Research Proposal

Shaobo Dang

Contents

1 Introduction

Data Mining has been an heating area in Machine Learning. It is especially important in the information age since the seemingly meaningless data often containing potential opportunity, as can be seen from large Internet companies such as Google, Facebook and Amazon [3]on their investment on large data processing project.

The primary task in data mining is to find the underlying distribution that generates the observed data to guide the computer to process the similar data automatically, saving the cost of both time and expense. As the data often consists of multi-modal distribution, modeling the distribution as a mixture of simpler distribution is found to be a well approximation. In parametric settings, the number of sub-distribution is fixed value which may cause the model fail to funtion as the complexity is not synchronized that of the data[**?**, 2, **?**]. Instead, Bayesian Nonparametric model, especially the Dirichlet Process employs a infinite countably number of components to handle the analysis by placing a prior distribution for mixing distribution. The nature of this set of problem bypasses the need to "determine" the number of comopnents in a finite mixture model[?].

Though a very promising and relatively simpler way to model, it is often difficult to process the following inference which is the main part of analysis. Except for placing conjugate prior, the desirable posterior distribution of most non-conjugate prior model is hard to obtain. Use of Dirichlet process mixture models has been computationally feasible with the development of Markov Chian methods for sampling from the posterior distribution of the parameters of the component distritution given observed data[?]. Though conjugate prior is preferable for it allows to introduce Gibbs sampling methods, it narrows the generalization of Dirichlet Process[?] to address more complicated problem. In most cases, the non-conjugate prior is placed, thus it should be researched on how to use MCMC under such circumstances.

2 Aim Of Research

Dirichlet Process mixture models(DPMMs) are widely used in machine learning. The theory behind it has extended *finite* mixture models to *infinite*, including automatic model selection in clusterin gproblems. The learning procedure, under any circumstances, could be simplified to a posterior inference. Variational Inference and Markov Chain Monte Carlo(MCMC) techniques are most commonly used to address the inference in two different ways, i.e. the approximation inference and sample-based inference.

Variational inference are often used for their parallelization and speed, but lack the limiting guarantees of MCMC while MCMC faces the difficulty of slow convergence. What we want to do is to develop a sampler, which is basically an updating version of MCMC that will:

- perseves limitin gguarantees
- improve the convergence rate
- can be parallelized to accommodate large datasets
- could be applicable to variety of DPMMs(conjugate and non-conjugate)

3 Literature Review

This section overview of the basic concept of Dirichlet Process, including the representation, the main application-Dirichlet Process Mixture Model and the inference methods.

3.1 Background

Large Data sets are often heterogeneous, cased by amalgams from underlying sub-populations[4]. When doing analysis on such data sets, it often involves grouping the data to make the original data more heterogeneous. Traditionally, parametric models using a fixed and finite number of parameters could be used, but they suffer from over- or under-fitting of data when there is a misfit between the complexity of the model and the amount of data available. Thus, model complexity selection is often an important issue in parametric modeling. But whether we use cross validation or marginal probabilities as the basis for selection, model selection is an operation that is fraught with difficulties[**?**], especially when the data set grows larger and larger.

The Bayesian Nonparametric approach is an alternative to parametric modeling and selection. Under this scheme, the model comes with an unbounded complexity, and under-fitting and over-fitting are mitigated[2]. Typically, we assume that the observed data set $x_1, x_2, \cdots, x_i$ are i.i.d(Independent Identical Distributed) sampled from some underlying unknown distribution $\mathbb{F}$. In Bayesian approach, a prior is placed over $\mathbb{F}$ then he posterior over $\mathbb{F}$ given data is computed. This prior over distributions is given by a parametric family. But constraining distributions to lie within parametric families limits the scope and type of inferences that can be made. Instead, the nonparametric approach used a prior over distributions with wide support, typically the support being the space of all distributions. The Dirichlet Process is currently one of the most popular Bayesian Nonparametric Models. It was first formalized in [**?**] for general Bayesian statistical modeling.,as a prior over distributions with wide support yet tractable posteriors. But the Dirichlet Process is limited by the fact that draws from it are often discrete distributions, and tractable posterior is hard to obtian when it generalized to more general non-conjugate priors until MCMC techniques became available in the area.

