Report on Dirichlet Process

Angie Shen

February 19, 2016

1 Bayesian Nonparametric Models

A model is **parametric** if it is indexed by a parameter with values on a finite parameter space. For example, in a regression problem, suppose the data show a linear trend. We can then use a linear function to model the data. The parameter space is the set of linear functions on $\mathbb{R}$. A linear function on $\mathbb{R}$ can be specified using two parameters, a coefficient β_0 for the intercept and a coefficient β_1 for the slope. In this case, the parameter space is two-dimensional, expressed as $\mathbb{R}^2$. The linear function is thus parametric. Suppose the data show a non-linear trend. The parameter space then becomes be the set of all continuous functions on $\mathbb{R}$. The parameter space is now infinite-dimensional, which means the model is nonparametric. A **Bayesian nonparametric model** is a Bayesian model on an infinite-dimensional parameter space, which means we have to define a probability distribution (the prior) on an infinite-dimensional space.

Traditionally, the 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. Compared to parametric models, Bayesian nonparametric models are more flexible in that they can adapt their complexity to the data: the number of parameters can grow with data size. This becomes particularly helpful in a clustering problem where the number of clusters can grow as new data points are observed.

2 Mixture Models

A **mixture model** corresponds to the probability distribution of a random variable that is derived from a collection of other random variables as follows: first, a random variable is selected from the collection according to given probabilities of selection, and then the value of the selected random variable is realized. The individual distributions that are combined to form the mixture distribution are called the **mixture components**, and the probabilities (or weights) associated with each component are called the **mixture proportions** or **mixture weights**. When the number of components in mixture model is finite, the model is a **finite mixture model**. When the mixture components are countably infinite, the model is an **infinite mixture model**.

To give a more concrete example, mixture models are used in clustering problems. We have observations $x_1, ..., x_n$, and the objective is to divide the sample into k subsets, the clusters. We want to group them in such a way that observations in the same cluster are more similar to each other than to those in other clusters. Each observation is assigned a unique cluster label 1, 2, ..., k. Such an assignment defines a partition of the observations $x_1, ..., x_n$ into k disjoint sets. Consider the problem of grouping college students based on their hobbies such as sports, music or reading. Each student belongs to a unique cluster. We need to figure out both both the identities of the clusters and the assignments of each student to them.

Mixture models can be used to understand the group structure of a data set. A finite mixture model assumes that there are K clusters, each associated with a parameter θ_k. Each observation x_i is assumed to be generated by first choosing a cluster k according to P_k and then generating the observation from its corresponding observation distribution parameterized by θ_k. Bayesian mixture models further contain a prior over P_k, and a prior over the cluster parameters θ_k.

2.1 Definition

A finite mixture model has the following components:

- N random variables corresponding to observations, each distributed according to a mixture of K components, with each component belonging to the same parametric family of distributions but with different parameters. It is assumed that it is unknown which mixture component underlies each particular observation.

- N corresponding unobserved(latent) indicator variables specifying the identity of the mixture component of each observation, each distributed according to a K-dimensional multinomial distribution.

- A set of K parameters, each specifying the parameter of the corresponding mixture component. Observations distributed according to a mixture of K-dimensional multinomial distributions will have a vector of K probabilities, collectively summing to 1.

- A set of K mixing proportions parametrizing the K-dimensional multinomial distribution, each of which is a probability (a real number between 0 and 1 inclusive), all of which sum to 1.

In a Bayesian setting, the mixing proportions and parameters of the mixture components will themselves be random variables, and prior distributions will be placed over the variables. In such a case, the mixing proportions are typically viewed as a K-dimensional random vector drawn from a Dirichlet distribution (the conjugate prior of the multinomial distribution), and the parameters will be distributed according to their respective conjugate priors.

Consider a Bayesian mixture model consisting of K components:

$$\begin{aligned} z_i|\pi_1...\pi_k &\sim \text{Mult}(\pi_1...\pi_k) \\ x_i|z_i,\theta_k &\sim F(\theta_{z_i)} \\ \pi_1...\pi_k|\alpha_1...\alpha_k &\sim \text{Dir}(\alpha_1...\alpha_k) \\ \theta_k|H &\sim H \end{aligned}$$

where x_i is the observation i, z_i is the indicator variable specifying the mixture component of observation i, π is the mixing proportion denoting the probability for each mixture component, α is the 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.

