COS 597C: Bayesian nonparametrics

Lecturer: David Blei
Scribes: Tao Yue

Lecture # 9
November 19, 2007

The Indian Buffet Process

Briefly, the Indian Buffet process is the Machine Learning attempt to cast Bayesian Non-parametrics into the latent feature model.

Latent Features

Some notation: let z_i be a latent feature vector on k features. z_i is simply a vector of indicators, specifying one of 2^k possible combinations of features. For example, for $k = 7$, we might have $[0011000]^T$.

Then x_i is just a distribution, e.g. $x_i \sim N\left(\eta^T z_i, \sigma^2\right)$. Notice the placement of z_i which selects a set of features.

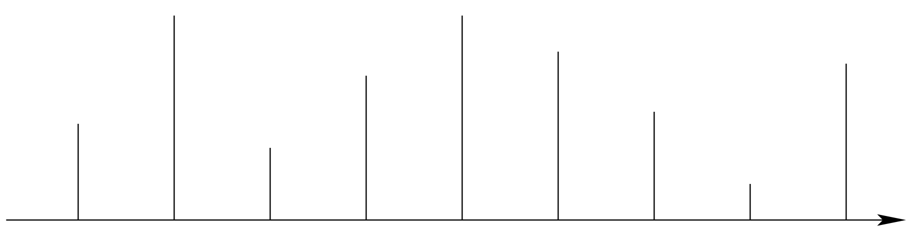
Latent features got started in factorial HMMs, which describe latent feature vectors evolving over time.

Finite Model

Let us look at the finite model, with K features, and where:

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

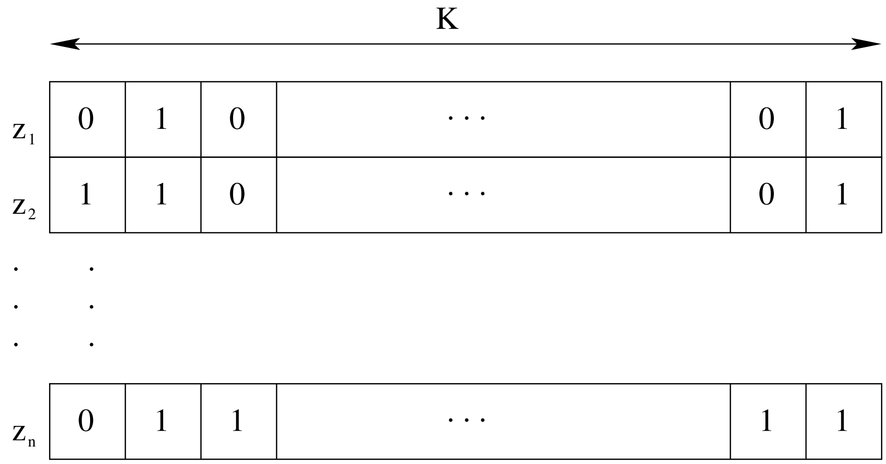
π_k is the relative probability of each feature being on, e.g.:



z_k are binary vectors, giving the latent structure that's used to generate the model. This is similar to Radford and Neal, letting $K \to \infty$.

The marginal probability of the binary vector matrix $\mathbf{Z}$ is:

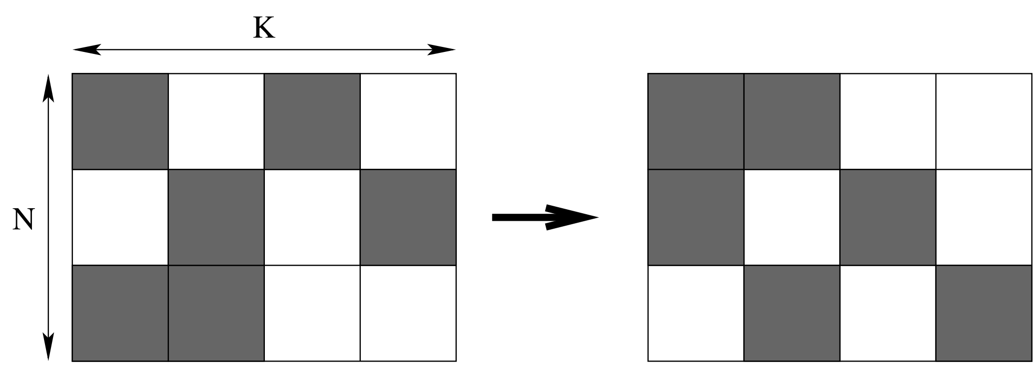
$$P(\mathbf{Z}) = \prod_{k=1}^{K} \frac{\frac{\alpha}{K}\Gamma(m_k + \frac{\alpha}{K})\Gamma(N - m_k + 1)}{\Gamma(N + 1 + \frac{\alpha}{K})}$$



