10-708: Probabilistic Graphical Models, Spring 2020

Lecture 23: Bayesian Nonparametrics: Dirichlet Proccesses

Lecturer: Eric P. Xing *Scribe: Hai Phan, Magesh Kanna, Stephen Tsou*

1 Introduction

Data can be represented using Parametric and non-parametric models. Parametric models are defined by a countable and fixed number of parameters. These models can be employed in the settings where the exact number of parameters to be used are known. Mixture of K-Gaussians, polynomial regression are few examples of parametric models. For non-parametric models, the number of parameters grows with the sample size. One example for a non-parametric model is Kernel density estimation. Here, the number of parameters can be random.
On the other hand, Bayesian nonparametrics models allow an infinite number of parameters *a priori* leading to infinite capacity. However, a finite dataset uses only a finite set of parameters and hence rest of the unused parameters are integrated out.

1.1 Mixture Models

Mixture Models are a class of parametric models which fixed number of parameters. In the scenario of clustered data, we can fit the data with K- Gaussian Mixture Models.

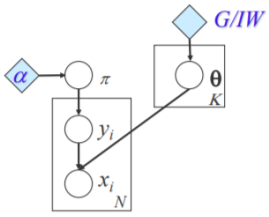

Figure 1: Bayesian Finite Mixture Model

$$p(x_1, x_2, ..., x_N|\pi, \{\mu_k\}, \{\Sigma_k\}) \;=\; \Pi_{n=1}^{\infty}\, \Sigma_{k=1}^{K}\, \pi_k\, \mathcal{N}(x_k|\mu_k, \Sigma_k)$$

We can choose the mixture weights and mixing parameters by using a Bayesian finite Mixture Modelling approach by putting a prior on them and integrating them out.

$$p(x_1, x_2, .., x_N) \;=\; \iiint (\Pi_{n=1}^{\infty}\, \Sigma_{k=1}^{K}\, \pi_k\, \mathcal{N}(x_k|\mu_k, \Sigma_k))\; p(\pi)\; p(\mu_{1:K})\; p(\Sigma_{1:K})\; d\pi\; d\mu_{1:K}\; d\Sigma_{1:K}$$

2 Dirichlet Distribution

The Dirichlet distribution is a distribution over the (K-1)-dimensional simplex. It is parametrized by a K-dimensional vector $(\alpha_1, \alpha_2, ..., \alpha_K)$ such that $\alpha_k \geq 0$ and $\sum_k \alpha_k > 0$. The Dirichlet Distribution is defined as,

$$Dirichlet(\alpha_1, \alpha_2, ..., \alpha_K) = \frac{\Pi_{k=1}^K \Gamma(\alpha_k)}{\Gamma(\Sigma_{k=1}^K \alpha_k)} \Pi_{k=1}^K \pi_k^{\alpha_k - 1}$$

If $\pi \sim$ Dirichlet$(\alpha_1, \alpha_2, ..., \alpha_K)$, then $\pi_k \geq 0$ for all k, and $\sum_k \pi_k = 1$. The expectation is given by :

$$\mathbb{E}[\pi_1, \pi_2, .., \pi_k] = \frac{(\alpha_1, \alpha_2..., \alpha_k)}{\Sigma_k \, \alpha_k}$$

Figure 2: 3-Component Dirichlet Distribution for different configurations.

2.1 Conjugacy to the multinomial distribution

Dirichlet distribution is the conjugate of multinomial distribution.Let $\pi \sim$ Dirichlet$(\alpha_1, \alpha_2, ..., \alpha_K)$ and $x_n \sim$ Multinomial(π), then we have,

$$\begin{aligned} p(\pi|x1, ..., x_n) &\propto p(x_1, x_2, .., x_n|\pi) \; p(\pi) \\ &= (\frac{\Pi_{k=1}^K \Gamma(\alpha_k)}{\Gamma(\Sigma_{k=1}^K \alpha_k)} \Pi_{k=1}^K \pi_k^{\alpha_k - 1})(\frac{n!}{m_1!, .., m_K!} \pi_1^{m_1} ....\pi_K^{m_K}) \\ &\propto \frac{\Pi_{k=1}^K \Gamma(\alpha_k + m_k)}{\Gamma(\Sigma_{k=1}^K \alpha_k + m_k)} \Pi_{k=1}^K \pi_k^{\alpha_k + m_k - 1} \\ &= Dirichlet(\alpha_1 + m_1, ..., \alpha_K + m_K) \end{aligned} \tag{1}$$

Here, m_k corresponds to the number of occurances of $x_n = k$ in the dataset.

