**Abstract**

By unveiling the latent variables that can generate observed properties of objects is one of the most fundamental issues in unsupervised learning. However, one of the crucial problems of unsupervised learning algorithms is to detect the latent structure. In K-means problem for example, we need to determine the number of clusters. One classic way is by performing model selection, while the other way is to use a Bayesian nonparametric method. One important method of Bayesian nonparametric method is the Indian buffet process (IBP), which is a stochastic process that provides a probability distribution over equivalence classes of binary matrices of bounded rows and potentially infinite columns. In this report, we first implement the Indian buffet process by Gibbs sampling and Metropolis-Hasting algorithm in python. For improvement in efficiency, we perform matrix calculation optimization and utilize JIT, Cython and parallel programming decrease the computation duration. Finally, we use Code Test for checking the validity and effectiveness of our acceleration and optimization.

The Indian Buffet process is very interesting in its approach to model objects using Bayesian Non-Parametrics, assuming the true dimensionality is unbounded. This concept is new to me and very intriguing at the same time. Statistical models exist, that provide latent structure in probabilistic models, but the critical question is the unknown dimesnionality of the representation, i.e. how many features are required to express the latent structure. Bayesian Non-Parametrics is an answer to this question. One way is to use the Chinese Restaurant Process which assigns each object to only one feature/class of the infinite array of features. The Indian Buffet Process extends this problem through its potential to assign an object (customer) to multiple features (dishes) [1]. As an example, we would prefer characterizing a person as married, atheist, female and democrat rather than simply assigning the person to one class.

We define a probability distribution over equivalence classes of binary matrices with a finite number of rows and an unbounded number of columns. This distribution is suitable for use as a prior in probabilistic models that represent objects using a potentially infinite array of features. We identify a simple generative process that results in the same distribution over equivalence classes, which we call the Indian buffet process. We illustrate the use of this distribution as a prior in an infinite latent fea- ture model, deriving a Markov chain Monte Carlo algorithm for inference in this model and applying the algorithm to an image dataset.

The Indian Buffet Process is a fundamental algorithm in Bayesian nonparametrics, and it models an infinite number of features to reveal the latent structure of the data. In this report, I implemented the algorithm in the paper ”Infinite Latent Feature Models and the Indian Buffet Process” [9], performed code testing to ensure correctness, optimized the Python code, and compared my work with some other existing algorithms and/or packages. A simulated image dataset is used for generating the posterior sampling of the linear-Gaussian binary latent feature model with the Indian Buffet Process as the prior.

Modeling objects is an interesting problem and one of the methods is unsupervised learning. The goal of unsupervised learning is to identify the underlying features that make up the object. Simplest way is to classify the objects into different subsets based on these features. It works well when the objects can be grouped into reasonable number of small groups [1]. When we cannot easily classify them into similar groups due to large range of features or when the objects might be better represented on their own with set of features that they have. For example, it might be more useful to describe a country as from Asia, predominantly Buddhist, tropical climate and low GDP instead of trying to classify it into some subset of countries [2]. When taking this approach, we need to consider the total features needed to represent the object. This is often treated as a model selection problem and some dimensionality is chosen based upon some measures [1]. An alternative used in nonparametric Bayesian models is to keep the dimensionality unbounded [3]. Chinese Restaurant Process uses unbounded dimensionality of classes and assigns objects to classes from infinite choice. Indian Buffet Process extends this to have unbounded dimensionality of latent features and assigns a subset of infinite features to the objects.

Adoption of the IBP as a prior distribution in Bayesian modeling has led to successful Bayesian nonparametric models of human cognition including models of latent feature learning (Austerweil & Griffiths, 2009), causal induction (Wood, Griffiths, & Ghahramani, 2006), similarity judgements (Navarro & Griffiths, 2008). This distribution is suitable for use as a prior in probabilistic models that represent objects using a potentially infinite array of features.

Both IBP and CRP model latent factors and perform dimensionality reduction (reduce the images or objects to latent features). They also both allow an infinite array of objects. Nevertheless, they solve different problems: IBP allows each customer to be assigned to multiple components (dishes), while CRP assigns each customer to a single component.

For example, factor analysis attempts to find a set of latent variables (or factors) that explain the correlations among the observed variables. A problem with factor analysis, however, is that the user has to specify the number of latent variables when using this technique. Should one use 5, 10, or 15 latent variables (or some other number)?

In general, Bayesian nonparametric methods grow the number of parameters as the size and complexity of the data set grow. An important example of a Bayesian nonparametric method is the Indian Buffet Process. (Another important example is the Dirichlet Process Mixture Model; see the Computational Cognition Cheat Sheet on Dirichlet Processes.)