where m_k is the number of times that the k^{th} feature is on, N is the number of items, and K is the number of features. This result falls out of the Beta/Bernoulli conjugate pair.

Infinite Model

Now take the limit as $K \to \infty$. Define the left-order-function lof($\mathbf{Z}$) to be the matrix that reorders columns according to their magnitude as binary numbers. For example:



Now let [$\mathbf{Z}$] be the set of matrices that are lof-equivalent. This is, of course, a many-to-one mapping, but it is a fine representation when feature order does not matter.

Then:

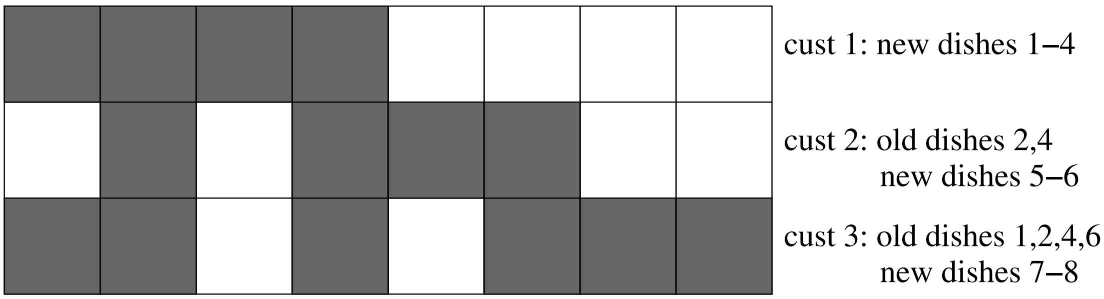
$$P([\mathbf{Z}]) = \frac{\alpha^{K_+}}{\prod_{h=1}^{2^N-1} K_h!} \exp\{-\alpha H_N\} \prod_{k=1}^{K_+} \frac{(N-m_k)!(m_k-1)!}{N!}$$

where K_+ is the number of features that are actually used. Refer to equations (29)-(34) in Griffiths and Ghahramani for the derivation.

Indian Buffet Process

Returning to our favorite restaurant analogy, an Indian Buffet Process is a process where:

- Customer 1 chooses the first $K_1^{(1)}$ dishes $\sim$ Poisson(α).
- Customer i chooses:
  - Each of the existing dishes with probability $\frac{m_k}{i}$.
  - $K_1^{(i)}$ additional dishes, where $K_1^{(i)} \sim$ Poisson$(\frac{\alpha}{i})$



Notice $\sum_i K_1^{(i)} = K_+$, and is finite. Notice also that this distribution is not exchangeable. $K_1^{(i)}$ depends on $\mathbf{Z}$, and we need $\mathbf{Z}$ to be well-formed.

Beta Process

The Beta process:

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

is a distribution on positive random measures. It is a type of Lévy process, which is described by a Lévy measure on space Ω of atoms and weight space $[0, 1]$:

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

Here, c is the concentration and B_0 is the base measure. When B_0 is discrete,

$$B_0 = \sum_i q_i \delta_{\omega_i}$$
$$B = \sum_i p_i \delta_{\omega_i}$$

where

$$p_i \sim \text{Beta}(c(\omega_i)q_i, c(\omega_i)(1-q_i))$$

If B_0 is a mix of continuous and discrete, we must account for the two parts separately. They are irreconcilable because the continuous case has zero probability of any one point.

