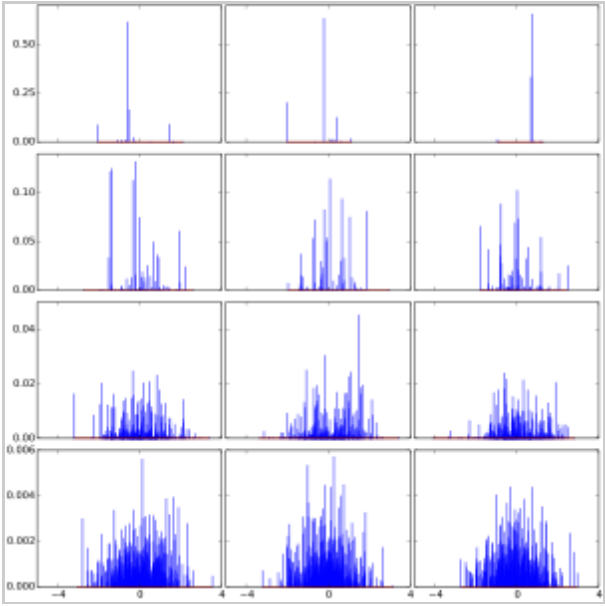
Dirichlet process

From Wikipedia, the free encyclopedia

In probability theory, **Dirichlet processes** (after Peter Gustav Lejeune Dirichlet) are a family of stochastic processes whose realizations are probability distributions. In other words, a Dirichlet process is a probability distribution whose range is itself a set of probability distributions. It is often used in Bayesian inference to describe the prior knowledge about the distribution of random variables—how likely it is that the random variables are distributed according to one or another particular distribution.



Draws from the Dirichlet process $\mathrm{DP}\left(N(0,1),\alpha\right)$. The four rows use different α (top to bottom: 1, 10, 100 and 1000) and each row contains three repetitions of the same experiment. As seen from the graphs, draws from a Dirichlet process are discrete distributions and they become less concentrated (more spread out) with increasing α. The graphs were generated using the stick-breaking process view of the Dirichlet process.

The Dirichlet process is specified by a base distribution H and a positive real number α called the concentration parameter (also known as scaling parameter). The base distribution is the expected value of the process, i.e., the Dirichlet process draws distributions "around" the base distribution the way a normal distribution draws real numbers around its mean. However, even if the base distribution is continuous, the distributions drawn from the Dirichlet process are almost surely discrete. The scaling parameter specifies how strong this discretization is: in the limit of $\alpha \to 0$, the realizations are all concentrated at a single value, while in the limit of $\alpha \to \infty$ the realizations become continuous. Between the two extremes the realizations are discrete distributions with less and less concentration as α increases.

The Dirichlet process can also 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 categorical distribution, the Dirichlet process is the conjugate prior for infinite, nonparametric discrete distributions. A particularly important application of Dirichlet processes is as a prior probability distribution in infinite mixture models.

The Dirichlet process was formally introduced by Thomas Ferguson in 1973[1] and has since been applied in data mining and machine learning, among others for natural language processing, computer vision and bioinformatics.

Contents

Introduction

Dirichlet processes are usually used when modeling data that tends to repeat previous values in a "rich get richer" fashion. Specifically, suppose that the generation of values $X_1, X_2, \dots$ can be simulated by the following algorithm.

> **Input:** H (a probability distribution called base distribution), α (a positive real number called scaling parameter)

1. Draw X_1 from the distribution H.
2. For $n > 1$:

   a) With probability $\frac{\alpha}{\alpha + n - 1}$ draw X_n from H.

   b) With probability $\frac{n_x}{\alpha + n - 1}$ set $X_n = x$, where n_x is the number of previous observations $X_j, j < n$, such that $X_j = x$.

At the same time, another common model for data is that the observations $X_1, X_2, \dots$ are assumed to be independent and identically distributed (i.i.d.) according to some distribution P. The goal in introducing Dirichlet processes is to

be able to describe the procedure outlined above in this i.i.d. model.

