Nonparametric Bayesian Models

-- Learning/Reasoning in Open Possible Worlds

Eric Xing
Lecture 19, November 12, 2014
Clustering
Image Segmentation

- **How to segment images?**
  - Manual segmentation (very expensive)
  - Algorithm segmentation
    - K-means
    - Statistical mixture models
    - Spectral clustering

- **Problems with most existing algorithms**
  - Ignore the spatial information
  - Perform the segmentation one image at a time
  - Need to specify the number of segments *a priori*
Object Recognition and Tracking

(1.8, 7.4, 2.3)  (1.9, 9.0, 2.1)  (1.9, 6.1, 2.2)
(0.9, 5.8, 3.1)  (0.7, 5.1, 3.2)  (0.6, 5.9, 3.2)

© Eric Xing @ CMU, 2014
The Evolution of Science

Research circles

Research topics

PNAS papers

1900

2000

© Eric Xing @ CMU, 2014
A Classical Approach

- Clustering as Mixture Modeling

- Then "model selection"
Model Selection vs. Posterior Inference

- Model selection
  - "intelligent" guess: ???
  - cross validation: data-hungry 😐
  - information theoretic:
    - AIC
    - TIC
    - MDL:
  - Bayes factor: need to compute data likelihood

- Posterior inference:
  we want to handle uncertainty of model complexity explicitly

\[ p(M | D) \propto p(D | M) p(M) \]

\[ M \equiv \{ \theta, K \} \]

- we favor a distribution that does not constrain \( M \) in a "closed" space!
Outline

- Motivation and challenge

- Dirichlet Process and Infinite Mixture
  - Formulation
  - Approximate Inference algorithm
  - Example: population clustering

- Hierarchical Dirichlet Process and Multi-Task Clustering
  - Formulation
  - Application: joint multiple population clustering

- Dynamic Dirichlet Process
  - Temporal DPM
  - Application: evolutionary clustering of documents

- Summary
Clustering

- How to label them?
- How many clusters???
Random Partition of Probability Space
Dirichlet Process

A CDF, $G$, on possible worlds of random partitions follows a Dirichlet Process if for any measurable finite partition $(\phi_1, \phi_2, \ldots, \phi_m)$:

$$(G(\phi_1), G(\phi_2), \ldots, G(\phi_m)) \sim \text{Dirichlet}(\alpha G_0(\phi_1), \ldots, \alpha G_0(\phi_m))$$

where $G_0$ is the base measure and $\alpha$ is the scale parameter.

Thus a Dirichlet Process $G$ defines a distribution of distribution.
$$G \sim \text{DP}(\alpha, G_0)$$

$$G = \sum_{k=1}^{\infty} \pi_k \delta(\theta_k)$$

$$\theta_k \sim G_0$$

$$\sum_{k=1}^{\infty} \pi_k = 1$$

$$\pi_k = \beta_k \prod_{j=1}^{k-1} (1 - \beta_j)$$

$$\beta_k \sim \text{Beta}(1, \alpha)$$

$$\Pi_{j=1}^{k-1} (1 - \beta_j)$$

$$\beta_k$$

$$\pi_k$$

<table>
<thead>
<tr>
<th>Location</th>
<th>Mass</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3</td>
<td>0.8</td>
</tr>
<tr>
<td>0.6</td>
<td>0.5</td>
</tr>
<tr>
<td>0</td>
<td>0.4</td>
</tr>
</tbody>
</table>

© Eric Xing @ CMU, 2014
Chinese Restaurant Process

\[ P(c_i = k \mid c_{-i}) = \begin{cases} 
\frac{1}{1+\alpha} & \text{if } k = 1 \\
\frac{1}{2+\alpha} & \text{if } k = 2 \\
\frac{1}{3+\alpha} & \text{if } k = 3 \\
\frac{m_1}{i+\alpha-1} & \text{if } k > 3 
\end{cases} \]

CRP defines an exchangeable distribution on partitions over an (infinite) sequence of samples, such a distribution is formally known as the Dirichlet Process (DP)
Graphical Model Representations of DP

- The CRP construction
- The Stick-breaking construction
Ancestral Inference

- Better recovery of the ancestors leads to better haplotyping results (because of more accurate grouping of common haplotypes)
- True haplotypes are obtainable with high cost, but they can validate model more subjectively (as opposed to examining saliency of clustering)
- Many other biological/scientific utilities

