10-708: Probabilistic Graphical Model **Spring 2015**

Lecture 18: Bayesian Nonparametrics: Dirichlet Processes

Lecturer: Avinava Dubey *Scribes: Chi Liu, Ji Oh Yoo, Ying Zhang*

18.1 Introduction

In parametric models, it is assumed that all the data can be represented by a fixed, finite number of parameters, thus the parameters do not grow with sample size. Typical examples for parametric model includes mixture of K Gaussians, polynomial regression, etc. In contrast, nonparametric models have number of parameters which can grow with the sample size, also the number of parameters can be a random variable. In 10702, statistical machine learning, nonparametric model deals with density estimation problems using techniques such as histograms, kernel density estimation.

In Bayesian nonparametrics, the number of parameters a priori can be infinite. For a finite data set, only a finite number of parameters will be used. The idea that unused parameters will be integrated out may seem counter intuitive, but it will become clear once we understand the Dirichlet process. For example,for clustering data of mixture of Gaussians, the parametric model will have fixed parameters if the number of mixture k is fixed. In the case of Bayesian inference, where the number of clusters K is unknown, K can be treated as a random variable. A prior is defined over an infinite dimensional model space, by selecting the number of parameters, inference can be performed. Such models have infinite capacity and an infinite number of parameters a priori;

18.1.1 Mixture model example

An example of a parametric model is the mixture K Gaussians with finite number of parameters μ_k, Σ_k:

$$p(x_1, \cdots, x_N|\pi, \{\mu_k\}, \{\Sigma_k\}) = \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)$$

In the Bayesian approach,the parameters would be assigned with a prior distribution and then integrated out:

$$p(x_1, \cdots, x_N) = \int \cdots \int \left(\prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k) \right) p(\pi) p(\mu_{1:K}) p(\Sigma_{1:K}) d\pi d\mu_{1:K} d\Sigma_{1:K}$$

To obtain better calculations, it is best to choose conjugate prior distributions to simplify posterior inference, as conjugate prior and posterior typically lies in the same family of distribution. For the Gaussian models, Gaussian and inverse Wishart distribution are the conjugate distributions for the mean and covariance. But how do we choose for the mixture weights? We use Dirichlet distribution to specify the weights.

18.2 The Dirichlet Distribution

The Dirichlet distribution is a distribution over the $(K-1)$-dimensional simplex, i.e. it is a distribution over the relative values of K components, with the constraint that the sum of all the components is 1. Thus, it is parametrized by a K-dimensional vector $(\alpha_1, \cdots, \alpha_K)$, such that $\alpha_k \geq 0 \; \forall \; k$ and $\sum_k \alpha_k > 0$. The Dirichlet distribution is given by,

$$\pi = (\pi_1, \cdots, \pi_K) \sim \text{Dirichlet}(\alpha_1, \cdots, \alpha_K) = \frac{\prod_{k=1}^K \Gamma(\alpha_k)}{\Gamma\left(\sum_{k=1}^K \alpha_k\right)} \prod_{k=1}^K \pi_k^{\alpha_k - 1}$$

If $\pi \sim \text{Dirichlet}(\alpha_1, \cdots, \alpha_K)$ then $\pi_k \geq 0$ for all k and $\sum_{k=1}^K \pi_k = 1$.

The expectation of the distribution is:

$$\mathbb{E}\left[(\pi_1, \cdots, \pi_K)\right] = \frac{(\alpha_1, \cdots, \alpha_k)}{\sum_k \alpha_k}$$

Figure 18.2 shows the density of a Dirichlet distribution for 3 components with different scaling parameters α. As the parameter values increase, the distribution becomes less peaky, i.e. more concentrated at the extremes. Also if the paramters have different values, the distribution is likely to be skewed.