3.2 Dirichlet Process

3.2.1 Dirichlet Distribution

A Dirichlet distribution is defined on the $(k - \mathit{1})$-dimensional probability simplex, whic is a surface in $\mathbb{R}^k$ denoted by Δ_k and defined to be the set of vectors whose k components are non-negative and sum to 1, that is

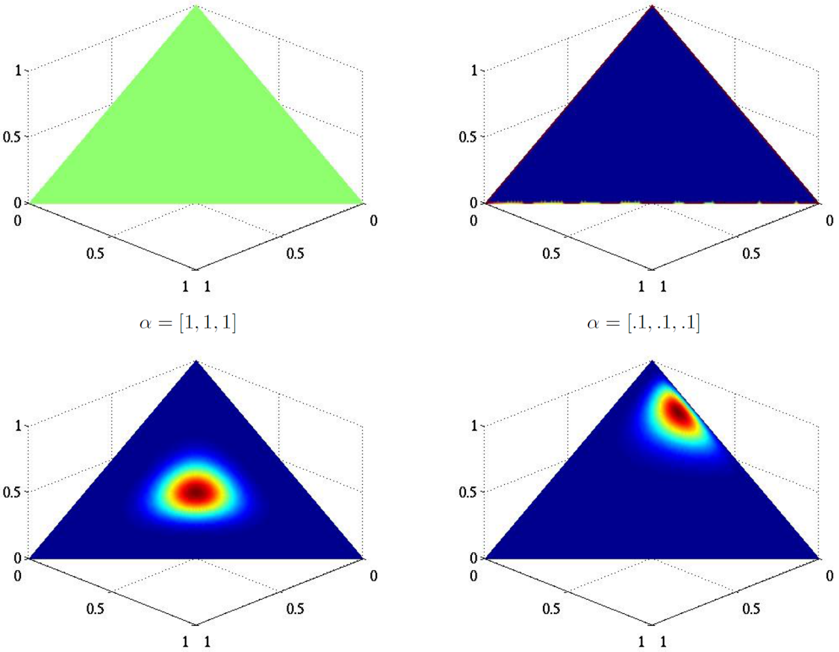
$$\Delta_k = \{q \in \mathbb{R}^k | \sum\nolimits_{i=1}^{k} q_i = 1, q_i \geq 0 \; for \; i = 1, 2, \cdots, k\} \tag{1}$$

While the set Δ_k lies in a k-dimensional space, Δ_k is itself a $(k-1)$-dimensional object. Each point q in the simplex can be thought of as a probability mass function(pmf) in its own right. The Dirichlet distribution can be thought of as a probability distribution over the $k - 1$-dimensional probability simplex Δ_k; that is, as a distribution over pmfs of length k.

Dirichlet Distribution: Let $Q = [Q_1, Q_2, \cdots, Q_k]$ be a random pmf, i.e. $Q_i \geq 0$ for $i = 1, 2, \cdots, k$ and $\sum_{i=1} kQ_i = 1$. In addition, suppose that $\alpha = [\alpha_1, \alpha_2, \cdots, \alpha_k]$ with $\alpha_i > 0$ for each i, and let $\alpha_0 = \sum_{i=1^k \alpha_i}$. Then Q is said to have a Dirichlet distribution with parameter α, which we denote by $Q \sim Dir(\alpha)$, if it has $f(a; \alpha) = 0$ then q is not a pmf, else if q is a pmf then

$$f(q; \alpha) = \frac{\Gamma(\alpha_0)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} q_i^{\alpha_i - 1} \tag{2}$$

where $\Gamma(s)$ denotes the gamma function, a generalization of the factorial function, for $s > 0, \Gamma(s+1) = s\Gamma(s)$, and for positive integers n, $\Gamma(n) = (n-1)!, \Gamma(1) = 1$. Fig. 1 shows the plots of the density fo the Dirichlet distribution over the two-dimensional probability simplex for $k = 3$ events lying in