In a clustering problem, we can express the cluster assignment as a random variable Z. Then $Z_i = k$ means X_i belongs to cluster k. We can obtain the distribution characterizing a single cluster k by conditioning on Z,

$$P_k(A) = P[X \in A|Z = k].$$

We define the probability for a newly generated observation to be in cluster k,

$$\pi_k = P(Z = k)$$

where $\sum_k \pi_k = 1$, since the π_k are probabilities of mutually exclusive events.

If we assume that all P_k are distributions with the conditional density $p(x|\theta_k)$, then the distribution of X has density

$$p(x) = \sum_{k \in \mathbb{N}} \pi_k p(x|\theta_k)$$

where π_k is the mixing proportion and θ_k are parameters associated with component k. A model with this density function is called a **mixture model**. If the number of clusters is finite, it is then a **finite mixture model**. The density can also be written in the form of an integral

$$p(x) = \sum_{k \in \mathbb{N}} \pi_k p(x|\theta_k) = \int p(x|\theta)\theta d\theta$$

where $\theta = \sum\limits_{k \in \mathbb{N}} \pi_k \delta_{\theta_k}$ is a discrete mixing measure encapsulating all the parameters of the mixture model and δ_θ is a point mass(indicator function) centered at θ. Probability measures are functions with certain special properties which allow them to be interpreted as distributions over some probability space Ω. A **Bayesian mixture model** is therefore a mixture model with a *random* mixing measure

$$\Theta = \sum_{k \in \mathbb{N}} \pi_k \delta_{\theta_k}$$

Bayesian infinite mixture models use mixing measures consisting of a countably infinite number of components

$$\Theta = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}$$

To define a Bayesian infinite mixture model, we have to choose the component densities $p(x|\theta)$, and we have to generate a random probability measure Θ. We need a prior over the infinite discrete mixing measure Θ, and the most common prior to use is a Dirichlet process. The resulting mixture model is called a DP mixture model (see Section 3.3.1).

2.2 Dirichlet Multinomial Distribution

2.2.1 Multinomial Distribution

The multinomial distribution is a generalization of the binomial distribution. Whereas the binomial distribution is the probability distribution of the number of successes for one of two outcomes (success or failure) in n trials, in a multinomial distribution, each trial results in exactly one of some fixed finite number k possible outcomes, with probabilities $p_1, ..., p_k$. While the trials are independent, their outcomes x are dependent because they must be summed to n. For example, in one trial, one of the elements x_i of the k-dimensional vector equals 1, and all remaining elements equal 0. Suppose $k = 6$, $i = 3$, then $x_3 = 1$, $\mathbf{x} = (0, 0, 1, 0, 0, 0)$, $\sum_{i=1}^{6} x_i = n = 1$.

If the random variables x_i indicate the number of times outcome i is observed over n trials, the vector $\mathbf{x} = (x_1, ..., x_k)$ follows a multinomial distribution with parameters n and $\mathbf{p}$, where $\mathbf{p} = (p_1, ..., p_k)$, $p_i \geq 0$ for $i = 1, ..., k$, $\sum_{i=1}^{k} p_i = 1$ and $\sum_{i=1}^{k} x_i = n$. The distribution of $\mathbf{x}$ is then

$$p(\mathbf{x}|\mathbf{p}) = \prod_{i=1}^{k} p_i^{x_i}$$

The joint distribution of $(x_1, ..., x_k)$ is

$$p(x_1, ..., x_k|\mathbf{p}) = \frac{n!}{x_1!...x_k!} \prod_{i=1}^{k} p_i^{x_i}$$

It can also be expressed using Gamma function as

$$p(x_1, ..., x_k|\mathbf{p}) = \frac{\Gamma(\sum\limits_{i=1}^{k} x_i + 1)}{\prod\limits_{i=1}^{k} \Gamma(x_i + 1)} \prod_{i=1}^{k} p_i^{x_i}$$

2.2.2 Dirichlet Distribution

Dirichlet distribution is a family of conjugate priors for parameters of the multinomial distribution. It's a distribution over the k-dimensional parameter vector $\mathbf{p}$ (in a mixture model they are the mixing proportions) of a multinomial distribution.

