Clustering

- How to label them?
- How many clusters???
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) (0.7, 5.1, 3.2) (0.6, 5.9, 3.2)

\( t=1 \) \( t=2 \) \( t=3 \)
The Evolution of Science

Research circles
Research topics
PNAS papers

Clustering as (nonparametric) Bayesian Inference

- Cluster = Mixture
- Each data point is generated from one Mixture

How Many clusters
- Cross Validation
- Data hungry!
- Information theoretic
- AIC, MDL, etc.
- Non-Parametric
  - Manage model uncertainty
  - Integrate over different clustering configurations
  - Base for many interesting extensions
Outline

- Dirichlet Process: From finite mixture to Infinite mixture model
  - The Chinese Restaurant Process
  - Example: heliotype inference

- Intro to Markov Chain Monte Carlo
  - Gibbs sampling

- Dynamic Dirichlet Process
  - The recurrent Chinese Restaurant Process
  - Example: Application: evolutionary clustering of documents

Clustering

- How to label them?
- How many clusters???
Mixture model: The Generative Process

- For data point $x_i$
  - Sample $c_i \sim \text{Multi}(\pi)$
  - Sample $x_i \sim f(\phi_{c_i})$

Cluster parameters DO NOT evolve over time
Number of clusters DOES NOT grow with the data

The Chinese Restaurant Process

- Customers correspond to data points
- Tables correspond to clusters/mixture components
- Dishes correspond to parameter of the mixtures
The Chinese Restaurant Process

For data point $x_i$:
- Choose table $j \propto N_j$ and Sample $x_i \sim f(\phi_j)$
- Choose a new table $K+1 \propto \alpha$
  - Sample $\phi_{K+1} \sim G_0$ and Sample $x_i \sim f(\phi_{K+1})$

The rich gets richer effect
CANNOT handle sequential data

More formally …

- This is a Dirichlet Process Mixture Model
- Three equivalent constructions:

Measure over Measures
Chinese Restaurant Process
The Stick-breaking construction

© Eric Xing @ CMU, 2006-2011
Dirichlet Process

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

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

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

Thus a Dirichlet Process $G$ defines a distribution of distribution.

Stick-breaking Process

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

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

$$\theta_k \sim G_0$$

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

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

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

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

$| \beta_k | \ | \pi_k |

0 | 0.4 | 0.4
0.6 | 0.5 | 0.3
0.3 | 0.8 | 0.24

$G_0$
Chinese Restaurant Process

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

\[
P(c_i = k \mid c_j) = \begin{cases} 
1 & \alpha \\
\frac{1}{1+\alpha} & 0 \\
\frac{1}{2+\alpha} & 0 \\
\frac{1}{3+\alpha} & 0 \\
m_i & \alpha \\
\frac{m_i}{i+\alpha-1} & \alpha \\
\end{cases}
\]

Case study: ancestral Inference

Essentially a clustering problem, but ...

- 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
Background: haplotype ambiguity

A Finite (Mixture of ) Allele Model

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

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

- Problem: \( K \) ? \( \mathcal{H} \) ?
Example: DP-haplotyper [Xing et al, 2004]

- Clustering human populations
  - Infinite mixture components (for population haplotypes)
  - Likelihood model (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)
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 \( h_{i_e} \)

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

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 \( c_{i_e}^{(j)} \) from

\[
p(c_{i_e}^{(j)} = k | c_{[-i_e]}, h, a) \propto p(c_{i_e}^{(j)} = k | c_{[-i_e]}) \cdot p(h_{i_e}^{(j)} | a_k, c, h_{[-i_e]}, a_k, \theta, \lambda, \alpha)
\]

2. Sample \( a_k \) from

\[
p(a_k | c, h) \propto \prod_{j: a_k^j = k} p(h_{i_e}^{(j)} | a_k, \theta, \lambda, \alpha)
\]

3. Sample \( h_{i_e}^{(j)} \) from

\[
p(h_{i_e}^{(j)} | c_{[-i_e]}^{(j)}, c, a, g)
\]

