let $A_1, \ldots, A_r$ be a finite measurable partition of Θ, and let $n_k = \#\{i : \theta_i \in A_k\}$ be the number of observed values in A_k. By (2) and the conjugacy between the Dirichlet and the multinomial distributions, we have:

$$(G(A_1), \ldots, G(A_r))|\theta_1, \ldots, \theta_n \sim \mathrm{Dir}(\alpha H(A_1) + n_1, \ldots, \alpha H(A_r) + n_r) \quad (3)$$

Since the above is true for all finite measurable partitions, the posterior distribution over G must be a DP as well. A little algebra shows that the posterior DP has updated concentration parameter $\alpha + n$ and base distribution $\frac{\alpha H + \sum_{i=1}^{n} \delta_{\theta_i}}{\alpha + n}$, where δ_i is a point mass located at θ_i and $n_k = \sum_{i=1}^{n} \delta_i(A_k)$. In other words, the DP provides a conjugate family of priors over distributions that is closed under posterior updates given observations. Rewriting the posterior DP, we

have:

$$G|\theta_1, \ldots, \theta_n \sim \text{DP}\left(\alpha + n, \frac{\alpha}{\alpha+n} H + \frac{n}{\alpha+n} \frac{\sum_{i=1}^n \delta_{\theta_i}}{n}\right) \tag{4}$$

Notice that the posterior base distribution is a weighted average between the prior base distribution H and the empirical distribution $\frac{\sum_{i=1}^n \delta_{\theta_i}}{n}$. The weight associated with the prior base distribution is proportional to α, while the empirical distribution has weight proportional to the number of observations n. Thus we can interpret α as the strength or mass associated with the prior. In the next section we will see that the posterior base distribution is also the predictive distribution of θ_{n+1} given $\theta_1, \ldots, \theta_n$. Taking $\alpha \to 0$, the prior becomes non-informative in the sense that the predictive distribution is just given by the empirical distribution. On the other hand, as the amount of observations grows large, $n \gg \alpha$, the posterior is simply dominated by the empirical distribution which is in turn a close approximation of the true underlying distribution. This gives a consistency property of the DP: the posterior DP approaches the true underlying distribution.

Predictive Distribution and the Blackwell-MacQueen Urn Scheme

Consider again drawing $G \sim \text{DP}(\alpha, H)$, and drawing an i.i.d. sequence $\theta_1, \theta_2, \ldots \sim G$. Consider the predictive distribution for θ_{n+1}, conditioned on $\theta_1, \ldots, \theta_n$ and with G marginalized out. Since $\theta_{n+1}|G, \theta_1, \ldots, \theta_n \sim G$, for a measurable $A \subset \Theta$, we have

$$\begin{aligned} P(\theta_{n+1} \in A|\theta_1, \ldots, \theta_n) &= E[G(A)|\theta_1, \ldots, \theta_n] \\ &= \frac{1}{\alpha + n}\left(\alpha H(A) + \sum_{i=1}^n \delta_{\theta_i}(A)\right) \end{aligned} \tag{5}$$

where the last step follows from the posterior base distribution of G given the first n observations. Thus with G marginalized out:

$$\theta_{n+1}|\theta_1, \ldots, \theta_n \sim \frac{1}{\alpha + n}\left(\alpha H + \sum_{i=1}^n \delta_{\theta_i}\right) \tag{6}$$

Therefore the posterior base distribution given $\theta_1, \ldots, \theta_n$ is also the predictive distribution of θ_{n+1}.

The sequence of predictive distributions (6) for $\theta_1, \theta_2, \ldots$ is called the Blackwell-MacQueen urn scheme [7]. The name stems from a metaphor useful in interpreting (6). Specifically, each value in Θ is a unique color, and draws $\theta \sim G$ are balls with the drawn value being the color of the ball. In addition we have an urn containing previously seen balls. In the beginning there are no balls in the urn, and we pick a color drawn from H, i.e. draw $\theta_1 \sim H$, paint a ball with that color, and drop it into the urn. In subsequent steps, say the $n+1$st, we will either, with probability $\frac{\alpha}{\alpha+n}$, pick a new color (draw $\theta_{n+1} \sim H$), paint a ball with that color and drop the ball into the urn, or, with probability $\frac{n}{\alpha+n}$,

reach into the urn to pick a random ball out (draw θ_{n+1} from the empirical distribution), paint a new ball with the same color and drop both balls back into the urn.