Essentially a clustering problem, but …
Example: DP-haplotyper [Xing et al, 2004]

- Clustering human populations

\[
\begin{align*}
\alpha & \quad G_0 \\
\alpha & \quad G \\
G & \quad A \\
G & \quad \theta \\
A & \quad H_{n1} \\
A & \quad H_{n2} \\
\theta & \quad G_n \\
N & \quad K 
\end{align*}
\]

\[
\text{DP} \quad \{ \text{infinite mixture components} \} \\
\text{Likelihood model} \quad \{ \text{for population haplotypes} \} \\
\{ \text{for individual haplotypes and genotypes} \}
\]

- Inference: Markov Chain Monte Carlo (MCMC)
  - Gibbs sampling
  - Metropolis Hasting
The DP Mixture of Ancestral Haplotypes

- The customers around a table in CRP form a cluster
  - associate a mixture component (i.e., a population haplotype) with a table
  - sample \(\{a, \theta\}\) at each table from a base measure \(G_0\) to obtain the population haplotype and nucleotide substitution frequency for that component

- With \(p(h|\{A, \theta\})\) and \(p(g|h_1,h_2)\), the CRP yields a posterior distribution on the number of population haplotypes (and on the haplotype configurations and the nucleotide substitution frequencies)
Inheritance and Observation Models

- Single-locus mutation model
  \[ A_{C_i} \rightarrow H_{i_e} \]
  \[ P_H(h_i | a_i, \theta) = \begin{cases} 
  \theta & \text{for } h_i = a_i \\
  \frac{1 - \theta}{|B| - 1} & \text{for } h_i \neq a_i 
  \end{cases} \]
  \[ \rightarrow h_i = a_i \text{ with prob. } \theta \]

- Noisy observation model
  \[ H_{i_1}, H_{i_2} \rightarrow G_i \]
  \[ P_G(g | h_1, h_2) : g_i = h_{1,i} \oplus h_{2,i} \text{ with prob. } \lambda \]
MCMC for Haplotype Inference

- Gibbs sampling for exploring the posterior distribution under the proposed model
  - Integrate out the parameters such as $\theta$ or $\lambda$, and sample $c_{i_e}$, $a_k$ and $h_{i_e}$

\[
p(c_{i_e} = k \mid c_{[-i_e]}, h, a) \propto p(c_{i_e} = k \mid c_{[-i_e]}) p(h_{i_e} \mid a_k, h_{[-i_e]}, c)
\]

- Gibbs sampling algorithm: draw samples of each random variable to be sampled given values of all the remaining variables
MCMC for Haplotype Inference

1. Sample $c_{ie}^{(j)}$, from

\[
p(c_{ie}^{(j)} = k | c^{[-j,i,e]}, h, a) \\
\propto p(c_{ie}^{(j)} = k | c^{[-j,i,e]}, m, n) p(h_{ie}^{(j)} | a_k, c, h^{[-j,i,e]}) \\
\propto (m_{jk}^{[-j,i,e]} + \tau \beta_k) p(h_{ie}^{(j)} | a_k, l_k^{[-j,i,e]}), \text{ for } k = 1, ..., K + 1
\]

2. Sample $a_k$ from

\[
p(a_k, t | c, h) \propto \prod_{j,i,e | c_{ie}^{(j)} = k} p(h_{ie,t}^{(j)} | a_k, t, l_{k,t}^{(j)}) \\
= \frac{\Gamma(\alpha_h + l_{k,t})\Gamma(\beta_h + l'_{k,t})}{\Gamma(\alpha_h + \beta_h + m_k)(|B| - 1)l'_{k,t}} R(\alpha_h, \beta_h)
\]

3. Sample $h_{ie}^{(j)}$ from

\[
p(h_{ie,t}^{(j)} | h^{[-i,e,t]}, c, a, g)
\]

- For DP scale parameter $\alpha$: a vague inverse Gamma prior
Convergence of Ancestral Inference
DP vs. Finite Mixture via EM

![Graph showing individual error for data sets 1 to 5, comparing DP and EM methods. The x-axis represents the data sets, and the y-axis represents the individual error. The graph indicates that DP has lower error than EM for data sets 1 and 2, while EM has lower error for data sets 3, 4, and 5.](image-url)
Outline