If we use $\mathbf{p} = (p_1, ..., p_k)$ to denote the k-dimensional parameter vector, where $p_i \geq 0$ for all i and $\sum_{i=1}^{k} p_i = 1$, then distribution of $\mathbf{p}|\boldsymbol{\alpha}$ is $\mathrm{Dir}(\alpha_1, ..., \alpha_k)$.

$$p(\mathbf{p}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum\limits_{i=1}^{k} \alpha_i)}{\prod\limits_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} p_i^{\alpha_i - 1}$$

2.2.3 Dirichlet Multinomial Distribution

Given

$$\begin{aligned} p_1, ..., p_k &\sim \text{Dir}(\alpha_1, ..., \alpha_k) \\ x_1, ..., x_k &\sim \text{Mult}(p_1, ..., p_k) \end{aligned}$$

The posterior is

$$\begin{aligned} f(\mathbf{p}|x_1, ..., x_k) &\propto p(x_1, ..., x_k|\mathbf{p})p(\mathbf{p}|\boldsymbol{\alpha}) \\ &\propto \prod_{i=1}^{k} p_i^{\alpha_i - 1} \prod_{i=1}^{k} p_i^{x_i} \\ &= \prod_{i=1}^{k} p_i^{\alpha_i + x_i - 1} \end{aligned}$$

which is $\text{Dir}(\alpha_1 + x_1, ..., \alpha_k + x_k)$.

3 Dirichlet Process

Dirichlet processes are a family of stochastic processes whose realizations are 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 Ω. A Dirichlet process is a probability distribution whose domain is itself a set of probability distributions; it's a distribution over random discrete mixing measure. Distributions drawn from a Dirichlet Process are discrete and cannot be described using a finite number of parameters. Dirichlet process can be seen as the infinite-dimensional generalization of the Dirichlet distribution. In the same way as the Dirichlet distribution is the conjugate prior for the multinomial distribution, the Dirichlet process is the conjugate prior for infinite random discrete distributions. A particularly important application of Dirichlet processes is as a prior probability distribution in infinite mixture models.

3.1 Formal Definition

Dirichlet process has Dirichlet distributed finite dimensional marginal distributions. Let Θ be a distribution over parameter space Ω and α be a positive real number. Then for any finite measurable partition $A_1, ..., A_k$ of Ω, the k-dimensional vector $(\Theta(A_1), ..., \Theta(A_k))$ is random since Θ is random. We say Θ is Dirichlet process distributed with base distribution G and concentration parameter α, $\Theta \sim \mathrm{DP}(\alpha, G)$, if

$$(\Theta(A_1), ..., \Theta(A_k)) \sim \mathrm{Dir}(\alpha G(A_1), ..., \alpha G(A_k))$$

for every finite measurable partition $A_1, ..., A_k$ of Ω.

3.2 Posterior Distribution and Predictive Distribution

Let $\Theta \sim \mathrm{DP}(\alpha, G)$. Since Θ is a random distribution, we can in turn draw samples from Θ itself. Let $\theta_1, ..., \theta_n$ be a sequence of independent draws from Θ. We are interested in the posterior distribution of Θ given observed values of $\theta_1, ..., \theta_n$, which is an updated DP

$$\Theta | \theta_1, ..., \theta_n \sim \mathrm{DP}(\alpha + n, \frac{\alpha}{\alpha + n} G + \frac{1}{\alpha + n} \sum_{k=1}^{n} \delta_{\theta_k})$$

The predictive distribution for θ_{n+1} conditioned on $\theta_1, ..., \theta_n$ and with Θ marginalized out is

$$\theta_{n=1} | \theta_1, ..., \theta_n \sim \frac{\alpha}{\alpha + n} G + \frac{1}{\alpha + n} \sum_{k=1}^{n} \delta_{\theta_k}$$

The posterior base distribution given $\theta_1, ..., \theta_n$ is also the predictive distribution of θ_{n+1}. Since the values of draws are repeated, let $\theta_1, ..., \theta_m$ be the unique values among $\theta_1, ..., \theta_n$, and n_k be the number of repeats of θ_k. The predictive distribution can be equivalently written as

$$\theta_{n=1} | \theta_1, ..., \theta_n \sim \frac{\alpha}{\alpha + n} G + \frac{1}{\alpha + n} \sum_{k=1}^{m} n_k \delta_{\theta_k}$$

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.

3.2.1 Sampling from a Dirichlet Process

Suppose that the generation of values $X_1, \ldots, X_n$ can be simulated by the following algorithm:

1. Draw X_1 from the distribution G.

2. For $i > 1$:

   (a) with probability $\frac{\alpha}{\alpha+n-1}$, draw X_i from G.

   (b) With probability $\frac{n_k}{\alpha+n-1}$, set $X_i = k$, where n_k is the number of previous observations $X_i, i < n$, such that $X_i = k$.