$\alpha = [1, 1, 1]$

$\alpha = [.1, .1, .1]$

$\alpha = [10, 10, 10]$

$\alpha = [2, 5, 15]$

Figure 1: Dirichlet distribution density plots(blue=low, red=high) over the probability simplex in $\mathbb{R}^3$ for various values of the parameter α

three-dimensional Euclidean space over a variety of parameter vector α. When $\alpha = [1, 1, 1]$, the Dirichlet distribution reduces to uniform distribution over the simplex. When all the components of α satisfy $\alpha_i > 1$, the density is monomodal with its mode somewhere in the interior of the simplex. And when $\alpha_i < 1$, the plots has sharp peaks at the vertices of the simplex. Another important feature should be mentioned, as can be seen from the definition of Dirichlet distribution, is that the support is open thus does not include the vertices or edge of the simplex, which in reality means that no component of a pmf drawn from a Dirichlet will ever be zero. Fig. 2 shows plots of samples drawn i.i.d from different Dirichlet distributions.

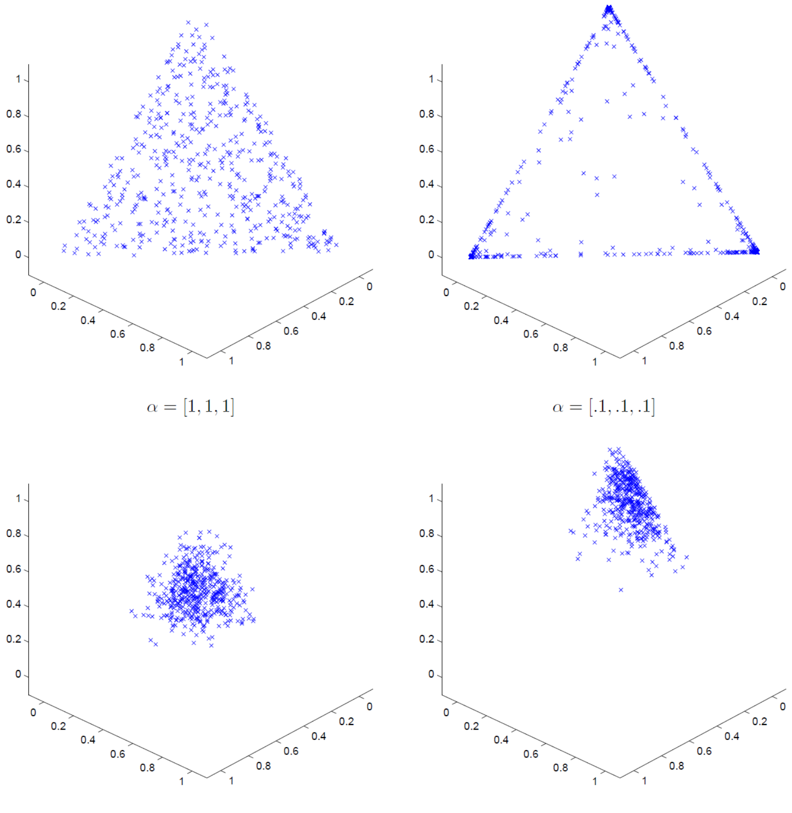
3.2.2 Dirichlet Process

A Dirichlet Process is a distribution over probability distributions. Suppose that G is a probability distribution over a measurable space Θ, then G is a probability distribution over Θ and a DP is a distribution over all such distributions. A DP is parameterized b a concentration parameter α and a base measure(base distribution) H. So, the formal definition of a DP would be

$$G \sim DP(\alpha, H) \tag{3}$$

it means for a finite set of measurable paritions $A_1 \cup \cdots \cup A_k = \Theta$,

$$(G(A_1) \cdots, G(A_k)) \sim Dir(\alpha H(A_1), \cdots, \alpha H(A_k)) \tag{4}$$



$\alpha = [1, 1, 1]$

$\alpha = [.1, .1, .1]$

$\alpha = [10, 10, 10]$

$\alpha = [2, 5, 15]$

Figure 2: Plots of sample pmfs drawn from Dirichlet distributions over the probability simplex in $\mathbb{R}^3$ *for various values of parameter* α

which means the probabilities that G assigns to any finite partition of Θ follow a Dirichlet distribution with parameters $(\alpha H(A_1), \cdots, \alpha H(A_k))$.

