Bayesian Non-parametrics

1 Introduction

[TBW]

2 Dirichlet prior on Multinomial

Suppose we have a bag of balls with K different colors. We pick a ball, and return it to the bag.

Example 1. *For example suppose our bag has only ball of 4 colors. Let's denote the observation from each of these colors to be denoted by N_1, N_2, N_3, N_4, where the sum of all observations is $\sum_i N_i = n$. The probability of different observations is the following:*

$$p(N_1 = n_1, N_2 = n_2, N_3 = n_3, N_4 = n_4) = \frac{n!}{n_1!n_2!n_3!n_4!}\theta_1^{n_1} \dots \theta_4^{n_4}$$

Like the above example, the distribution of different observations, given n total observation is the following:

$$p(N_1 = n_1, \dots, N_K = n_K) = \frac{n!}{n_1!n_2! \dots n_4!}\theta_1^{n_1} \dots \theta_K^{n_K}$$

The parameter of interest is $\Theta = (\theta_1, \dots, \theta_K)$ which lies in the following simplex:

$$\mathcal{S} = \left\{ \Theta = (\theta_1, \dots, \theta_K) \,|\, \sum_i \theta_i = 1, \theta_i \geq 0 \right\}$$

One example distribution on such simplex like $\mathcal{S}$ is the Dirichlet distribution:

$$\Theta \sim \text{Dir}(\alpha_1, \dots, \alpha_K) \Leftrightarrow p(\Theta|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_i \alpha_i)}{\Gamma(\alpha_1) \dots \Gamma(\alpha_K)} \prod_{i=1}^{K} \theta^{\alpha_i - 1}$$

Dirichlet distribution is a conjugate prior for the Multinomial distribution, i.e. given a Dirchlet prior and a Multinomial likelihood, the posterior will be of Dirichlet distribution.

Given the above likelihood and prior distributions, the posteior over the parameters will be the following:

$$p(\Theta|(x_1,\ldots,x_n)) \propto \prod_{i=1}^{n} \theta_i^{n_i+\alpha_i-1}$$

which is $\text{Dir}(\mathbf{n}+\boldsymbol{\alpha})$, given $\mathbf{n} = (n_1,\ldots,n_K)$.

2.1 Polya's urn scheme

Definition 1 (Exchangeability)**.** *A random process $x_1,\ldots,x_n$ is called infinitely exchangeable, if for any $n \in \mathbb{N}$ and any permutation function $\sigma(.)$,*

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

Theorem 1 (de Finetti's Theorem)**.** *Suppose a random process $x_1, x_2, \ldots, x_n$ is infinitely exchangeable, then there exists an unknown distribution π such that*

$$p(x_1,\ldots,x_n) = \int \left[\prod_{i=1}^{n} p(x_i|\theta)\right] \pi(\theta) d\theta$$

2.2 Taking the limit to infinity

The construction here is from [Griffiths and Ghahramani(2011)]. Now assume that we have the same construction as before for Multinomial likelihood with Dirichlet prior. This time, we want to create a model, with $K < \infty$ number of outcomes for Multinomial, and after the construction taking its limit into infinity:

$$\mathbf{x} = (x_1,\ldots,x_n)|\boldsymbol{\Theta} \sim \text{Multinomial}(\theta_1,...,\theta_K)$$
$$\boldsymbol{\Theta} = (\theta_1,\ldots,\theta_n)|\boldsymbol{\alpha} \sim \text{Dirichlet}(\frac{\alpha}{K},\ldots,\frac{\alpha}{K})$$

Note that assuming the parameters of the Dirichlet distribution to be $\alpha_i = \frac{\alpha}{K}$ is to make sure that the definition of the Dirichlet is proper when $K \to \infty$. The probability over the observations, by averaging over the latent

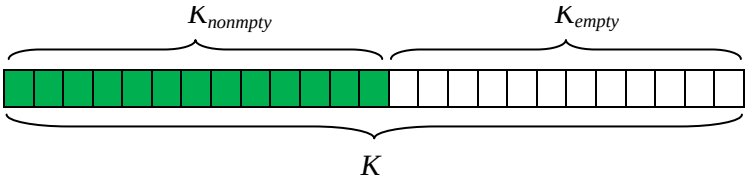

Figure 1: Splitting classes into empty and nonempty.