An analogy to this process is the Polya Urn Scheme. Specifically, each value in Ω is a unique color, and draws X_i are balls with the value being the color of the ball. In addition we have an urn that contains previously seen balls. In the beginning there are no balls in the urn, and we pick a color drawn from G (draw $X_i \sim G$), 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 $\frac{\alpha}{n+\alpha-1}$, pick a new color (draw $X_{n+1} \sim G$), paint a ball with that color and drop the ball into the urn, or, with probability $\frac{n}{n+\alpha-1}$, reach into the urn to pick a random ball out (draw X_{n+1} from the empirical distribution), paint a new ball with the same color and drop both balls back into the urn.

In the context of a clustering problem, this means that, with probability proportional to the number of data points we have already seen, we randomly pick a point we have seen before. Points reoccur and clusters form. The clusters have a "rich gets richer" property: picking from a cluster that is already large is more probable than picking from one that is small. With probability proportional to α, we draw a new point from G. We create a new cluster. If G is is a continuous function, such as a Gaussian distribution, we will never draw the same point again.

3.3 Stick-Breaking Process

To define a Bayesian infinite mixture model, we have to generate a random probability measure $\Theta = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}$ where $\sum_{k=1}^{\infty} \pi_k = 1$. We can sample θ_k independently from a distribution G, but we cannot sample the weights π_k independently because $\sum_{k=1}^{\infty} \pi_k = 1$. After we sample π_1 from probability distribution H on $[0, 1]$,

π_2 is no longer distributed on [0,1]; it can only take values in $[0, 1 - \pi_1]$. We can think of $I_k = [0, 1 - (\pi_1 + ... + \pi_k)]$ as the remaining probability mass after the first k probabilities have been determined.

Instead of sampling the weights π_k, we sample independent proportions V_k from distribution H and then calculate the weights as

$$\pi_k = I_k \cdot V_k$$

Since I_k changes from step to step as $I_k = (1 - V_k) \cdot I_{k-1}$, we can generate the sequence as

$$V_1, V_2, ... \sim_{iid} H, \qquad \pi_k = V_k \prod_{i=1}^{k-1} (1 - V_i)$$

Such sampling procedure is called **stick-breaking process**. We can think of the interval as a stick from which pieces $(1 - V_k)$ are repeatedly broken off. We can now generate a random discrete measure $\Theta = \sum\limits_{k=1}^{\infty} \pi_k \delta_{\theta_k}$ by sampling θ_k from G and independent proportions from H on $[0, 1]$. The random discrete measure generated using beta distribution as H is a **Dirichlet process**.

3.3.1 Stick-Breaking Definition of Dirichlet Process

If $\alpha > 0$ and G is a probability distribution on Ω_θ, the random discrete measure $\Theta = \sum\limits_{k=1}^{\infty} \pi_k \delta_{\theta_k}$ generated by

$$\begin{aligned} &V_1, V_2, ... \sim_{iid} \text{Beta}(1, \alpha), \qquad \pi_k = V_k \prod_{i=1}^{k-1} (1 - V_i) \\ &\theta_1, \theta_2, ... \sim_{iid} G \end{aligned}$$

is a **Dirichlet Process** with **base distribution** G and **concentration** α, denoted by $\text{DP}(\alpha, G)$.

If we integrate a parametric density $p(x|\theta)$ against a random measure Θ generated by a Dirichlet process, we obtain a mixture model

$$p(x) = \sum_{k \in \mathbb{N}} \pi_k p(x|\theta_k)$$

called a **Dirichlet process mixture model**. Observations $X_1, X_2, ...$ are generated from a DP mixture model according to

$$\begin{aligned} \Theta &\sim \mathrm{DP}(\alpha, G) \\ \theta_1, \theta_2, ...|\Theta &\sim_{iid} \Theta \\ X_i &\sim p(x|\theta_i) \end{aligned}$$

The model represents a population subdivided into an infinite number of clusters. For a finite sample $X_1 = x_1, ..., X_n = x_n$, we can observe at most n of these clusters.

3.4 Random Partition

Assigning each observation in a sample to a unique cluster induces a random partition of the sample. For example, suppose we record observations $X_1, ..., X_1$ and compute a clustering solution that subdivides the data into three clusters $(\{X1, X2, X4, X7, X10\}, \{X3, X5\}, \{X6, X8, X9\})$. We can encode this solution as a partition of the index set $[10] = \{1, ..., 10\} : (\{1, 2, 4, 7, 10\}, \{3, 5\}, \{6, 8, 9\})$.