Since G is a random distribution we can inturn draw samples from G itself. Let $\theta_1, \theta_2, \cdots, \theta_n$ be a sequence of independent draws from G. What is the posterior distribution of G given observed values of $\theta_1, \theta_2, \cdots, \theta_n$. In other words, what is the posterior predictive distribution for a new item $p(\theta_{N+1}|\theta_1, \theta_2, \cdots, \theta_n) = \int p(\theta_{N+1}|G)p(G|\theta_1, \theta_2, \cdots, \theta_n)dG$? It is shown in ... that

$$(G(A_1), \cdots, G(A_k))|(\theta_1, \theta_2, \cdots, \theta_n) \sim Dir((\alpha H(A_1) + n_1, \cdots, \alpha H(A_k) + n_k) \tag{5}$$

where $n_k = \#\{i : \theta_i \in A_k\}$ is the number of observed values in A_k. Rewriting the posterior DP, we have:

$$G|(\theta_1, \theta_2, \cdots, \theta_n) \sim DP(\alpha + n, \frac{\alpha}{\alpha + n} H + \frac{n}{\alpha + n} \frac{\sum_{i=1}^n \delta(\theta_i)}{n}) \tag{6}$$

and for the posterior distribution,

$$\begin{aligned} p(\theta_{N+1}|\theta_1, \theta_2, \cdots, \theta_n) &= E[G(A)|\theta_1, \theta_2, \cdots, \theta_n] \\ &= \frac{1}{\alpha + n} \left(\alpha H + \sum_{i=1}^n \delta_{\theta_i} \right) \end{aligned} \tag{7}$$

Therefore the posterior base distribution given $\theta_1, \theta_2, \cdots, \theta_n$ is also the predictive distribution of θ_{n+1}. The sequence of predictive distribution 7 for $\theta_1, \theta_2, \cdots, \theta_n$ is very important and different interpretations on it indicate different properties of the result. To better understand it, we will introduce the most famous three representations of 7: **Pólya urn, Chinese Restaurant Process, Stick-breaking Process**.

- **Pólya urn**
  In this analogy, suppose we are drawing colored balls from an urn, θ_i represents the color of the i-th ball drawn . For each ball drawn, we place it back and add another one in the same color into the urn. In the beginning, we pick a color drawn from H, paint a ball with that color and drop it into the urn. In the following nth step, we will either, pick a new color with probability$\frac{\alpha}{\alpha+n}$, or with probability $\frac{n}{\alpha+n}$ pick a random ball out of the urn. This process induces a "rich get richer" property on the frequencies of colors inside the urn. Also, it should be noticed that the predictive distribution has point masses located at the previous draws $\theta_1, \theta_2, \cdots, \theta_n$. Thus the distribution G itself has point masses. When sample size grows larger, the value of any draw will be repeated by another draw, implying that G is composed only of a weighted sum of point masses and thus it is a discrete distribution.

- **Chinese Restaurant Process**
  Another representation of Dirichlet Process-Chinese Restaurant Process(CRP) implies a clustering property. Equation 7 could be rewrite as :

$$p(\theta_{N+1}|\theta_1, \theta_2, \cdots, \theta_n) = \frac{1}{\alpha + n} \left(\alpha H + \sum_{i=1}^n n_k \delta_{\theta_i^\star} \right) \tag{8}$$

where $\theta_i^\star$ is the unique values among $\theta_1, \theta_2, \cdots, \theta_n$, and n_k is the number of repeats of $\theta_i^\star$. $\theta_i^\star$ will be repeated by θ_{n+1} with probability proportional to n_k. The larger n_k is, the hight the probability that it will grow. This is similar to the "rich get richer" scheme in **Pólya urn**, where large clusters grow larger. We can see that the unique values of $\theta_1, \theta_2, \cdots, \theta_n$ induce a partitioning of the set $[n] = \{1, 2, \cdots, n\}$ into clusters such that within some cluster k, the θ_i's take on the same value $\theta_k^\star$. The random partion encapsulates all the properties of the DP. If we invert the generative process, we can reconstruct the joint distribution over $\theta_1, \theta_2, \cdots, \theta_n$ by first drawin ga random partion on $[n]$, then each cluster $k$in the partition draw a $\theta_k^\star \sim H$, and finally assign $\theta_i = \theta_k^\star$ for each i in cluster k.

