The Indian Buffet Process

Eli Bingham[1] Matt Dickenson[2]

[1]University of North Carolina

[2]Duke University

February 10, 2014

Outline

1. Introduction
2. Latent feature models
3. Dirichlet and Chinese Restaurant Processes
4. Finite latent feature models
5. Beta and Indian Buffet Processes
6. Demonstration/visualization
7. Alternative derivations
8. Gibbs sampling and inference
9. Applications: Choice Behavior, Topic Models, and Cascading IBP
10. Discussion

Latent feature models

- Feature model: N items described by K features
- Dense feature model: every feature is present in every item, e.g. PCA
- Sparse feature model: only some features present in each item, and we can assume feature values and presence are independent:

$$\begin{aligned} \mathbf{F} &= \mathbf{A} \otimes \mathbf{Z} \\ P(\mathbf{F}) &= P(\mathbf{A})P(\mathbf{Z}) \end{aligned}$$

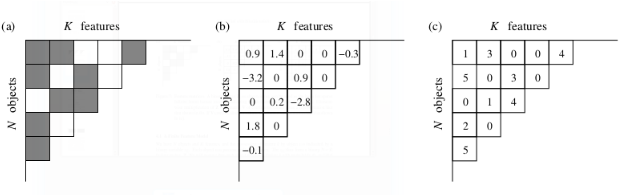


Figure: Griffiths and Ghahramani (2011) Figure 3

Example

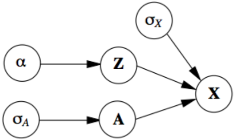


Figure: Griffiths and Ghahramani (2011) Figure 7

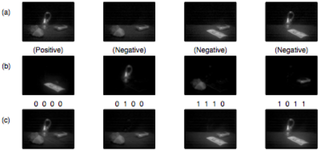


Figure: Griffiths and Ghahramani (2011) Figure 9

Motivation

- Problem with finite latent feature model: K is fixed
- Goal: construct nonparametric prior on $\mathbf{Z}$ so that K grows with the complexity of the dataset
- As with DPMMs, we can try to build one by taking $K \to \infty$ in a finite feature model

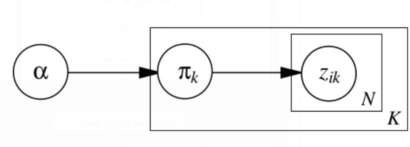


Figure: Griffiths and Ghahramani (2011) Figure 4

Background: Dirichlet Process

Finite version (Dirichlet distribution):

- Assignment of an object to a class is independent of all other assignments: $P(c|\theta) = \prod_{i=1}^{N} P(c_i|\theta) = \prod_{i=1}^{N} \theta_{c_i}$
- $\theta|\alpha \sim \text{Dirichlet}(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K})$ (if symmetric)
- $c_i|\theta \sim \text{Discrete}(\theta)$, where Discrete : Bernoulli :: Multinomial : Binomial

Integrating out θ: $P(c) = \frac{\prod_{k=1}^{K} \Gamma(m_k + \frac{\alpha}{K})}{\Gamma(\frac{\alpha}{K})^K} \frac{\Gamma(\alpha)}{\Gamma(N+\alpha)}$

As $K \to \infty$, we get the CRP

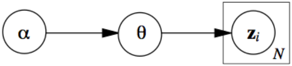


Figure: Griffiths and Ghahramani (2011) Figure 1

Background: Chinese Restaurant Process

1. N customers enter (in sequence) a restaurant with an infinite number of tables, each with infinite seating
2. First customer sits at first table with probability $\frac{\alpha}{\alpha} = 1$
3. i^{th} customer sits at the k^{th} table with probability $\frac{m_k}{i+\alpha-1}$, where m_k is the number of previous customers who sat at table k, or a new table with probability $\frac{\alpha}{i+\alpha-1}$

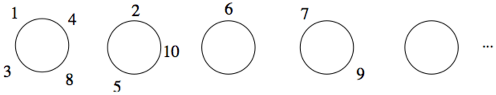


Figure: Griffiths and Ghahramani (2011) Figure 2

Limitation: each object (customer) can only belong to one class (table).