variables $\mathbf{\Theta}$:

$$p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{\Theta})p(\mathbf{\Theta}|\boldsymbol{\alpha})d\mathbf{\Theta} \tag{1}$$

$$= \int \frac{\prod_{i=1}^{K} \theta_i^{m_i+\frac{\alpha}{K}-1}}{D(\frac{\alpha}{K},\ldots,\frac{\alpha}{K})} d\mathbf{\Theta} \tag{2}$$

$$= \frac{D(n_1+\frac{\alpha}{K},\ldots,n_k+\frac{\alpha}{K})}{D(\frac{\alpha}{K},\ldots,\frac{\alpha}{K})} \tag{3}$$

$$= \frac{\prod_{k=1}^{K}\Gamma(n_k+\frac{\alpha}{K})}{\Gamma(\frac{\alpha}{K})^K}\frac{\Gamma(\alpha)}{\Gamma(n+\alpha)} \tag{4}$$

Now use the property of the Gamma function, $\Gamma(x+1) = x\Gamma(x)$:

$$p(\mathbf{x}) = \left(\frac{\alpha}{K}\right)^K \left(\prod_{i=1}^{K}\prod_{j=1}^{n_k-1}(j+\frac{\alpha}{K})\right)\frac{\Gamma(\alpha)}{\Gamma(n+\alpha)}$$

Now if you set $K \to \infty$ this distribution will be zero! This actually makes sense: if you have infinite number of classes, the probability that you have at least one element from each class, which needs infinite number of samples, is zero. So what?! All this to find out that the distribution of any observation from infinite cluster is zero?! The trick is that, in practice no one cares about the clusters which have no element inside. In other words, we want to create a model, which has potentially infinite number of classes, but the data might come only from a subset of the classes. For that, we change the notation. We split the classes into *empty* and *nonempty* classes, as shown in Figure 1. Note that, there is no ordering between the classes, and we can specify the labels to the classes based on any order that we want.

Given this notation, we create a modified distribution based on $p(\mathbf{x})$ which is the probability of observing finite data, from finite classes, given that there are potentially infinite number of classes, and we denote it with

$p([\mathbf{x}])$:

$$p([\mathbf{x}]) = \frac{K!}{K_{empty}!}\left(\frac{\alpha}{K_{empty}}\right)^{K_{empty}}\left(\prod_{i=1}^{K_{empty}}\prod_{j=1}^{n_k-1}(j+\frac{\alpha}{K_{empty}})\right)\frac{\Gamma(\alpha)}{\Gamma(n+\alpha)}$$

where $\frac{K!}{K_{empty}!}$ the number of possible assignments of $K_{nonempty}$ objects into K classes. Doing so, and taking the limit into infinity, we have:

$$\lim_{K\to\infty} p([\mathbf{x}]) = \alpha^{K_{empty}}\left(\prod_{i=1}^{K_{empty}}\prod_{j=1}^{n_k-1}(j+\frac{\alpha}{K_{empty}})\right)\frac{\Gamma(\alpha)}{\Gamma(n+\alpha)} \tag{5}$$

2.3 Chinese Restaurant Process (CRP)

We can write the conditional distribution of the the i-th observation in the following form:

$$p(X_i = x_i|x_{i-1}, x_{i-2}, \ldots, x_1) = \frac{p(X_i = x_i, x_{i-1}, x_{i-2}, \ldots, x_1)}{p(x_{i-1}, x_{i-2}, \ldots, x_1)}$$

If the observation $X_i = x_i$ belongs to one of the classes which is already occupied, we can use the closed form in Equation 4:

$$p(X_i = x_i|x_{i-1}, x_{i-2}, \ldots, x_1) = \frac{\frac{\Gamma(n_k+\frac{\alpha}{K}+1)}{\Gamma(i+\alpha)}}{\frac{\Gamma(n_k+\frac{\alpha}{K})}{\Gamma(i-1+\alpha)}} = \frac{n_K+\frac{\alpha}{K}}{i-1+\alpha}$$