Dirichlet Distribution is a distribution over positive vectors that sum to one. We can associate each entry with a set of parameters. For instance, in the case of Gaussian Mixture Models, the parameters would be mean and covariance of each cluster. In a Bayesian setting, these parameters are random. We can combine the distribution over probability vectors with a distribution over parameters to get a distribution over distributions over parameters.

2.2 Properties of Dirichlet Distribution

Collapsing Property: Lets examine the relationship between the Dirichlet distribution and Gamma distributions. Let $\eta_k \sim$Gamma$(\alpha_k, , 1)$ represent K independent Gamma distributed variables. The normalized

sum is represented by a Dirichlet distribution as follows:

$$\frac{(\eta_1, ..., \eta_K)}{\Sigma_k \eta_k} \sim \text{ Dirichlet}(\alpha_1, \alpha_2, ..., \alpha_K)$$

If $\eta_1 \sim$ Gamma$(\alpha_1, 1)$ and $\eta_2 \sim$ Gamma$(\alpha_2, 1)$, then $\eta_1 + \eta_2 \sim$ Gamma($\alpha_1 + \alpha_2$,1).
Therefore, if $(\pi_1, .., \pi_K) \sim$ Dirichlet$(\alpha_1, .., \alpha_K)$ then,

$$(\pi_1 + \pi_2, .., \pi_K) \sim \text{ Dirichlet}(\alpha_1 + \alpha_2, .., \alpha_K)$$

This property is called the collapsing property and is used to reduce the dimensionality of the Dirichlet distribution.

Splitting Property: The beta distribution is a Dirichlet distribution on the 1-simplex. Let $(\pi_1, .., \pi_K) \sim$ Dirichlet$(\alpha_1, .., \alpha_K)$ and $\theta \sim Beta(\alpha_1\, b, \alpha_1\, (1-b))$ for $0 < b < 1$.Then,

$$(\pi_1\theta, \pi_1(1-\theta), \pi_2.., \pi_K) \sim \text{ Dirichlet}(\alpha_1 b_1, \alpha_1(1-b_1), \alpha_2., \alpha_K)$$

More generally, if $\theta \sim$ Dirichlet$(\alpha_1 b_1, \alpha_1 b_2, .., \alpha_1 b_N)$, $\Sigma_i b_i = 1$, then,

$$(\pi_1\theta_1, ..., \pi_1\theta_N, \pi_2, .., \pi_K) \sim \text{Dirichlet}(\alpha_1\theta_1, ..., \alpha_1\theta_N, ..., \alpha_K)$$

This property is called the splitting property.

Re-normalization Property: Dirichlet Distribution follows the renormalization property.
Let $(\pi_1, .., \pi_K) \sim$ Dirichlet$(\alpha_1, .., \alpha_K)$,

$$\frac{(\pi_2, ..\pi_K)}{\Sigma_{k=2}^K \pi_k} \sim \text{Dirichlet}(\alpha_2, ..., \alpha_K)$$

2.3 Infinite Dimensional Prior

In clustering, determining the number of clusters a priori is very hard. Therefore, we allow infinite number of clusters as priors thereby ensuring to have more clusters than required.
An infinite mixture model is defined by:

$$p(x_1|\pi, \{\mu_k\}, \{\Sigma_k\}) \;=\; \Sigma_k\; \pi_k\, \mathcal{N}(x_k|\mu_k, \Sigma_k)$$

We start with Dirichlet distribution with two components $\pi^{(K)} = (\pi_1^{(2)}, \pi_2^{(2)}) \sim Dirichlet(\frac{\alpha}{2}, \frac{\alpha}{2})$ and split each component with the help of the splitting property.

On splitting according to this rule,

$$\theta_1^{(2)}, \theta_2^{(2)} \sim \text{ Beta}(\frac{\alpha}{2}.\frac{1}{2}, \frac{\alpha}{2}.\frac{1}{2})$$

On repeating this process, we get,

$$\pi^{(K)} \sim \text{Dirichlet}(\frac{\alpha}{K}, .., \frac{\alpha}{K})$$

As K in limit tends to infinity, we get a vector with infinitely many components

3 Dirichlet Process

Let define a Dirichlet process. Let H be a base measure distribution on some space Ω. This space can be a Gaussian distribution. We define the following distribution.

$$\pi \sim \lim_{K \to \infty} \text{Dirichlet}\left(\frac{\alpha}{K}, \cdots, \frac{\alpha}{K}\right)$$

For $k = 1 \cdots \infty$, denote

$$\theta_k \sim H$$

Then:

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

is an infinite distribution over base measure H. We can re-write Dirichlet process as follow.