Finite latent feature models

The basic finite distribution on $z_{i,k}$s:

$$\begin{aligned}\pi_k|\alpha &\sim \text{Beta}(\frac{\alpha}{K},1)\\ z_{i,k}|\pi_k &\sim \text{Bernoulli}(\pi_k)\end{aligned}$$

As with DPMMs, we can marginalize out latent feature presence probabilities π_k to obtain a distribution on matrices $\mathbf{Z} \in \{0,1\}^{N\times K}$:

$$\begin{aligned}P(\mathbf{Z}) &= \prod_{k=1}^{K}\int\left(\prod_{i=1}^{N}P(z_{ik}|\pi_k)\right)P(\pi_k)d\pi_k\\ &= \prod_{k=1}^{K}\frac{\frac{\alpha}{K}\Gamma(m_k+\frac{\alpha}{K})\Gamma(N-m_k+1)}{\Gamma(N+1+\frac{\alpha}{K})}\end{aligned}$$

The $K \to \infty$ limit

- $lof(\mathbf{Z})$ is the matrix obtained by ordering the columns of $\mathbf{Z}$ as N-digit binary numbers
- To define a probability over infinitely wide binary matrices using de Finetti's Theorem, we need exchangeable symmetry, so we define *lof* equivalence classes by modding out column order:

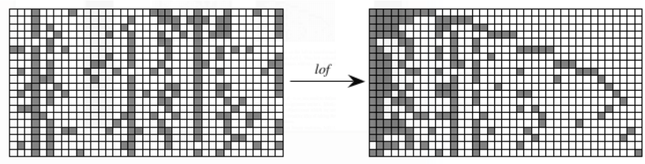


Figure: Griffiths and Ghahramani (2011) Figure 5

Indian Buffet Process

Indian Buffet Process:

1. N customers enter (in sequence) a buffet restaurant with an infinite number of dishes
2. First customer fills her plate with Poisson(α) number of dishes
3. i^{th} customer samples dishes in proportion to their popularity, with probability $\frac{m_k}{i}$, where m_k is the number of previous customers who sampled dish k
4. i^{th} customer then samples $K_1^{(i)} \sim$ Poisson($\frac{\alpha}{i}$) number of new dishes

Resulting probability distribution on matrices:

$$P(\mathbf{Z}) = \frac{\alpha^{K_+}}{\prod_{i=1}^{N} K_1^{(i)}!} \exp(\alpha H_N) \prod_{k=1}^{K_+} \frac{(N-m_k)!(m_k-1)!}{N!}$$

Alternative derivation: Beta Process

The CRP is obtained by marginalizing over a Dirichlet process:

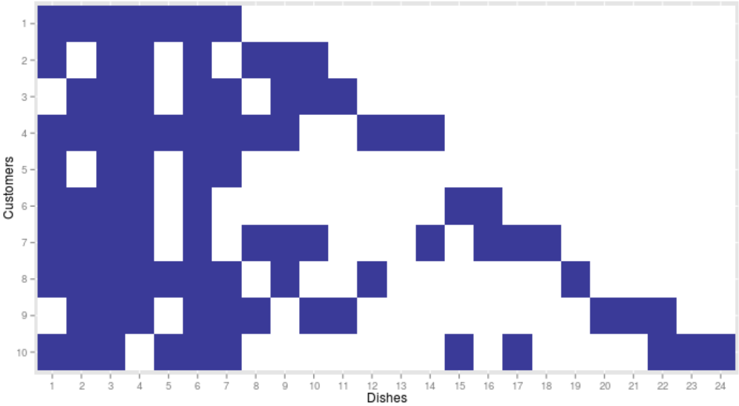
$$\begin{aligned}
\theta|\alpha, G_0 &\sim \mathrm{DP}(\alpha, G_0) \\
C|\theta &\sim \mathrm{Multinomial}(\theta) \\
CRP(\alpha) &\sim \int P(C|\theta)P(\theta|\alpha, G_0)d\theta
\end{aligned}$$

