Applied Bayesian Nonparametrics

Special Topics in Machine Learning
Brown University CSCI 2950-P, Fall 2011

November 1: Hierarchical Dirichlet Process
Hidden Markov Models & Hidden Markov Trees
Static Clustering

- How many clusters are there?
- How should model complexity grow as more data is observed?

Mixture of Gaussians
Dirichlet Process (DP) Mixtures

\[ p(y) = \sum_{k=1}^{\infty} \pi_k \mathcal{N}(y \mid \mu_k, \Lambda_k) \]

- Dirichlet processes define a prior distribution on weights assigned to mixture components:

\[
\pi_k = v_k \left(1 - \sum_{\ell=1}^{k-1} \pi_\ell\right) = v_k \prod_{\ell=1}^{k-1} (1 - v_\ell)
\]

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

Ferguson 1973, Sethuraman 1994
Temporal Segmentation

- Markov switching models for time series data
- Cluster based on underlying mode dynamics

Hidden Markov Model

True mode sequence

Observations

$z_1 \rightarrow z_2 \rightarrow z_3 \rightarrow \cdots \rightarrow z_T$

$y_1 \rightarrow y_2 \rightarrow y_3 \rightarrow \cdots \rightarrow y_T$

modes

observations
Hidden Markov Models

\[ P = \begin{bmatrix}
\pi_1 \\
\pi_2 \\
\vdots \\
\pi_K
\end{bmatrix} \]

\[ z_t \sim \pi_{z_{t-1}} \]

\[ y_t \sim F(\theta_{z_t}) \]

\[
\begin{array}{c}
\text{Time} \\
1 & 2 & 3 & \ldots \\
1 & \circ & \circ & \circ & \circ & \circ & \circ \\
2 & \circ & \circ & \circ & \circ & \circ & \circ \\
3 & \circ & \circ & \circ & \circ & \circ & \circ \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
K & \circ & \circ & \circ & \circ & \circ & \circ \\
\end{array}
\]

\[
\begin{array}{c}
\text{Mode} \\
1 & 2 & 3 & \ldots \\
1 & \circ & \circ & \circ & \circ & \circ & \circ \\
2 & \circ & \circ & \circ & \circ & \circ & \circ \\
3 & \circ & \circ & \circ & \circ & \circ & \circ \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
K & \circ & \circ & \circ & \circ & \circ & \circ \\
\end{array}
\]
Hidden Markov Models

\[ \pi_j \quad \theta_k \]

\[ \alpha \quad \lambda \]

\[ y_1 \quad y_2 \quad y_3 \quad \cdots \quad y_T \]

\[ z_1 \quad z_2 \quad z_3 \quad \cdots \quad z_T \]

\[ Z_+ \]

\[ \pi_1 \quad \pi_2 \quad \pi_3 \quad \cdots \quad \pi_K \]

\[ 1 \quad 2 \quad 3 \quad \cdots \]

Time

\[ 1 \quad 2 \quad 3 \quad \cdots \]

\[ 2 \quad 3 \quad \cdots \quad \cdots \]

\[ K \quad \cdots \quad \cdots \quad \cdots \]

Modes

Observations
Hidden Markov Models

\[ \pi_j \quad \theta_k \]

\[ z_1 \quad z_2 \quad \cdots \quad z_T \]

\[ y_1 \quad y_2 \quad y_3 \quad \cdots \quad y_T \]

\[ \alpha \quad \lambda \]

\[ \pi_1 \quad \pi_{12} \quad \pi_{13} \quad \pi_{1K} \]

\[ \mathbb{Z}_+ \]

\[ 1 \quad 2 \quad 3 \quad \cdots \]

Time

\[ 1 \quad 2 \quad 3 \quad \cdots \]

\[ 1 \quad 2 \quad 3 \quad \cdots \quad K \]
Hidden Markov Models

\[ \pi_j, K \]
\[ \theta_k, K \]
\[ z_1, z_2, \ldots, z_T \text{ modes} \]
\[ y_1, y_2, y_3, \ldots, y_T \text{ observations} \]

\[ \pi_3, \pi_{32}, \pi_{33}, \ldots, \pi_{3K} \]
\[ z_1, z_2, z_3, \ldots, K \]

Time

1 2 3 \ldots

1
2
3
\vdots
\vdots
\vdots
K
Issue 1: How many modes?

