Infinite latent feature models and the Indian Buffet Process

Tom Griffiths
Brain and Cognitive Sciences
MIT

Joint work with Zoubin Ghahramani
Beyond categorization

- Categorization: associate objects with latent class

- Many domains require richer representations
  - membership in multiple latent classes
  - more generally, latent features

- How do we choose the dimensionality of representations?
  - problem of *model selection*
Perspectives on model selection

- Compare multiple models of different dimensionality
  - Bayes factors, cross-validation, etc.
  - (usually) false assumption of fixed dimension
  - hard to apply to large model spaces

- Define a single model of unbounded dimensionality
  - posterior on dimensionality via posterior on parameters
  - allows dimensionality to grow with new data
  - pursued in nonparametric Bayesian density estimation
    (e.g., Antoniak, 1974; Escobar & West, 1995)
Latent class models

• Associate each datapoint $x_i$ with a latent class $z_i$
  – data matrix $X = [x_1 \ldots x_N]^T$
  – class vector $z = [z_1 \ldots z_N]^T$

• Model defined by
  – prior on class assignments $P(z)$
  – likelihood $p(X|z)$

• How do we choose the number of classes?

• Nonparametric Bayes: define $P(z)$ to allow infinitely many classes, of which a finite subset are used
Chinese restaurant process (CRP)

- Chinese restaurant with infinitely many infinite tables
- $N$ customers sit down
  - the first customer sits at the first table
  - the $i$th customer chooses a table at random

\[
\begin{align*}
P(\text{occupied table } k | \text{previous customers}) &= \frac{m_k}{\alpha + i - 1} \\
P(\text{next unoccupied table} | \text{previous customers}) &= \frac{\alpha}{\alpha + i - 1}
\end{align*}
\]
Chinese restaurant process (CRP)

- Defines a distribution over partitions
- e.g., \((1 \ 3 \ 4 \ 8) \ (2 \ 5 \ 10) \ (6) \ (7 \ 9)\)
- Customers are exchangeable (Aldous, 1985; Pitman, 2002)

\[
P(z) = \alpha^K \left( \prod_{k=1}^{K} (m_k - 1)! \right) \frac{\Gamma(\alpha)}{\Gamma(N + \alpha)}
\]
Beyond categorization

- The CRP allows number of classes to be inferred
  - a prior on class assignments of unbounded dimension
  - distribution over partitions

- Can we apply a parallel strategy with other representations?

- Infinite latent feature models
  - a prior on feature assignments of unbounded dimension
  - distribution over binary matrices
Different feature representations

- Binary features
Different feature representations

<table>
<thead>
<tr>
<th>N objects</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Binary features
- Factorial features
Different feature representations

- Binary features
- Factorial features
- Continuous features
Latent feature models

• Associate each datapoint \( x_i \) with a latent feature vector \( z_i \)
  – data matrix \( X = [x_1 \ldots x_N]^T \)
  – feature matrix \( Z = [z_1 \ldots z_N]^T \)

• Model defined by
  – prior on feature assignments \( P(Z) \)
  – likelihood \( p(X|Z) \)

• How do we choose the number of features?

• Nonparametric Bayes: define \( P(Z) \) to allow infinitely many features, of which a finite subset are used
Priors on binary matrices

• Start with priors on $N \times K$ matrices, take $K \to \infty$

• Two cases:
  – “class matrices”: one 1 per row
  – “feature matrices”: general binary matrices

• Two priors:
  – the Chinese restaurant process
  – the Indian buffet process
Class matrices

\[
z_i | \theta \sim \text{Discrete}(\theta)
\]

\[
\theta \sim \text{Dirichlet}\left(\frac{\alpha}{K}\right)
\]
Class matrices

\[ P(\mathbf{Z}) = \int_{\Delta} \prod_{i=1}^{N} P(\mathbf{z}_i|\theta)p(\theta) \, d\theta \]
Left-ordered form

- History $h$ of each class: binary column vector
- $lof$ orders columns by values of binary histories
$lof$ equivalence classes

- $X$ and $Y$ are $lof$ equivalent iff $lof(X) = lof(Y)$
- Class matrices: $lof$ equivalence classes are partitions
lof equivalence classes

\[
\lim_{k \to \infty} P([\mathbf{Z}]) = \alpha^K \left( \prod_{k=1}^{K_+} (m_k - 1)! \right) \frac{\Gamma(\alpha)}{\Gamma(N + \alpha)}
\]

(see also Green & Richardson, 2001; Neal, 1992)
Feature matrices

• For general binary matrices

$$z_{ik} \sim \text{Bernoulli}(\theta_k)$$

$$\theta_k \sim \text{Beta}(\frac{\alpha}{K}, 1)$$
Feature matrices

• For general binary matrices

\[ z_{ik} \sim \text{Bernoulli}(\theta_k) \]
\[ \theta_k \sim \text{Beta}\left(\frac{\alpha}{K}, 1\right) \]

• For a finite matrix \( Z \)

\[ P(Z) = \int_0^1 \cdots \int_0^1 P(Z|\theta_1, \ldots, \theta_k) \prod_{k=1}^K p(\theta_k) \, d\theta_k \]
Feature matrices

• For general binary matrices

\[ z_{ik} \sim \text{Bernoulli}(\theta_k) \]
\[ \theta_k \sim \text{Beta}\left(\frac{\alpha}{K}, 1\right) \]