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

Samples from the DP are discrete. We call the point masses in the distribution outcome, atoms. Figure 3 illustrate this. Intuitively speaking, The values of point masses are defined by the length of the bars. The

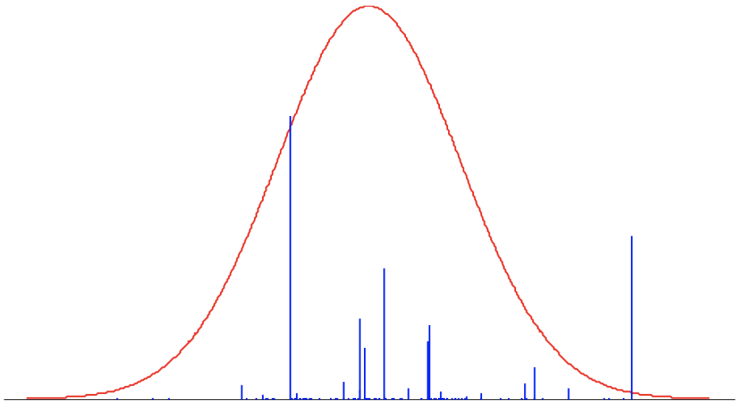

Figure 3: The point masses, atoms, in the resulting distribution. H determines the positions of these atoms concentration parameter α determines the distribution over atom sizes. The smaller value of α is, the sparser distribution is.

Properties of the Dirichlet process. For partitions $A_1, \cdots, A_K$ of Ω, e.g. different color values in an infinite color space. The total mass for each partition is presented in Figure 4.

A Dirichlet process is the unique distribution over probability distributions on spaces Ω for any finite partition $A_1, \cdots, A_K \in \Omega$.

$$(P(A_1), \cdots, P(A_K)) \sim \text{Dirichlet}(\alpha H(A_1), \cdots, \alpha H(A_K))$$

Conjugacy of the Dirichlet process. Denote $P(A_k)$ the mass assigned by $G \sim \text{DP}(\alpha, H)$. Again, we have as follow.

$$(P(A_1), \cdots, P(A_K)) \sim \text{Dirichlet}(\alpha H(A_1), \cdots, \alpha H(A_K))$$

If we see an observation in the J^{th} segment, then:

$$(P(A_1), \cdots, P(A_K) | X_1 \in A_j) \sim \text{Dirichlet}(\alpha H(A_1), \cdots, \alpha H(A_j) + 1, \cdots, \alpha H(A_K))$$

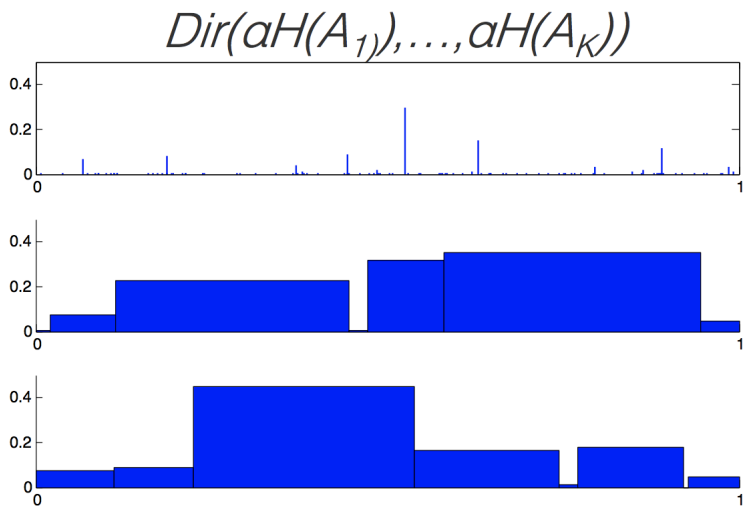

Figure 4: The point masses for each partition.