This representation follows from de Finetti's Theorem, which gives the existence of conditionally independent representations of infinite exchangeable joint distributions. The IBP is obtained by marginalizing over a Beta process:

$$\begin{aligned}
\theta|\alpha, \beta, G_0 &\sim \mathrm{BP}(\alpha, \beta, G_0) \\
z_i|\theta &\sim \mathrm{BeP}(\theta) \\
\mathrm{IBP}(\alpha, \beta) &\sim \int \prod_i P(z_i|\theta)P(\theta|\alpha, \beta, G_0)d\theta
\end{aligned}$$

Demo

mcdickenson.shinyapps.io/ibp-demo



Alternative derivation: Stick-Breaking

1. Recursively break (an initially unit-length) stick, breaking off a Beta$(\alpha, 1)$ portion at each step
2. Let each portion of the "stick", π_k represent the probability of each feature (sorted from largest to smallest)

This helps to show the relation between the Dirichlet process and the IBP. The stick-breaking construction is also useful for defining inference algorithms.

Properties of the Resulting Distribution

- The "effective" dimension $K_+ \sim \text{Poisson}(\alpha H_N)$
- The number of dishes on each customer's plate is distributed Poisson(α) (by exchangeability)
- $\mathbf{Z}$ remains sparse as $K \to \infty$: effective dimensions of $\mathbf{Z}$ are $N \times K_+$, and the expected number of entries is $N\alpha$

Inference by Gibbs Sampling

Given some data $\mathbf{X}$, we want to sample from a marginal posterior

$$P(z_{i,k} = 1|\mathbf{Z}_{-(ik)}, \mathbf{X}) \propto P(\mathbf{X}|\mathbf{Z})P(z_{i,k} = 1|\mathbf{Z}_{-(ik)})$$

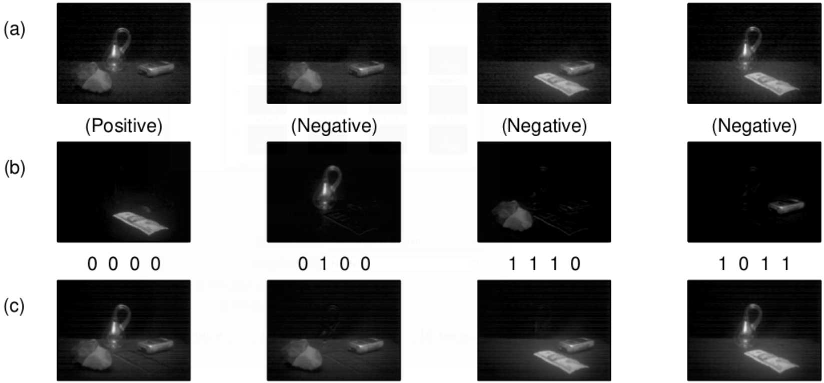
Iterate continuously over the rows $\mathbf{z}_i$ $(i = 1 \ldots N)$ of $\mathbf{Z}$:

1. For each column k of $\mathbf{Z}$, if $m_{-i,k} = 0$ (i.e. the rest of the column is empty) delete the column; otherwise set $z_{i,k} = 1$ with probability

