

The Dirichlet Process Mixture (DPM) Model

Ananth Ranganathan

20th September 2004

1 The Dirichlet Distribution

The Dirichlet distribution forms our first step toward understanding the DPM model. The Dirichlet distribution is a multi-parameter generalization of the Beta distribution and defines a distribution over distributions, i.e. the result of sampling a Dirichlet is a distribution on some discrete probability space. Let $\Theta = \{\theta_1, \theta_2, \dots, \theta_n\}$ be a probability distribution on the discrete space $\mathcal{X} = \{\mathcal{X}_1, \mathcal{X}_2, \dots, \mathcal{X}_n\}$ s.t. $P(X = \mathcal{X}_i) = \theta_i$ where X is a random variable in the space \mathcal{X} . The Dirichlet distribution on Θ is given by the formula

$$P(\Theta | \alpha, M) = \frac{\Gamma(\alpha)}{\prod_{i=1}^n \Gamma(\alpha m_i)} \prod_{i=1}^n \theta_i^{\alpha m_i - 1} \quad (1)$$

where $M = \{m_1, m_2, \dots, m_n\}$ is the *base measure* defined on \mathcal{X} and is the mean value of Θ , and α is a precision parameter that says how concentrated the distribution is around M . Both Θ and M are normalized, i.e. sum to unity, since they are proper probability distributions. α can be regarded as the number of pseudo-measurements observed to obtain M , i.e. the number of events relating to the random variable X observed a priori. The greater the number of pseudo-measurements the more our confidence in M , and hence, the more the distribution is concentrated around M .

To make the above discussion concrete, consider the example of a 6-faced die. A Dirichlet distribution can be defined on the space of possible observations from the die, i.e. the space $\mathcal{X} = \{1, 2, 3, 4, 5, 6\}$. If we consider the die to be fair *a priori*, then M can be defined as $M = \{\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}\}$ and we arbitrarily set $\alpha = 6$ (which can be understood as corresponding to the case of our having observed every outcome of the die once a priori). The Dirichlet distribution defined by these values of α and M can now be sampled to yield, for example, $\Theta = \{0.113767, 0.179602, 0.273959, 0.153161, 0.169832, 0.109679\}$.

Clearly, the distribution used in the above example is not the only one possible on die observations. We could, for instance, consider a Dirichlet distribution on the space $\mathcal{X}' = \{\{1, 2\}, \{3, 4\}, \{5, 6\}\}$, so that $M = \{m_1, m_2, m_3\}$ and $\Theta = \{\theta_1, \theta_2, \theta_3\}$ are vectors of length 3. Θ is then a distribution on the random variable X taking a value from one of the sets in \mathcal{X}' , i.e. $P(X \in \{1, 2\}) = \theta_1$ and so on. More generally, for any partition of a discrete space \mathcal{X} into n sets $\mathcal{X}' = \{\mathcal{X}_1, \mathcal{X}_2, \dots, \mathcal{X}_n\}$ s.t. $\mathcal{X}_i \cap \mathcal{X}_j = \Phi \quad \forall \mathcal{X}_1, \mathcal{X}_2 \in \mathcal{X}'$ and $\bigcup_{i=1}^n \mathcal{X}_i = \mathcal{X}$, we can define a Dirichlet distribution $Dir(\Theta; \alpha, M)$ on \mathcal{X}' , where $P(X \in \mathcal{X}_i) = \theta_i$ for $1 \leq i \leq n$. We now introduce new notation replacing θ_i by $\Theta(\mathcal{X}_i)$ (and, correspondingly, m_i by $M(\mathcal{X}_i)$, so that the Dirichlet distribution on \mathcal{X} can be written as

$$\Theta(\mathcal{X}_1), \Theta(\mathcal{X}_2), \dots, \Theta(\mathcal{X}_n) \sim Dir(\Theta; \alpha, M) \quad (2)$$

where $Dir(\cdot)$ is the Dirichlet density function. The intuition behind (2) is important as it forms the definition of the Dirichlet process in continuous spaces.