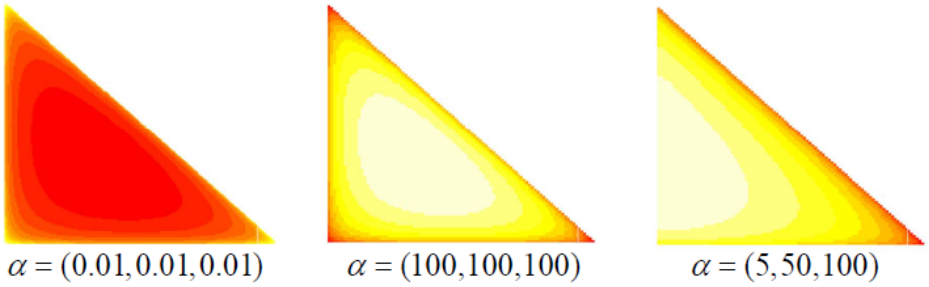

Figure 18.1: Density of the 3-component Dirichlet distribution for different parameters. Red indicates higher density.

18.2.1 Conjugacy to the multinomial distribution

Dirichlet distribution is actually the conjugate of the multinomial distribution. If $\pi \sim \text{Dirichlet}(\alpha_1, \cdots, \alpha_K)$ and x_n are *iid* samples from the distribution,the posterior distribution is

$$\begin{aligned}
p(\pi | x_1, \cdots, x_n) &\propto p(x_1, \cdots, x_n | \pi) p(\pi) \\
&= \left(\frac{\prod_{k=1}^K \Gamma(\alpha_k)}{\Gamma\left(\sum_{k=1}^K \alpha_k\right)} \prod_{k=1}^K \pi_k^{\alpha_k - 1} \right) \left(\frac{n!}{m_1! \cdots m_K!} \pi_1^{m_1} \cdots \pi_K^{m_K} \right) \\
&\propto \frac{\prod_{k=1}^K \Gamma(\alpha_k + m_k)}{\Gamma\left(\sum_{k=1}^K \alpha_k + m_k\right)} \prod_{k=1}^K \pi_k^{\alpha_k + m_k - 1} \\
&= \text{Dirichlet}(\alpha_1 + m_1, \cdots, \alpha_K + m_K)
\end{aligned}$$

where m_k are the counts of instances of $x_n = k$ in the data set.

The Dirichlet distribution is a distribution over positive vectors that sum up to 1. Each entry can be associated with a set of parameters, e.g. in finite mixture model, each entry is associated with a mean and variance. Because in Bayesians setting, these parameters should be treated as random, thus the Bayesian distribution can be deemed as a distribution over finite-dimensional distributions, where the intermediate distribution originates from parameters.

A sample from a Gaussian is a real-valued number whereas a sample from Dirichlet distribution is a probability vector. Each element of a Dirichlet distributed vector is associated with a parameter value drawn from some distribution. Sample from a Dirichlet prior is a probability distribution over parameters.

18.2.2 Properties of the Dirichlet distribution

From the relation to the Gamma distribution: if $\eta_k \sim \text{Gamma}(\alpha_k, 1)$,

$$\frac{(\eta_1, \cdots, \eta_K)}{\sum_k \eta_k} \sim \text{Dirichlet}(\alpha_1, \cdots, \alpha_K)$$

If $\eta_1 \sim \text{Gamma}(\alpha_1, 1)$,if $\eta_2 \sim \text{Gamma}(\alpha_2, 1)$,then

$$\eta_1 + \eta_2 \sim \text{Gamma}(\alpha_1 + \alpha_2, 1)$$

Therefore, if $(\pi_1, \cdots, \pi_K) \sim \text{Dirichlet}(\alpha_1, \cdots, \alpha_K)$then,Dirichlet distribution satisfies the merging rule:

$$(\pi_1 + \pi_2, \pi_3, \cdots, \pi_K) \sim \text{Dirichlet}(\alpha_1 + \alpha_2, \alpha_3, \cdots, \alpha_K)$$

This property of merging rule allows the reduction of the dimensionality of the Dirichlet distribution.