The Blackwell-MacQueen urn scheme has been used to show the existence of the DP [7]. Starting from (6), which are perfectly well-defined conditional distributions regardless of the question of the existence of DPs, we can construct a distribution over sequences $\theta_1, \theta_2, \ldots$ by iteratively drawing each θ_i given $\theta_1, \ldots, \theta_{i-1}$. For $n \geq 1$ let

$$P(\theta_1, \ldots, \theta_n) = \prod_{i=1}^{n} P(\theta_i | \theta_1, \ldots, \theta_{i-1}) \tag{7}$$

be the joint distribution over the first n observations, where the conditional distributions are given by (6). It is straightforward to verify that this random sequence is infinitely exchangeable. That is, for every n, the probability of generating $\theta_1, \ldots, \theta_n$ using (6), in that order, is equal to the probability of drawing them in any alternative order. More precisely, given any permutation σ on $1, \ldots, n$, we have

$$P(\theta_1, \ldots, \theta_n) = P(\theta_{\sigma(1)}, \ldots, \theta_{\sigma(n)}) \tag{8}$$

Now de Finetti's theorem states that for any infinitely exchangeable sequence $\theta_1, \theta_2, \ldots$ there is a random distribution G such that the sequence is composed of i.i.d. draws from it:

$$P(\theta_1, \ldots, \theta_n) = \int \prod_{i=1}^{n} G(\theta_i)\, dP(G) \tag{9}$$

In our setting, the prior over the random distribution $P(G)$ is precisely the Dirichlet process $\mathrm{DP}(\alpha, H)$, thus establishing existence.

A salient property of the predictive distribution (6) is that it has point masses located at the previous draws $\theta_1, \ldots, \theta_n$. A first observation is that with positive probability draws from G will take on the same value, regardless of smoothness of H. This implies that the distribution G itself has point masses. A further observation is that for a long enough sequence of draws from G, the value of any draw will be repeated by another draw, implying that G is composed only of a weighted sum of point masses, i.e. it is a discrete distribution. We will see two sections below that this is indeed the case, and give a simple construction for G called the stick-breaking construction. Before that, we shall investigate the clustering property of the DP.

Clustering, Partitions and the Chinese Restaurant Process

In addition to the discreteness property of draws from a DP, (6) also implies a clustering property. The discreteness and clustering properties of the DP play crucial roles in the use of DPs for clustering via DP mixture models, described in the application section. For now we assume that H is smooth, so that all

repeated values are due to the discreteness property of the DP and not due to H itself[2]. Since the values of draws are repeated, let $\theta_1^*, \ldots, \theta_m^*$ be the unique values among $\theta_1, \ldots, \theta_n$, and n_k be the number of repeats of θ_k^*. The predictive distribution can be equivalently written as:

$$\theta_{n+1}|\theta_1, \ldots, \theta_n \sim \frac{1}{\alpha + n}\left(\alpha H + \sum_{k=1}^{m} n_k \delta_{\theta_k^*}\right) \tag{10}$$

Notice that value θ_k^* will be repeated by θ_{n+1} with probability proportional to n_k, the number of times it has already been observed. The larger n_k is, the higher the probability that it will grow. This is a rich-gets-richer phenomenon, where large clusters (a set of θ_i's with identical values θ_k^* being considered a cluster) grow larger faster.

We can delve further into the clustering property of the DP by looking at partitions induced by the clustering. The unique values of $\theta_1, \ldots, \theta_n$ induce a partitioning of the set $[n] = \{1, \ldots, n\}$ into clusters such that within each cluster, say cluster k, the θ_i's take on the same value θ_k^*. Given that $\theta_1, \ldots, \theta_n$ are random, this induces a random partition of $[n]$. This random partition in fact encapsulates all the properties of the DP, and is a very well studied mathematical object in its own right, predating even the DP itself [2, 9, 10]. To see how it encapsulates the DP, we simply invert the generative process. Starting from the distribution over random partitions, we can reconstruct the joint distribution (7) over $\theta_1, \ldots, \theta_n$, by first drawing a random partition on $[n]$, then for each cluster k in the partition draw a $\theta_k^* \sim H$, and finally assign $\theta_i = \theta_k^*$ for each i in cluster k. From the joint distribution (7) we can obtain the DP by appealing to de Finetti's theorem.

