

Hierarchical Beta Processes and the Indian Buffet Process

Romain Thibaux

Computer Science Division
University of California, Berkeley
Berkeley, CA 94720

Michael I. Jordan

Computer Science Division and Department of Statistics
University of California, Berkeley
Berkeley, CA 94720

Technical Report 719
Department of Statistics

November 5, 2006

Abstract

We show that the beta process is the de Finetti mixing distribution underlying the Indian buffet process of [2]. This result shows that the beta process plays the role for the Indian buffet process that the Dirichlet process plays for Chinese restaurant process, a parallel that guides us in deriving analogs for the beta process of the many known extensions of the Dirichlet process. In particular we define Bayesian hierarchies of beta processes and use the connection to the beta process to develop posterior inference algorithms for the Indian buffet process. We also present an application to document classification, exploring a relationship between the hierarchical beta process and smoothed naive Bayes models.

1 Introduction

Mixture models provide a well-known probabilistic approach to clustering in which the data are assumed to arise from an exchangeable set of choices among a finite set of mixture components. *Dirichlet process mixture models* provide a nonparametric Bayesian approach to mixture modeling that does not require the number of mixture components to be known in advance [1]. The basic idea is that the Dirichlet process induces a prior distribution over partitions of the data, a distribution that is readily combined with a prior distribution over parameters and a likelihood. The distribution over partitions can be generated incrementally using a simple scheme known as the *Chinese restaurant process*.

As an alternative to the multinomial representation underlying classical mixture models, *factorial models* associate to each data point a set of latent Bernoulli variables. The factorial representation has several advantages. First, the Bernoulli variables may have a natural interpretation as “featural” descriptions of objects. Second, the representation of objects in terms of sets of Bernoulli variables provide a natural way to define interesting topologies on clusters (e.g., as the number of features that two clusters have in common). Third, the number of clusters representable with m features is 2^m , and thus the factorial approach may be appropriate for situations involving large numbers of clusters.

As in the mixture model setting, it is desirable to consider nonparametric Bayesian approaches to factorial modeling that remove the assumption that the cardinality of the set of features is known a priori. An important first step in this direction has been provided by Griffiths and Ghahramani [2], who defined a stochastic process on features that can be viewed as a factorial analog of the Chinese restaurant process. This process, referred to as the *Indian buffet process*, involves the metaphor of a sequence of customers tasting dishes in an infinite buffet. Let Z_i be a binary vector where $Z_{i,k} = 1$ if customer i tastes dish k . Customer i tastes dish k with probability m_k/i , where m_k is the number of customers that have previously tasted dish k ; that is, $Z_{i,k} \sim \text{Ber}(m_k/i)$. Having sampled from the dishes previously sampled by other customers, customer i then goes on to taste an additional number of new dishes determined by a draw from a $\text{Poisson}(\alpha/i)$ distribution. Modulo a reordering of the features, the Indian buffet process can be shown to generate an exchangeable distribution over binary matrices (that is, $P(Z_1, \dots, Z_n) = P(Z_{\sigma(1)}, \dots, Z_{\sigma(n)})$ for any permutation σ).

Given such an exchangeability result, it is natural to inquire as to the underlying distribution that renders the sequence conditionally independent. Indeed, De Finetti’s theorem states that the distribution of any infinitely exchangeable sequence can be written

$$P(Z_1, \dots, Z_n) = \int \left[\prod_{i=1}^n P(Z_i|B) \right] dP(B),$$

where B is the random element that renders the variables $\{Z_i\}$ conditionally independent and where we will refer to the distribution $P(B)$ as the “de Finetti mixing distribution.” For the Chinese restaurant process, the underlying de Finetti mixing distribution is known—it is the Dirichlet process. As this result suggests, identifying the de Finetti mixing distribution behind a given exchangeable sequence is important; it greatly extends the range of statistical applications of the

exchangeable sequence.

In this paper we make the following three contributions:

1. We identify the de Finetti mixing distribution behind the Indian buffet process. In particular, in Sec. 4 we show that this distribution is the *beta process*. We also show that this connection yields a two-parameter generalization of the Indian buffet process. While the beta process has been previously studied for its applications in survival analysis, this result shows that it is also the natural object of study in nonparametric Bayesian factorial modeling.
2. In Sec. 5 we exploit the link between the beta process and the Indian buffet process to provide a new algorithm to sample beta processes.
3. In Sec. 6 we define the *hierarchical beta process*, an analog for factorial modeling of the hierarchical Dirichlet process [9]. The hierarchical beta process makes it possible to specify models in which features are shared among a number of groups. We present an example of such a model in an application to document classification in Sec. 7, where we also explore the relationship of the hierarchical beta process to naive Bayes models.

2 The beta process

The beta process was defined by Hjort [3] for applications in survival analysis. In those applications, the beta process plays the role of a distribution on functions (cumulative hazard functions) defined on the positive real line. In our applications, the sample paths of the beta process need to be defined on more general spaces. We thus develop a nomenclature that is more suited to these more general applications.