This must be true for all possible partitions of Ω. This is only possible if the posterior of G, given an observation x, is given by

$$G|X_1 = x \sim \text{DP}\left(\alpha + 1, \frac{\alpha H + \delta_x}{\alpha + 1}\right)$$

Predictive distribution. The Dirichlet process clusters observations because of being a prior in mixture models. A new data point can either join an existing cluster, or start a new cluster.

Assume H is a continuous distribution on Ω. This means for every point θ in Ω, $P_H(\theta) = \theta$. With first data point, Dirichlet process starts a new cluster samples a cluster parameter θ_1. We have now split our parameter space in two: the singleton θ_1, and everything else. Let π_1 be the size of atom at θ_1. The combined mass of all the other atoms is $\pi^* = 1 - \pi_1$.

$$\text{priori, } (\pi_1, \pi_*) \sim \text{Dirichlet}(0, \alpha)$$

$$\text{posteriori, } \quad (\pi_1, \pi_*)|X_1 = \theta_1 \sim \text{Dirichlet}(1, \alpha)$$

If we integrate out π_1 we get.

$$\begin{aligned} P(X_2 = \theta_k | X_1 = \theta_1) &= \int P(X_2 = \theta_k | (\pi_1, \pi_*)) P((\pi_1, \pi_* | X_1 = \theta_1) d\pi_1 \\ &= \int \pi_k \text{Dirichlet}((\pi_1, 1 - \pi)|1, \alpha) d\pi_1 \\ &= \mathbb{E}_{\text{Dirichlet}(1,\alpha)}[\pi_k] \\ &= \begin{cases} \frac{1}{1+\alpha} & \text{if } k = 1 \\ \frac{\alpha}{1+\alpha} & \text{for new k} \end{cases} \end{aligned} \tag{2}$$

The probability $\frac{1}{1+\alpha}$ is for the old cluster (e.g. parameter θ stay at cluster $k = 1$) while $\frac{\alpha}{1+\alpha}$ is probability for a new cluster. Lets say we choose to start a new cluster, and sample a new parameter $\theta_2 \sim H$. Let π_2 be the size of the atom at θ_2.

$$\text{posteriori, } (\pi_1, \pi_2, \pi_*)|X_1 = \theta_1, X_2 = \theta_2 \sim \text{Dirichlet}(1, \alpha)$$

If we integrate out $\pi = (\pi_1, \pi_2, \pi_*)$, we achieve.

$$\begin{aligned} P(X_3 = \theta_k | X_1 = \theta_1, X_2 = \theta_2) &= \int P(X_3 = \theta_k|\pi) P(\pi | X_1 = \theta_1, X_2 = \theta_2) d\pi \\ &= \mathbb{E}_{\text{Dirichlet}(1,1,\alpha)}[\pi_k] \\ &= \begin{cases} \frac{1}{2+\alpha} & \text{if } k = 1, 2 \\ \frac{\alpha}{2+\alpha} & \text{for new k} \end{cases} \end{aligned} \tag{3}$$

In general, if m_k is the number of times we have seen $X_i = k$, and K is the total number of observed values.

$$\begin{aligned} P(X_{n+1} = \theta_k | X_1, \cdots, X_n) &= \int P(X_{n+1} = \theta_k|\pi) P(\pi | X_1, \cdots, X_n) d\pi \\ &= \mathbb{E}_{\text{Dirichlet}(m_1, m_2, \cdots, m_K, \alpha)}[\pi_k] \\ &= \begin{cases} \frac{m_k}{n+\alpha} & \text{if } k \leq K \\ \frac{\alpha}{n+\alpha} & \text{for new cluster} \end{cases} \end{aligned} \tag{4}$$

This is a general form for predictive distribution for the next observation. This distribution have rich-get-richer property, which involve more rich observations.

3.1 Interpretations Of Dirichlet Process

DP – a Pólya urn/ Hoppe Urn Process.