Note that this probability only with the assumption that the observation $X_i = x_i$ in one of the occupied (*nonempty*) classes. If we take limit $K \to \infty$, we will have:

$$\lim_{K\to\infty} p(X_i = x_i|x_{i-1}, x_{i-2}, \ldots, x_1) = \lim_{K\to\infty}\frac{n_K+\frac{\alpha}{K}}{i-1+\alpha} = \frac{n_K}{i-1+\alpha}$$

If the $X_i = x_i$ does not belong to the the set of occupied classes, the probability of this event is the complement of the above probability. Therefore, the CRP distribution is the following:

$$\tilde{p}(X_i = x_i|x_{i-1}, x_{i-2}, \ldots, x_1) = \begin{cases} \frac{n_k}{i-1+\alpha} & k \leq K_{nonempty} \\ \frac{\alpha}{i-1+\alpha} & k = K_{nonempty}+1 \end{cases}$$

An example of this construction is represented in Figure 2.

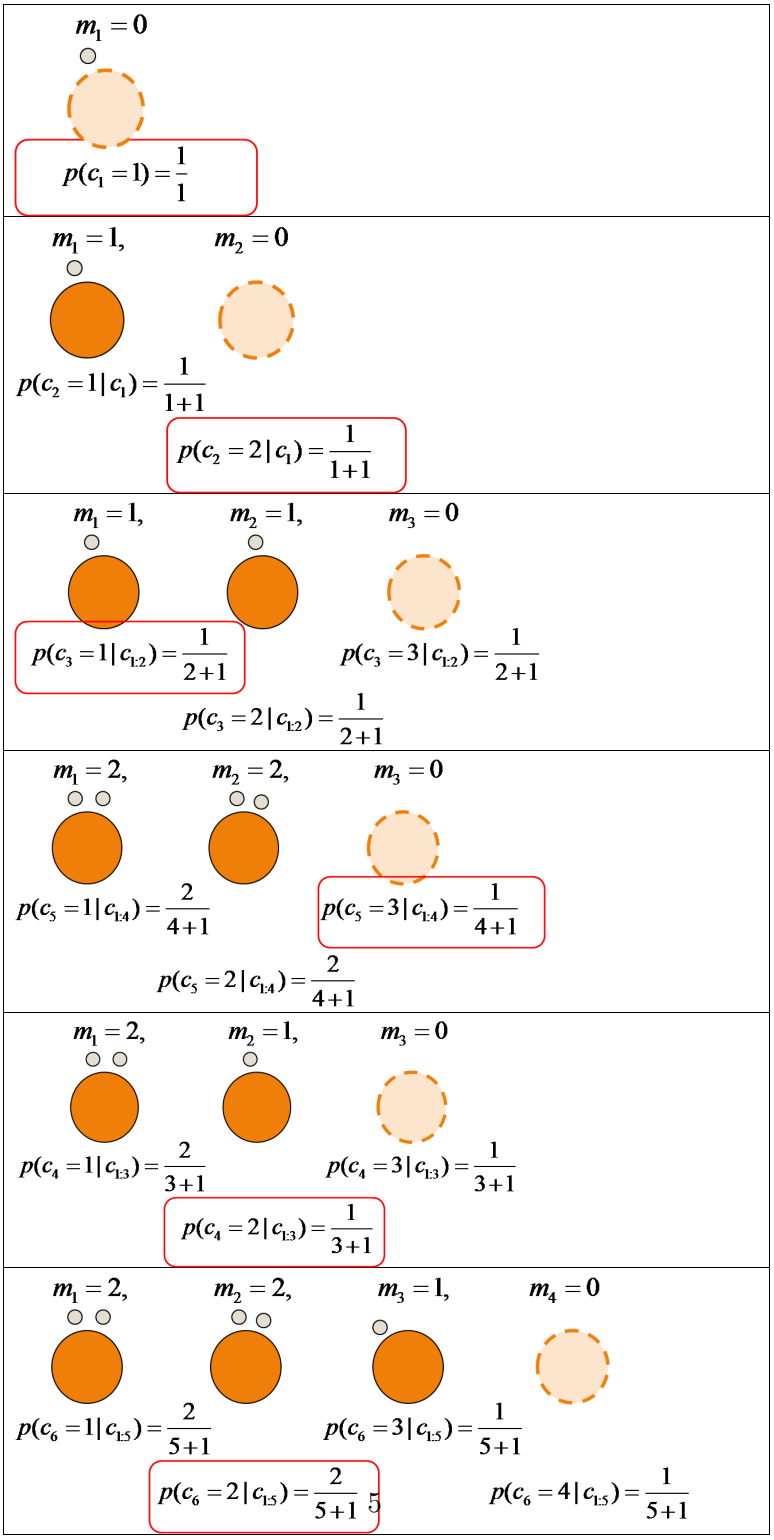

Figure 2: Example run of Chinese Restaurant Process.

Remark 1. *CRP is an exchangeable process:*

$$\mathbb{P}(122) = 1 \times \frac{\alpha}{1+\alpha} \times \frac{1}{2+\alpha}$$

If you change the order you will get the same distribution:

$$\mathbb{P}(212) = 1 \times \frac{\alpha}{1+\alpha} \times \frac{1}{2+\alpha}$$

Even if you change the labeling you will get the same distribution, since the labeling is just the way we denote different classes, and are matter of choice:

$$\mathbb{P}(112) = 1 \times \frac{1}{1+\alpha} \times \frac{\alpha}{2+\alpha}$$

3 A mixture model with potentially infinite number of components

In this section using the Chinese Restaurant Process we want to create a mixture model which has potentially infinite number of components. This model is called *Dirichlet Process Mixture* (DPM) or *Mixture of Dirichlet Processes* (MDP) model.

A mixture model could be represented in the following form:

$$x_1, \dots, x_n \overset{i.i.d.}{\sim} \sum_{k=1}^{K} w_k f_{\theta_k}$$

In this model the number of the mixture is fixed and is denoted by K. We want to create a model in which the number of the components is a random variable. Thus, from now on, though we use K to denote the number of the mixture components, but this number if not fixed, and is a random variable in our model. For this example we assume the Gaussian distribution for each component, though later we will show that the sample could come from any distribution, unless there is a computational barrier:

$$x_i | z_i = k \sim \mathcal{N}(\mu_k, \mathbf{1})$$

in which z_i denoted the cluster number for each sample x_i. Thus, the set of the unknowns are the following:

$$\left.\begin{array}{r} \text{Latent variables: } z_1, \dots, z_n \\ \text{Parameters: } \mu_1, \dots, \mu_K \end{array}\right\} \quad := \boldsymbol{\Theta}$$

which we all denote by $\mathbf{\Theta}$. Also assume that the distribution over $\mu_i \sim G_0 = \mathcal{N}(0, \tau^2)$. We denote the prior over z_i by CRP: $z_i \sim \text{CRP}(\alpha)$. The translation of this prior is the following process:

1. For the first sample, create a class and name it 1: $z_1 = 1, \mu_1 \sim G_0$.
2. For the second sample,

$$\begin{cases} \text{choose class one } z_2 = 1 & \text{with probability } \frac{1}{1+\alpha} \\ \text{create a new class: } z_2 = 2 & \text{with probability } \frac{\alpha}{1+\alpha} \text{ and sample a new } \mu \sim G_0 \end{cases}$$

$\vdots$

More formally, we denote the above procedure with the following shorthand:

$$\begin{cases} z_1, \ldots, z_n \sim \text{CRP})(\alpha) \\ \mu_1, \ldots, \mu_K \sim \mathcal{N}(0, \tau^2) \end{cases}$$

Remark 2. *In the above representation we we debited the set of unique mean values with $(\mu_1, \mu_2, \ldots)$. Depending on the data, the best model might be obtained from (μ_1), or (μ_1, μ_2), or (μ_1, μ_2, μ_3), or etc. This the parameter representation is proportional to:*

$$(\mu_1) \times (\mu_1, \mu_2) \times (\mu_1, \mu_2, \mu_3) \times \ldots \times (\mu_1, \mu_2, \ldots, \mu_n)$$

which is exponential in terms of n (the number of samples). Since the inference algorithm needs to find the best model for the data, by some sort of search in the parameters space, inference with this model is probably very hard, or needs strong approximations. For that, we change the representation.