Definition. A *beta process* $B \sim \text{BP}(c, B_0)$ is a distribution on positive random measures over a space Ω (e.g., \mathbb{R}). The beta process has two parameters: c is a positive function over Ω that we call the *concentration function*, and B_0 is a fixed measure on Ω , called the *base measure*. In the special case where c is a constant it will be called the *concentration parameter*.

A beta process is a particular kind of independent increment process, or *Lévy process*:

$$S \cap R = \emptyset \implies B(S) \text{ and } B(R) \text{ are independent.}$$

The Lévy-Khinchine theorem [4, 6] states that a Lévy process is characterized by its *Lévy measure*. The Lévy measure of the beta process $\text{BP}(c, B_0)$ is:

$$\nu(d\omega, dp) = c(\omega)p^{-1}(1-p)^{c(\omega)-1}dpB_0(d\omega). \tag{1}$$

The Lévy measure has the following elegant interpretation. It is a measure on $\Omega \times [0, 1]$, where Ω is a space of atoms, and $[0, 1]$ is the space of weight associated with these atoms. To draw B from the beta process distribution, draw a set of points $(\omega_i, p_i) \in \Omega \times [0, 1]$ from a Poisson process with base measure ν (see Fig. 2), and let:

$$B = \sum_i p_i \delta_{\omega_i}$$

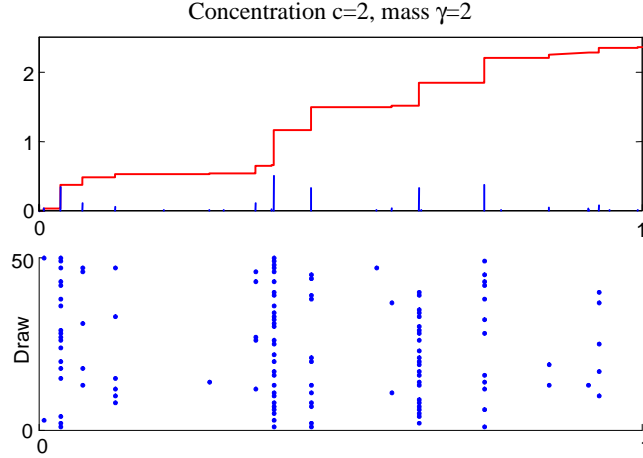


Figure 1: **Top.** A beta process is a random discrete measure, a collection of atoms with random locations and weight. The top figure shows a measure sampled from a beta process, along with the corresponding cumulative distribution function. **Bottom.** 100 samples from this measure, one per line. Samples are *sets* of points, obtained by including each point independently with probability given by the weight of the measure at that point.

which implies

$$B(S) = \sum_{i:\omega_i \in S} p_i$$

for all $S \in \Omega$. As this representation shows, B is discrete (with probability one). Note, however, that because ν has infinite mass near $b = 0$, the Poisson process will generate infinitely many points with small weight.

When the base measure B_0 is discrete: $B_0 = \sum_i q_i \delta_{\omega_i}$, then B has atoms at the same locations $B = \sum_i p_i \delta_{\omega_i}$ with $p_i \sim \text{Beta}(c(\omega_i)q_i, c(\omega_i)(1 - q_i))$. This imposes $q_i \in [0, 1]$. If B_0 is mixed discrete-continuous, B is the sum of the two independent contributions.

3 The Bernoulli process

Definition. Let B be a measure on Ω . We define a *Bernoulli process* with *hazard measure* B , written $X \sim \text{BeP}(B)$, as follows. If B is continuous, X is simply a Poisson process with intensity B , which we represent as a sum of delta functions at the jumps of the sample path. If B is discrete, then $B = \sum_i p_i \delta_{\omega_i}$ then $X = \sum_i b_i \delta_{\omega_i}$ where the b_i are independent Bernoulli variables with the probability that $b_i = 1$ equal to p_i . A Bernoulli process is also a particular kind of Lévy process. As for the beta process, a Bernoulli process with mixed discrete-continuous measure is the sum of the two independent contributions.

We can intuitively think of Ω as a space of potential “features,” and X as an object defined by the features it possesses. The random measure B encodes the probability that X possesses each

particular feature. In the Indian buffet metaphor, X is a customer and its features are the dishes it tastes.

Conjugacy. Let $B \sim \text{BP}(c, B_0)$, and let $X_i|B \sim \text{BeP}(B)$ for $i = 1, \dots, n$ be n independent Bernoulli process draws from B . Let $X_{1\dots n}$ denote the set of observations $\{X_1, \dots, X_n\}$. Reformulating a result of Hjort [3] using the language of Bernoulli processes, the posterior distribution of B after observing $X_{1\dots n}$ is still a beta process with modified parameters:

$$B|X_{1\dots n} \sim \text{BP} \left(c + n, \frac{c}{c+n} B_0 + \frac{1}{c+n} \sum_{i=1}^n X_i \right). \quad (2)$$

That is, the beta process is conjugate to the Bernoulli process.