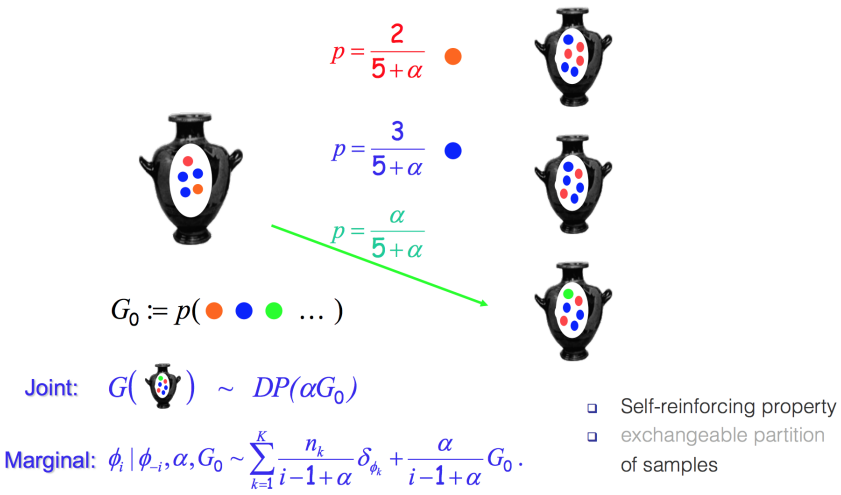

Figure 5: Visualization of pólya urn

The pólya urn scheme is a way of imagining the predictive distribution of new samples under the dirichlet process. It is one way which can be used to represent a dirichlet process. It is seen as an opposite model of sampling without replacement. In this version, they represent the dirichlet as a visual metaphor of a non-transparent urn that contains colored balls that can be drawn randomly. The example in class is actually the hoppe urn or a mutator urn which is more closely related to the Chinese Restaurant Process. In it, a singular ball is initially placed into the urn. It has a fixed $\frac{\alpha}{n+\alpha}$ probability of being drawn where n is the number of other balls in the urn. Every time this initial colored ball is drawn, a new colored ball is put into the urn. If the second colored ball is drawn, a second ball of the same color is put into the urn. If the first

colored ball is drawn again, a ball of a third color is put into the urn. For any color that is not the first, the probability of drawing that color (i.e. blue) is $\frac{k}{n+\alpha}$ where k is the number of ball of that color (i.e. that are already in the urn. Here, the joint distribution of all of the filled in urns after sampling n balls is known as the Dirichlet Process. In this example, the DP Prior is a prior distribution over the colors. The pólya urn is the procedure that defines how to draw colors for every new ball given the previous ones. The classic example of pólya urn has two balls of two different colors initially in the urn.

DP - Stick Breaking

In the non-parametric model case, stick-breaking is a way of generating a dirichlet process by generating a recursive series of beta distributions on a remaining finite mass. It is defined by a Griffiths Engen McCloskey Distribution (GEM). $\forall k$, a_k= 1 and $b_k = \alpha$, with clusters k = $1 : \infty$, atoms $\pi = (\pi_1, \pi_2, ...\pi_k) \sim GEM(\alpha)$:

$$V_k = Beta(a_k, b_k)$$

$$\pi_k = [\Pi_{j=1}^{k-1}(1 - V_j)]V_k$$

When the GEM distribution above is coupled with a parameter distribution $\mu_k \sim \mathcal{N}(\mu_0, \sigma_0)$ for instance, a dirichlet process can be defined as

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\mu,k} = DP(\alpha, \mathcal{N}(\mu_0, \sigma_0)$$

DP - Chinese Restaurant Process

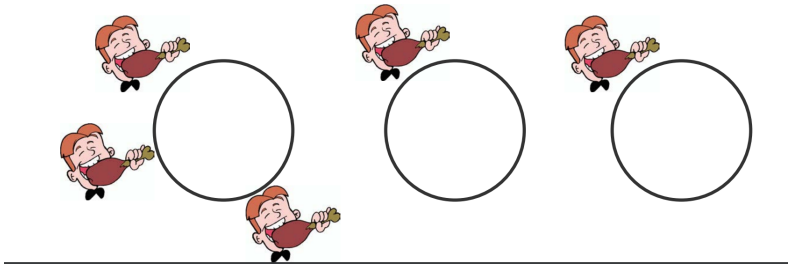

Figure 6: Visualization of Chinese Restaurant Process of 5 Customers

If one integrates all of the π_k out of the above dirichlet process, this is called the Chinese Restaurant Process. Metaphorically, it is akin to seating up to an infinite amount of customers at an up to an infinite amount of tables. It is an index set on a set of permutations.

One interesting aspect of it is both the hoppe urn and Chinese Restaurant Process are invariant to permutation. In other words, the clustering of N customers does not depend on the order in which they arrive and hence exchangeable.