Comparatively, the Dirichlet distribution also satisfies the expansion rule, which allows the increase of the dimensionality. The Dirichlet distribution over the 1-dimensional simplex is the Beta distribution. Let $(\pi_1, \cdots, \pi_K) \sim \text{Dirichlet}(\alpha_1, \cdots, \alpha_K)$ and $\theta \sim \text{Beta}(\alpha_1 b, \alpha_1(1-b))$ for $0 < b < 1$. Then one dimensional Dirichlet distribution can be split into two dimensions,

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

More generally, if $\theta \sim \text{Dirichlet}(\alpha_1 b_1, \alpha_1 b_2, \cdots, \alpha_1 b_N)$ and $\sum_i b_i = 1$, then,

$$(\pi_1\theta_1, \cdots, \pi_1\theta_N, \pi_2, \cdots, \pi_K) \sim \text{Dirichlet}(\alpha_1 b_1, \cdots, \alpha_1 b_N, \alpha_2, \cdots, \alpha_K)$$

Finally, the Dirichlet distribution also possesses the renormalization property. That is, If $(\pi_1, \cdots, \pi_K) \sim \text{Dirichlet}(\alpha_1, \cdots, \alpha_K)$,then,

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

18.2.3 Constructing an infinite-dimensional prior

In clustering data, the number of clusters is usually not known a priori. To define a prior for the mixture weights, the solution is to set an infinite number of clusters a priori. In this way,the number of clusters is always larger than needed for a specific task. An infinite mixture model has the form,

$$p(x_n|\pi, \{\mu_k\}, \{\Sigma_k\}) = \sum_{k=1}^{\infty} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)$$

Since the aim is to obtain something like a Dirichlet prior, except it should have the property of infinite number of components, to define the appropriate prior, the following scheme is plausible. Start with a two-component Dirichlet distribution, with scaling parameter α divided equally between both components, to have form of symmetry:

$$\pi^{(2)} = (\pi_1^{(2)}, \pi_2^{(2)}) \sim \text{Dirichlet}\left(\frac{\alpha}{2}, \frac{\alpha}{2}\right)$$

Then, split off components according to the expansion rule:

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

$$\pi^{(4)} = (\theta_1^{(2)}\pi_1^{(2)}, (1-\theta_1^{(2)})\pi_1^{(2)}, \theta_2^{(2)}\pi_2^{(2)}, (1-\theta_2^{(2)})\pi_2^{(2)}) \sim \text{Dirichlet}\left(\frac{\alpha}{4}, \frac{\alpha}{4}, \frac{\alpha}{4}, \frac{\alpha}{4}\right)$$

Repeat this process until it satisfies the following condition,

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

As K goes to infinity, a prior over an infinite-dimensional space is obtained. In practice, it would not be necessary to use all of these components, finitely many components are enough to reflect the data.

18.3 The Dirichlet Process

The previous scheme, though not a derivation, motivate us to define the Dirichlet Process in a similar fashion. Let the base measure H be a distribution over some space Ω (e.g. a Gaussian distribution on the real line). Let

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

For each point in this Dirichlet distribution, perform a draw from the base measure:

$$\theta_k \sim H \text{ for } k = 1, ..., \infty$$

Then,

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

is an infinite discrete distribution over the continuous space H. We write this as a Dirichlet Process:

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

Samples from the Dirichlet Process are discrete. The point masses in the resulting distribution are named as atoms; the locations of atoms in Ω are drawn from the base measure H, and their weights are drawn from an infinite-dimensional Dirichlet distribution.

The concentration parameter α determines the distribution over atom sizes(weights); smaller values of α yields sparser distributions, with greater weights on each atom.

A Dirichlet Process is the unique distribution over probability distributions on some space Ω such that for any finite partition $A_1, \cdots, A_K$ of Ω, the total mass assigned to each partition is distributed according to the following relation

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

18.3.1 Conjugacy of the Dirichlet process

Let $A_1, \cdots, A_k$ be a partition of Ω, and let H be a measure on Ω. Denote $P(A_k)$ as the mass assigned by $G \sim \text{DP}(\alpha, H)$ to partition A_k. Then,

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

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

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

Since this must be true for all possible partitions of Ω, it is only possible if the posterior for G is given by:

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

18.3.2 Predictive distribution

The Dirichlet distribution can be modeled as a prior for mixture models, thus the Dirichlet process could be further utilized to cluster observations. A new data point can either join an existing cluster or start a new cluster. The question is what is the predictive distribution for a new data point?