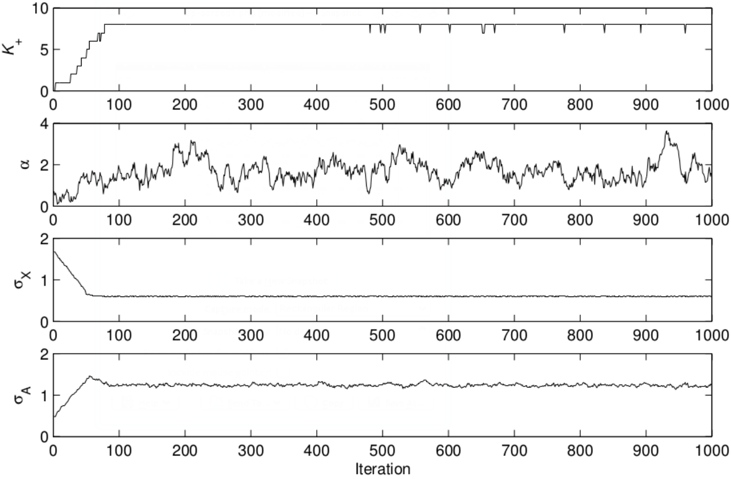
$$P(\mathbf{X}|\mathbf{Z})P(z_{i,k} = 1|\mathbf{z}_{-i,k}) = P(\mathbf{X}|\mathbf{Z})\frac{m_{-i,k}}{N}$$

2. At the end of the row, add $K_1^{(i)} \sim P(\mathbf{X}|\mathbf{Z})\text{Poisson}(\frac{\alpha}{N})$ new columns with ones in row i

Inference by Gibbs sampling



Inference by Gibbs sampling



Application 1: Choice Behavior

"A Choice Model with Infinitely Many Latent Features"
(Görür, Jäkel, and Rasmussen, ICML 2006)

- Customers compare items (e.g. cell phones) based on the (binary) features of each; more features are better
- Number of features is potentially infinite and ordering is not important, so IBP is used
- Celebrity example: "With whom would you prefer to spend an hour of conversation?"

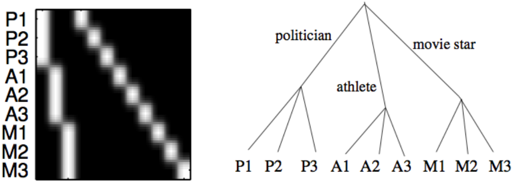


Figure: Görür, Jäkel, and Rasmussen (2006) Figure 3

Application 2: Topic Modeling

“The IBP Compound Dirichlet Process
and its Application to Focused Topic Modeling”
Williamson, Wang, Heller, and Blei (2010)

Stick-breaking construction:

$$\begin{aligned} \mu_k &\sim \mathrm{Beta}(\alpha, 1) \\ \pi_k &= \prod_{j=1}^{k} \mu_j \\ b_{m,k} &\sim \mathrm{Bernoulli}(\pi_k) \end{aligned}$$

Application 2: Topic Modeling

Focused topic model:

1. for $k = 1, 2, \ldots$
   - Sample stick length π_k
   - Sample relative mass $\phi_k \sim \text{Gamma}(\gamma, 1)$
   - Draw topic distribution over words: $\beta_k \sim \text{Dirichlet}(\eta)$
2. for $m = 1, \ldots, M$
   - Sample binary vector b_m
   - Draw total number of words $n^{(m)} \sim NB(\sum_k b_{m,k}\phi_k, 1/2)$
   - Sample distribution over topics $\theta_m \sim \text{Dirichlet}(b_m \cdot \phi)$
   - For each word $w_{m,i}, i = 1, \ldots, n^{(m)}$
     1. Draw topic index $z_{m,i} \sim \text{Discrete}(\theta_m)$
     2. Draw word $w_{m,i} \sim \text{Discrete}(\beta_{z_{m_i}})$

Application 2: Topic Modeling

An advantage of the focused topic model is that it separates the global topic proportions from the distribution over topics within a document. A rare topic within the corpus can be dominant within a document (e.g. baseball), and a frequent topic can be a small proportion of many documents.