• For a finite matrix \( Z \)

\[
P(Z) = \int_0^1 \cdots \int_0^1 P(Z|\theta_1, \ldots, \theta_k) \prod_{k=1}^{K} p(\theta_k) \, d\theta_k
\]

• Taking the limit as \( K \to \infty \)

\[
P([Z]) = \exp\left\{-\alpha \sum_{i=1}^{N} \frac{1}{i} \right\} \frac{\alpha^{K+}}{\prod_{h>0} K_h!} \prod_{k \leq K_+} \frac{(N - m_k)!}{N!} \frac{(m_k - 1)!}{K_h!}
\]
Indian buffet process (IBP)

• Indian restaurant with infinitely many infinite dishes

• $N$ customers serve themselves
  – the first customer samples $\text{Poisson}(\alpha)$ dishes
  – the $i$th customer
    samples a previously sampled dish with probability $\frac{m_k}{i+1}$
    then samples $\text{Poisson}\left(\frac{\alpha}{i}\right)$ new dishes
Indian buffet process (IBP)

- Indian restaurant with infinitely many infinite dishes
- \( N \) customers serve themselves
  - the first customer samples \( \text{Poisson}(\alpha) \) dishes
  - the \( i \)th customer
    - samples a previously sampled dish with probability \( \frac{m_k}{i+1} \)
    - then samples \( \text{Poisson}\left(\frac{\alpha}{i}\right) \) new dishes
Indian buffet process (IBP)

- Indian restaurant with infinitely many infinite dishes
- $N$ customers serve themselves
  - the first customer samples $\text{Poisson}(\alpha)$ dishes
  - the $i$th customer
    samples a previously sampled dish with probability $\frac{m_k}{i+1}$
    then samples $\text{Poisson}\left(\frac{\alpha}{i}\right)$ new dishes
Indian buffet process (IBP)

- Indian restaurant with infinitely many infinite dishes
- $N$ customers serve themselves
  - the first customer samples $\text{Poisson}(\alpha)$ dishes
  - the $i$th customer
    samples a previously sampled dish with probability $\frac{m_k}{i+1}$
    then samples $\text{Poisson}(\frac{\alpha}{i})$ new dishes
Indian buffet process (IBP)

- Indian restaurant with infinitely many infinite dishes
- $N$ customers serve themselves
  - the first customer samples $\text{Poisson}(\alpha)$ dishes
  - the $i$th customer samples a previously sampled dish with probability $\frac{m_k}{i+1}$ then samples $\text{Poisson}(\frac{\alpha}{i})$ new dishes
Properties of the IBP

- Customers are exchangeable

- Total number of dishes $K_+ \sim \text{Poisson}(\alpha \sum_{i=1}^{N} \frac{1}{i})$
  - i.e. $K_+ \to \infty$ as $N \to \infty$, as with the CRP

- Number of dishes sampled by each customer $\sim \text{Poisson}(\alpha)$
  - sparsity is coupled to dimension

- Expected number of non-zero entries in $Z$ is $N\alpha$
A linear-Gaussian model

- Likelihood $P(X|Z)$ specified by
  
  $x_i \sim \text{Gaussian}(z_iA, \sigma_X I)$
  
  $A \sim \text{Gaussian}(0, \sigma_A I)$

- For $Z \sim \text{CRP}(\alpha)$, spherical Gaussian mixture model

- For $Z \sim \text{IBP}(\alpha)$, binary latent factor model

- Compute posterior distribution $P(Z|X)$
Gibbs sampling

• Sequentially sample feature assignments

\[ P(z_{ik}|X, z_{(-i)k}) \propto p(x_i|X_{-i}, Z)P(z_{ik}|z_{(-i)k}) \]

• IBP provides

\[ P(z_{ik}|X, z_{(-i)k}) = \frac{m_{k,-i}}{N} \]

• Draw \( \text{Poisson}(\frac{\alpha}{N}) \) new features
Example: seven segment LED displays

- 256 images, each $9 \times 5$ pixels
- 16 instances of 16 characters, corrupted with noise
- Range of binary representations
  - most sparse: 1 bit per character, 16 dimensions
  - lowest dimension: $\approx 5$ bits per character, 7 dimensions
Example: seven segment LED displays
Conclusion

- Strategy for model selection from nonparametric Bayes: prior over combinatorial structures of variable dimension

<table>
<thead>
<tr>
<th>Structure</th>
<th>Distribution</th>
<th>Models</th>
</tr>
</thead>
<tbody>
<tr>
<td>partitions</td>
<td>CRP</td>
<td>infinite mixture models</td>
</tr>
<tr>
<td>binary matrices</td>
<td>IBP</td>
<td>infinite binary latent factors</td>
</tr>
<tr>
<td>($\times \mathbb{Z}$)</td>
<td></td>
<td>infinite CVQ</td>
</tr>
<tr>
<td>($\times \mathbb{R}$)</td>
<td></td>
<td>infinite sparse PCA</td>
</tr>
</tbody>
</table>

- Other exchangeable distributions?
Equivalent processes...

- The Indonesian Rijstafel Process
- The Mongolian Barbeque Process
- The Swedish Smorgasbord Process
- The All-You-Can-Eat Seafood Process
  - justifies appearance of Poisson