1.1 Posterior update using the Multinomial distribution

Consider N observations X_1, X_2, \dots, X_N that are multinomially distributed according to Θ . If n_i is the number of times the event \mathcal{X}_i is observed in the N observations, the posterior probability on Θ can be obtained simply using Bayes Law as follows

$$P(\Theta | \alpha, M, X_{1:N}) = kP(X_{1:N} | \alpha, M, \Theta)P(\Theta | \alpha, M)$$

$$\begin{aligned}
&= k \prod_{i=1}^n \theta_i^{n_i} \times \prod_{i=1}^n \theta_i^{\alpha m_i - 1} \\
&= k \prod_{i=1}^n \theta_i^{\alpha m_i + n_i - 1} \\
&= \text{Dir}(\Theta; \alpha^*, M^*)
\end{aligned}$$

where k is a normalization constant and

$$\begin{aligned}
\alpha^* &= \alpha + N \\
M^* &= \frac{\alpha M + N \hat{F}}{\alpha + N}
\end{aligned} \tag{3}$$

where \hat{F} is the empirical distribution (i.e, simply the proportion of occurrence) of the n events in the observations. The posterior is again a Dirichlet distribution with altered parameters and so the Dirichlet distribution is a conjugate prior to the Multinomial distribution.

Now consider the probability of the $(N + 1)$ th observation X_{N+1} , given all the previous observations and the Dirichlet distribution parameters, $P(X_{N+1} | X_{1:N}, \alpha, M)$. Specifically, we want to calculate the probability that X_{N+1} is the event \mathcal{X}_j in the space \mathcal{X} , i.e. $P(X_{N+1} \in \mathcal{X}_j | X_{1:N}, \alpha, M)$. The calculation is performed by marginalizing over Θ

$$\begin{aligned}
P(X_{N+1} \in \mathcal{X}_j | X_{1:N}, \alpha, M) &= \int_{\Theta} P(X_{N+1} \in \mathcal{X}_j | \Theta) P(\Theta | X_{1:N}, \alpha, M) \\
&= \int_{\Theta} \theta_j \text{Dir}(\Theta | \alpha^*, M^*) \\
&= E(\theta_j) \\
&= \frac{\alpha m_j^*}{\sum_{i=1}^n \alpha m_i^*} = m_j^*
\end{aligned}$$

where α^* and $M^* = \{m_1^*, m_2^*, \dots, m_n^*\}$ are as defined in (3) so that $m_j^* = \frac{\alpha m_j + \sum_{i=1}^N \delta(X_i = \mathcal{X}_j)}{\alpha + N}$. Hence, we get

$$P(X_{N+1} \in \mathcal{X}_j | X_{1:N}, \alpha, M) = \frac{\alpha m_j + \sum_{i=1}^N \delta(X_i = \mathcal{X}_j)}{\alpha + N} \tag{4}$$

Note that the derivation above uses the property of the Dirichlet distribution that $E(\theta_j) = \frac{M(\mathcal{X}_j)}{M(\mathcal{X})}$.

1.2 The Dirichlet distribution through the Polya Urn Model

Many probability distributions can be obtained using urn models [6]. The urn model that corresponds to the Dirichlet distribution is the Polya Urn model.

Consider a bag with α balls of which initially αm_j are of color j , $1 \leq j \leq n$ (assuming for now that all the αm_j s are integers). We draw balls at random from the bag and at each step, replace the ball that we drew by two balls of the same color. Then, if we denote probability of the obtaining a ball of color j at the i th step $P(X_i = j)$, it is easy to obtain

$$\begin{aligned}
P(X_1 = j) &= \frac{\alpha m_j}{\sum_{i=1}^n \alpha m_i} \\
P(X_2 = j | X_1) &= \frac{\alpha m_j + \delta(X_1 = j)}{\sum_{i=1}^n \alpha m_i}
\end{aligned}$$

and so on, till we get

$$P(X_{N+1} = j | X_{1:N}) = \frac{\alpha m_j + \sum_{i=1}^N \delta(X_i = j)}{\alpha + N} \tag{5}$$

which is the same as (4). Hence, a Polya urn process gives rise to the Dirichlet distribution in the discrete case. In fact, this is trivially true from the definition of the Polya Urn model.