4 Connection to the Indian buffet process

We now present the connection between the beta process and the Indian buffet process. The first step is to marginalize out B to obtain the marginal distribution of X_1 . Independence of X_1 on disjoint intervals is preserved, so X_1 is still a Bernoulli process, and its expectation is $E(X_1) = E(E(X_1|B)) = E(B) = B_0$, so its hazard measure is B_0 .¹

Combining this with Eq. (2) and using $P(X_{n+1}|X_{1\dots n}) = E_{B|X_{1\dots n}} P(X_{n+1}|B)$ gives us the following formula, which we rewrite using the notation $m_{n,j}$, the number of customers among $X_{1\dots n}$ having tried dish ω_j :

$$X_{n+1}|X_{1\dots n} \sim \text{BeP} \left(\frac{c}{c+n} B_0 + \frac{1}{c+n} \sum_{i=1}^n X_i \right) \quad (3)$$

$$= \text{BeP} \left(\frac{c}{c+n} B_0 + \sum_j \frac{m_{n,j}}{c+n} \delta_{\omega_j} \right). \quad (4)$$

To make the connection to the Indian buffet process let us first assume that c is a *constant* and B_0 is *continuous* with finite total mass $B_0(\Omega) = \gamma$. Observe what happens when we generate $X_{1\dots n}$ sequentially using Eq. (4). Since $X_1 \sim \text{BeP}(B_0)$ and B_0 is continuous, X_1 is a Poisson process with intensity B_0 . In particular, the total number of features of X_1 is $X_1(\Omega) \sim \text{Poi}(\gamma)$. This corresponds to the first customer trying a $\text{Poi}(\gamma)$ number of dishes.

Separating the base measure of Eq. (3) into its continuous and discrete parts, we see that X_{n+1} is the sum of two independent Bernoulli processes: $X_{n+1} = U + V$ where $U \sim \text{BeP}(\sum_j \frac{m_{n,j}}{c+n} \delta_{\omega_j})$ has an atom at ω_j (tastes dish j) with independent probability $\frac{m_{n,j}}{c+n}$ and $V \sim \text{BeP}(\frac{c}{c+n} B_0)$ is a Poisson process with intensity $\frac{c}{c+n} B_0$, generating a $\text{Poi} \left(\frac{c\gamma}{c+n} \right)$ number of new features (new dishes).

This is a two-parameter (c, γ) generalization of the Indian buffet process of Griffiths and Ghahramani, which we recover when we let $(c, \gamma) = (1, \alpha)$. We call c the *concentration parameter* and γ

¹We show that $E(B) = B_0$ in the Appendix.

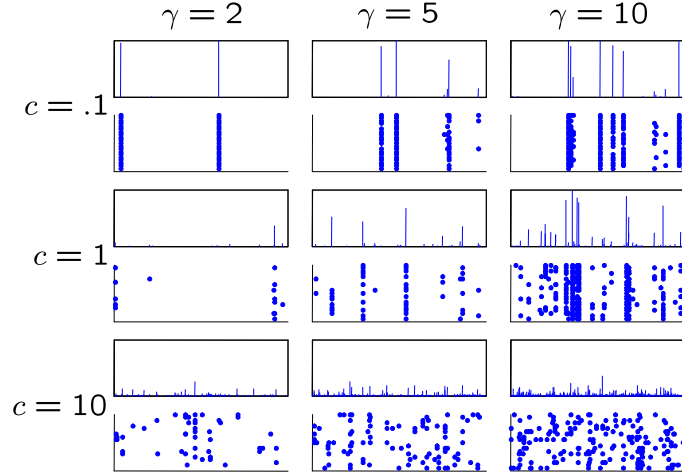


Figure 2: Draws from a beta process with concentration c and uniform base measure with mass γ , as we vary c and γ . For each draw, 20 samples are shown from the corresponding Bernoulli process, one per line.

the *mass parameter*. The customers together try a $\text{Poi}(n\gamma)$ number of dishes, but because they tend to try the same dishes the number of unique dishes is only $\text{Poi}\left(\gamma \sum_{i=0}^{n-1} \frac{c}{c+i}\right)$, roughly

$$\text{Poi}\left(\gamma + \gamma c \log\left(\frac{c+n}{c+1}\right)\right). \quad (5)$$

This quantity becomes $\text{Poi}(\gamma)$ if $c \rightarrow 0$ (all customers share the same dishes) or $\text{Poi}(n\gamma)$ if $c \rightarrow \infty$ (no sharing), justifying the name *concentration*. The effect of c and γ is illustrated in Fig. 2.