Now, we change the previous representation to the following equivalent representation:

$$\begin{cases} x_i | \theta_i \sim \mathcal{N}(\theta_i, \mathbf{1}) \\ \theta_i \sim \mathcal{N}(0, \tau^2) \end{cases}$$

and similar to the previous representation we define the parameters $\mathbf{\Theta} = (\theta_1, \ldots, \theta_n)$, and the size of the clusters is represented by $K = [\mathbf{\Theta}]$ which is the criminality of unique θ_i's. One can find that there is a one-to-one correspondence between parameters of this parameterization $(\theta_1, \ldots, \theta_n)$ and the previous one, $(\mu_1, \ldots, \mu_K, z_1, \ldots, z_n)$.

The sampling procedure in this formulation, equivalent to the previous form, and could written as following:

1. Sample the first mean: $\theta_1 \sim \mathcal{N}(0, \tau^2)$

2. Sample the second mean: $\theta_2|\theta_1 \sim \frac{1}{1+\alpha}\delta_{\theta_1} + \frac{\alpha}{1+\alpha}\mathcal{N}(0, \tau^2)$

   $\vdots$

k+1. Sample $(k+1)$-th mean from:

$$\theta_{k+1}|\theta_{1:k} \sim \frac{1}{k+\alpha}\sum_{i=1}^{k}\delta_{\theta_i} + \frac{\alpha}{\alpha+k}\mathcal{N}(0, \tau^2) \tag{6}$$

4 Gibbs sampling on the posterior of DPM

These algorithms are from the great paper [Neal(2000)] (Section 3), which only need conjugacy between the distribution of the mixture function and the prior over its parameters. First, before moving the sampling we show a closed form for the joint distribution of the θ_i, $p(\theta_1, \theta_2, \ldots, \theta_n)$.

Remark 3. *First, for practical reasons assume that in Equation 6, instead of a continuous distribution, we have a discretized version of it (we have a probability mass function), and denote it with $\tilde{G}_0$. Now we want to find the probability of the following mean values (for simplicity suppose the values are in the domain of the PMF):*

$$p(1.2, 1.2, 2.3, 2.3, 1.2) = \left(\tilde{G}_0(1.2) \times \frac{1}{1+\alpha}\right) \times \frac{\alpha}{2+\alpha} \times \left(\tilde{G}_0(2.3) \times \frac{1}{3+\alpha}\right) \times \frac{2}{4+\alpha}$$

Using this example, one can write the following general formula:

$$p(\theta_1, \theta_2, \ldots, \theta_n) = \frac{\left[\prod_{j=1}^{K} \tilde{G}_0(\mu_j)(n_j - 1)!\right] \alpha^{K-1}}{(\alpha+1)\ldots(\alpha+n-1)}$$

given that $\mu_1, \ldots, \mu_K$ represent the unique values of $\theta_1, \ldots, \theta_n$ and the frequency of μ_j is denoted by n_j, where $\sum_{j=1}^{K} n_j = n$.

Now, we continue the sampling on the posterior. The posterior distribution is $\pi(\mathbf{\Theta}|x_{1:n})$. Since sampling from this joint is hard, we sample each θ_i

separately:

$$\pi(\theta_i|\mathbf{\Theta}_{\setminus i}, x_{1:n}) \propto p(x_{1:n}|\theta_{1:n})\pi(\theta_{1:n}) \tag{7}$$

$$\propto \sum_{j=1}^{n} \frac{1}{\alpha+n-1}\phi(x_i-\theta_j)\delta_{\theta_j}(\theta_i) + \frac{\alpha}{\alpha+n-1}\phi(x_i-\theta_i)\mathcal{N}(0,\tau^2) \tag{8}$$

where $\phi(.)$ is the mixture component.

Example 2. *How would you sample from the following distribution?*

$$W = \frac{1}{3}\delta(x-x_0) + \frac{2}{3}\mathcal{N}(0,1)$$

The answer is very simple. First, sample from a uniform distribution $\mathcal{U}(0,1)$*. If the result is less than,* $1/3$ *choose* x_0 *to be the sample, if not, sample from the uniform distribution.*

Thus, in order to have a successful sampling from the above distribution we need to be able to sample from the distribution $\phi(.)\mathcal{N}(0,\tau^2)$ in the Equation 8, we need to have closed form for $\phi(.)\mathcal{N}(0,\tau^2)$. If $\phi(.)$ is from a normal the resulting distribution will have a closed form.