For DP scale parameter \( \alpha \): a vague inverse Gamma prior
Convergence of Ancestral Inference

DP vs. Finite Mixture via EM
Outline

- Dirichlet Process: From finite mixture to Infinite mixture model
  - The Chinese Restaurant Process
  - Example: heliotype inference
- Intro to Markov Chain Monte Carlo
  - Gibbs sampling
- Dynamic Dirichlet Process
  - The recurrent Chinese Restaurant Process
  - Example: Application: evolutionary clustering of documents

Monte Carlo methods

- Draw random samples from the desired distribution
- Yield a stochastic representation of a complex distribution
  - Marginals and other expectations can be approximated using sample-based averages
    \[ E[f(x)] = \frac{1}{N} \sum_{i=1}^{N} f(x^{(i)}) \]
- Asymptotically exact and easy to apply to arbitrary models
- Challenges:
  - How to draw samples from a given dist. (not all distributions can be trivially sampled)?
  - How to make better use of the samples (not all samples are useful, or equally useful, see an example later)?
  - How to know we’ve sampled enough?
Example: naive sampling

- Sampling: Construct samples according to probabilities given in a BN.

Alarm example: (Choose the right sampling sequence)
1) Sampling: $P(B) = 0.001, 0.999$ suppose it is false, $B_0$. Same for $E_0$. $P(A|B_0, E_0) = 0.001, 0.999$ suppose it is false...

2) Frequency counting: In the samples right, $P(J|A_0) = P(J,A_0)/P(A_0) = 1/9, 8/9$.

Example: naive sampling

- Sampling: Construct samples according to probabilities given in a BN.

Alarm example: (Choose the right sampling sequence)
1) Sampling: $P(B) = 0.001, 0.999$ suppose it is false, $B_0$. Same for $E_0$. $P(A|B_0, E_0) = 0.001, 0.999$ suppose it is false...

2) Frequency counting: In the samples right, $P(J|A_0) = P(J,A_0)/P(A_0) = 1/9, 8/9$. 
Example: naive sampling

- Sampling: Construct samples according to probabilities given in a BN.

Alarm example: (Choose the right sampling sequence)

3) what if we want to compute $P(J|A1)$? we have only one sample ...

4) what if we want to compute $P(J|B1)$? No such sample available!
$P(J|A1)=P(J,B1)/P(B1)$ can not be defined.

For a model with hundreds or more variables, rare events will be very hard to garner enough samples even after a long time or sampling ...

Monte Carlo methods (cond.)

- Direct Sampling
  - We have seen it.
  - Very difficult to populate a high-dimensional state space

- Rejection Sampling
  - Create samples like direct sampling, only count samples which is consistent with given evidences.

- ....

- Markov chain Monte Carlo (MCMC)
Markov chain Monte Carlo

- Samples are obtained from a **Markov chain** (of sequentially evolving distributions) whose **stationary distribution** is the desired \( p(x) \).

- Construct a Markov chain whose stationary distribution is the target density \( P(X|e) \).
- Run for \( T \) samples (burn-in time) until the chain converges/mixes/reaches stationary distribution.
- Then collect \( M \) (correlated) samples \( x_m \).
- Key issues:
  - Designing proposals so that the chain mixes rapidly.
  - Diagnosing convergence.

Gibbs sampling

- Gibbs sampling is an MCMC algorithm that is especially appropriate for inference in graphical models.