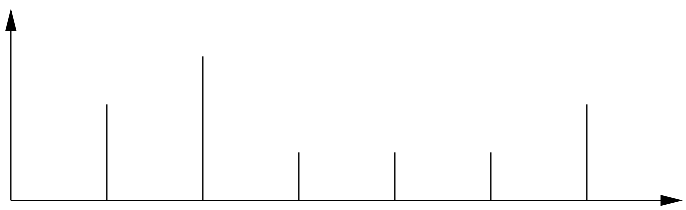
Bernoulli Process

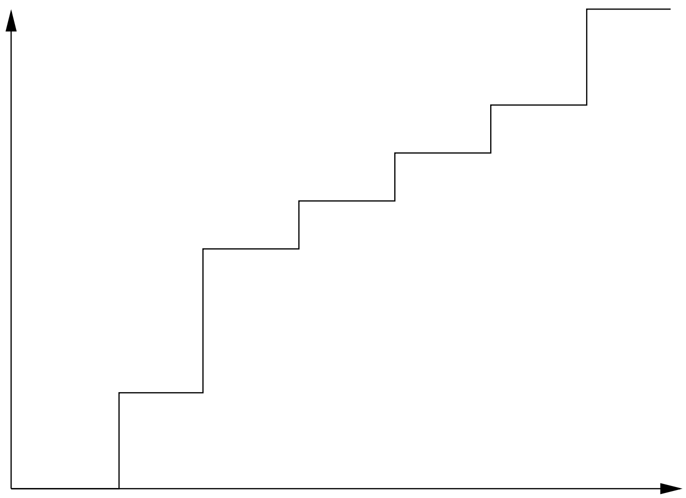
Define Bernoulli process with hazard measure B:

$$X \sim \mathrm{BeP}(B)$$

where B is a measure on Ω.

If B is continuous, then X is a Poisson process with intensity B, with steps at each of the delta functions.





If B is discrete, then:

$$X = \sum_i b_i \delta_{\omega_i}$$
$$b_i \sim \mathrm{Bernoulli}(p_i)$$

Indian Buffet Process

Given:

$$B \sim \mathrm{BP}(c, B_0)$$
$$X \sim \mathrm{BeP}(B)$$

then the Beta process is conjugate to the Bernoulli process:

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

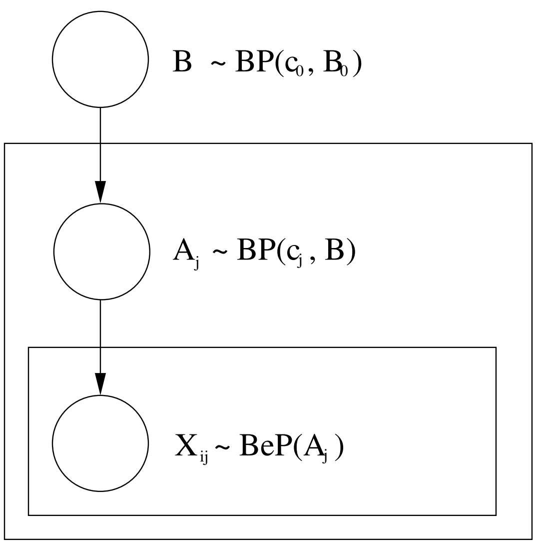
and:

$$\begin{aligned} X_{n+1}|X_{1:n} &\sim \text{BeP}\left(\frac{c\gamma}{c+n}B_0 + \frac{1}{c+n}\sum_i x_i\right) \\ &= \text{BeP}\left(\frac{c\gamma}{c+n}B_0 + \sum_j \frac{m_{n,j}}{c+n}\delta_{\omega_j}\right) \end{aligned}$$

with scaling parameter γ. Notice the correspondance to the Indian Buffet Process. We generate Poisson($\frac{c\gamma}{c+n}$) new features, i.e. taste a Poisson number of new dishes. And we taste each existing dish j with probability $\frac{m_{n,j}}{c+n}$.

Hierarchical Beta Process

The graphical model for the hBP is simply:



Thibaux and Jordan then apply a Monte Carlo inference procedure on document classification and obtain 58% accuracy. For document classification, this accuracy rate is not super-impressive. Notice that they ended up at the boundaries of the grid, which suggests that the whole space hasn't actually been searched. It's unclear what the motivating application for this would be. But factorial models are very important, so there must be some application area.