The distribution over partitions is called the Chinese restaurant process (CRP) due to a different metaphor[3]. In this metaphor we have a Chinese restaurant with an infinite number of tables, each of which can seat an infinite number of customers. The first customer enters the restaurant and sits at the first table. The second customer enters and decides either to sit with the first customer, or by herself at a new table. In general, the $n+1$st customer either joins an already occupied table k with probability proportional to the number n_k of customers already sitting there, or sits at a new table with probability proportional to α. Identifying customers with integers $1, 2, \ldots$ and tables as clusters, after n customers have sat down the tables define a partition of $[n]$ with the distribution over partitions being the same as the one above. The fact that most Chinese restaurants have round tables is an important aspect of the CRP. This is because it does not just define a distribution over partitions of $[n]$, it also defines a distribution over permutations of $[n]$, with each table corresponding to a cycle of the permutation. We do not need to explore this aspect further and refer the interested reader to [9, 10].

This distribution over partitions first appeared in population genetics, where it was found to be a robust distribution over alleles (clusters) among gametes

[2]Similar conclusions can be drawn when H has atoms, there is just more bookkeeping.

[3]The name was coined by Lester Dubins and Jim Pitman in the early 1980's [9].

(observations) under simplifying assumptions on the population, and is known under the name of Ewens sampling formula [2]. Before moving on we shall consider just one illuminating aspect, specifically the distribution of the number of clusters among n observations. Notice that for $i \geq 1$, the observation θ_i takes on a new value (thus incrementing m by one) with probability $\frac{\alpha}{\alpha+i-1}$ independently of the number of clusters among previous θ's. Thus the number of cluster m has mean and variance:

$$\begin{aligned} E[m|n] &= \sum_{i=1}^{n} \frac{\alpha}{\alpha+i-1} = \alpha(\psi(\alpha+n) - \psi(\alpha)) \\ &\simeq \alpha \log\left(1+\frac{n}{\alpha}\right) \qquad \text{for } N, \alpha \gg 0, \qquad (11) \\ V[m|n] &= \alpha(\psi(\alpha+n) - \psi(\alpha)) + \alpha^2(\psi'(\alpha+n) - \psi'(\alpha)) \\ &\simeq \alpha \log\left(1+\frac{n}{\alpha}\right) \qquad \text{for } n > \alpha \gg 0, \qquad (12) \end{aligned}$$

where $\psi(\cdot)$ is the digamma function. Note that the number of clusters grows only logarithmically in the number of observations. This slow growth of the number of clusters makes sense because of the rich-gets-richer phenomenon: we expect there to be large clusters thus the number of clusters m has to be smaller than the number of observations n. Notice that α controls the number of clusters in a direct manner, with larger α implying a larger number of clusters a priori. This intuition will help in the application of DPs to mixture models.

Stick-breaking Construction

We have already intuited that draws from a DP are composed of a weighted sum of point masses. [8] made this precise by providing a constructive definition of the DP as such, called the stick-breaking construction. This construction is also significantly more straightforward and general than previous proofs of the existence of DPs. It is simply given as follows:

$$\begin{aligned} \beta_k &\sim \text{Beta}(1, \alpha) & \theta_k^* &\sim H \\ \pi_k &= \beta_k \prod_{l=1}^{k-1}(1-\beta_k) & G &= \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*} \end{aligned} \qquad (13)$$

Then $G \sim \text{DP}(\alpha, H)$. The construction of π can be understood metaphorically as follows. Starting with a stick of length 1, we break it at β_1, assigning π_1 to be the length of stick we just broke off. Now recursively break the other portion to obtain π_2, π_3 and so forth. The stick-breaking distribution over π is sometimes written $\pi \sim \text{GEM}(\alpha)$, where the letters stand for Griffiths, Engen and McCloskey [10]. Because of its simplicity, the stick-breaking construction has lead to a variety of extensions as well as novel inference techniques for the Dirichlet process [11].

Applications

Because of its simplicity, DPs are used across a wide variety of applications of Bayesian analysis in both statistics and machine learning. The simplest and most prevalent applications include: Bayesian model validation, density estimation and clustering via mixture models. We shall briefly describe the first two classes before detailing DP mixture models.