2 The Dirichlet Process

The Dirichlet process is simply an extension of the Dirichlet distribution to continuous spaces. Referring back, we see that (2) implies the existence of a Dirichlet distribution on every partition of any (possibly continuous) space \mathcal{X} , since the partition is itself a discrete space. The Dirichlet Process $\mathcal{D}\mathcal{P}(\Theta; \alpha, M)$ is the unique distribution over the space of all possible distributions on \mathcal{X} , such that the relation

$$\Theta(\mathcal{X}_1), \Theta(\mathcal{X}_2), \dots, \Theta(\mathcal{X}_n) \sim \text{Dir}(\alpha, M) \quad (6)$$

holds for every natural number n and every n -partition $\{\mathcal{X}_1, \mathcal{X}_2, \dots, \mathcal{X}_n\}$ of \mathcal{X} [5], where we denote the Dirichlet process as $\mathcal{D}\mathcal{P}(\cdot)$.

At first glance, it may seem that Θ is a continuous distribution since M is continuous. However, Blackwell [1] showed that Dirichlet Processes are discrete as they consist of countably infinite point probability masses. This gives rise to the important property that values observed from a Dirichlet process previously have a non-zero probability of occurring again.

All the properties of the Dirichlet distribution, including the equivalence with the Polya urn scheme, also hold for the Dirichlet process. Indeed, an alternate method for obtaining the Dirichlet process is to consider the limit as the number of colors in the Polya urn scheme tends to a continuum [2]. This limit yields an important formula called the Blackwell-MacQueen formula that forms the basis of the majority of algorithms for performing inference over Dirichlet processes. The formula is analogous to (5) in continuous spaces, and is given as

$$P(X_{N+1} = j | X_{1:N}) = \begin{cases} \frac{1}{\alpha+N} \sum_{i=1}^N \delta(X_i = j) & \exists k \leq N, \text{ s.t. } X_k = j \\ \frac{\alpha}{\alpha+N} M(j) & X_k \neq j, \forall 1 \leq k \leq N \end{cases} \quad (7)$$

3 The Dirichlet Process Mixture Model

Consider a mixture model of the form $y_i \sim \sum_{i=1}^k \pi_i f(y | \theta_i)$. Hence y is distributed as a mixture of distributions having the same parametric form f but differing in their parameters. Also let all the parameters θ_i be drawn from the same distribution G_0 . This mixture model can be expressed hierarchically as follows-

$$\begin{aligned} y_i | c_i, \Theta &\sim f(y | \theta_{c_i}) \\ c | \pi_{1:k} &\sim \text{Discrete}(\pi_1, \pi_2, \dots, \pi_k) \\ \theta_i &\sim G_0(\theta) \\ \pi_1, \pi_2, \dots, \pi_K &\sim \text{Dir}(\alpha, M) \end{aligned} \quad (8)$$

Here c_i are the indicators or labels that assign the measurements y_i to a parameter value θ_{c_i} and π_i are the mixture coefficients that are drawn from a Dirichlet distribution. Given the mixture coefficients, the indicator variables are distributed multinomially (an individual label is discretely distributed, see (??)). It is to be noted that the latent indicator variables are used here only to simplify notation. If the number of components in the mixture is known a priori, the parameters for each component can be drawn from G_0 beforehand, and then the Dirichlet distribution would be on the probability of selection of these parameters i.e., the set $\{\theta_1, \theta_2, \dots, \theta_k\}$.

Let us now consider the limit of this model as $k \rightarrow \infty$. It can be seen that in the limit, the Dirichlet distribution becomes a Dirichlet process with base measure M . For each indicator c_i drawn conditioned on all the previous $(i-1)$ indicators from the Multinomial distribution, there is a corresponding θ_i that is drawn from G_0 . In the limit $k \rightarrow \infty$, the labels lose their meaning as the space of possible labels becomes continuous. We can discard the use of labels in the model and let the parameters be drawn from a Dirichlet process with base measure G_0 instead.