- The procedure
  - we have variable set \( X = \{x_1, x_2, \ldots, x_n\} \) for a GM
  - at each step one of the variables \( X_i \) is selected (at random or according to some fixed sequences), denote the remaining variables as \( X_{-i} \), and its current value as \( x_{-i}(t-1) \)
  - Using the "alarm network" as an example, say at time \( t \) we choose \( X_E \) and we denote the current value assignments of the remaining variables, \( X_{-E} \), obtained from previous samples, as \( x_{-E}^{(t-1)} = \{x_d^{(t-1)}, x_d^{(t-1)}, x_d^{(t-1)}, x_{-d}^{(t-1)}\} \)
  - the conditional distribution \( p(X_i|X_{-i}^{(t-1)}) \) is computed
  - a value \( x_i^{(t)} \) is sampled from this distribution
  - the sample \( x_i^{(t)} \) replaces the previous sampled value of \( X_i \) in \( X \).
  - i.e., \( x^{(t)} = x_{-E}^{(t-1)} \cup x_E^{(t)} \)
Why Gibbs Sampling is Simple

- **Markov-Blanket**
  - A variable is independent from others, given its parents, children and children’s parents. d-separation.
  \( \Rightarrow p(X|X_i) = p(X|MB(X_i)) \)

- **Gibbs sampling**
  - Create a random sample.
  Every step, choose one variable and sample it by \( P(X|MB(X)) \) based on previous sample.

Example: alarm network again

- To calculate \( P(J|B_1,M_1) \)
- Choose \( (B_1,E_0,A_1,M_1,J_1) \) as a start
- **Evidences** are \( B_1, M_1 \), **variables** are \( A, E, J \).
- Choose next variable as \( A \)
- Sample \( A \) by \( P(A|MB(A)) = P(A|B_1, E_0, M_1, J_1) \) suppose to be false.
- \( (B_1, E_0, A_0, M_1, J_1) \)
- Choose next random variable as \( E \), sample \( E \sim P(E|B_1, A_0) \)
- ...
Example

\[ P(J_1 | B_1, M_1) = 0.90 \]
\[ P(J_1 | E_1, M_0) = 0.14 \]
\[ P(E_1 | J_1) = 0.01 \]
\[ P(E_1 | M_1) = 0.04 \]
\[ P(E_1 | M_1, J_1) = 0.17 \]
Outline

- Dirichlet Process: From finite mixture to Infinite mixture model
  - The Chinese Restaurant Process
  - Example: heliotype inference

- Intro to Markov Chain Monte Carlo
  - Gibbs sampling

- Dynamic Dirichlet Process
  - The recurrent Chinese Restaurant Process
  - Example: Application: evolutionary clustering of documents

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

© Eric Xing @ CMU, 2006-2011
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.

Generative Process

- Customers at time \( T=1 \) are seated as before:
  - Choose table \( j \propto N_{1,1} \) and Sample \( x_i \sim f(\phi_{1,i}) \)
  - Choose a new table \( K+1 \propto \alpha \)
  - Sample \( \phi_{K+1,1} \sim G_0 \) and Sample \( x_i \sim f(\phi_{K+1,i}) \)

Dish eaten at table 3 at time epoch 1
OR the parameters of cluster 3 at time epoch 1
\[ T_1 = 1 \]
\[ T_2 = 2 \]

\[ N_{1,1} = 2 \quad N_{2,1} = 3 \quad N_{3,1} = 1 \]
Sample $\phi_{1,2} \sim P(\cdot | \phi_{1,1})$
At the end of epoch 2

Newly born cluster

Died out cluster

$\phi_{1,1}$ $\phi_{2,1}$ $\phi_{3,1}$

$T=1$

$\phi_{1,2}$ $\phi_{2,2}$ $\phi_{3,1}$ $\phi_{4,2}$

$N_{1,1}=2$ $N_{2,1}=3$ $N_{3,1}=1$

$T=2$

$N_{1,2}=2$ $N_{2,2}=2$ $N_{4,2}=1$

$T=3$
Temporal DPM

- Can be extended to model higher-order dependencies
- Can decay dependencies over time
  - Pseudo-counts for table \( k \) at time \( t \) is

\[
\sum_{w=1}^{W} \left( e^{\frac{-w}{\lambda}} N_{k,t-w} \right)
\]