We can define CRP(α, N) as a distribution over all partitions of the labeled set [N] := $\{1, 2, ..., N\}$. For example, $\pi_{[5]} = \{\{1, 3, \}, \{2\}, \{4, 5\}\}$ is a partition of [5]. This distribution is defined recursively. Given a partition $\pi_{[n]}$, the destination of the next person n + 1 has the following distribution.

$$P(n + 1 \text{ joins table c } |\pi_{[n]}) = \frac{|c|}{n + \alpha}$$

$$P(n+1 \text{ starts a new table } |\pi_{[n]} = \frac{\alpha}{n+\alpha}$$

The probability of the example partition $\{\{1,3,\},\{2\},\{6,4,5\}\}$ under CRP(α, 6) is:

$$\frac{\alpha}{\alpha} \cdot \frac{\alpha}{\alpha+1} \cdot \frac{1}{\alpha+2} \cdot \frac{\alpha}{\alpha+3} \cdot \frac{1}{\alpha+4} \cdot \frac{2}{\alpha+5}.$$

Without loss of generality, we can see that the probability of a given partition $\pi_{[N]} \sim$ CRP(α, N) is:

$$P(\pi_{[N]}) = \frac{1}{\alpha(\alpha+1)\cdots(\alpha+N-1)} \Pi_{c \epsilon \pi_{[N]}} \alpha(|c|-1)!)$$

$$= \frac{\alpha^k}{\alpha(\alpha+1)\cdots(\alpha+N-1)} \Pi_{c \epsilon \pi_{[N]}} (|c|-1)!$$

where K represents the number of clusters c in $\pi_{[N]}$. From the above we can see that CRP is exchangeable because only the sizes of the clusters affect the probability.

4 Graphical Model Representations of DP

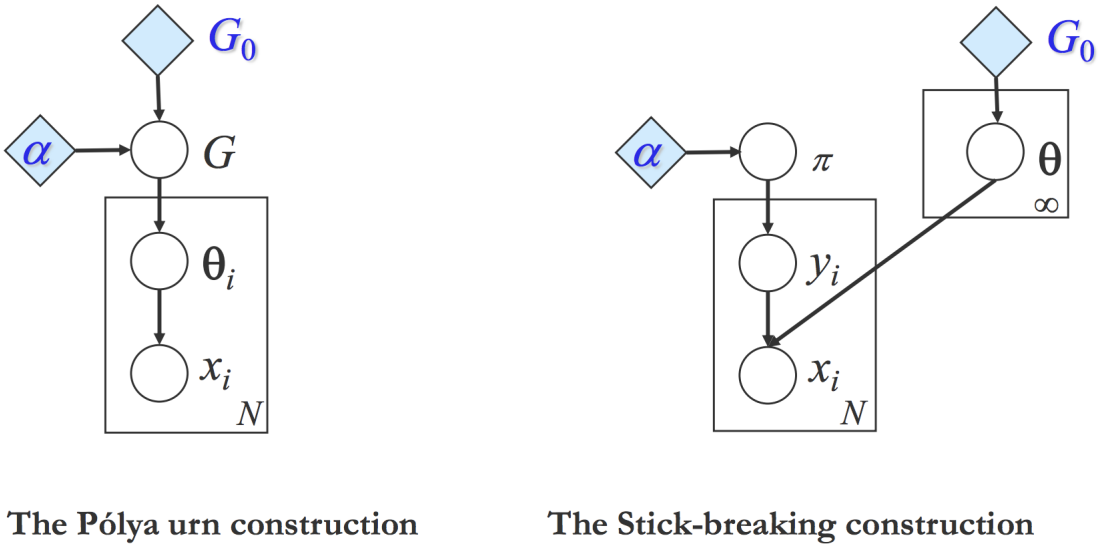

Figure 7: Two typical graphical models of Dirichlet process.

There are two different way to present the same distribution as graphical models. For Pólya urn construction, samples x_i are directly sampled from centroid θ_i, which is sampled from prior of a discrete Dirichlet Process.

In the stick-breaking construction, sample x_i is an indicator function of the cluster y_i , which is computed from multinomial distribution π.