The distribution over partitions is called the Chinese Restaurant Process(CRP) in which we have a Chinese restaurant with infinite tables, each of which can seat an infinite number of customers. The first customer enters the restaurant and sits at the first table. And in the following, when the $n+1$th customer comes, he will either joins a tale k with probability propotional to the number n_k of people already sitting there or sts at a new table with probability propotional to α. The CRP define a distribution over partitions of $[n]$ and a distribution over permutations of $[n]$.

- **Stick-breaking Construction** The third representation of DP is very intuitive and the most widely used one to generate a sample from it. By knowing the fact that draws from a DP consists actually of a weighted sum of point masses, ...provides a constructive and straightforward definition of the DP. It simply following the flowwin steps.

$$\beta_k \sim Beta(1, \alpha) \qquad \theta_k^\star \sim H, \tag{9}$$

$$\pi_k = \beta_k \prod_{l=1}^{k-1} (1 - \beta_k) \qquad G = \sum_{k=1}^{\infty} \pi_k \delta_k^\star, \tag{10}$$

$$\tag{11}$$

Then we have $G \sim DP(\alpha, H)$. The construction of π could be interpreted as breaking a stick of length 1 step by step. First break it at β_1, assigning π_1 to be the length of segment we just got. Then recursively breaking the remaining portion to obtain $\pi_2, \pi_3, \cdots, \pi_{n-1}$. THis is a very straightforward and simple procedure.

3.2.3 Dirichlet Process Mixture Models

In this section, we will briefly describe the Dirichlet Process Mixture Model Most machine learning problem is targeted to learn a set of parameters describing the model from training data set. This sort of learning is often evaluated in two terms, with the first one being how well the model fits the data, expressed as accuracy or squared error, and the second one being a complexity penalty(favoring simpler models)[**?**, 1], also referred to as Ocam's Razor. In practical problem, the model complexity is hard to evaluate. Improper model complexity will lead to over-fitting or under-fi tting[2, 4], which will then affect the model generalization, i.e to be applied to practical use. Although through rigid training, desirable models could be trained but this is basically a trial and error process[**?**]. That is the movivation of discovering adaptive model complexity selection methods, among which Bayesian Nonparametric Method is the most widely used.

DP, as the most widely used methods among Bayesian Nonparametric Models, has found applications in both statistics and machine learning, including Bayesian model validation, density estimation and clustering via mixture models, among which the last one is the most salient when talking about DP.

Model validation is to evaluate whether a model gives a good fit to observed data. Under Bayesian approach, we would usually compute the marginal probability of the data under the model and compare the marginal probability to that of other candidate model. The one with the highest probability will be chosen as the best fitting to the observed data. Here arises an issue that how to choose the models to be compared. Usually, a set of candidate models as large as possible would be desirable. But it would be easier if we re-think this in a Bayesian Nonparametric way, i.e. to use the space of all possible distribution as our comparison class, with a prior over distributions. The DP is always the first priority for its similar nature to this problem. The approach is to use the given parametric model as the base distribution of the DP, with DP serving as a nonparametric relaxation around this parametric model. If the parametric model performs as well or better than the DP relaxed model, we are convinced that the model is valid.

For density estimation, the aim is to modeling the latent density from which the observed data is drawn. To avoid the poor performance caused by the limitation in parametric model, we again employ a Nonparametric prior over all densities. If we drawn samples from a DP, which is distribution over distribution, we will obtain a random distribution which is discrete,thus has no densities. The solution is to smooth out draws form the DP with a kernel. Let $G \sim DP(\alpha, H)$ and $f(x|\theta)$ be a family of densities indexed by θ. Then

$$p(x) = \int f(x|\theta)G(\theta)d\theta \tag{12}$$

The most common application of the Dirichlet process is to cluster data using mixture models. The traditional finite mixture model assumes that there are K clusters, each associated with a parameter θ_k. Each observation y_n is assumed to be generated by first choosing a sluster c_n according to $P(c_n)$ and then generating the observation from its corresponding observation destribution parameterized by θ_{c_n}. Finite model can accomodate many kinds of data by changing the data generating distribution.