The $X_1, X_2, \ldots$ observations are not independent, since we have to consider the previous results when generating the next value. They are, however, exchangeable. This fact can be shown by calculating the joint probability distribution of the observations and noticing that the resulting formula only depends on which x values occur among the observations and how many repetitions they each have. Because of this exchangeability, de Finetti's representation theorem applies and it implies that the observations $X_1, X_2, \ldots$ are conditionally independent given a (latent) distribution P. This P is a random variable itself and has a distribution. This distribution (over distributions) is called Dirichlet process (DP). In summary, this means that we get an equivalent procedure to the above algorithm:

1. Draw a distribution P from $\mathrm{DP}(H, \alpha)$
2. Draw observations $X_1, X_2 \ldots$ independently from P.

In practice, however, drawing a concrete distribution P is impossible, since its specification requires an infinite amount of information. This is a common phenomenon in the context of Bayesian non-parametric statistics where a typical task is to learn distributions on function spaces, which involve effectively infinitely many parameters. The key insight is that in many applications the infinite dimensional distributions appear only as an intermediary computational device and are not required for either the initial specification of prior beliefs or for the statement of the final inference. The Dirichlet process can be used to circumvent infinite computational requirements as described above.

Formal definition

Given a measurable set S, a base probability distribution H and a positive real number α, the Dirichlet process $\mathrm{DP}(H, \alpha)$ is a stochastic process whose sample path (or realization, i.e. an infinite set of random variates drawn from the process) is a probability distribution over S and the following holds. For any measureable finite partition of S, say $\{B_i\}_{i=1}^{n}$,

$$\begin{aligned} &\text{if } X \sim \mathrm{DP}(H, \alpha) \\ &\text{then } (X(B_1), \ldots, X(B_n)) \sim \mathrm{Dir}(\alpha H(B_1), \ldots, \alpha H(B_n)) \end{aligned},$$

where Dir denotes the Dirichlet distribution and the notation $X \sim D$ means that the random variable X is distributed according to the distribution D.

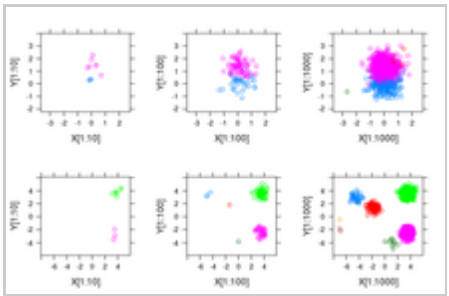
Alternative views

There are several equivalent views of the Dirichlet process. Besides the definition

above, the Dirichlet process can be defined implicitly through de Finetti's theorem as described in the first section; this is often called the Chinese restaurant process. A third alternative is the stick-breaking process, which defines the Dirichlet process constructively by writing a distribution sampled from the process as $f(x) = \sum_{k=1}^{\infty} \beta_k \delta_{x_k}(x)$, where $\{x_k\}_{k=1}^{\infty}$ are samples from the base distribution H, δ_{x_k} is an indicator function centered on x_k (zero everywhere except for $\delta_{x_k}(x_k) = 1$) and the β_k are defined by a recursive scheme that repeatedly samples from the beta distribution $\mathrm{Beta}(1, \alpha)$.

Use in Dirichlet mixture models

To understand what Dirichlet processes are and the problem they solve we consider the example of data clustering. It is a common situation that data points are assumed to be distributed in a hierarchical fashion where each data point belongs to a (randomly chosen) cluster and the members of a cluster are further distributed randomly within that cluster.



Simulation of 1000 observations drawn from a Dirichlet mixture model. Each observation within a cluster is drawn independently from the multivariate normal distribution $N(\mu_k, 1/4)$. The cluster means μ_k are drawn from a distribution G which itself is drawn from a Dirichlet process with concentration parameter $\alpha = 0.5$ and base distribution $H = N(2, 16)$. Each row is a new simulation.