How does one validate that a model gives a good fit to some observed data? The Bayesian approach would usually involve computing the marginal probability of the observed data under the model, and comparing this marginal probability to that for other models. If the marginal probability of the model of interest is highest we may conclude that we have a good fit. The choice of models to compare against is an issue in this approach, since it is desirable to compare against as large a class of models as possible. The Bayesian nonparametric approach gives an answer to this question: use the space of all possible distributions as our comparison class, with a prior over distributions. The DP is a popular choice for this prior, due to its simplicity, wide coverage of the class of all distributions, and recent advances in computationally efficient inference in DP models. The approach is usually to use the given parametric model as the base distribution of the DP, with the DP serving as a nonparametric relaxation around this parametric model. If the parametric model performs as well or better than the DP relaxed model, we have convincing evidence of the validity of the model.

Another application of DPs is in density estimation [12, 5, 3, 6]. Here we are interested in modeling the density from which a given set of observations is drawn. To avoid limiting ourselves to any parametric class, we may again use a nonparametric prior over all densities. Here again DPs are a popular. However note that distributions drawn from a DP are discrete, thus do not have densities. The solution is to smooth out draws from the DP with a kernel. Let $G \sim \mathrm{DP}(\alpha, H)$ and let $f(x|\theta)$ be a family of densities (kernels) indexed by θ. We use the following as our nonparametric density of x:

$$p(x) = \int f(x|\theta)G(\theta)\,d\theta \tag{14}$$

Similarly, smoothing out DPs in this way is also useful in the nonparametric relaxation setting above. As we see below, this way of smoothing out DPs is equivalent to DP mixture models, if the data distributions $F(\theta)$ below are smooth with densities given by $f(x|\theta)$.

Dirichlet Process Mixture Models

The most common application of the Dirichlet process is in clustering data using mixture models [12, 5, 3, 6]. Here the nonparametric nature of the Dirichlet process translates to mixture models with a countably infinite number of components. We model a set of observations $\{x_1, \ldots, x_n\}$ using a set of latent parameters $\{\theta_1, \ldots, \theta_n\}$. Each θ_i is drawn independently and identically from

G, while each x_i has distribution $F(\theta_i)$ parametrized by θ_i:

$$\begin{aligned} x_i|\theta_i &\sim F(\theta_i) \\ \theta_i|G &\sim G \\ G|\alpha, H &\sim \mathrm{DP}(\alpha, H) \end{aligned} \tag{15}$$

Because G is discrete, multiple θ_i's can take on the same value simultaneously, and the above model can be seen as a mixture model, where x_i's with the same value of θ_i belong to the same cluster. The mixture perspective can be made more in agreement with the usual representation of mixture models using the stick-breaking construction (13). Let z_i be a cluster assignment variable, which takes on value k with probability π_k. Then (15) can be equivalently expressed as

$$\begin{aligned} \pi|\alpha &\sim \mathrm{GEM}(\alpha) & \theta_k^*|H &\sim H \\ z_i|\pi &\sim \mathrm{Mult}(\pi) & x_i|z_i, \{\theta_k^*\} &\sim F(\theta_{z_i}^*) \end{aligned} \tag{16}$$

with $G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*}$ and $\theta_i = \theta_{z_i}^*$. In mixture modeling terminology, π is the mixing proportion, θ_k^* are the cluster parameters, $F(\theta_k^*)$ is the distribution over data in cluster k, and H the prior over cluster parameters.

The DP mixture model is an *infinite* mixture model—a mixture model with a countably infinite number of clusters. However, because the π_k's decrease exponentially quickly, only a small number of clusters will be used to model the data a priori (in fact, as we saw previously, the expected number of components used a priori is logarithmic in the number of observations). This is different than a finite mixture model, which uses a fixed number of clusters to model the data. In the DP mixture model, the actual number of clusters used to model data is not fixed, and can be automatically inferred from data using the usual Bayesian posterior inference framework (see [4] for a survey of MCMC inference procedures for DP mixture models). The equivalent operation for finite mixture models would be model averaging or model selection for the appropriate number of components, an approach which is fraught with difficulties. Thus infinite mixture models as exemplified by DP mixture models provide a compelling alternative to the traditional finite mixture model paradigm.