Assume H is a continuous distribution on Ω and θ in Ω are parameters to model the observed data points. As the distribution is continuous, for every point θ in Ω, $H(\theta) = 0$. Once a first data point is obtained, we can start a new cluster with a sample parameter θ_1. Splitting the parameter space in two: the singleton θ_1, and everything else. Let π_1 be the atom at θ_1. The combined mass of all the other atoms is $\pi_* = 1 - \pi_1$, and the prior and posterior can be written as,

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

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

If π_1 is integrated out,then

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

In general, if we have observed n samples, and m_k is the number of times we have seen $X_i = k$ with K being the total number of observed values, then:

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

From the equation above, we note that the observed value k will be repeated with the probability propotional to m_k, while the probability of a new sample being a new cluster is propotional to α. More samples belonging to cluster k means high probability it will grow. This phenomenon is called rich-gets-richer property.

18.4 Several understandings of Dirichlet process

18.4.1 Chinese restaurant process (CRP)

The distribution over partitions can be described in terms of the following restaurant metaphor. We assume that a Chinese restaurant has infinite tables, each of which can seat infinite customers. In addition, there is only one dish on each table.

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 choose a new table. In the general case, the $n+1st$ customer either choose to join an already occupied table k with the probability propotional to the number of customers n_k already sittting there, or sit at a new table with the probability propotional to α. In this metaphor, customers are identified with the intergers 1,2,3,..and tables as clusters. When all the n customers have sat down the tables, they are partitioned into clusters, which exhibits the clustering property of the Dirichlet process above.

The number of cluster m has the following mean and variance:

$$\begin{aligned} E[m|n] &= \sum_{i=1}^{n} \frac{\alpha}{\alpha + i - 1} \\ &\simeq \alpha log(1 + \frac{n}{\alpha}) \\ V[m|n] &\simeq \alpha log(1 + \frac{n}{\alpha}) \end{aligned}$$

We note that the number of clusters is logarithmically propotional to the number of observations because of the rich-gets-richer property. The parameter α also controls the number of clusters. Larger α implies a larger number of clusters.

Another important property is that the distribution over the clustering of the first n customers does not depend on the order in which they arrived.

18.4.2 Polya urn scheme

Polya urn scheme is another metaphor to understand the predictive distribution (conditional posterior distribution). Polya urn scheme produces a sequence $x_1, x_2...$ with the following conditions:

$$x_n|x_{1:n-1} \sim \frac{\sum_{i=1}^{n-1} \delta_{x_i} + \alpha H}{n - 1 + \alpha}$$

Imaging picking balls of different colors from an urn:

1. start with no balls in the urn;

2. with the probability $\propto \alpha$, draw $x_i \sim H$, and add a ball of that color into the urn;

3. with the probability $\propto n - 1$, pick a ball at random from the urn, record x_n to its color, return the ball into the urn and place a second ball of the same color into the urn.

From the process above, we note that Polya urn scheme is similar to the Chinese restaurant process (CRP). In addition, polya urn scheme has the exchangeability property that the joint probability is invariant of the permutations of $x_1, x_2, ...$.

18.4.3 Stick breaking construction

Since draws from a DP are composed of a weighted sum of point masse, stick breaking construction provides a constructive definition of the Dirichlet process. It is simply defined as follows:

$$\beta_k \sim Beta(1,\alpha) \qquad \theta_k^* \sim H$$
$$\pi_k = \beta_k \prod_{l=1}^{k-1} 1-\beta_l \quad G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*}$$

Then $G \sim DP(\alpha, H)$. The construction of π can be understood as follows:

1. starting with a stick with legth 1, and break it at β_1, assigning π_1 to be the length of the stick we broke off;

2. recursively break the other portion to obtain π_2, π_3 and so forth.

