Infinite Mixture and Dirichlet Process

Probabilistic Graphical Models (10-708)

Lecture 20, Nov 28, 2007

Eric Xing

Reading:

Clustering

p(IK)
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)

Modeling The Mind ...

Latent brain processes:

fMRI scan:

Read sentence
View picture
Decide whether consistent
The Evolution of Science

Partially Observed, Open and Evolving Possible Worlds

- Unbounded # of objects/trajectories
- Changing attributes
- Birth/death, merge/split
- Relational ambiguity

The parametric paradigm:

- Finite
- Structurally unambiguous

How to open it up?
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: Parsimony, Ockam's Razor
  - Bayes factor: need to compute data likelihood

- Posterior inference:
  - we want to handle uncertainty of model complexity explicitly
    \[ p(\mathcal{M} | D) \propto p(D | \mathcal{M}) p(\mathcal{M}) \]
    \[ \mathcal{M} = \{\theta, K\} \]
  - we favor a distribution that does not constrain \( \mathcal{M} \) in a "closed" space!
Two "Recent" Developments

- First order probabilistic languages (FOPLs)
  - Examples: PRM, BLOG …
  - Lift graphical models to "open" world (#rv, relation, index, lifespan …)
  - Focus on complete, consistent, and operating rules to instantiate possible worlds, and formal language of expressing such rules
  - Operational way of defining distributions over possible worlds, via sampling methods

- Bayesian Nonparametrics
  - Examples: Dirichlet processes, stick-breaking processes …
  - From finite, to infinite mixture, to more complex constructions (hierarchies, spatial/temporal sequences, …)
  - Focus on the laws and behaviors of both the generative formalisms and resulting distributions
  - Often offer explicit expression of distributions, and expose the structure of the distributions --- motivate various approximate schemes

Clustering

- How to label them?
- How many clusters???
Genetic Demography

- Are there genetic prototypes among them?
- What are they?
- How many? (how many ancestors do we have?)

Genetic Polymorphisms

The ABO Blood System

<table>
<thead>
<tr>
<th>Blood Type (genotype)</th>
<th>Type A (AA, AD)</th>
<th>Type B (BB, BD)</th>
<th>Type AB (AB)</th>
<th>Type O (OO)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Red Blood Cell Surface Proteins (phenotype)</td>
<td>A, agglutinogens only</td>
<td>B, agglutinogens only</td>
<td>A and B, agglutinogens</td>
<td>No agglutinogens</td>
</tr>
<tr>
<td>Plasma Antibodies (phenotype)</td>
<td>A agglutinin only</td>
<td>B agglutinin only</td>
<td>No agglutinin</td>
<td>A and B agglutinins</td>
</tr>
</tbody>
</table>
Biological Terms

- **Genetic polymorphism**: a difference in DNA sequence among individuals, groups, or populations

- **Single Nucleotide Polymorphism (SNP)**: DNA sequence variation occurring when a single nucleotide - A, T, C, or G - differs between members of the species
  - Each variant is called an “allele”
  - Almost always bi-allelic
  - Account for most of the genetic diversity among different (normal) individuals, e.g. drug response, disease susceptibility

From SNPs to Haplotypes

- Alleles of adjacent SNPs on a chromosome form **haplotypes**

- Powerful in the study of **disease association or genetic evolution**
Haplotype and Genotype

- A collection of alleles derived from the same chromosome

<table>
<thead>
<tr>
<th>Genotypes</th>
<th>Haplotypes</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 13</td>
<td>2 13</td>
</tr>
<tr>
<td>1 6</td>
<td>6 1</td>
</tr>
<tr>
<td>9 15</td>
<td>9 15</td>
</tr>
<tr>
<td>4 17</td>
<td>17 4</td>
</tr>
<tr>
<td>1 9</td>
<td>1 9</td>
</tr>
<tr>
<td>2 6</td>
<td>6 2</td>
</tr>
<tr>
<td>9 17</td>
<td>9 17</td>
</tr>
<tr>
<td>2 12</td>
<td>2 12</td>
</tr>
<tr>
<td>7 12</td>
<td>12 7</td>
</tr>
<tr>
<td>6 14</td>
<td>14 6</td>
</tr>
<tr>
<td>1 7</td>
<td>7 1</td>
</tr>
<tr>
<td>18 18</td>
<td>18 18</td>
</tr>
<tr>
<td>1 4</td>
<td>1 4</td>
</tr>
<tr>
<td>10 10</td>
<td>10 10</td>
</tr>
</tbody>
</table>