5 An algorithm to generate beta processes

Eq. (2) shows that the weight p_j of an atom at location ω_j that has been sampled at least once, that is for which $m_{n,j} > 0$, is beta-distributed:

$$p_j | X_{1..n} \sim \text{Beta}(m_{n,j}, c + n - m_{n,j})$$

If we draw p_j as soon as we observe ω_j , that is when $m_{n,j} = 1$, we obtain the following algorithm to build B . Call \hat{B}_n the approximation of B obtained after n steps of this algorithm, starting with $\hat{B}_0 = 0$. At each step $n \geq 1$:

- sample $K_n \sim \text{Poi}\left(\frac{c\gamma}{c+n-1}\right)$,
- sample K_n new locations ω_j from $\frac{1}{\gamma}B_0$ independently,

- sample their weight $p_j \sim \text{Beta}(1, c + n - 1)$ independently,
- $\hat{B}_n = \hat{B}_{n-1} + \sum_{j=1}^{K_n} p_j \delta_{\omega_j}$.

\hat{B}_n is justified as an approximation of B since $\lim_{n \rightarrow \infty} \hat{B}_n = B$ with probability one. The expected mass added at step n is $E(\hat{B}_n(\Omega) - \hat{B}_{n-1}(\Omega)) = \frac{c\gamma}{(c+n)(c+n-1)}$, and the expected remaining mass after step n is $E(B(\Omega) - \hat{B}_n(\Omega)) = \frac{c\gamma}{c+n}$.

Other algorithms exist to build approximations of beta processes. The Inverse Levy Measure algorithm of Wolpert and Ickstadt [10] is very general and can generate atoms in decreasing order of weight, but requires inverting the incomplete beta function at each step, which is computationally intensive. The algorithm of Lee and Kim [5] bypasses this difficulty by approximating the beta process by a compound Poisson process but requires a fixed approximation level. This means that their algorithm only converges in distribution.

Our algorithm is a simple and efficient alternative. It is closely related to the stick-breaking construction of Dirichlet processes [8], in that it generates the atoms of B in a size-biased order.

6 The hierarchical beta process

The parallel with the Dirichlet Process leads us to consider hierarchies of beta processes in a manner akin to the hierarchical Dirichlet processes of [9]. To motivate our construction, let us consider the following application to document classification (to which we return in Sec. 7).

Suppose that our training data X is a list of documents, where each document is classified by one of n topics. We model a document by the set of words it contains. In particular we do not take the number of appearances of each word into account. We assume that document $X_{i,j}$ is generated by including each word ω independently with a probability p_ω^j specific to topic j . These probabilities form a discrete measure A_j over the space of words Ω , and we put a beta process $\text{BP}(c_j, B)$ prior on A_j .

If B is a continuous measure, implying that Ω is infinite, with probability one the A_j 's will share no atoms, an undesirable result in the practical application to documents. For topics to share words, B must be discrete. On the other hand, B is unknown a priori, so it must be random. This suggests that B should itself be generated as a realization of a beta process. We thus put a beta process prior $\text{BP}(c_0, B_0)$ on B . This allows sharing of statistical strength among topics.

In summary we have the following model, whose graphical representation is shown in Fig. 3.

$$\begin{array}{lll}
 \text{Baseline} & B & \sim \text{BP}(c_0, B_0) \\
 \text{Topics} & A_j & \sim \text{BP}(c_j, B) \quad \forall j \leq n \\
 \text{Documents} & X_{i,j} & \sim \text{BeP}(A_j) \quad \forall i \leq n_j
 \end{array} \tag{6}$$

We want to perform posterior inference in this model. In particular to classify a new document Y we need to compare its probability under each topic: $P(X_{n_j+1,j} = Y|X)$ where $X_{n_j+1,j}$ is a new document in topic j . The next two subsections give a Monte Carlo inference algorithm to do this for hierarchies of arbitrarily many levels.

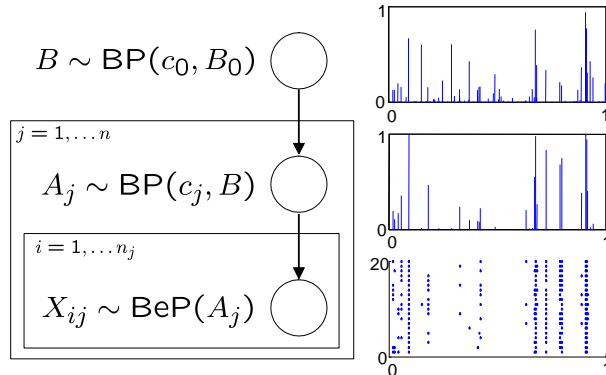


Figure 3: **Left.** Graphical model for the hierarchical beta process. **Right.** Example draws from this model with $c_0 = c_j = 1$ and B_0 uniform on $[0, 1]$ with mass $\gamma = 10$. From top to bottom are shown a sample for B_0 , A_j and 25 samples $X_{1,j}, \dots, X_{25,j}$.

6.1 The discrete part

Since all elements of our model are Lévy processes, we can choose a partition of the space Ω and perform inference separately on each part. In particular, we choose the partition that has cells $\{\omega\}$ for each of the features ω that have been observed at least once, and has a single (large) cell containing the rest of the space. We first consider inference with respect to the singletons $\{\omega\}$, and return to inference over the remaining cell in the following subsection.