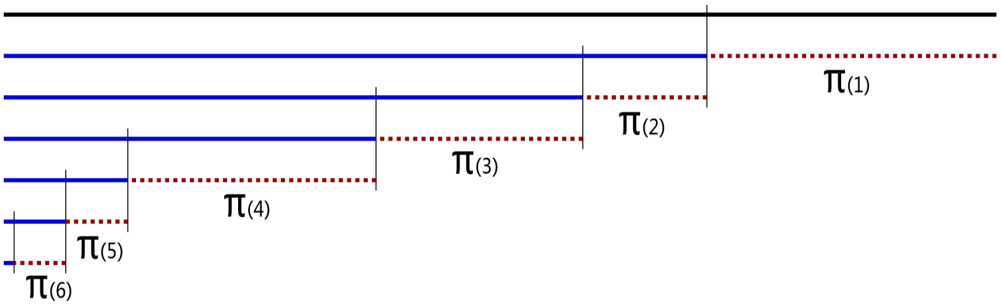

Figure 18.2: Illustration of stick breaking construction

18.5 Inference in DP mixture model

18.5.1 Collapsed Sampler

Collapsed Sampler integrates out G to get the Chinese Restaurant Process. As the observations in CRP are exchangeable, we can rearrange the ordering so that just sampled data point is the last data point. Then,

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

where z_n is the cluster allocation of the nth data point and K is the total number of instantiated clusters. But as we are only updating one datapoint at a time, the mixing can be slow. Also, if the likelihood is not conjugate, integrating out parameter values for new features can be difficult.

18.5.2 Blocked Gibbs Sampler

In Blocked Gibbs Sampler, instead of integrating out G, we instantiate G. As G is infinite-diemensional, we can approxmiate it with a truncated stick-breaking process:

$$
\begin{aligned}
G^K &= \sum_{k=1}^{K} \pi_k \delta_{\theta_k} \\
\pi_k &= b_k \prod_{j=1}^{k-1} (1 - b_j) \\
b_k &\sim Beta(1, \alpha), \text{ where } k = 1, ..., K-1 \\
b_K &= 1
\end{aligned}
\tag{18.1}
$$

Then, we can sample the cluster indicators as:

$$p(z_n = k| \text{ rest }) \propto \pi_k f(x_n|\theta_k)$$

$$b_k|\text{rest } \sim Beta\Big(1 + m_k, \alpha + \sum_{j=k+1}^{K} m_j\Big)$$

But this fixed truction introduces errors.

18.5.3 Slice Sampler

In Slice sampler, we introduce a uniform random variable u_n for each data point and sample indicator z_n using:

$$p(z_n = k| \text{ rest }) = I(\pi_k > u_n) f(x_n|\theta_k)$$

where the conditional distribution of u_n on the others is just Uniform$[0, \pi_{z_n}]$.

Conditioned on u_n and z_n, we can sample π_k according to the block Gibbs sampler. And we only need to represent a finite number of components, K, such that

$$1 - \sum_{k=1}^{K} \pi_k < \min(u_n)$$

18.6 Hierarchical Dirichlet Process

18.6.1 Infinite Topic Model

Hierarchical Dirichlet process (HDP) is a non-parametric, Bayesian approach to cluster the given data with countably infinite number of cluster idnetities using multiple Dirichlet process. Consider a topic model using LDA with infinite number of topics. Remember that topic models describe documents using a distribution over features. Here are important assumptions in the topic model:

- Each feature is a distribution over words.

- Each document is represented as a collection of words (usually bag of words).
- The words within a document are distributed according to a document-specific mixture model i.e. Each word in a document is associated with a feature.
- The features are shared between documents.
- The features learned tend to give high probability to semantically related words i.e. words with same topic.

Encoding the infinite number of topics in LDA is tricky, as each distribution is associated with a distribution over K topics, which is fixed and finite. To make the number of topics infinite, we can replace the Dirichlet distribution over the topics with a Dirichlet process, but we need to make sure what happens when we simply replace it. Even though a single topic is shared in document A and B, each atom of each Dirichlet process will pick a topic *independently* of the other topics. Thus, for choosing the base measure for our model, we need discrete (as continuous measure will give zero probability of picking the same topic twice) and infinite, and random base measure. And this can be done by sampling the base measure from a Dirichlet process as in Figure 18.6.1.

