
Hierarchical Beta Processes and the Indian Buffet Process

Romain Thibaux
Dept. of EECS
University of California, Berkeley
Berkeley, CA 94720

Michael I. Jordan
Dept. of EECS and Dept. of Statistics
University of California, Berkeley
Berkeley, CA 94720

Presented by Mike Hughes
14 November 2011

Brown University CS2950p Applied Bayesian Nonparametrics

OVERVIEW

Contributions of the paper covered in this talk

- 1) Introduction of the Beta Process (and Bernoulli Process)
- 2) Connections to the Indian Buffet Process
- 3) Hierarchical Beta Process
- 4) Experiments

Contributions NOT thoroughly covered

- 1) new algorithm for generating samples from BP
- 2) posterior inference for the BP

BETA PROCESS vs. DIRICHLET PROCESS

$$B \sim \text{BP}(c, \gamma B_0)$$

$$B = \sum_{k=1}^{\infty} b_k \delta_{\theta_k}$$

$$G \sim \text{DP}(\alpha G_0)$$

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

STICK-BREAKING VIEW

$$\theta_k \sim B_0$$

$$b_k \sim \text{Beta}(c b_{0,k}, c(1 - b_{0,k}))$$

$$\text{where } b_{0,k} = \gamma B_0(\{\theta_k\})$$

$$\theta_k \sim G_0$$

$$\pi \sim \text{GEM}(\alpha) \quad \sum_k \pi_k = 1$$

DISTRIBUTION ON PARTITIONS

independent

Dirichlet

PREDICTIVE DISTRIBUTION / CULINARY METAPHOR

Indian Buffet process

Chinese Restaurant process

BETA / BERNOULLI PROCESS

Generative Model for infinite binary vectors X_i given DISCRETE base measure B_0

$$B \sim \text{BP}(c, \gamma B_0)$$

$$X_i \sim \text{BernP}(B)$$

$$X_i = \sum_{k=1}^K f_{i,k} \delta_{\theta_k} \quad \text{where } f_{i,k} \sim \text{Bern}(b_k)$$

Posterior Conjugacy

$$B|X_{1\dots n} \sim \text{BP} \left(c + n, \frac{c}{c+n} B_0 + \frac{1}{c+n} \sum_{i=1}^n X_i \right)$$