Hence, the DPM model is

$$\begin{aligned} y_i | \theta_i &\sim f(y | \theta_i) \\ \theta_i | G &\sim G(\theta) \\ G &\sim \mathcal{D}\mathcal{P}(\alpha G_0(\theta)) \end{aligned} \quad (9)$$

where $\mathcal{D}\mathcal{P}(\alpha_0 G_0)$ is the Dirichlet Process with base measure G_0 and spread α , and G is a random distribution drawn from the DP.

The alternate way to express the above argument is as follows. Using (4), we obtain the marginal distribution of c_i given $c_{1:i-1}$ as

$$P(c_i = c \mid c_1, c_2, \dots, c_{i-1}) = \frac{n_{i,c} + m_c}{m_c + i - 1} \quad (10)$$

where m_c is the prior expectation of c using the measure M , and $n_{i,c}$ is the number of occurrences of c in the first $i - 1$ indicator variables. As $K \rightarrow \infty$, the prior expectation of any one specific label tends to zero (the probability of any point in a continuous distribution is zero) and hence, the limit of the above prior reaches the following

$$P(c_i = c \mid c_1, c_2, \dots, c_{i-1}, \alpha, M) = \begin{cases} \frac{n_{i,c}}{\alpha + i - 1} & \exists j < i, s.t. c_j = c \\ \frac{\alpha}{\alpha + i - 1} & \forall j < i, c_j \neq c \end{cases} \quad (11)$$

Now from (9), it can be seen that the marginal probability of θ_i given $\theta_{1:i-1}$ is given by the Blackwell-MacQueen Polya Urn formula (7).

$$P(\theta_i = \theta \mid \theta_1, \theta_2, \dots, \theta_{i-1}, \alpha, G_0) = \begin{cases} \frac{1}{\alpha + i - 1} \sum_{j=0}^{i-1} \delta(\theta - \theta_j) & \exists j < i, s.t. \theta_j = \theta \\ \frac{\alpha}{\alpha + i - 1} G_0 & \forall j < i, \theta_j \neq \theta \end{cases} \quad (12)$$

Due to the correspondence between equations (11) and (12), it can be seen that in the limit $k \rightarrow \infty$, the model (8) and (9) are the same.

A mechanical though unintuitive method for testing the applicability of the DPM to a problem is as follows. Any parametric model for the measurements y_i described hierarchically as

$$\begin{aligned} y_i \mid \theta_i &\sim f(y \mid \theta_i) \\ \theta_i \mid \psi &\sim h(\theta \mid \psi) \end{aligned} \quad (13)$$

can be replaced with a DPM model of the form (9) by removing the assumption of the parametric prior h at the second stage and instead replacing it with a general distribution G that has a Dirichlet process prior [5]

4 Sampling using a DPM

Escobar [4] first showed that MCMC techniques, specifically Gibbs sampling, could be brought to bear on posterior density estimation if the Blackwell-MacQueen Polya Urn formulation of the DP is used. Consider (12) again, but now, we condition on not only $\theta_{1:i-1}$ but on all θ indexed from 1 to n except i , where n is some whole number. We denote this vector by $\theta^{(i-)}$. (Note: We can only do this because samples from the DP are exchangeable, meaning that the joint distribution of the variables does not depend on the order in which they are considered).