Generalizations and Extensions

The DP is the canonical distribution over probability measures and a wide range of generalizations have been proposed in the literature. First and foremost is the *Pitman-Yor process* [13, 11], which has recently seen successful applications modeling data exhibiting power-law properties [14, 15]. The Pitman-Yor process includes a third parameter $d \in [0, 1)$, with $d = 0$ reducing to the DP. The various representations of the DP, including the Chinese restaurant process and the stick-breaking construction, have analogues for the Pitman-Yor process. Other generalizations of the DP are obtained by generalizing one of

its representations. These include Pólya trees, normalized random measure, Poisson-Kingman models, species sampling models and stick-breaking priors.

The DP has also been used in more complex models involving more than one random probability measure. For example, in nonparametric regression we might have one probability measure for each value of a covariate, and in multi-task settings each task might be associated with a probability measure with dependence across tasks implemented using a hierarchical Bayesian model. In the first situation the class of models is typically called dependent Dirichlet processes [16], while in the second the appropriate model is a hierarchical Dirichlet process [17].

Future Directions

The Dirichlet process, and Bayesian nonparametrics in general, is an active area of research within both machine learning and statistics. Current research trends span a number of directions. Firstly there is the issue of efficient inference in DP models. [4] is an excellent survey of the state-of-the-art in 2000, with all algorithms based on Gibbs sampling or small-step Metropolis-Hastings MCMC sampling. Since then there has been much work, including split-and-merge and large-step auxiliary variable MCMC sampling, sequential Monte Carlo, expectation propagation, and variational methods. Secondly there has been interest in extending the DP, both in terms of new random distributions, as well as novel classes of nonparametric objects inspired by the DP. Thirdly, theoretical issues of convergence and consistency are being explored to provide frequentist guarantees for Bayesian nonparametric models. Finally there are applications of such models, to clustering, transfer learning, relational learning, models of cognition, sequence learning, and regression and classification among others. We believe DPs and Bayesian nonparametrics will prove to be rich and fertile grounds for research for years to come.

Cross References

Bayesian Methods, Prior Probabilities, Bayesian Nonparametrics.

Recommended Reading

In addition to the references embedded in the text above, we recommend the book [18] on Bayesian nonparametrics.

[1] T. S. Ferguson. A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1(2):209–230, 1973.

[2] W. J. Ewens. The sampling theory of selectively neutral alleles. *Theoretical Population Biology*, 3:87–112, 1972.

[3] M. D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90:577–588, 1995.

[4] R. M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9:249–265, 2000.

[5] R. M. Neal. Bayesian mixture modeling. In *Proceedings of the Workshop on Maximum Entropy and Bayesian Methods of Statistical Analysis*, volume 11, pages 197–211, 1992.

[6] C. E. Rasmussen. The infinite Gaussian mixture model. In *Advances in Neural Information Processing Systems*, volume 12, 2000.

[7] D. Blackwell and J. B. MacQueen. Ferguson distributions via Pólya urn schemes. *Annals of Statistics*, 1:353–355, 1973.

[8] J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.

[9] D. Aldous. Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII–1983*, pages 1–198. Springer, Berlin, 1985.

[10] J. Pitman. Combinatorial stochastic processes. Technical Report 621, Department of Statistics, University of California at Berkeley, 2002. Lecture notes for St. Flour Summer School.

[11] H. Ishwaran and L. F. James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173, 2001.

[12] A.Y. Lo. On a class of bayesian nonparametric estimates: I. density estimates. *Annals of Statistics*, 12(1):351–357, 1984.

[13] J. Pitman and M. Yor. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Annals of Probability*, 25:855–900, 1997.

[14] S. Goldwater, T.L. Griffiths, and M. Johnson. Interpolating between types and tokens by estimating power-law generators. In *Advances in Neural Information Processing Systems*, volume 18, 2006.

[15] Y. W. Teh. A hierarchical Bayesian language model based on Pitman-Yor processes. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 985–992, 2006.

[16] S. MacEachern. Dependent nonparametric processes. In *Proceedings of the Section on Bayesian Statistical Science*. American Statistical Association, 1999.

[17] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[18] N. Hjort, C. Holmes, P. Müller, and S. Walker, editors. *Bayesian Nonparametrics*. Number 28 in Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2010.