Example 1

For example, we might be interested in how people will vote on a number of questions in an upcoming election. A reasonable model for this situation might be to classify each voter as a liberal, a conservative or a moderate and then model the event that a voter says “Yes” to any particular question as a Bernoulli random variable with probability dependent on which political cluster they belong to. By looking at how votes were cast in previous years on similar pieces of legislation one could fit a predictive model using a simple clustering algorithm such as k-means. That algorithm, however, requires knowing in advance the number of clusters that generated the data. In many situations it is not possible to determine this ahead of time, and even when we can reasonably assume a number of clusters we would still like to be able to check this assumption. For example, in the voting example above the division into liberal, conservative and moderate might not be finely tuned enough; attributes such as a religion, class or race could also be critical for modeling voter

behavior.

Example 2

As another example, we might be interested in modeling the velocities of galaxies using a simple model assuming that the velocities are clustered, for instance by assuming each velocity is distributed according to the normal distribution $v_i \sim N(\mu_k, \sigma^2)$, where the ith observation belongs to the kth cluster of galaxies with common expected velocity. In this case it is far from obvious how to determine a priori how many clusters (of common velocities) there should be and any model for this would be highly suspect and should be checked against the data. By using a Dirichlet process prior for the distribution of cluster means we circumvent the need to explicitly specify ahead of time how many clusters there are, although the concentration parameter still controls it implicitly.

We consider this example in more detail. A first naive model is to presuppose that there are K clusters of normally distributed velocities with common known fixed variance σ^2. Denoting the event that the ith observation is in the kth cluster as $z_i = k$ we can write this model as:

$$\begin{aligned}
(v_i \mid z_i = k, \mu_k) &\sim N(\mu_k, \sigma^2) \\
\mathrm{P}(z_i = k) &= \pi_k \\
(\boldsymbol{\pi} \mid \alpha) &\sim \mathrm{Dir}\left(\frac{\alpha}{K} \cdot \mathbf{1}_K\right) \\
\mu_k &\sim H(\lambda)
\end{aligned}$$

That is, we assume that the data belongs to K distinct clusters with means μ_k and that π_k is the (unknown) prior probability of a data point belonging to the kth cluster. We assume that we have no initial information distinguishing the clusters, which is captured by the symmetric prior $\mathbf{Dir}\left(\alpha/K \cdot \mathbf{1}_K\right)$. Here $\mathbf{Dir}$ denotes the Dirichlet distribution and $\mathbf{1}_K$ denotes a vector of length K where each element is 1. We further assign independent and identical prior distributions $H(\lambda)$ to each of the cluster means, where H may be any parametric distribution with parameters denoted as λ. The hyper-parameters α and λ are taken to be known fixed constants, chosen to reflect our prior beliefs about the system. To understand the connection to Dirichlet process priors we rewrite this model in an equivalent but more suggestive form:

$$
\begin{aligned}
(v_i \mid \tilde{\mu}_i) &\sim N(\tilde{\mu}_i, \sigma^2) \\
\tilde{\mu}_i &\sim G = \sum_{k=1}^{K} \pi_k \delta_{\mu_k}(\tilde{\mu}_i) \\
(\boldsymbol{\pi} \mid \alpha) &\sim \operatorname{Dir}\left(\frac{\alpha}{K} \cdot \mathbf{1}_K\right) \\
\mu_k &\sim H(\lambda)
\end{aligned}
$$

Instead of imagining that each data point is first assigned a cluster and then drawn from the distribution associated to that cluster we now think of each observation being associated with parameter $\tilde{\mu}_i$ drawn from some discrete distribution G with support on the K means. That is, we are now treating the $\tilde{\mu}_i$ as being drawn from the random distribution G and our prior information is incorporated into the model by the distribution over distributions G.