5 Inference

The inference process is to find the potential association between points with infinite number of parameters. Here is the Dirichlet Process mixture model for this.

$$G := \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k} \sim \mathrm{DP}(\alpha, H)$$

$$\phi_n \sim G$$

$$x_n \sim f(\phi_n)$$

Collapsed sampler. The collapsed sampler is integrated to Chinese restaurant process (CRP). When sampling any data point, we can always rearrange the ordering so that it is the last data point. Denote z_n be the cluster allocation of n^{th} data point and K be the total number of instantiated clusters. We have as follow.

$$p(z_n = k | x_n, z_{-n}, \phi_{1:K} \propto \begin{cases} m_k f(x_n|\phi_k) & k \leq K \\ \alpha \int_\Omega f(x_n|\phi) H(d\phi) & k = K+1 \end{cases}$$

There are several problems with collapsed sampler. First of all, we can only sample one data point at a time, leading to the slow performance of mixing. Second, if the likelihood is not conjugate, integrating out parameter values for new features can be difficult.

6 Topic models

Topic models describe documents using a distribution over features. Each feature is a distribution over words. Each document is represented as a collection of words (usually unordered – "bag of words" assumption). The words within a document are distributed according to a document-specific mixture model. The features are shared among documents. The features learned tend to give high probability to semantically related words – "topics".

6.1 Latent Dirichlet allocation

For each topic $k = 1, \cdots, K$, we sample a distribution over words, $\beta \sim \mathrm{Dir}(\eta_1, \cdots, \eta_v)$. For each document $m = 1, \cdots, M$, we sample a distribution over topics, $\theta_m \sim \mathrm{Dir}(a_1, \cdots, a_K)$. For each word, $n = 1, \cdots, N_m$, we sample a topic $z_{nn} \sim \mathrm{Discrete}(\theta_m)$, a word $w_{mk} \sim \mathrm{Discrete}(\beta_z)$.

6.2 Constructing a topic model with infinitely many topics

Latent Dirichlet Allocation (LDA) is one of popular topic models used to classify text in document for a specific topic. In LDA, each distribution is associated with a distribution over K topics. The problem is that

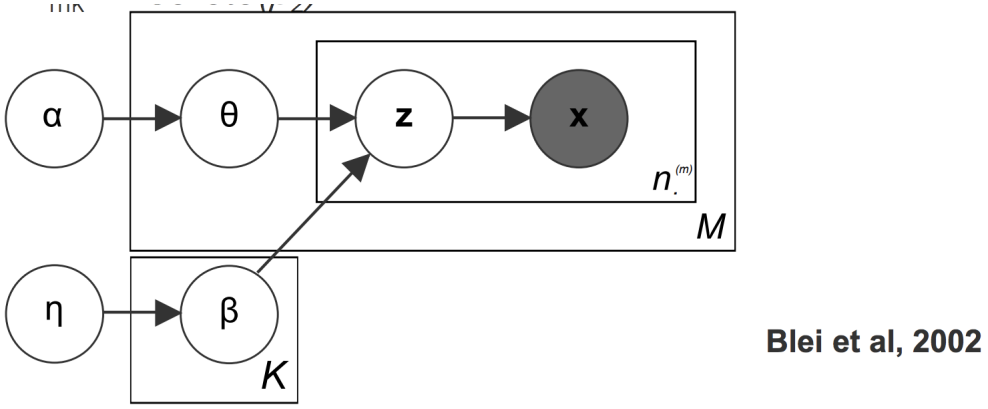

Figure 8: Topic model.

how can we choose the number of topics? So we can apply a Dirichlet process to find number of clusters which is topics. Also, we want to make sure the topics are shared between documents.

Sharing topics. In LDA, we have M independent samples from a Dirichlet distribution. The weights are different, but the topics are fixed to be the same. If we replace the Dirichlet distributions with Dirichlet processes, each atom of each Dirichlet process will pick a topic independently of the other topics. When we sample the same two topics, the probability is zero because of continuous property of the base measure. So, we need to use a discrete base feature, e.g. consider a base measure $H = \sum_{k=1}^{K} \alpha_k \delta_{\beta_k}$, we have LDA to choose number of topics. With infinite number of topics and the locations, we want an infinite, discrete base measure.

7 Hierarchical Dirichlet Process (Teh et al, 2006)

Sample the base measure from a Dirichlet process.

$$G_0 \sim \mathrm{DP}(\gamma, H)$$

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

7.1 Chinese restaurant franchise

Chinese restaurant franchise is an analog for Hierarchical Dirichlet Process. Imagine a franchise of restaurants, serving an infinitely large, global menu. Each table in each restaurant orders a single dish. We denote the following terms.

- Let n_{rt} be the number of customers in restaurant r sitting at table t.
- Let m_{rd} be the number of tables in restaurant r serving dish d.
- Let m_d be the number of tables, across all restaurants, serving dish d.