Hierarchical Dirichlet Process HMM

- Dirichlet process (DP):
  - Mode space of unbounded size
  - Model complexity adapts to observations

- Hierarchical:
  - Ties mode transition distributions
  - *Shared* sparsity

Hierarchical Dirichlet Process HMM

Infinite HMM: Beal, et. al., *NIPS* 2002
HDP-HMM: Teh, et. al., *JASA* 2006
Hierarchical Dirichlet Process HMM

• Global transition distribution:
  \[ \beta \sim \text{Stick}(\gamma) \]

• Mode-specific transition distributions:
  \[ \pi_j \sim \text{DP}(\alpha|\beta) \quad j = 1, 2, 3, \ldots \]

\text{sparsity of } \beta \text{ is shared} \quad \implies \quad E[\pi_{jk}] = \beta_k
Issue 2: Temporal Persistence

HDP-HMM inferred mode sequence

True mode sequence

Hidden Markov Model
“Sticky” HDP-HMM
“Sticky” HDP-HMM

\[ \beta \sim \text{Stick}(\gamma) \]
\[ \pi_j \sim \text{DP}(\alpha \beta + \kappa \delta_j) \]

mode-specific base measure

\[ E[\pi_{jk}] = \beta_k \]
\[ E[\pi_{jk}] = \frac{\alpha \beta_k + \kappa \delta(j, k)}{\alpha + \kappa} \]

Infinite HMM: Beal, et.al., NIPS 2002
Direct Assignment Sampler

- **Marginalize:**
  - Transition densities
  - Emission parameters

- **Sequentially sample:**
  \[ z_t^{(n)} \sim p(z_t | z_t^{(n-1)}, \alpha, \kappa) p(y_t | z, y_t, \lambda) \]

Collapsed Gibbs Sampler

Chinese restaurant prior

Conjugate base measure \( \Rightarrow \) closed form

Splits true mode, hard to merge
Block Resampling

- Approximate HDP-HMM weak limit approximation

- Average transition density
  \[ m_{t,t-1}(z_{t-1}) \propto \sum p(z_t | \pi_{z_{t-1}}^{(n)}) p(y_t | \theta_{z_t}^{(n)}) m_{t+1,t}^{(n)}(z_t) \]
  \[ (\Rightarrow \text{transition densities}) \]

- Sample:
  - Block sample \( z_{1:T}^{(n)} \) as:
    \[ z_{t}^{(n)} \sim p(z_t | \pi_{z_{t-1}}^{(n)}) p(y_t | \theta_{z_t}^{(n)}) m_{t+1,t}^{(n)}(z_t) \]

\[ \beta \sim \text{Dir}(\gamma / L, \ldots, \gamma / L) \]
\[ \pi_j \sim \text{Dir}(\alpha \beta_1, \ldots, \alpha \beta_j + \kappa, \ldots, \alpha \beta_L) \]
Results: Gaussian Emissions

- Blocked sampler
- Sequential sampler

Graphs showing normalized Hamming distance over iterations for HDP-HMM and Sticky HDP-HMM.
Results: Fast Switching

- Observations
- True mode sequence
- Normalized Hamming Distance
- Time
- Iteration

Sticky HDP-HMM
HDP-HMM
Results: Discrete Data

Test Predictions
Results: Discrete Data

Sticky HDP-HMM

Non-sticky HDP-HMM
Why a Global Base Measure?
Why a Global Base Measure?
Hyperparameters

- Place priors on hyperparameters and infer them from data
- Weakly informative priors
- All results use the same settings

Related self-transition parameter:
Beal, et.al., NIPS 2002
Speaker Diarization
Results: 21 meetings

<table>
<thead>
<tr>
<th></th>
<th>Overall DER</th>
<th>Best DER</th>
<th>Worst DER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sticky HDP-HMM</td>
<td>17.84%</td>
<td>1.26%</td>
<td>34.29%</td>
</tr>
<tr>
<td>Non-Sticky HDP-HMM</td>
<td>23.91%</td>
<td>6.26%</td>
<td>46.95%</td>
</tr>
<tr>
<td>ICSI</td>
<td>18.37%</td>
<td>4.39%</td>
<td>32.23%</td>
</tr>
</tbody>
</table>
Results: Meeting 1

Sticky DER = 1.26%
ICSI DER = 7.56%
Results: Meeting 18

Sticky DER = 20.48%

ICSI DER = 22.00%
**HDP-HMM: Multimodal Emissions**

- Approximate multimodal emissions with DP mixture
- Temporal mode persistence disambiguates model
Why Complex Emissions?
Results: Mixture Emissions