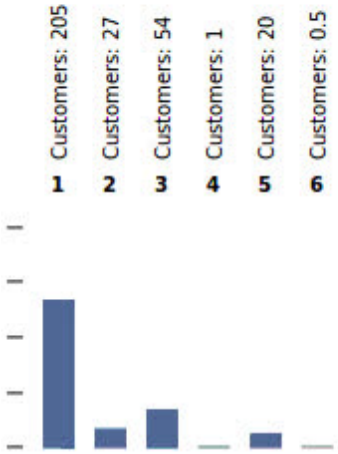
We would now like to extend this model to work without pre-specifying a fixed number of clusters K. Mathematically, this means we would like to select a random prior distribution $G(\tilde{\mu}_i) = \sum_{k=1}^{\infty} \pi_k \delta_{\mu_k}(\tilde{\mu}_i)$ where the values of the clusters means μ_k are again independently distributed according to $H(\lambda)$ and the distribution over π_k is symmetric over the infinite set of clusters. This is exactly what is accomplished by the model:

$$
\begin{aligned}
(v_i \mid \tilde{\mu}_i) &\sim N(\tilde{\mu}_i, \sigma^2) \\
\tilde{\mu}_i &\sim G \\
G &\sim \operatorname{DP}(H(\lambda), \alpha)
\end{aligned}
$$

With this in hand we can better understand the computational merits of the Dirichlet process. Suppose that we wanted to draw n observations from the naive model with exactly K clusters. A simple algorithm for doing this would be to draw K values of μ_k from $H(\lambda)$, a distribution π from $\operatorname{Dir}(\alpha/K \cdot \mathbf{1}_K)$ and then for each observation independently sample the cluster k with probability π_k and the value of the observation according to $N\left(\mu_k, \sigma^2\right)$. It is easy to see that this algorithm does not work in case where we allow infinite clusters because this would require sampling an infinite dimensional parameter $\boldsymbol{\pi}$. However, it is still possible to sample observations v_i. One can e.g. use the Chinese restaurant representation described below and calculate the probability for used clusters and a new cluster to be created. This avoids having to explicitly specify $\boldsymbol{\pi}$. Other solutions are based on a truncation of clusters: A (high) upper bound to the true number of clusters is introduced and cluster numbers higher than the lower bound are treated as one cluster.

Fitting the model described above based on observed data D means finding the posterior distribution $p(\pi, \mu \mid D)$ over cluster probabilities and their associated means. In the infinite dimensional case it is obviously impossible to write down the posterior explicitly. It is, however, possible to draw samples from this posterior using a modified Gibbs sampler.[2] This is the critical fact that makes the Dirichlet process prior useful for inference.

The Chinese restaurant process



ıinese restaurant process with scaling parameter $\alpha = 0.5$. Tables are hidden once the ble can not be displayed anymore; however, every table has infinitely many seats. interactive animation.[3])

A widely employed metaphor for the Dirichlet process is based on the so-called **Chinese restaurant process**. The name stems from the impression that Chinese restaurants would have infinitely many tables. The metaphor is as follows:

Imagine a Chinese restaurant in which customers enter. A new customer sits down at a table with a probability proportional to the number of customers already sitting there. Additionally, a customer opens a new table with a probability proportional to the scaling parameter α. After infinitely many customers entered, one obtains a probability distribution over infinitely many tables to be chosen. This probability distribution over the tables is a random sample of the probabilities of observations drawn from a Dirichlet process with scaling parameter α.

If one associates draws from the base measure H with every table, the resulting distribution over the sample space S is a random sample of a Dirichlet process. The Chinese restaurant process is related to the Pólya urn sampling scheme which yields samples from finite Dirichlet distributions.

Because customers sit at a table with a probability proportional to the number of customers already sitting at the table, two properties of the DP can be deduced:

1. The Dirichlet process exhibits a self-reinforcing property: The more often a given value has been sampled in the past, the more likely it is to be sampled again.
2. Even if H is a distribution over an uncountable set, there is a nonzero probability that two samples will have exactly the same value, because the probability mass will concentrate on a small number of tables.