Inference for $\{\omega\}$ deals only with the values $b_0 = B_0(\{\omega\})$, $b = B(\{\omega\})$, $a_j = A_j(\{\omega\})$ and $x_{ij} = X_{i,j}(\{\omega\})$. Let x denote the set of all x_{ij} and let a denote the set of all a_j . These variables form a slice of the hierarchy of their respective processes, and they have the following distributions:

$$\begin{aligned}
 b &\sim \text{Beta}(c_0 b_0, c_0(1 - b_0)) \\
 a_j &\sim \text{Beta}(c_j b, c_j(1 - b)) \\
 x_{ij} &\sim \text{Ber}(a_j).
 \end{aligned} \tag{7}$$

Strictly speaking, if B_0 is continuous, we have $b_0 = 0$ and so the prior over b is improper. We treat this by considering b_0 to be non-zero and taking the limit as b_0 approaches zero. This is justified under the limit construction of the beta process (see Theorem 3.1 of [3]).

For a fixed value of b , we can average over a using conjugacy. To average over b , we use Metropolis-Hastings; that is, we sample b from an approximation of its conditional distribution and correct for the difference by rejection.

Let $m_j = \sum_{i=1}^{n_j} x_{ij}$. Marginalizing out a in Eq. (7) and using $\Gamma(x+1) = x\Gamma(x)$, the log posterior

distribution of b given x is (up to a constant):

$$\begin{aligned}
f(b) &= (c_0 b_0 - 1) \log(b) + (c_0(1 - b_0) - 1) \log(1 - b) \\
&\quad + \sum_{j=1}^n \sum_{i=0}^{m_j-1} \log(c_j b + i) \\
&\quad + \sum_{j=1}^n \sum_{i=0}^{n_j-m_j-1} \log(c_j(1 - b) + i). \tag{8}
\end{aligned}$$

This posterior is log concave and has a maximum at $b^* \in (0, 1)$ which we can obtain by binary search. Using the concavity of $\log(c_j b + i)$ for $i > 0$, $\log(c_j(1 - b) + i)$ for $i \geq 0$ and $\log(1 - b)$ tangentially at b^* we obtain the following upper bound on f (up to a constant):

$$\begin{aligned}
g(b) &= (\alpha - 1) \log(b) - \frac{b}{\beta} \\
\text{where } \alpha &= c_0 b_0 + \sum_{j=1}^n \mathbf{1}_{m_j > 0} \\
\text{and } \frac{1}{\beta} &= \frac{c_0(1 - b_0) - 1}{1 - b^*} - \sum_{j=1}^n \sum_{i=1}^{m_j-1} \frac{c_j}{c_j b^* + i} \\
&\quad + \sum_{j=1}^n \sum_{i=0}^{n_j-m_j-1} \frac{c_j}{c_j(1 - b^*) + i}.
\end{aligned}$$

The function g is, up to a constant, the log density of a Gamma(α, β) variable. We can therefore use g as a Metropolis-Hastings proposal distribution; that is, sample $b' \sim \text{Gamma}(\alpha, \beta)$, and correct for the difference between f and g by accepting the move with probability $\min(1, e^\rho)$ where

$$\rho = f(b') - f(b) - (g(b') - g(b)).$$

The fact that g is, up to a constant, an upper bound on f means that f/g is bounded. This ensures that using reweighted samples of g as an estimate of f has finite variance. The fact that it is tangent to f at its maximum makes g a good approximation of f , maximizing the acceptance probability.

Setting $u = 1 - b$ in Eq. (8) maintains the form of f , only exchanging the coefficients. Therefore we can choose instead to approximate $1 - b$ by a gamma variable. We may pick the best of these two possible approximations. We choose the one with the lowest variance $\alpha\beta^2$.

This algorithm gives us samples from $P(b|x)$. In particular, with T samples, b_1, \dots, b_T , we can compute the following approximation:

$$\begin{aligned}
P(x_{n_j+1,j} = 1|x) &= E(E(a_j|b, x)|x) \\
&= \frac{c_j E(b|x) + m_j}{c_j + n_j} \\
\text{where } E(b|x) &\approx \frac{1}{T} \sum_t b_t.
\end{aligned}$$

In the end if we want samples of a we can obtain them easily. By conjugacy of Bernoulli and beta, the conditional distribution of a_j is beta, from which we can sample easily:

$$a_j|b, x \sim \text{Beta}(c_j b + m_j, c_j(1 - b) + n_j - m_j).$$

6.2 The continuous part

Let's now look at the rest of the space, where all observations are equal to zero. We choose to approximate B on that part of the space by \hat{B}_N obtained after N steps of the algorithm of Sec. 5. That is \hat{B}_N consists of $K_k \sim \text{Poi}(\frac{c_0 \gamma}{c_0 + k - 1})$ atoms at each level k where $k = 1, \dots, N$. For each atom (ω, b) out of the K_k atoms of level k , we have the following hierarchy, which is similar to Eq. (7) except that it refers to a random location ω chosen in size-biased order.