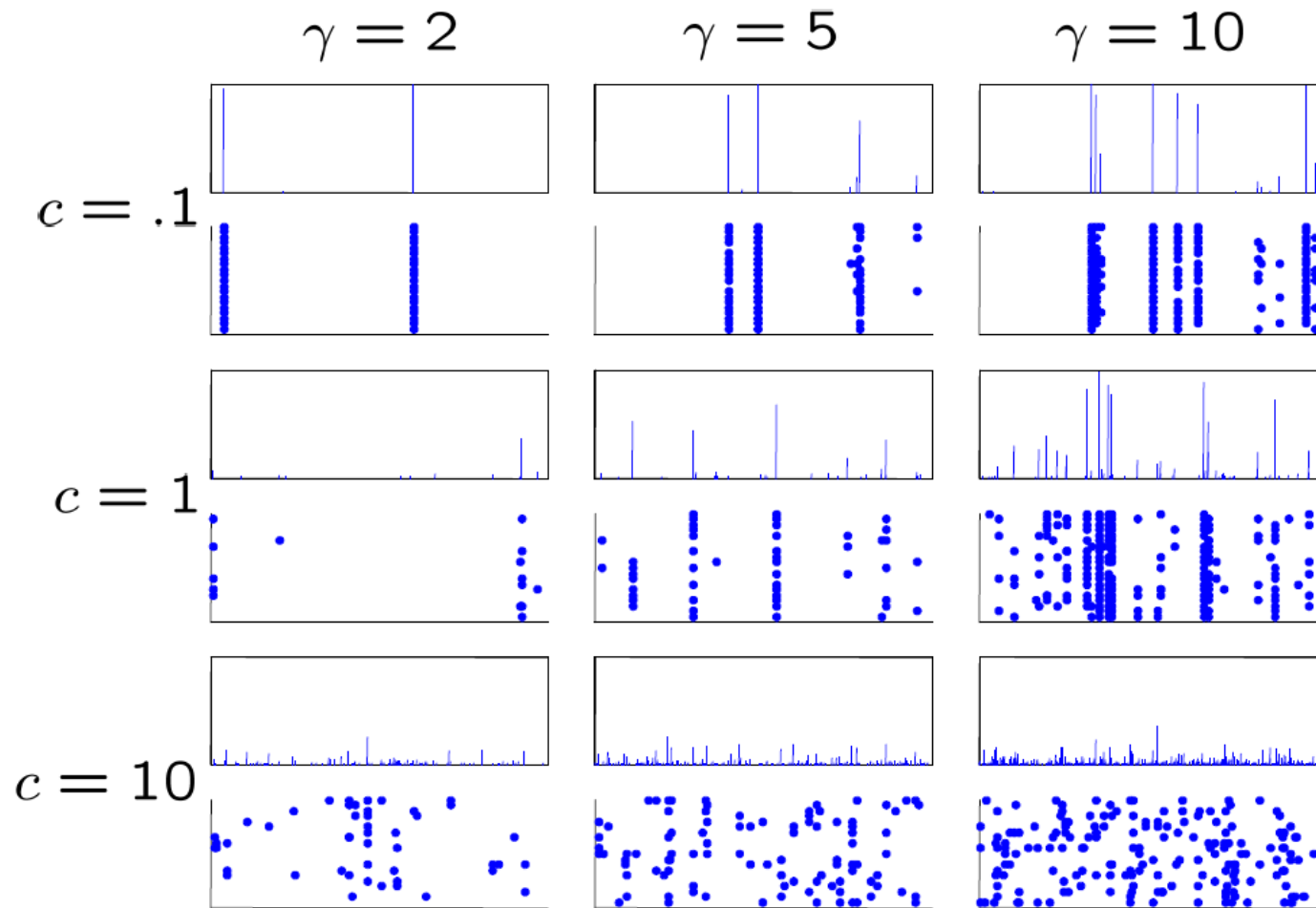
CONNECTION TO INDIAN BUFFET

Marginalizing over “latent feature weights” b_k

$$\begin{aligned} X_{n+1} | X_{1\dots n} &\sim \text{BeP} \left(\frac{c}{c+n} B_0 + \frac{1}{c+n} \sum_{i=1}^n X_i \right) \\ &= \text{BeP} \left(\frac{c}{c+n} B_0 + \sum_j \frac{m_{n,j}}{c+n} \delta_{\omega_j} \right) \end{aligned}$$

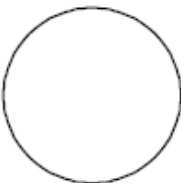
... just like marginalizing over the cluster frequencies in the CRP

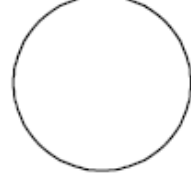
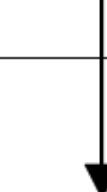
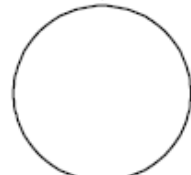
HYPERPARAMETERS



of unique dishes across all customers = $\text{Poi}(\gamma)$ if $c \rightarrow 0$ (everybody shares)
 $\text{Poi}(n\gamma)$ if $c \rightarrow \infty$ (no sharing)