Bayesian mixture models further contain a prior over the mixing distributions. The nature of Dirichlet process will translate the mixing model to a countably infinite number of components. we model a set of observations $\{x_1, \cdots, x_n\}$ using a set of latent parameters $\{\theta_1, \cdots, \theta_n\}$, each θ_i is drawn I.I.d from G, while each x_i has distribution $F(\theta_i)$ parameterized by θ_i:

$$\begin{aligned} x_i|\theta_i &\sim F(\theta_i) \quad (13) \\ \theta_i|G &\sim G \quad (14) \\ G|\alpha, H &\sim DP(\alpha, H) \quad (15) \end{aligned}$$

Because G is discrete, multiple θ_i's can take on the same value simultaneously, and the model above can be seen as a mixture model, where x_i's with the same value of θ_i belong to the same cluster. The mixture perspective can be made more in agreement with the usual representation of mixture models using the stick-breaking construction. Let z_i be a cluster assignment variable, which takes on value k with probability π_k. Then equation 3.2.3 could be expressed as:

$$\pi|\alpha \sim GEM(\alpha) \qquad \theta^\star_k|H \sim H \tag{16}$$

$$z_i|\pi \sim Mult(\pi) \qquad x_i|z_i, \{\theta^\star_k\} \sim F(\theta^\star_{z_i}) \tag{17}$$

with $G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta^\star_k}$, $\theta_i = \theta^\star_{z_i}$, π being the mixing proportion, $\theta^\star_k$ being the cluster parameters, $F(\theta^\star_k)$ being the distribution over data in cluster k and H the prior over cluster parameters.

From above expression, it is seen that DP mixture model is an infinite mixture model-a mixture model with a countably infinite number of clusters. Different from finite mixture model using a fixed number of clusters, π_k's decrease exponentially quickly, and only a small number of clusters will be used to model the data a priori. In the DP mixture model, the actual number of clusters used to model data is not fixed, and can be automatically inferred from data using the usual Bayesian posterior inference framework. The equivalent operation for finite mixture models would be model averaging or model selection for the appropriate number of components.

3.3 Inference

We have described three different representations of Dirichlet Process, and all of them posit a generative probabilistic process of a collection of observed (and future) data that includes hidden structure. And the observed data is analyzed by examining the posterior distribution of the hidden structure given the observations; this gives us a distribution over which latent structure likely generated our data.

Thus, the basic computational problem in DP modeling (as in most of Bayesian statistics) is computing the posterior[**?**]. Unfortunately, for many models posterior is not available in closed form. But there are several ways to approximate it. The most widely used posterior inference methods in Bayesian nonparametric models are Markov Chain Monte Carlo (MCMC) methods. The idea MCMC methods is to define a Markov chain on the hidden variables that has the posterior as its equilibrium distribution [**?**]. By drawing samples from this Markov chain, one eventually obtains samples from the posterior. A simple form of MCMC sampling is Gibbs sampling, where the Markov chain is constructed by considering the conditional distribution of each hidden variable given the others and the observations. CRP mixtures are particularly amenable to Gibbs sampling due to the exchangability property, and each observation can be considered to be the "last" one and the distribution of Equation 2 can be used as one term of the conditional distribution. [**?**] provides an excellent survey of Gibbs sampling and other MCMC algorithms for inference in CRP mixture models.

An alternative approach to approximating the posterior is variational inference [**?**]. This approach is based on the idea of approximating the posterior with a simpler family of distributions and searching for the member of that family that is closest to it. Although variational methods are not guaranteed to recover the true posterior (unless it belongs to the simple family of distributions), they are typically faster than MCMC [**?**]and convergence assessment is straightforward. These methods have been applied to CRP mixture models.