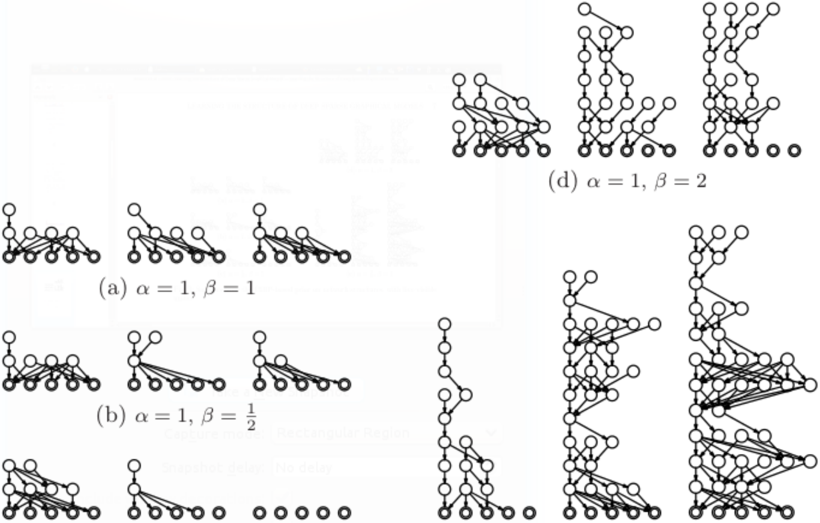
Application 3: CIBPs for Belief Network Structure Learning

- Adams et al construct a prior on infinitely wide, infinitely deep sparse belief network structures
- The Cascading Indian Buffet Process is an infinite sequence of binary matrices

$$\mathbf{Z}^{(m)} \sim \mathrm{IBP}(\alpha, \beta)$$

  where $Z^{(m+1)}$ has the same number of rows as columns in $Z^{(m)}$
- Culinary analogy: Each "dish" in the restaurant also corresponds to a "customer" in the next restaurant
- Theorem: the CIBP generative process stops growing at finite "depth"
- They also present MCMC algorithms for inference and test on images

CIBP prior samples



(a) $\alpha = 1,\ \beta = 1$

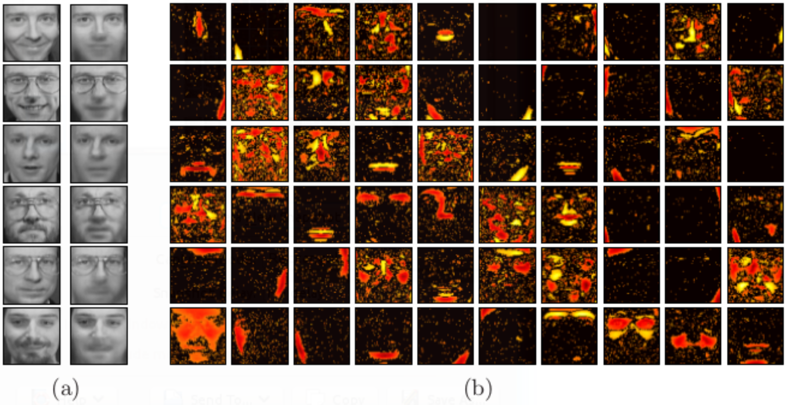
(b) $\alpha = 1,\ \beta = \frac{1}{2}$

(c) $\alpha = \frac{1}{2},\ \beta = 1$

(d) $\alpha = 1,\ \beta = 2$

(e) $\alpha = \frac{3}{2},\ \beta = 1$

CIBP Facial reconstruction



Discussion

Limitations of IBP:

1. Coupling of average number of features α and total number of features $N\alpha$ (can be overcome with a two-parameter generalization)
2. Computationally complex, can be time-consuming

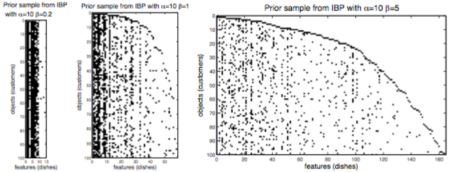


Figure: Griffiths and Ghahramani (2011) Figure 10