The stick-breaking process

A third approach to the Dirichlet process is the so-called stick-breaking process view. Remember that draws from a Dirichlet process are distributions over a set S. As noted previously, the distribution drawn is discrete with probability 1. In the stick-breaking process view, we explicitly use the discreteness and give the probability mass function of this (random) discrete distribution as:

$$f(\theta) = \sum_{k=1}^{\infty} \beta_k \cdot \delta_{\theta_k}(\theta)$$

where δ_{θ_k} is the indicator function which evaluates to zero everywhere, except for $\delta_{\theta_k}(\theta_k) = 1$. Since this distribution is random itself, its mass function is parameterized by two sets of random variables: the locations $\{\theta_k\}_{k=1}^{\infty}$ and the corresponding probabilities $\{\beta_k\}_{k=1}^{\infty}$. In the following, we present without proof what these random variables are.

The locations θ_k are independent and identically distributed according to H, the base distribution of the Dirichlet process. The probabilities β_k are given by a procedure resembling the breaking of a unit-length stick (hence the name):

$$\beta_k = \beta'_k \cdot \prod_{i=1}^{k-1} \left(1 - \beta'_i\right)$$

where β'_k are independent random variables with the beta distribution $\mathrm{Beta}(1, \alpha)$. The resemblance to 'stick-breaking' can be seen by considering β_k as the length of a piece of a stick. We start with a unit-length stick and in each step we break off a portion of the remaining stick according to β'_k and assign this broken-off piece to β_k. The formula can be understood by noting that after the first $k - 1$ values have

their portions assigned, the length of the remainder of the stick is $\prod_{i=1}^{k-1}\left(1-\beta_i'\right)$ and this piece is broken according to β_k' and gets assigned to β_k.

The smaller α is, the less of the stick will be left for subsequent values (on average), yielding more concentrated distributions.

The Pólya urn scheme

Yet another way to visualize the Dirichlet process and Chinese restaurant process is as a modified Pólya urn scheme. Imagine that we start with an urn filled with α black balls. Then we proceed as follows:

1. Each time we need an observation, we draw a ball from the urn.
2. If the ball is black, we generate a new (non-black) color uniformly, label a new ball this color, drop the new ball into the urn along with the ball we drew, and return the color we generated.
3. Otherwise, label a new ball with the color of the ball we drew, drop the new ball into the urn along with the ball we drew, and return the color we observed.

The resulting distribution over colors is the same as the distribution over tables in the Chinese restaurant process. Furthermore, when we draw a black ball, if rather than generating a new color, we instead pick a random value from a base distribution H and use that value to label the new ball, the resulting distribution over labels will be the same as the distribution over values in a Dirichlet process.

Applications of the Dirichlet process

Dirichlet processes are frequently used in *Bayesian nonparametric statistics*. "Nonparametric" here does not mean a parameter-less model, rather a model in which representations grow as more data are observed. Bayesian nonparametric models have gained considerable popularity in the field of machine learning because of the above-mentioned flexibility, especially in unsupervised learning. In a Bayesian nonparametric model, the prior and posterior distributions are not parametric distributions, but stochastic processes.[4] The fact that the Dirichlet distribution is a probability distribution on the simplex of sets of non-negative numbers that sum to one makes it a good candidate to model distributions over distributions or distributions over functions. Additionally, the nonparametric nature of this model makes it an ideal candidate for clustering problems where the distinct number of clusters is unknown beforehand.