$$\begin{aligned} b &\sim \text{Beta}(1, c_0 + k - 1) \\ a_j &\sim \text{Beta}(c_j b, c_j(1 - b)) \\ x_{ij} &\sim \text{Ber}(a_j) \end{aligned} \tag{9}$$

We want to infer the distribution of the next observation $X_{n_j+1,j}$ from group (or topic) j , given that all other observations from all other groups are zero, that is $X = 0$. The locations of the atoms of $X_{n_j+1,j}$ will be $\frac{B_0}{\gamma}$ distributed so we only need to know the distribution of $X_{n_j+1,j}(\Omega)$, the number of atoms. $X = 0$ implies that all levels k have generated zero observations. Since each level is independent, we can reason on each level separately, where we want the posterior distribution of K_k and of the variables in Eq. (9).

Let $x = \{x_{ij}|j \leq n, i \leq n_j\}$ and $a = \{a_j|j \leq n\}$. Let P_k be the probability over b, a and x defined by Eq. (9) and let $q_k = P_k(x = 0)$. The posterior distribution of K_k is

$$\begin{aligned} P(K_k = m|X = 0) &\propto q_k^m \text{Poi}\left(\frac{c_0 \gamma}{c_0 + k - 1}\right)(m) \\ \text{so } K_k|X = 0 &\sim \text{Poi}\left(\frac{c_0 \gamma}{c_0 + k - 1} q_k\right). \end{aligned}$$

Let $p_k = P_k(x_{n_j+1,j} = 1, x = 0)$, then $P_k(x_{n_j+1,j} = 1|x = 0) = \frac{p_k}{q_k}$. Let D_k be the number of atoms of $X_{n_j+1,j}$ from level k

$$D_k \sim \text{Poi}\left(\frac{c_0 \gamma}{c_0 + k - 1} q_k \frac{p_k}{q_k}\right) = \text{Poi}\left(\frac{c_0 \gamma}{c_0 + k - 1} p_k\right).$$

Adding the contributions of all levels we get the following result (which is exact for $N = \infty$):

$$X_{n_j+1,j}(\Omega) \sim \text{Poi}\left(\sum_{k=1}^N \frac{c_0 \gamma}{c_0 + k - 1} p_k\right)$$

Using T samples $b_{k,1}, \dots, b_{k,T}$ from Eq. (9) we can compute p_k :

$$\begin{aligned} \text{Let } r(b) &= E_k \left(a_{jt} \prod_{j'} (1 - a_{j't})^{n_{j'}} \middle| b \right) \\ &= \frac{c_j b}{c_j + n_j} \prod_{j'=1}^n \frac{\Gamma(c_{j'}) \Gamma(c_{j'}(1-b) + n_{j'})}{\Gamma(c_{j'}(1-b)) \Gamma(c_{j'} + n_{j'})} \\ \text{then } p_k &= E_k[r(b)] \approx \frac{1}{T} \sum_{t=1}^T r(b_{k,t}). \end{aligned} \quad (10)$$

6.3 Larger hierarchies

We now extend model Eq. (6) to larger hierarchies such as:

$$\begin{aligned} \text{Baseline} \quad B &\sim \text{BP}(c_0, B_0) \\ \text{Topics} \quad A_j &\sim \text{BP}(c_j, B) \quad \forall j \leq n \\ \text{Subtopics} \quad S_{l,j} &\sim \text{BP}(c_{l,j}, A_j) \quad \forall l \leq n_j \\ \text{Documents} \quad X_{i,l,j} &\sim \text{BeP}(S_{l,j}) \quad \forall i \leq n_{l,j} \end{aligned}$$

To extend the algorithm of Sec. 6.2 we draw samples of a from Eq. (9) and replace b by a in Eq. (10). Extending the algorithm of Sec. 6.1 is less immediate since we can no longer use conjugacy to integrate out a . The Markov chain must now instantiate a and b . Sec. 6.1 lets us sample $a|b, x$, leaving us with the task of sampling $b|b_0, a$.