- 5-mode HMM
  - # emission components
    - \( n_k \sim \text{Uniform}[1, 10] \)
  - Equal mixture weights
- Distance between observations not direct factor in grouping observations within mode
Results: Mixture Emissions

HDP-HMM DP emissions

Sticky HDP-HMM DP emissions
Results: Mixture Emissions

- Improves *predictive probability* of test sequences
- Likely to see larger improvement in higher dimensions
Is it Mixing?
Issue 3: Complex Local Dynamics

- Discrete clusters may not accurately capture high-dimensional data
- Autoregressive HMM: Discrete-mode switching of smooth observation dynamics
Linear Dynamical Systems

- **State space LTI model:**
  \[
  x_t = Ax_{t-1} + e_t \\
  y_t = Cx_t + w_t
  \]
  \[e_t \sim \mathcal{N}(0, \Sigma) \quad w_t \sim \mathcal{N}(0, R)\]

- **Vector autoregressive (VAR) process:**
  \[
  y_t = \sum_{i=1}^{r} A_i y_{t-i} + e_t
  \]
  \[e_t \sim \mathcal{N}(0, \Sigma)\]
Linear Dynamical Systems

- **State space LTI model:**
  \[ x_t = Ax_{t-1} + e_t \]
  \[ y_t = Cx_t + w_t \]
  \[ e_t \sim \mathcal{N}(0, \Sigma) \quad w_t \sim \mathcal{N}(0, R) \]

- **Vector autoregressive (VAR) process:**
  \[ x_t = \begin{bmatrix} A_1 & A_2 & \cdots & A_T \\ I & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & I & 0 \end{bmatrix} x_{t-1} + \begin{bmatrix} I \\ 0 \\ \vdots \\ 0 \end{bmatrix} e_t \]
  \[ y_t = \begin{bmatrix} I & 0 & \cdots & 0 \end{bmatrix} x_t. \]
Switching Dynamical Systems

Switching linear dynamical system (SLDS):

\[ z_t \sim \pi_{z_{t-1}} \]
\[ x_t = A^{(z_t)} x_{t-1} + e_t(z_t) \]
\[ y_t = C x_t + w_t \]
\[ e_t \sim \mathcal{N}(0, \Sigma^{(z_t)}) \quad w_t \sim \mathcal{N}(0, R) \]

Switching VAR process:

\[ z_t \sim \pi_{z_{t-1}} \]
\[ y_t = \sum_{i=1}^{r} A_i^{(z_t)} y_{t-i} + e_t(z_t) \]
\[ e_t \sim \mathcal{N}(0, \Sigma^{(z_t)}) \]
HDP-AR-HMM and HDP-SLDS

\[ \theta_k = \{ A^{(k)}_{1:r}, \Sigma^{(k)} \} \]

\[ z_t \sim \pi_{z_{t-1}} \]

\[ y_t = \sum_{i=1}^{r} A_i^{(z_t)} y_{t-i} + e_t(z_t) \]

\[ \theta_k = \{ A^{(k)}, \Sigma^{(k)}, R \} \]

\[ z_t \sim \pi_{z_{t-1}} \]

\[ x_t = A^{(z_t)} x_{t-1} + e_t(z_t) \]

\[ y_t = C x_t + w_t \]

\[ C = [I \ 0] \]
Results: IBOVESPA

- Data: Sao Paolo stock index
- Goal: detect changes in volatility
- Compare inferred change-points to 10 cited world events


*Hong Kong stock index falls 10.4%*
Dancing Honey Bees
Honey Bee Results: HDP-AR(1)-HMM

Sequence 1

HDP-AR-HMM: 88.1%
SLDS [Oh]: 93.4%

Sequence 2

HDP-AR-HMM: 92.5%
SLDS [Oh]: 90.2%

Sequence 3

HDP-AR-HMM: 88.2%
SLDS [Oh]: 90.4%
Low-level Image Analysis

Goals:

- Accurately model the statistics of natural images
- Exploit the availability of large digital image collections
Wavelet Decompositions

- Bandpass decomposition of images into multiple scales & orientations
- Multiscale dependencies captured via latent quadtree structure
Wavelets: Marginal Statistics