To make things more precise, a partition

$$\Pi = (B_1, B_2, ...)$$

of $\mathbb{N}$ is a subdivision of $\mathbb{N}$ into a (possibly infinite) number of subsets $B_i \in \mathbb{N}$, such that each $i \in \mathbb{N}$ is contained in exactly one set B_k. The sets B_k are called the **blocks** of the partition. We used variables Z_i to encode cluster assignments by setting $Z_i = k$ if X_i is in cluster k). In terms of partitions, this means

$$Z_i = k \Leftrightarrow i \in B_k$$

and a random sequence $(Z_1, Z_2, ...)$ is hence equivalent to a random partition $(B_1, B_2, ...)$.

Given a random discrete probability measure $\Theta = \sum \pi_k \delta_{\theta_k}$, we can generate a random partition Π of $\mathbb{N}$ by sampling the variables $Z_i, Z_2 ...$ with probabilities $P(Z_i = k) = \pi_k$. Any discrete probability measure Θ hence parametrizes a distribution $P_\Theta(\Pi)$ on random partitions. If Θ is a random discrete probability measure with distribution Q, we can define a distribution on partitions by integrating out Θ

$$P(\Pi) = \int P_\Theta(\Pi) Q d\Theta$$

We can sample from this distribution in two steps, as $\Theta \sim Q$ and $\Pi|\Theta \sim P_\Theta$.

3.5 Chinese Restaurant Process

In a clustering problem, we want to group the observations into clusters. Whereas the parameter in Bayesian mixture is a discrete random probability measure, the partition represents the actual subdivision of the sample. We could then argue that a more appropriate choice for the model parameter would be a partition, in which case the prior should be a distribution on partitions.

The **Chinese restaurant process** with concentration α is the distribution $P(\Pi)$ on partitions that we obtain if we choose a Dirichlet process with parameters (α, G) to be Q, the distribution on the random discrete probability measure Θ.

The CRP and DP are closely related. To substitute a random partition prior (CRP) for a random measure prior (DP), we start from the distribution over random partitions P_Π, draw a random partition from P_Π, draw a $\theta_k \sim G$ for each cluster k in the partition, and finally sample $X_i|\theta_k \sim p(x|\theta_k)$ each i in cluster k.

Here is how the analogy works. 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 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 data points 1, 2, . . . and tables as clusters, after n customers have sat down, the tables define a partition of the data points distributed according to a CRP(α) as decribed above. The process can be represented by:

For observations indexed by $i = 1, 2, ...,$

1. put i into an existing block B_k with probability $\frac{n_k}{a+n-1}$
2. create a new block containing only i with probability $\frac{\alpha}{a+n-1}$

4 Code

4.1 Procedure for Generating Datasets

1. generate a sequence from 1 to n
2. randomly sample $n \cdot p$ entries from the n entries without replacement
3. randomly sample $(n - n \cdot p)$ entries from the $n \cdot p$ entries without replacement
4. concatenent the two samples from step 2 and 3
5. create a table of the data frame

4.2 Function for Generating Datasets

Function for simulating datasets with parameters n (number of total initial data points) and p (proportion of data points resampled)

Algorithm 1: Generating Datasets

```
Input: Number of total initial data points $n$; Proportion of data points
       resampled $p$
Output: Table of simulated dataset
$sampAll \leftarrow sample(n, n * p, replace = FALSE)$
$sampRep \leftarrow sample(sampAll, n - n * p, replace = FALSE)$
$sampComb \leftarrow c(sampAll, sampRep)$
$dfComb \leftarrow data.frame(sample = sampComb)$
$tab.sim \leftarrow table(dfComb)$
return $tab.sim$
```

4.3 CRP

1. create a vector of table assignments

2. the i-th customer assigned to new table with probability a / (i + a)

3. the i-th customer assigned to existing table with probability n.j / (i + a), where n.j is the number of customers currently sitting at table j

Algorithm 2: CRP

```
**Input**: Total number of customers $N$; Dispersion parameter $\alpha$
**Output**: Table assignments for the $N$ customers
$assign \leftarrow c(rep(0, N))$
$assign[1] \leftarrow 1$
**for** $i \leftarrow 2$ ***to*** $N$ **do**
    $prob.new \leftarrow \alpha/(\alpha + i)$
    **if** $runif(1) < prob.new$ **then**
        $assign[i] \leftarrow max(assign) + 1$
    **else**
        $freqs \leftarrow c(rep(0, (i - 1)))$
    **for** $j \leftarrow 1$ ***to*** $(n - 1)$ **do**
        $n.j \leftarrow sum(assign == j)$
        $freqs[j] \leftarrow n.j/(i + a)$
    $assign[i] \leftarrow sample(1 : (i - 1), 1, prob = freqs)$
**return** $assign$
```