Notes on Dirichlet Processes

Daniel Jiwoong Im
School of Engineering
University of Guelph
imj@uoguelph.ca

Abstract

This note was made after studying the Dirichlet process (DP) using various resources. This is written as a simple summary of the DP and should not to be used as a study guide. The code for Chinese restaurant process is included.

1 Basic Distributions

Before diving into the Dirichlet process, we will go over some of the basic distributions that may help us understand or recall.

1.1 Beta-Bernoulli Model

The Bernoulli distribution is a probability distribution of a random variable that has either success or failures with probability θ and $1-\theta$. It is simple as tossing a coin with one side of the face being probability θ. After tossing N times, the probability of getting the same toss is

$$p(\mathcal{D}|\theta) = \theta^{\sum_i \mathbb{I}(x_i=1)}(1-\theta)^{\sum_i \mathbb{I}(x_i=0)}$$

where $\mathcal{D}$ is the result from tossing a coin N times and $\mathbb{I}$ is the indicator function.

The Beta distribution is a probability distribution over probabilities θ.

$$p(\theta|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \tag{1}$$

$$\mathbb{E}[\text{Beta}(\theta|\alpha,\beta)] = \frac{\alpha}{\alpha+\beta} \tag{2}$$

where α and β are the positive scaling parameters (hyper-parameters) and $\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$ is the normalization term. Of course, the distribution is over the probabilities, so θ must be between 0 and 1. Depending on parameter α and β, the function varies in $[0,1]$.

We will consider the problems from the Bayesian perspective. For example, consider the batting average of a baseball player. The probability of the base hit is

$$\theta = \frac{\text{\# of times a player gets a base hit}}{\text{\# of times he goes up to bat}}.$$

This requires the record of a player's previous hits. However, suppose that we do not have enough data due to this player being a rookie. Then, we would like to input some prior knowledge about the batting

probability. This task can be modelled using the Beta-Bernoulli distribution.

$$\begin{aligned} p(\theta|\mathcal{D}) &\propto p(\mathcal{D}|\theta)p(\theta) \\ &= \text{Bernoulli}(\mathcal{D}|\theta)\text{Beta}(\theta|\alpha,\beta) \\ &= [\theta^{N_1}(1-\theta)^{N_0}][\theta^{\alpha-1}(1-\theta)^{\beta-1}] \\ &= \theta^{N_1+\alpha-1}(1-\theta)^{N_0+\beta-1} \\ &\propto \text{Beta}(\theta|N_1+\alpha-1, N_0+\beta-1) \end{aligned}$$

where $N_1 = \sum_i \mathbb{I}(x_i = 1)$ and $N_0 = \sum_i \mathbb{I}(x_i = 0)$. The posterior probability of θ is based on the likelihood and prior, which are measured by the Bernolli and Beta distribution. We can see from the derivation that the Beta distribution is a conjugate-prior of Bernoulli distribution.

Back to the batting problem. In order to input prior knowledge on the batting probability, we can look at the average batting among the rookie baseball players. Suppose the average batting is $\frac{\alpha}{\alpha+\beta}$, then we can use the Beta distribution with paramter α and β.

1.2 Dirichlet-Categorical Model

The Multinomial distribution is a multivariate generalization of the binomial distribution. They are also known as the Categorical distribution. The idea is that given k categories and one of the k categories are chosen for each trial.

$$\text{Multi}(\mathcal{D}|\theta) \propto \prod_k \theta_k^{\sum_j \mathbb{I}(x_j=1)}$$

where $\sum_k \theta_k = 1$. For example, we throw a die with k faces n times. Each face of a die has θ_k probability to be chosen.

The Dirichlet distribution is also a multivariate generalization of the Beta distribution. It can be viewed as a distribution over the distribution θ.

$$\text{Dir}(\theta|\alpha) = \frac{\prod \Gamma(\alpha_i)}{\Gamma(\alpha_0)} \prod_k \theta_k^{\alpha_k - 1}$$

where $\alpha_0 = \sum_k \alpha_k$ and $\frac{1}{\beta(\alpha)} = \frac{\prod \Gamma(\alpha_i)}{\Gamma(\alpha_0)}$ is the normalization term. The Dirichlet distribution is defined over the $(K-1)$ simplex $\mathcal{S} - \{\theta \in \mathbb{R}^k | \theta_k \geq 0, \sum_k \theta_k = 1\}$.