Both MCMC and variational strategies for posterior inference provide a data-directed mechanism for simultaneously searching the space of models and

nding optimal parameters. This is convenient mixture modeling because we avoid needing to

t models for each candidate number of components. It is essential in more complex settings where the algorithm searches over a space that is diffcult to effciently enumerate and explore.

3.4 Summary

In this section, we briefly reviewed the basics in Dirichlet process, including their three different representations-Pólya urn, Chinese Restaurant Process and Stick-breaking Process, and each one shows the different properties of Dirichlet Process. Also, we delve into the inference strategies of Dirichlet Process, including Markov Chain Monte Carlo and Variational Inference.

Generally, Dirichlet Process, as the most representative modeling in Bayesian Nonparametric analysis, are an emerging trend for buildin gflexible models whose strcture grows and adapts to data. But the easy and flexible modeling procedure does not guaranteen a relaxed inference procedure especially when the data set grows larger. Thus, the poposal is mainly focus on the inference procedure, aiming at developing a more efficient mechenism to inference.

4 Methodology

This section will describe the methodology to fulfill the research target. As mentioned in the 2, the primary goal of this research is to introduce the bayesian learning method into active learning framework to handling the learning problem with variant size of data set and enable the model flexible enough to adapt the model complexity.

Fig3 is the main flowchart. briefing the detailed process, and in each main stage the major work should be:

1. **Literature Review**

   This part should not be restricted just to the two main mentioned topics, instead during this procedure, a comprehensive review on the machine learning literature should be carried out. Besides the classical learning theories, extensive reading on analysis of their characteristics and application should also be focused on. The goal is to gain a general understanding of the theory framework, relation to other learning theory, their history, development and current research.

2. **Develop a revised MCMC inference for DP**

   MCMC methods, although guaranteed to converge to the posterior with enough samples, have two drawbacks, first being that the samplers must be run for many iterations before convergence and second it is diffcult to assess convergence. And for large data sets, the computation may be more