HIERARCHICAL BETA PROCESS

$$B \sim \text{BP}(c_0, B_0)$$


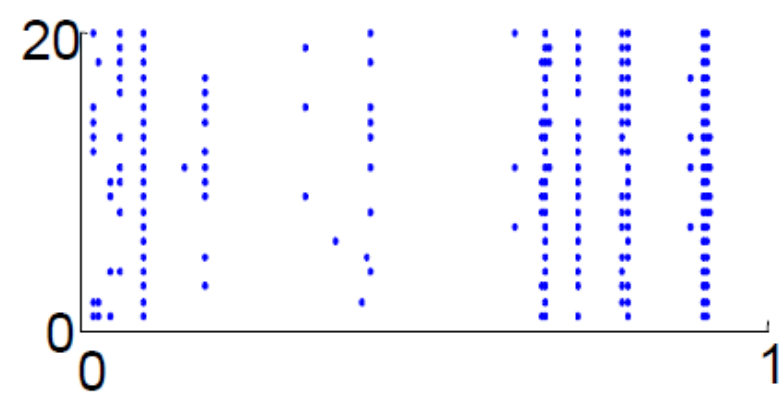
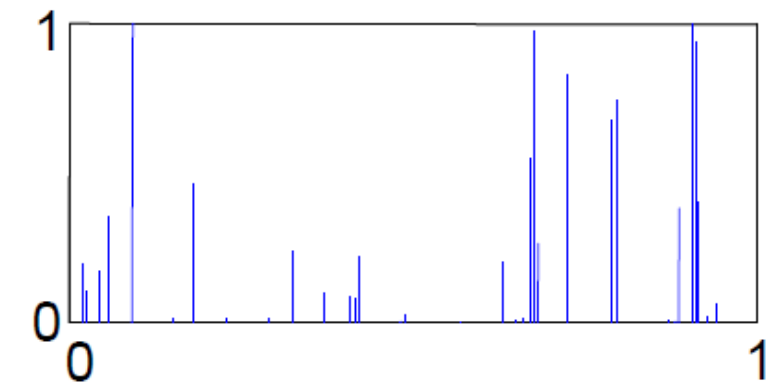
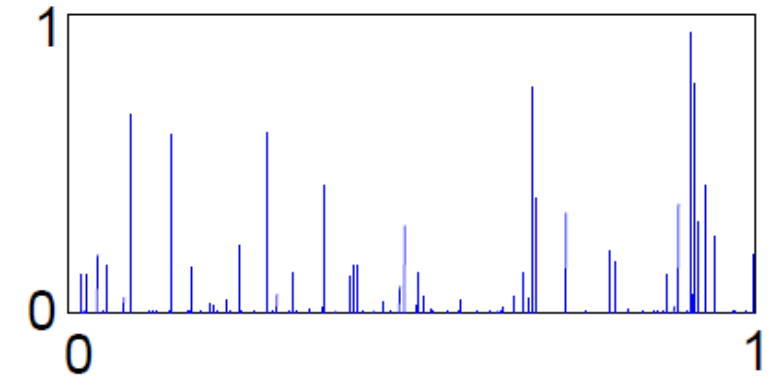


$j = 1, \dots, n$

$$A_j \sim \text{BP}(c_j, B)$$

$i = 1, \dots, n_j$

$$X_{ij} \sim \text{BeP}(A_j)$$



INFERENCE

Marginalize over category specific weights $a_{j,k}$

Learn values b_k for instantiated (observed) features

(sketch on board)

Flat vs. Hierarchical Models

Flat category-specific models (naïve Bayes) have bad properties for unbalanced data:

$$\hat{p}_{j,k} = \frac{m_{j,k} + a}{n_j + a + b} \rightarrow \frac{a}{a + b} \quad \text{when } m_{j,k} = 0, n_j \text{ small}$$

... but what if feature k is very rare in other categories where we have loads of data?

Hierarchical modeling (HBP) allows shrinking towards probabilities from other categories

$$\hat{p}_{j,k} = \frac{m_{j,k} + c_j b_k}{n_j + c_j} \rightarrow b_k \quad \text{when } m_{j,k} = 0, n_j \text{ small}$$

EXPERIMENTS

DATA: 20 Newsgroups (each a separate category)

documents / category ranged from 100, 94, ... 8 , 2

All words were used without any pruning or feature selection.

HBP: 58% accuracy

Naïve BAYES: 50% accuracy

I would have liked to see more thorough experiments...

is there any benefit when the data is **not** unbalanced?

how expensive is feature selection relative to training the HBP?