- History size
- Decay factory
- Number of customers sitting at table \( k \) at time epoch \( t-w \)

\[
N_{2,3} = \sum_{w=1}^{W} \left( e^{\frac{-w}{\lambda}} N_{k,t-w} \right)
\]
Results: NIPS 12

- Building a simple dynamic topic model
- Chain dynamics is as before
- Emission model for document $x_{k,t}$ is:
  - Project $\phi_k$ over the simplex
  - Sample $x_{k,t} | c_{t,i} \sim \text{Multinomial}(\text{Logistic}(\phi_{k,i}))$
- Unlike LDA here a document belongs to one topic
- Use this model to analyze NIPS12 corpus
  - Proceeding of NIPS conference 1987-1999
Theory of MCMC (optional)

- **Definition:** Markov Chains
  - Given an n-dimensional state space
  - Random vector $X = (x_1, ..., x_n)$
  - $x^{(0)} = x$ at time-step $t$
  - $x^{(0)}$ transitions to $x^{(t+1)}$ with prob
    \[ P(x^{(t+1)} | x^{(0)}, ..., x^{(t)}) = T(x^{(t+1)} | x^{(t)}) = T(x^{(t)} \to x^{(t+1)}) \]
- **Homogenous:** chain determined by state $x^{(0)}$, fixed transition kernel $T$ (rows sum to 1)
- **Equilibrium:** $\pi(x)$ is a **stationary (equilibrium) distribution** if
  \[ \pi(x') = \sum_x \pi(x) T(x \to x') \]
  i.e., is a left eigenvector of the transition matrix $\pi^T = \pi T$.
Markov Chains

- An MC is **irreducible** if transition graph connected
- An MC is **aperiodic** if it is not trapped in cycles
- An MC is **ergodic** (regular) if you can get from state $x$ to $x'$ in a finite number of steps.
- **Detailed balance**: $\text{prob}(x(t) \rightarrow x(i-1)) = \text{prob}(x(i-1) \rightarrow x(t))$
  \[
p(x(t))T(x(i-1) \mid x(t)) = p(x(i-1))T(x(t) \mid x(i-1))
\]
summing over $x(i-1)$
  \[
p(x(t)) = \sum_{x(i-1)} p(x(i-1))T(x(t) \mid x(i-1))
\]
- Detailed balance $\rightarrow$ stationary dist exists

MCMC Via Metropolis-Hastings

- Treat the target distribution as stationary distribution
- Sample from an easier proposal distribution, followed by an acceptance test
- This induces a transition matrix that satisfies detailed balance
  - MH proposes moves according to $Q(x' \mid x)$ and accepts samples with probability $A(x' \mid x)$.
  - The induced transition matrix is $T(x \rightarrow x') = Q(x' \mid x)A(x' \mid x)$
  - Detailed balance means
    \[
    \pi(x)Q(x' \mid x)A(x' \mid x) = \pi(x')Q(x \mid x')A(x \mid x')
    \]
  - Hence the acceptance ratio is
    \[
    A(x' \mid x) = \min \left( 1, \frac{\pi(x')Q(x' \mid x')}{\pi(x)Q(x \mid x')} \right)
    \]
Gibbs sampling

- Gibbs sampling is a special case of MH
- The transition matrix updates each node one at a time using the following proposal:
  \[ Q((x_j, x_{-j}) \rightarrow (x_j', x_{-j})) = p(x_j' | x_{-j}) \]
- This is efficient since for two reasons
  - It leads to samples that is always accepted
    \[
    \alpha((x_j, x_{-j}) \rightarrow (x_j', x_{-j})) = \min\left( \frac{p(x_j', x_{-j})q(x_j' \mid x_{-j})}{p(x_j, x_{-j})q(x_j \mid x_{-j})}, 1 \right)
    \]
    Thus
    \[
    T((x_j, x_{-j}) \rightarrow (x_j', x_{-j})) = p(x_j' | x_{-j})
    \]
  - It is efficient since \( p(x_j' | x_{-j}) \) only depends on the values in \( x_j \)’s Markov blanket