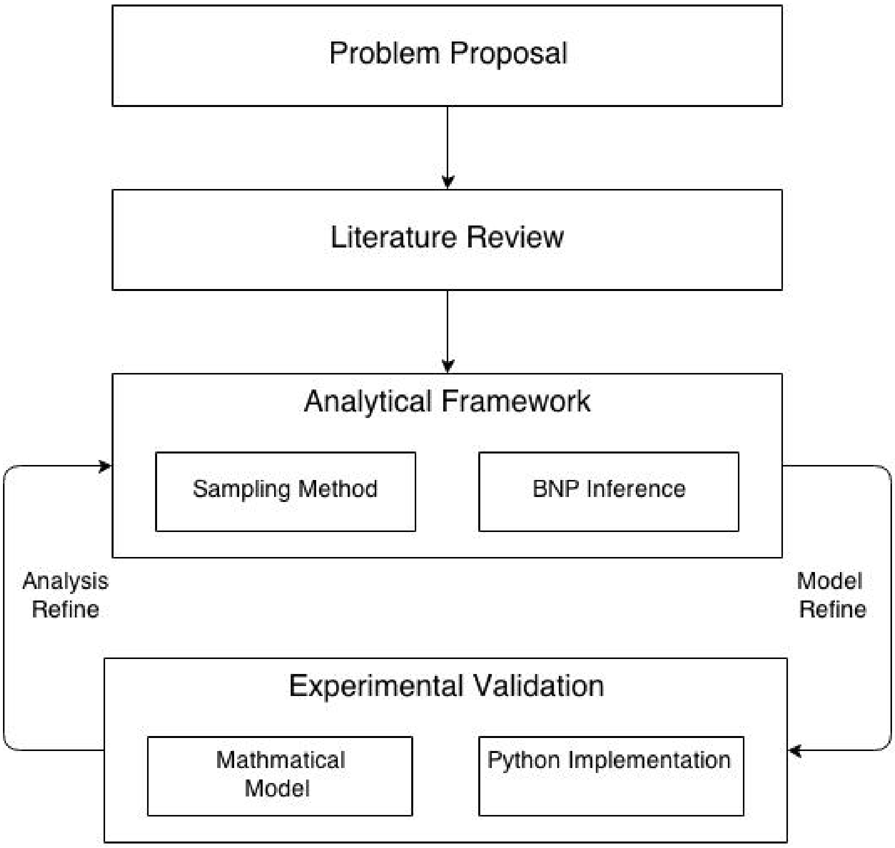


Figure 3: Research Flowchart

expensive. This lies in the nature of MCMC that it is computationally expensive and not scalable[**?**]. This is conflated with the observation that the Markov nature of such inference techniques necessarily require the computations to be sequential. We plan to try a novel approach that reparameterize the random measure by exploits invariance properties of the Dirichlet process and contidional independence is introduced between sets of atoms. This will enables transition operators on different parts of the posterior to be simulated in parallel, with minimum communication. It also should be mentioned that we should try not to alter the prior or require an approximating target distribution.

3. **Research on Bayesian Non-parametric Learning scheme**

   As the second main research topic, it is also very important to develop the overall learning scheme. BNP model, although very flexible and covering multiple learning tasks. The main focus on this area is how to overcome the low convergency rate in traditional BNP learning. Derived from traditional nonparametric model, Bayesian Non-parametric method is also constrained by the common shortcoming of nonparametric family, which is the high cost on resources[**?**, 5]. Every sample should be kept for the model updating, which will be a disaster as dimension of data set grows. In this research, we aim to bring active learning procedure into the BNP model. During the model updating, instead of storing all training samples, only those is most representative and informative will be kept. The key problem is how to make most of the selected samples in BNP learning process as all the parameters will be updated by using them.

4. **Implementation and Experiment**

   This steps include implementing the proposed framework and test the algorithm on different data set regarding the specific problem this framework will be used to evaluate its performance. The validation should contain two aspects, both on algorithm performance and computation efficiency.

   Most of the programming will be in Python, as it is free, with many open-sourced mathimatical toolbox, and is easy to use. And the programming will be run on regular PC.

   The result will be compared to the state-of-art methods,in the area of the problem to be solved, such as classification, regression or density estimation, on the basis of both the efficiency and accuracy. The accuracy is to test how our proposed method performs regarding to the specific target of the problem. And the efficiency is to check whether this framework will decrease the rely on calculation resource, both time and space. The performance will be the feedback for analytical refinement.

5 Research Plan

In this section, a research plan is listed to achieve the previous proposed objective as well as some resources required for the research.

5.1 Research Timetable

In this section, a timetable of the research plan will be described. Basically, this research is planned to span the first one and a half year of my research.

Semester-1 2014

- COMP9417 Data Mining and Machine Leanring course
- Basic Bayesian theory learning
- Review Bayesian Non-parametric Theory
- Implenmentation of BNP model

Semester-2 2014

- Literature Review
- GSOE9400 Research Course
- Review MCMC Theory
- Review Bayesian Nonparametric Inference methods
- Implementing the general coding framework and test on public dataset

Semester-1 2015

- Develop the proposed sampling methods and implementation
- Experiment of the proopsed learning theory and testing
- Paper writing and publishing

5.2 Resource Required

In addition to the usual work space and equipment resources, some practical data set to test the performace is required. This could be supplied by NICTA as part of this research is related to one of its research project.

The only potential risk is that during the initial stage, many mathmatical hypothesis could be placed on the model to facilitate the theoretical inference, and the performance of the framework could be unsatisfied. But this could be address by modelling refine process. Finally, this framework will be upgraded for more complex problem thus the two main part should be loose-coupled to each other.

References

[1] M. D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of the american statistical association*, 90(430):577–588, 1995.

[2] S. J. Gershman and D. M. Blei. A tutorial on bayesian nonparametric models. *Journal of Mathematical Psychology*, 56(1):1–12, 2012.

[3] A. McAfee, E. Brynjolfsson, T. H. Davenport, D. Patil, and D. Barton. Big data. *The management revolution. Harvard Bus Rev*, 90(10):61–67, 2012.

[4] P. Müller and F. A. Quintana. Nonparametric bayesian data analysis. *Statistical science*, pages 95–110, 2004.

[5] V. N. Vapnik and V. Vapnik. *Statistical learning theory*, volume 1. Wiley New York, 1998.