The steps in Chinese restaurant franchise are as follow.

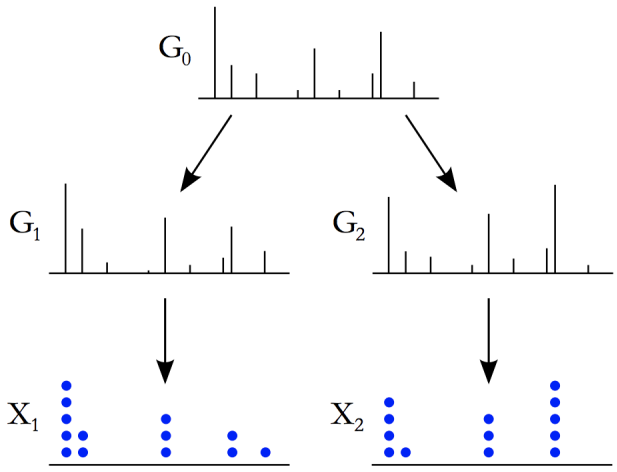

Figure 9: Hierarchical Dirichlet Process.

- The first customer enters a restaurant, and picks a table.
- The n^{th} customer enters the restaurant. He sits at an existing table with probability $\frac{m_k}{(n-1+\alpha)}$, where m_k is the number of people sat at table k. He starts a new table with probability $\frac{\alpha}{n-1+\alpha}$

Each table in each restaurant picks a dish, with probability proportional to the number of times it has been served across all restaurants.

$$p(\text{table } t \text{ chooses dish } d|\text{previous tables}) = \begin{cases} \frac{m_d}{T+\gamma} & \text{for an existing table} \\ \frac{\gamma}{T+\gamma} & \text{for a new table} \end{cases}$$

8 An infinite topic model

In the above mentioned metaphors, to relate to the problem of topic model we can consider restaurants, dishes as documents, topics respectively. Let H be a V-dimensional Dirichlet distribution, so a sample from H is a distribution over a vocabulary of V words.

$$G_0 := \sum_{k=1}^{\infty} \pi_k \delta_{\beta_k} \sim \text{DP}(\alpha, H)$$

For each document $m = 1, \cdots, M$,

- Sample a distribution over topics, $G_m \sim \text{DP}(\gamma, G_0)$.
- For each word $n = 1, \cdots, N_m$, we sample a topic $\phi_{mn} \sim \text{Discrete}(G_0)$ and a word $w_{mk} \sim \text{Discrete}(\phi_{mn}$

9 Experiment: The "right" number of topics

In Figure 10 (left), the perplexity is evaluated by mixture models range between 10 and 120. In Figure 10 (right) the posterior over the number of topics is computed by the hierarchical Dirichlet Process mixture models of the optimal LDA models.

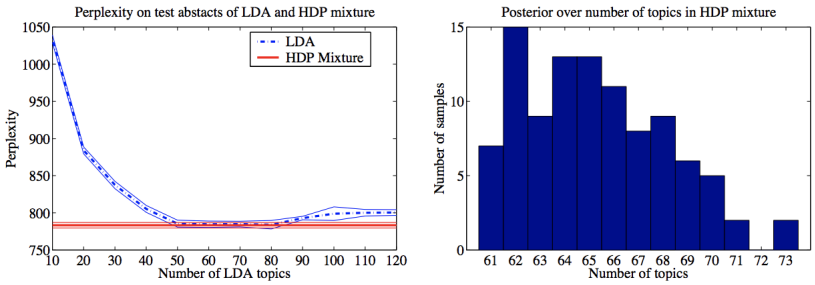

Figure 10: (Left) Comparison of latent Dirichlet allocation and the hierarchical Dirichlet process mixture.Results are averaged over 10 runs; the error bars are one standard error. (Right) Histogram of the number oftopics for the hierarchical Dirichlet process mixture over 100 posterior samples.

10 Summary

Bayesian approach only follow finite-component clustering while nonparametric Bayesian approach involves infinite number of clusters. Dirichlet Process provides a conjugate infinite-dimensional prior. DP can be considered as an infinite limit of a Dirichlet distribution. DP is very powerful in particular metaphor applications: Stick-breaking process, Polya urn process, and Chinese restaurant process.

Hierarchical DP utilize discrete base measure, allowing to reuse atoms. This leads to a powerful application for topic model: topic sharing.