Chromosome phase is unknown | Chromosome phase is known

Haplotype Re-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 …
A Finite (Mixture of ) Allele Model

- The probability of a genotype $g$:
  
  $$p(g) = \sum_{h_1, h_2 \in \mathcal{H}} p(h_1, h_2) p(g \mid h_1, h_2)$$

- Standard settings:
  - $\mathcal{H} = K << 2^J$ fixed-sized population haplotype pool
  - $p(h_1, h_2) = p(h_1)p(h_2) = f_1f_2$ Hardy-Weinberg equilibrium

- Problem: $K$? $\mathcal{H}$?

A Infinite (Mixture of ) Allele Model

- How?
  - Via a nonparametric hierarchical Bayesian formalism!
### Stick-breaking Process

\[ G \sim \text{DP}(\alpha, G_0) \]

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

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

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

<table>
<thead>
<tr>
<th>( k )</th>
<th>( \beta_k )</th>
<th>( \pi_k )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.4</td>
<td>0.4</td>
</tr>
<tr>
<td>1</td>
<td>0.6</td>
<td>0.5</td>
</tr>
<tr>
<td>2</td>
<td>0.3</td>
<td>0.8</td>
</tr>
</tbody>
</table>

\[ \text{Location} \]

### Graphical Model

\[ \beta_k \sim P(\nu) \]

\[ \theta \sim \text{Gamma}(\alpha, \beta) \]

\[ \mathcal{G} \]

\[ H_n \]

\[ A_k \]

\[ \theta \]

\[ N \]

\[ Z \]

\[ x \]

\[ \alpha \]

\[ \beta \]
The DP Mixture of Ancestral Haplotypes

- The customers around a table 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)
DP-haplotyper

- Inference: Markov Chain Monte Carlo (MCMC)
  - Gibbs sampling
  - Metropolis Hasting

\[ G_0 \sim \text{Unif}(a) \cdot \prod_j \text{Beta}(\theta_j) \]

- Choice of base measure:

- Nucleotide-substitution model:
  \[ p(h_i | a_i, \theta) = \prod_j p(h_j | a_j, \theta_j) \]
  where \[ p(h_j | a_j, \theta_j) = \begin{cases} \theta_j & \text{if } h_j = a_j \\ 1 - \theta_j & \text{if } h_j = a_j \end{cases} \]

- Noisy genotyping model:
  \[ p(g_i | h_i, h_j) = \prod_j p(g_j | h_j, h_i) \]
  where \[ p(g_j | h_j, h_i) = \begin{cases} \gamma & \text{if } h_j \oplus h_i = g_j \\ 1-\gamma & \text{if } h_j \oplus h_i \neq g_j \end{cases} \]
Gibbs sampling

Starting from some initial haplotype reconstruction $H^{(0)}$, pick a first table with an arbitrary $a_i^{(0)}$, and form initial population-hap pool $A^{(0)} = \{a_i^{(0)}\}$:

i) Choose an individual $i$ and one of his/her two haplotypes $t$, uniformly and at random, from all ambiguous individuals;

ii) Sample $c_i^{(t+1)}$ from $p(c_i^{(t+1)} | c_i^{(t)}, H^{(t)}, A^{(t)})$, update $c^{(t+1)}$;

iii) Sample $a_k^{(t+1)}$, where $k = c_i^{(t+1)}$, from $p(a_k^{(t+1)} | \forall h_i^{(t)}$, s.t. $c_i^{(t+1)} = k$); update $A^{(t+1)}$;

iii) Sample $h_i^{(t+1)}$ from $p(h_i^{(t+1)} | c_i^{(t+1)}, H^{(t)}, A^{(t+1)})$, update $H^{(t+1)}$.

Convergence of Ancestral Inference

Eric Xing 26
Haplotyping Error

The Gabriel data