Our aim is to find the posterior on θ_i , given a data instance y_i . The posterior can be calculated using Bayes law as

$$P(\theta_i \mid \theta^{(i-)}, y_i) \propto P(y_i \mid \theta_i) P(\theta_i \mid \theta^{(i-)}) \quad (14)$$

where all the probabilities are implicitly conditioned on the Dirichlet process parameters. The prior on θ_i is obtained from (12) as

$$P(\theta_i = \theta \mid \theta^{(i-)}) = \frac{\alpha}{\alpha + n - 1} G_0(\theta) + \frac{1}{\alpha + n - 1} \sum_{\substack{j=1 \\ j \neq i}}^n \delta(\theta - \theta_j) \quad (15)$$

while the likelihood of the data is simply $f(y_i; \theta_i)$ from (9). The posterior is thus

$$P(\theta_i \mid \theta^{(i-)}, y_i) = b \alpha G_0(\theta_i) f(y_i; \theta_i) + b \sum_{\substack{j=1 \\ j \neq i}}^n f(y_i; \theta_j) \delta(\theta_j) \quad (16)$$

$$\begin{aligned}
b &= \left(\alpha q_0 + \sum_{\substack{j=1 \\ j \neq i}}^n f(y_i; \theta_j) \right)^{-1} \\
q_0 &= \int_{\theta} G_0(\theta) f(y_i | \theta)
\end{aligned} \tag{17}$$

where b is a normalizing constant, and $\delta(\theta_i - \theta_j)$ is a point probability mass at θ_j .

It can be observed that q_0 is actually the marginal distribution of y_i and hence, is the inverse of the normalizing term in (14). (16) is often written in terms of the posterior $h(\theta_i | y_i) = \frac{G_0(\theta_i) f(y_i; \theta_i)}{\int_{\theta} G_0(\theta) f(y_i; \theta)}$ as

$$P(\theta_i | \theta^{(-i)}, y_i) = b \alpha q_0 h(\theta_i | y_i) + b \sum_{\substack{j=1 \\ j \neq i}}^n f(y_i; \theta_j) \delta(\theta_i - \theta_j) \tag{18}$$

This can also be written in a form that demonstrates the mixture nature of the marginal posterior on θ_i and also gives a simple algorithm for sampling from $\theta_i | \theta^{(i-)}, y_i$

$$P(\theta_i | \theta^{(-i)}, y_i) = \begin{cases} \theta_j & \text{with probability } b f(y_i; \theta_j) \\ \sim h(\theta | y_i) & \text{with probability } b \alpha q_0 \end{cases} \tag{19}$$

A Gibbs sampling algorithm using (19) can be easily designed to perform sampling on the space of θ s.

DPMs can be categorized as being conjugate models or non-conjugate models. In a conjugate model, the distributions f and G_0 are conjugate and hence, the integration in the calculation of q_0 can be performed analytically. If this is not the case, then the DPM is said to have a non-conjugate prior and inference becomes much harder. Only recently has a satisfactory solution to this problem been found [3, 7].

5 The Partitioning Problem as inference over a mixture model

Consider a situation where we have N measurements $Y = \{y_i | i \in [1, N]\}$ that are distributed as a mixture density $P(y_i) = \sum_{i=1}^k \pi_i f(y; \theta_i)$ where the θ are the parameters of the distribution f and the π are the mixing coefficients. The number of components in the mixture, k , is unknown. However, it is known that each measurement y_i is generated from only one of the components of the mixture, i.e. given a specific set of parameters θ_i^* , $y_i | \theta_i^* \sim f(y_i; \theta_i^*)$. The parameters θ are in turn modeled hierarchically as $\theta \sim h(\psi)$. The problem is to classify or cluster the measurements with regard to the mixture component that generated it (or to the mixture component that it “belongs” to). Hence, each mixture component is associated with a disjoint subset of the set of measurements and the mixture components give rise to a partition of the set of measurements.

This problem could be solved with Reversible Jump MCMC as it involves inferring a mixture density [8]. However, when using this technique (or many others), the distribution h has to be specified, and the parameters θ and hyper-parameters ψ have to be inferred. The parameter estimation, in particular, adds significantly to the complexity of the problem. Non-parametric estimation overcomes this problem by eliminating the need for parameters. In addition, DPMs do not assume any particular parametric form for h , but instead replace it with a general distribution with a Dirichlet process prior as explained in the next section.