5 Dirichlet Process

Dichlet Processes(DP) is a crucial element in Bayesian nonparametric modelling; since it is able to model infinite number of components. A DP is infact generalization of Dirichlet Distribution to infinite random variables. To make it clear, if we have set B of disjoint partitions on a set, $\{B_1, ..., B_k\}$, we have a DP,

$$G|\alpha, G_0 \sim \mathrm{DP}(\alpha, G_0).$$

in which α is scaling parameter and G_0 is base distribution, such that values of random measures have joint Dirichlet distribution with scaling parameter $\alpha G_0(.)$,

$$(G(B_1), ..., G(B_k)) \sim \mathrm{Dir}\left(\alpha G_0(B_1), ..., \alpha G_0(B_k)\right).$$

This definition has different interpretations, such as Chinese resturant process, or stick breaking process which are inherently the same. Using the clustering behaviour of DPs one could definite unlimited clusters and use it in infinite mixtor model, similar to conventional mixture models,

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

$$\eta_n | G \sim G,$$

$$X_n | \eta_n \sim p(x_n | \eta_n).$$

Generally these Dirichlet mixture models are trained using MCMC methods. While one could do faster, but approximate learning using mean-field variational approximation. To make everything simple, it is assumed that data is sampled from an exponentional family. In that case, one could assume that G_0, the base distribution of Dirichlet is conjugate prior of the aforementioned exponential family. The approximation is based on the constructive view of DP, i.e. stick breaking process, in which we have an infinite sequence of mixture weights $\pi = \{\pi_i\}_{i=1}^{\infty}$ derived from the following:

$$\beta_i \sim \text{Beta}(1, \alpha)$$

$$\pi_i = \beta_i \prod_{l=1}^{i-1} (1 - \beta_l) = \beta_i \left(1 - \sum_{l=1}^{i-1} \pi_l \right)$$

and we denote it by $\pi \sim \text{GEM}(\alpha)$. If we define $G(\boldsymbol{\theta}) = \sum_{k=1}^{\infty} \pi_k \delta_{\boldsymbol{\theta}_k}(\boldsymbol{\theta})$, where $\pi \sim \text{GEM}(\alpha)$ and $\boldsymbol{\theta}_k \sim H$, then one can show that $G \sim \text{DP}(\alpha, H)$.

5.1 Variational inference for DPM

In order to find the details of the variational inference for DPM go to the related chapter **??** [1]

6 Other nonparametric models

6.1 Infinite Hidden Markov Model (iHMM)

In [Beal et al.(2002)Beal, Ghahramani, and Rasmussen] they've shown an HMM with countably infinite states using Dirichlet processes. They have only three parameters that need to be learnt from data. Conventionally HMM could be trained using Baum-Welch algorithm by taking int account counts of outputs, after specifying the set of possible hidden states. While this parameter optimization would result in over/under fitting the model; thus it might seem a reasonable idea to provide a Bayesian model of HMMs. They use a two-level hierarchical Dirichlet Process model to create infinite state structure.To do inference they do Gibbs sampling which takes quite a long time.

[1] Or here: `http://web.engr.illinois.edu/~khashab2/learn/variational.pdf`

7 Bibliographical notes

Some of the explanations are based on Feng Liang's "nonparametric" course at UIUC.

References

[Beal et al.(2002)Beal, Ghahramani, and Rasmussen] Matthew J Beal, Zoubin Ghahramani, and Carl Edward Rasmussen. The infinite hidden markov model. *Advances in neural information processing systems*, 14: 577–584, 2002.

[Griffiths and Ghahramani(2011)] Thomas L Griffiths and Zoubin Ghahramani. The indian buffet process: An introduction and review. *Journal of Machine Learning Research*, 12:1185–1224, 2011.

[Neal(2000)] Radford M Neal. Markov chain sampling methods for dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265, 2000.