- Motivation and challenge
- Dirichlet Process and Infinite Mixture
  - Formulation
  - Approximate Inference algorithm
  - Example: population clustering
- Hierarchical Dirichlet Process and Multi-Task Clustering
  - Formulation
  - Application: joint multiple population clustering
- Dynamic Dirichlet Process
  - Temporal DPM
  - Application: evolutionary clustering of documents
- Summary
Multi-population Genetic Demography

- Pool everything together and solve 1 hap problem?
  - --- ignore population structures
- Solve 4 hap problems separately?
  - --- data fragmentation
- Co-clustering … solve 4 *coupled* hap problems jointly
Hierarchical Dirichlet Process

- Two level Pólya urn scheme
  - At the $i$-th step in $j$-th "group",

- Choose $\theta_k$ with prob. $\frac{m_{jk}}{\sum_k m_{jk} + \alpha_0}$
- Go to the upper level DP with prob. $\frac{\alpha_0}{\sum_k m_{jk} + \alpha_0}$
- Choose $\theta_k$ with prob. $\frac{n_k}{\sum n_k + \gamma}$
- Draw a new sample with prob. $\frac{\gamma}{\sum n_k + \gamma}$
Results - Simulated Data

- 5 populations with 20 individuals each (two kinds of mutation rates)
- 5 populations share parts of their ancestral haplotypes
- the sequence length = 10

Haplotype error
Results - International HapMap DB

- Different sample sizes, and different # of sub-populations
Topic Models for Images

“beach”

Latent Dirichlet Allocation (LDA)
Infinite Topic Model for Image

1. **A single image with \( k \) topics**
   - \( \alpha \)
   - \( H \)
   - \( \pi_j \)
   - \( \theta \)
   - \( z_i \)
   - \( w_i \)
   - **An LDA**

2. **A single image with inf-topic**
   - \( \alpha \)
   - \( H \)
   - \( \pi_j \)
   - \( \theta \)
   - \( z_i \)
   - \( w_i \)
   - **A DP**

3. **\( J \) images with inf-topic**
   - \( \gamma \)
   - \( \beta \)
   - \( \alpha \)
   - \( H \)
   - \( \pi_j \)
   - \( z_i \)
   - \( w_i \)
   - **An HDP**
Outline

- Motivation and challenge
- Dirichlet Process and Infinite Mixture
  - Formulation
  - Approximate Inference algorithm
  - Example: population clustering
- Hierarchical Dirichlet Process and Multi-Task Clustering
  - Formulation
  - Application: joint multiple population clustering
- Dynamic Dirichlet Process
  - Temporal DPM
  - Application: evolutionary clustering of documents
- Summary
Evolutionary Clustering

- Adapts the number of mixture components over time
  - Mixture components can die out
  - New mixture components are born at any time
  - Retained mixture components parameters evolve according to a Markovian dynamics
● The Recurrent Chinese Restaurant Process

- The restaurant operates in epochs
- The restaurant is closed at the end of each epoch
- The state of the restaurant at time epoch $t$ depends on that at time epoch $t-1$
  - Can be extended to higher-order dependencies.
\[ N_{2,3} = \sum_{w=1}^{W} \left( e^{\frac{-w}{\lambda}} N_{k,t-w} \right) \]
TDPM Generative Power

DPM
\[ W = T \]
\[ \lambda = \infty \]

TDPM
\[ W = 4 \]
\[ \lambda = 0.4 \]

Independent DPMs
\[ W = 0 \]
\[ \lambda = ? \] (any)
The Big Picture

K-means

Fixed-dimensions Dynamic clustering

DPM

TDPM

© Eric Xing @ CMU, 2014
Summary

- A non-parametric Bayesian model for Pattern Uncovery
  - Finite mixture model of latent patterns (e.g., image segments, objects)
    - infinite mixture of prototypes: alternative to model selection
    - hierarchical infinite mixture
    - temporal infinite mixture model

- Applications in general data-mining …
Shortcomings of Hidden Markov Model

- HMM models capture dependences between each state and only its corresponding observation
  - NLP example: In a sentence segmentation task, each segmental state may depend not just on a single word (and the adjacent segmental stages), but also on the (non-local) features of the whole line such as line length, indentation, amount of white space, etc.

- Mismatch between learning objective function and prediction objective function
  - HMM learns a joint distribution of states and observations $P(Y, X)$, but in a prediction task, we need the conditional probability $P(Y|X)$