Up to a constant the log conditional probability of b is

$$\begin{aligned} f_2(b) &= (c_0 b_0 - 1) \log(b) + (c_0(1 - b_0) - 1) \log(1 - b) \\ &\quad - \sum_{j=1}^n [\log(\Gamma(c_j b)) + \log(\Gamma(c_j(1 - b)))] \\ &\quad + \sum_{j=1}^n c_j b \log \left(\frac{a_j}{1 - a_j} \right). \end{aligned} \quad (11)$$

Since $-\log(\Gamma(x)) - \log x$ is concave, f_2 itself is concave with a maximum b^* in $(0, 1)$ which we can obtain by binary search. Tangentially at a point x^* we can again use this concavity to obtain this upper bound, tight at $x = x^*$:

$$\begin{aligned} -\log \left(\frac{\Gamma(x)}{\Gamma(x^*)} \right) &\leq \log \left(\frac{x}{x^*} \right) - \left(\psi(x^*) + \frac{1}{x^*} \right) x \\ \text{where } \psi(x) &= \frac{\Gamma'(x)}{\Gamma(x)}. \end{aligned}$$

We apply this bound to Eq. (11) with $x^* = c_j b^*$ and $x^* = c_j(1 - b^*)$, and also upper bound $\log(1 - b)$ with its tangent at b^* , obtaining the following upper bound g_2 of f_2 (omitting the constant):

$$g_2(b) = (\alpha - 1) \log(b) - \frac{b}{\beta} \quad \text{where}$$

$$\begin{aligned}
\alpha &= c_0 b_0 + n \\
\frac{1}{\beta} &= \frac{c_0(1 - b_0) - n - 1}{1 - b^*} + \frac{n}{b^*} - \sum_{j=1}^n c_j \log \left(\frac{a_j}{1 - a_j} \right) \\
&\quad + \sum_{j=1}^n [c_j \psi(c_j b^*) - c_j \psi(c_j(1 - b^*))].
\end{aligned}$$

We can use this $\text{Gamma}(\alpha, \beta)$ variable, or the one obtained by setting $u = 1 - b$ in Eq. (11) as a proposal distribution as in Sec. 6.1. The case of $b|b_0, a$ is the general case for nodes of large hierarchies, so this algorithm can handle hierarchies of arbitrary depth.

7 Application to document classification

Naive Bayes is a very simple yet powerful probabilistic model used for classification. It models documents as lists of features, and assumes that features are independent given the topic. Bayes' rule can then be used on new documents to infer their topic. This method has been used successfully in many domains despite its simplistic assumptions.

Naive Bayes does suffer however from several known shortcomings. Consider a binary feature ω and let $p_{j,\omega}$ be the probability that a document from topic j has feature ω . Estimating $p_{j,\omega}$ as its maximum likelihood $\frac{m_{j,\omega}}{n_{j,\omega}}$ leads to many features having probability 0 or 1 and makes inference impossible. To prevent such extreme values, $p_{j,\omega}$ is generally estimated with *Laplace smoothing*, which can be interpreted as placing a common $\text{Beta}(a, b)$ prior on $p_{j,\omega}$:

$$\hat{p}_{j,\omega} = \frac{m_{j,\omega} + a}{n_{j,\omega} + a + b}$$

Laplace smoothing also corrects for unbalanced training data by imposing greater smoothing on the probabilities of small topics, for which we have low confidence. Nonetheless, Laplace smoothing can lead to paradoxes, in particular with unbalanced data [7]. Consider the situation where most topics u have enough data to show with confidence that $p_{u,\omega}$ is close to a very small value \bar{p} . The impact on classification of $p_{j,\omega}$ is *relative* to $p_{u,\omega}$ for $u \neq j$ so if topic j has little data, we expect $p_{j,\omega}$ to be close to \bar{p} for it to have little impact. Laplace smoothing however brings it close to $\frac{a}{a+b}$, very far from \bar{p} , where it will have an enormous impact. This inconsistency makes rare features hurt performance, and leads to the practice of combining naive Bayes with feature selection, potentially wasting information.

We propose instead to use a hierarchical beta process (hBP) as a prior over the probabilities $p_{j,\omega}$. Such a hierarchical Bayesian model allows sharing among topics by shrinking the maximum likelihood probabilities $\hat{p}_{1,\omega}, \dots, \hat{p}_{1,\omega}$ towards *each other* rather than towards $\frac{a}{a+b}$.

Such an effect could in principle be achieved using a finite model with a hierarchical beta prior; however, such an approach would not permit new features that do not appear in the training data. The model in Eq. (6) allows the number of known features to grow with data, and the number of unknown features to serve as evidence for belonging to a poorly known topic, one for which we have

little training data. A hBP gives a consistent prior for varying amounts of data, whereas Laplace smoothing amounts to changing the prior every time a new feature appears.

We compared the performance of hBP and naive Bayes on 1000 posts from the *20 Newsgroups* dataset², grouped into 20 topics. We chose an unbalanced dataset, with the number of documents per topic decreasing linearly from 100 to 2. We randomly selected 40% of these papers as a test set for the task of classifying them into their correct area. We encoded the documents X_{ij} as a set of binary features representing the presence of words. All words were used without any pruning or feature selection.