As draws from a Dirichlet process are discrete, an important use is as a prior probability in infinite mixture models. In this case, S is the parametric set of

component distributions. The generative process is therefore that a sample is drawn from a Dirichlet process, and for each data point in turn a value is drawn from this sample distribution and used as the component distribution for that data point. The fact that there is no limit to the number of distinct components which may be generated makes this kind of model appropriate for the case when the number of mixture components is not well-defined in advance. For example, the infinite mixture of Gaussians model [5], as well as associated mixture regression models, e.g. [6]

The infinite nature of these models also lends them to natural language processing applications, where it is often desirable to treat the vocabulary as an infinite, discrete set.

The Dirichlet Process can also be used for nonparametric hypothesis testing, i.e. to develop Bayesian nonparametric versions of the classical nonparametric hypothesis tests, e.g. sign test, Wilcoxon rank sum test, Wilcoxon signed-rank test, etc. For instance, Bayesian nonparametric versions of the Wilcoxon rank sum test and the Wilcoxon signed-rank test have been developed by using the imprecise Dirichlet process, a prior ignorance Dirichlet process.

Related distributions

- The Pitman–Yor process is a generalization of the Dirichlet process to accommodate power-law tails
- The hierarchical Dirichlet process extends the ordinary Dirichlet process for modelling grouped data.

References

1. Ferguson, Thomas (1973). "Bayesian analysis of some nonparametric problems". *Annals of Statistics*. **1** (2): 209–230. doi:10.1214/aos/1176342360. MR 350949.
2. Sudderth, Erik (2006). *Graphical Models for Visual Object Recognition and Tracking* (PDF) (Ph.D.). MIT Press.
3. http://topicmodels.west.uni-koblenz.de/ckling/tmt/crp.html?parameters=0.5&dp=1#
4. Nils Lid Hjort, Chris Holmes, Peter Müller and Stephen G. Walker (2010). *Bayesian Nonparametrics*. Cambridge University Press. ISBN 0-521-51346-4.
5. Rasmussen, Carl (2000). "The Infinite Gaussian Mixture Model" (PDF). *Advances in Neural Information Processing Systems*. **12**: 554–560.
6. Sotirios P. Chatzis, Dimitrios Korkinof, and Yiannis Demiris, “A nonparametric Bayesian approach toward robot learning by demonstration,” Robotics and Autonomous Systems, vol. 60, no. 6, pp. 789–802, June 2012. [1] (http://www.sciencedirect.com/science/article/pii/S0921889012000334)

External links

- Introduction to the Dirichlet Distribution and Related Processes by Frigyik, Kapila and Gupta (https://www.ee.washington.edu/techsite/papers/documents/UWEETR-2010-0006.pdf)
- Yee Whye Teh's overview of Dirichlet processes (http://www.gatsby.ucl.ac.uk/~ywteh/research/npbayes/Teh2010a.pdf)
- Webpage for the NIPS 2003 workshop on non-parametric Bayesian methods (https://web.archive.org/web/20070524045420/http://www.cs.toronto.edu:80/~beal/npbayes/)
- Michael Jordan's NIPS 2005 tutorial: *Nonparametric Bayesian Methods: Dirichlet Processes, Chinese Restaurant Processes and All That* (http://www.cs.berkeley.edu/~jordan/nips-tutorial05.ps)
- Peter Green's summary of construction of Dirichlet Processes (http://www.maths.bris.ac.uk/~maxvd/cribsheet.pdf)
- Peter Green's paper on probabilistic models of Dirichlet Processes with implications for statistical modelling and analysis (http://www.stats.bris.ac.uk/~peter/papers/GreenCDP.pdf)
- Zoubin Ghahramani's UAI 2005 tutorial on Nonparametric Bayesian methods (http://learning.eng.cam.ac.uk/zoubin/talks/uai05tutorial-b.pdf)
- GIMM software for performing cluster analysis using Infinite Mixture Models (http://ClusterAnalysis.org)
- A Toy Example of Clustering using Dirichlet Process. (https://archive.is/20121215093339/http://www.ece.sunysb.edu/~zyweng/dpcluster.html) by Zhiyuan Weng

Categories: Stochastic processes | Nonparametric Bayesian statistics

---