6 An Example

I will illustrate partition sampling using DPMs using the example of partitioning N real numbers $R = \{y_i | i \in [1, N]\}$ that are distributed normally, i.e. $P(r) = \sum_{i=1}^k \pi_i \mathcal{N}(\mu_i, 1)$. Each number in R is generated from the one of the components of the mixture and hence, each set in the partition corresponds to a particular normal distribution. The mean of the normal distribution corresponding to a set in the partition can be seen as the “true” value which is represented by

the (noisy) measurements that make up the set. The problem can also be viewed as that of finding the clustering distribution of R given that the elements in R are distributed normally (but with different parameters).

Comparing with (9), it can be observed that in this case f is a univariate normal distribution with unknown mean but known, constant variance equal to unity. The base measure G_0 is taken to be the standard normal distribution $\mathcal{N}(0, 1)$. We can then define our model to be the following

$$\begin{aligned} y_i | \mu_i &\sim \mathcal{N}(\mu_i, 1) \\ \mu_i &\sim G(\mu) \\ G &\sim \mathcal{DP}(\alpha G_0(\mu)) \\ G_0 &= \mathcal{N}(0, 1) \end{aligned} \tag{20}$$

Note that it is possible to extend the model to include parametrized distributions for the case of unknown variance, α , and G_0 . This is not done here for reasons of simplicity.

Performing the calculations using (16), we find

$$q_0 = \frac{1}{2\sqrt{\pi}} \exp -\frac{y_i^2}{4}$$

and

$$h(\mu | y_i) = \mathcal{N}\left(\frac{y_i}{2}, \frac{1}{2}\right)$$

and hence, our Gibbs sampler becomes (from (19))

$$P(\mu_i | \mu^{(-i)}, y_i) = \begin{cases} \mu_j & \text{with probability proportional to } f(y_i; \mu_j) \\ \sim h(\mu | y_i) & \text{with probability proportional to } \alpha q_0 \end{cases}$$

We initialize the Gibbs sampler by consider each of the n input instances $y_{1:n}$ as being in its own set, i.e. $\mu_i^{(0)} = y_i$. Subsequently, the j th step of the Gibbs sampling is done in the following way

$$\begin{aligned} \text{Sample } \mu_1^{(j)} \text{ from } & \mu_1 | \mu_2 = \mu_2^{(j-1)}, \mu_3 = \mu_3^{(j-1)}, \dots, \mu_n = \mu_n^{(j-1)} \\ \text{Sample } \mu_2^{(j)} \text{ from } & \mu_2 | \mu_1 = \mu_1^{(j)}, \mu_3 = \mu_3^{(j-1)}, \dots, \mu_n = \mu_n^{(j-1)} \\ & \vdots \\ \text{Sample } \mu_n^{(j)} \text{ from } & \mu_n | \mu_1 = \mu_1^{(j)}, \mu_2 = \mu_2^{(j)}, \dots, \mu_{n-1} = \mu_{n-1}^{(j)} \end{aligned}$$

References

- [1] D. Blackwell. Discreteness of Ferguson selections. *Annals of Statistics*, 1:356–358, 1973.
- [2] D. Blackwell and J.B. MacQueen. Ferguson distributions via polya urn schemes. *Annals of Statistics*, 1:353–355, 1973.
- [3] P. Damien, J. C. Wakefield, and S. G. Walker. Gibbs sampling for Bayesian nonconjugate and hierarchical models using auxiliary variables. *Journal of the Royal Statistical Society Series B*, 61:331–344, 1999.
- [4] M. D. Escobar. Estimating the means of several normal populations by nonparametric estimation of the distribution of the means. Unpublished dissertation, Yale University, 1988.
- [5] T. S. Ferguson. A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1:209–230, 1973.
- [6] N. L. Johnson and S. Kotz. *Urn Models and their Applications*. John Wiley and Sons, 1977.
- [7] S. N. MacEachern and P. Muller. Estimating mixture of dirichlet process models. *Journal of Computational and Graphical Statistics*, 7:223–238, 1998.
- [8] S. Richardson and P. J. Green. On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society. Series B (Methodological)*, 59:731–792, 1997.