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

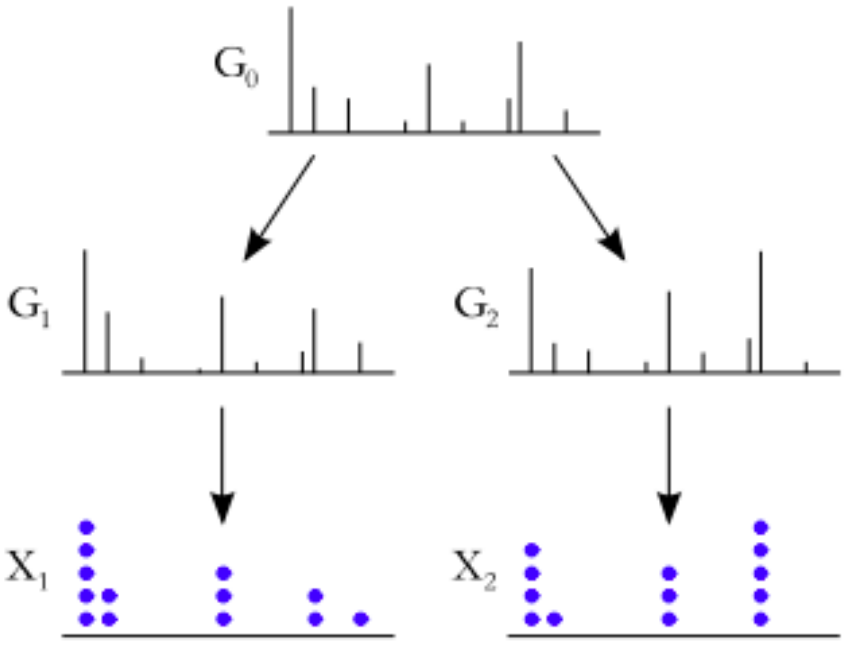

Figure 18.3: Base measure sampling in Hierarchical Dirichlet Process.

18.6.2 Chinese Restaurant Franchise

HDP can be interpreted as a metaphor in Chinese Restaurant Franchise, an extension of Chinese Restaurant Process. Instead of a single Chinese restaurant, there is a franchise of restaurants, serving an infinitely many but global menu, shared in the franchise. Each table in each restaurant orders a single dish. Let n_{rt} be the number of customers in restaurant r sitting at table t, and m_{rd} be the number of tables in restaurant r serving dish d, and m_d be the number of tables across all restaruants serving dish d. Custermers enter the restaurants and sit at tables as in the CRP: nth customer sits at an existing table with probability $\frac{m_k}{n-1+\alpha}$, where m_k is the number of peaple in table k, and starts a new table with probability $\frac{\alpha}{n-1+\alpha}$. And, each table in each restaurant picks a dish with probability proportional to the number of times that it has been searved across *all* restaurants, i.e.,

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

While there is no right answer for the number of topics to choose in our problem domain, but the Figure 18.6.2 shows that we can use HDP to go through a wide range of number of topics and find the appropriate number of topics for our data based on the perplexity measure.

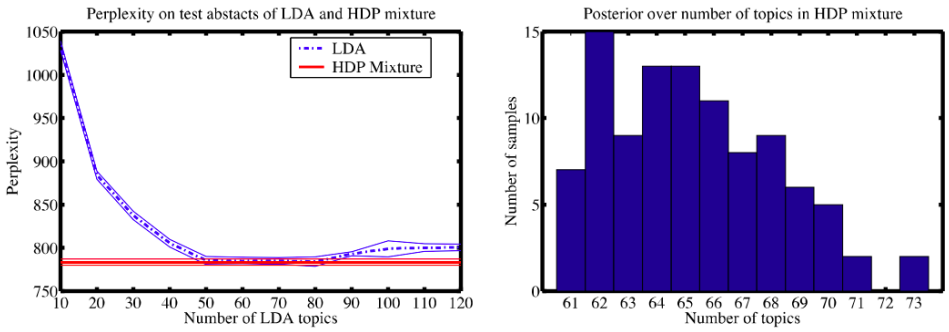

Figure 18.4: Perplexity and posterior samples to find the "right" number of topics.