We used the model in Eq. (6), setting the parameters c_0 , γ and c_j a priori in the following way. Since we expect a lot of commonalities between topics, with differences concentrated on a few words only, we take c_j to be small. Therefore documents drawn from Eq. (6) are close to being drawn from a common $\text{BP}(c_0, \gamma)$ prior, under which the expected number of features per document is γ . Estimating this value from the data gives $\gamma = 150$. Knowing γ we can solve for c_0 by matching the expectation of Eq. (5) to the number of unique features N in the data. This can be done by interpreting Eq. (5) as a fixed point equation:

$$c_0 \leftarrow \frac{F - \gamma}{\gamma} \left(\log \left(\frac{c_0 + n}{c_0 + 1} \right) \right)^{-1}$$

leading to $c_0 = 70$. Finally we selected the value of c_j by cross-validation on a held-out portion of the training set, giving us $c_j = 10^{-4}$.

We then ran our Monte Carlo inference procedure and classified each document Y of the test set by finding the class for which $P(X_{n_j+1,j} = Y|X)$ was highest, obtaining 58% accuracy. The acceptance rate of the Metropolis-Hastings moves was above 90%, showing the quality of the gamma approximation.

By comparison, we performed a broad grid search over the space of Laplace smoothing parameters $a, b \in [10^{-11}, 10^7]$ for the best performing naive Bayes model. For $a = 10^{-8}$ and $b = 10^7$, naive Bayes reached 50% accuracy.

The classification of documents is often tackled with a multinomial models under the bag-of-words assumption. The advantage of a feature-based model is that it becomes very natural to add other non-text features such as “presence of a header,” “the text is right-justified,” “the font is red,” etc. The favorable properties of the hBP with regards to rare features implies that we can safely include a much larger set of features than would be possible for a naive Bayes model.

8 Conclusions

In this paper we have shown that the beta process—originally developed for applications in survival analysis—is the natural object of study for nonparametric Bayesian factorial modeling. Representing data points in terms of sets of features, factorial models provide substantial flexibility relative to the multinomial representations associated with classical mixture models. We have shown that

²The data are available at <http://people.csail.mit.edu/jrennie/20Newsgroups/>.

the beta process is the de Finetti mixing distribution underlying the Indian buffet process, a distribution on sparse binary matrices. This result parallels the relationship between the Dirichlet process and the Chinese restaurant process.

We have also shown that the beta process can be extended to a recursively-defined hierarchy of beta processes. This representation makes it possible to develop nonparametric Bayesian models in which unbounded sets of features can be shared among multiple nested groups of data.

Compared to the Dirichlet process, the beta process has the potential advantage of being an independent increments process (the Gaussian process is another example of an independent increments process). However, some of the simplifying features of the Dirichlet process do not carry over to the beta process, and in particular we have needed to design new inference algorithms for beta process and hierarchical beta process mixture models rather than simply borrowing from Dirichlet process methods. Our inference methods are elementary and additional work on inference algorithms will be necessary to fully exploit beta process models.

9 Appendix

9.1 Moments of the beta process

Hjort [3] derives the following moments for any set $S \subset \Omega$. If B_0 is continuous,

$$\begin{aligned} EB(S) &= \int_{S \times [0,1]} b\nu(d\omega, db) \\ &= \int_S c(\omega)B_0^c(d\omega) \int_{[0,1]} (1-b)^{c(\omega)-1} db \\ &= B_0(S). \end{aligned}$$

If B_0 is discrete:

$$EB(S) = \sum_{i:\omega_i \in S} E(b_i)\delta_{\omega_i} = \sum_{i:\omega_i \in S} b_0\delta_{\omega_i} = B_0(S)$$

$$\text{and } \text{Var } B(S) = \int_S \frac{B_0(d\omega)(1-B_0(d\omega))}{c(\omega)+1}.$$

Thus despite being discrete, B can be viewed as an approximation of B_0 , with fluctuations going to zero as $c \rightarrow \infty$.

References

- [1] C. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 2(6):1152–1174, 1974.
- [2] T. Griffiths and Z. Ghahramani. Infinite latent feature models and the Indian buffet process. In *Advances in Neural Information Processing Systems (NIPS) 18*, 2005.

- [3] N. L. Hjort. Nonparametric Bayes estimators based on Beta processes in models for life history data. *The Annals of Statistics*, 18(3):1259–1294, 1990.
- [4] A. Y. Khinchine. Korrelationstheorie der stationären stochastischen prozesse. *Mathematische Annalen*, 109:604–615, 1934.
- [5] J. Lee and Y. Kim. A new algorithm to generate beta processes. *Computational Statistics and Data Analysis*, 47:441–453, 2004.
- [6] P. Lévy. Théorie de l’addition des variables aléatoires. Gauthiers-Villars, Paris, 1937.
- [7] J. Rennie, L. Shih, J. Teevan, and D. Karger. Tackling the poor assumptions of naïve Bayes text classifiers. In *International Conference on Machine Learning (ICML) 20*, 2003.
- [8] J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.
- [9] Y. W. Teh, M. I. Jordan, M. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 2006. To appear.
- [10] R. Wolpert and K. Ickstadt. Simulation of Lévy random fields. In D. Dey, P. Muller, and D. Sinha, editors, *Practical Nonparametric and Semiparametric Bayesian Statistics*. Springer-Verlag, New York, 1998.