In Bayesian statistics, Dirichlet distributions are used as prior of the multinomial distribution. Similar to Beta-bernoulli model, Dirichlet-categorical model becomes

$$\begin{aligned} p(\theta|\mathcal{D}) &\propto p(\mathcal{D}|\theta)p(\theta) \\ &= \prod^N \prod^K \theta_k^{\mathbb{I}_k(x_i=1)} \prod^K \theta_k^{\alpha_k - 1} \\ &= \prod^K \theta_k^{\sum^N \mathbb{I}_k(x_i=1)} \prod^K \theta_k^{\alpha_k - 1} \\ &= \prod^K \theta_k^{n_k + \alpha_k - 1} \\ &\propto \text{Dir}(\theta|\mathbf{N}+\alpha) \end{aligned}$$

where $n_k = \sum^N \mathbb{I}_k(x_i = 1)$ and $\mathbf{N} = \{(n_1, n_2, \cdots, n_K)\}$. At the test time,

$$\begin{aligned} p(\mathbf{x}|\mathbf{D}) &= \int p(\mathbf{x}|\mathcal{D},\theta)p(\theta|\mathcal{D})d\theta \\ &= \int \theta_x p(\theta|\mathcal{D})d\theta \\ &= \mathbb{E}[\theta_x] = \frac{\mathbf{n_x} + \alpha_\mathbf{x}}{N + \alpha_0} \end{aligned}$$

2 Dirichlet Process

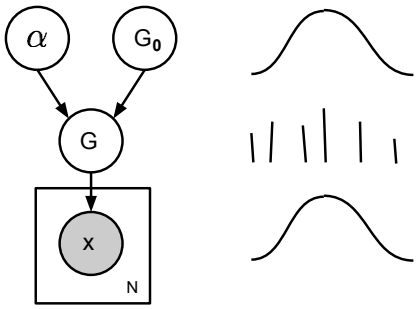

Figure 1: The Dirichlet Process

The Dirichlet Process is a stochastic process that has a collection of probability distributions.

$$G \sim \text{DP}(\alpha, G_0)$$

where G_0 is the base distribution and α is positive scaling parameter. Figure 1 shows the diagram of the Dirichlet Process. We can see that base distribution G_0 is over a continous domain and G is over a discrete space. Note that the likelihood of drawing two equal samples from G_0 is zero since it is in continuous domain, but the chances of drawing two of the same samples in G is not zero. The joint distribution over the samples is

$$p(x_1, \cdots, x_n | G_0) = \int p(G_0) p(G|G_0) \prod_i^N p(x_i|G) dG$$

Generating a new sample based on the previous sample gives us

$$x_n | x_{n-1}, \cdots, x_1 = \begin{cases} x_n = x_i & \text{with probability } \frac{\#\text{ of } x_i\text{s}}{N+\alpha-1} \\ \text{new } x_n & \text{with probability } \frac{\alpha}{N+\alpha-1} \end{cases}$$

This sampling strategy has the effect of richer get richer, because the likelihood term is the funtion of the number of samples.

2.1 The Chinese Restaurant Process

This sampling strategy is closely related the Chinese restaurant process. Consider a restaurant with possibly infinitely many tables x_i. Whenever a customer enters the restaurant, there is a chance that a customer will join his friend's table and another chance that a customer will sit at a new table. As there are more friends at the table, there is a higher chance of a new customer sitting in that table. Hence, if we think of each table as cluster, then the Chinese restaurant process has the clustering effect.

In summary, we generate the table assignments G_i accroding to the Chinese restaurant process by generating the table parameters ϕ_i from the base distribution G_0. Given the table assignments and table parameters, we generate each data points with the distribution $F(\phi_{G_i})$ (like a Gaussian distribution with the mean and variance ϕ_{G_i}).

2.2 The Polya Urn Model

In the Polya Urn Model, we assume that there is a distribution G_0 over the colours. We start with an empty urn and we are supposed to draw the coloured balls from the urn. But first, we pick a coloured ball from the base distribution G_0 and place the ball into the urn. Then, we pick a ball from the urn.

In summary, we generate the colours ϕ_i from the base distirbution G_0 according to the Polya Urn model. Then, we sample the color balls from the distribution $F(\phi_i)$.

2.3 The Stick-Breaking Process

The idea of stick-breaking process goes like this. Suppose we have a long thin 1 metre chocolate bar. First, we cut the chocolate bar. Then, pick a friend based on how good of friends they are and give it

to that friend. Similarly, that friend will repeat the process again. This process can go on forever, but the chocolate bar was originally a metre long and it will keep on shrinking. In here, the length of the chocolate you received is the probability of your feelings towards that friend, which is G.

In summary, we can generate a stick length w_i according to the stick process and stick parameter ϕ_i from the base distribution G_0. Then, we generate G_i from the multinomial distribution using stick length as the parameter w_i. Then, we sample the data point x_i from the distribution $F(\phi_{G_i})$.

References

Edwin Chen, Infinite Mixture Models with Nonparametric Bayes and the Dirichlet Process, *http://blog.echen.me/2012/03/20/infinite-mixture-models-with-nonparametric-bayes-and-the-dirichlet-process/*

Dirichlet Process Wikipedia,*http://en.wikipedia.org/wiki/Dirichlet_process*

Yee Whye Teh, Dirichlet Process, *http://www.gatsby.ucl.ac.uk/ ywteh/research/npbayes/dp.pdf*

Yee Whye Teh, Bayesian Nonparametrics (2011) *http://videolectures.net/mlss2011_teh_nonparametrics/*