- Wavelet Coefficient
- Smooth Surfaces
- Occlusion Boundaries & Texture

[Diagram showing log probability against wavelet coefficient with annotations]
Gaussian Mixture Models

\[ x_i = v_i u_i \]
\[ v_i \geq 0 \quad u_i \sim \mathcal{N}(0, \Lambda) \]

**Gaussian Scale Mixture (GSM)**

Wainwright & Simoncelli, 2000

\[ x_i \sim \pi \mathcal{N}(0, \Lambda_0) + (1 - \pi)\mathcal{N}(0, \Lambda_1) \]

**Binary Gaussian Mixture**

Computational advantages…
Wavelets: Joint Statistics

Pairwise Joint Histograms:

- Orientation
- Scale
- Vertical
- Horizontal

Pairwise Conditional Histograms:

- Orientation
- Scale
- Vertical
- Horizontal

Large magnitude wavelet coefficients...

- *Persist* across multiple scales
- *Cluster* at adjacent spatial locations
Binary Hidden Markov Trees

*Crouse, Nowak, & Baraniuk, 1998*

- Coefficients marginally distributed as mixtures of two Gaussians
- Markov dependencies between hidden states capture persistence of image contours across locations and scales
- Each orientation is modeled independently

\[ \pi_k \rightarrow \text{state transition distributions} \]

\[ z_{ti} \sim \pi_{z_{Pa}(ti)} \]

\[ \Lambda_k \rightarrow \text{state-specific emission covariances} \]

\[ x_{ti} \sim \mathcal{N}(0, \Lambda_{z_{ti}}) \]

\[ z_{ti} \rightarrow \text{hidden state or cluster assignment} \]

\[ z_{ti} \in \{0, 1\} \]

\[ x_{ti} \rightarrow \text{observed wavelet coefficient} \]
Validation: Image Denoising

Original

Corrupted by Additive White Gaussian Noise
(PSNR = 24.61 dB)
Denoising with Binary HMTs

Noisy Input

Denoised (EM algorithm)

- Is two states per scale sufficient? How many is enough?
- Should states be shared the same way for all images, or for all wavelet decompositions?
Hierarchical Dirichlet Process Hidden Markov Trees

- \( z_{ti} \rightarrow \) indexes *infinite* set of hidden states
  \( z_{ti} \in \{1, 2, 3, \ldots\} \)

- \( x_{ti} \rightarrow \) observed *vector* of wavelet coefficients

- \( \pi_k \rightarrow \) infinite set of state *transition* distributions

\[
 z_{ti} \sim \pi_{z_{Pa(ti)}}
\]

- \( \Lambda_k \rightarrow \) state-specific *emission* covariances

\[
 x_{ti} \sim \mathcal{N}(0, \Lambda_{z_{ti}}) \\
 \Lambda_k \sim H
\]
Why a Hierarchical DP? (Teh et al. 2004)

- Hierarchical DP prior allows us to learn a potentially infinite set of *appearance patterns* from natural images.
- Hierarchical coupling ensures, with high probability, that a common set of *child* states are reachable from each *parent*.

\[
\pi_{k}^{d_{t_i}}(\ell) = \Pr \left[ z_{t_i} = \ell \mid z_{Pa(t_i)} \right] \\
\beta \sim \text{Stick}(\gamma)
\]

**Average state frequencies**

\[
\mathbb{E} \left[ \pi_{k}^{d} \right] = \beta
\]

**Sparsity & variability of transition distributions**

\[
\alpha \rightarrow \text{Transition distributions}
\]

\[
\pi_{k}^{d} \sim \text{DP}(\alpha, \beta)
\]

**Global classes**
Denoising: Input

24.61 dB
Denoising: Binary HMT

29.35 dB

Crouse, Nowak, & Baraniuk, 1998
Denoising: HDP-HMT

32.10 dB
Denoising: Local GSM

Portilla et. al., 2003
Estimating Clean Images

Empirical Bayesian approach estimates model parameters from the noisy image

Transfer denoising approach reuses multiscale hidden state patterns of clean images for making robust predictions
Denoising Einstein

Noisy
10.60 dB, 0.057

Original

BLS-GSM
26.38 dB, 0.647

BM3D
26.49 dB, 0.659

HDP-HMT (Emp. Bayes)
25.64 dB, 0.564

HDP-HMT (Transfer)
26.80 dB, 0.664
Natural Scene Denoising

<table>
<thead>
<tr>
<th>Method</th>
<th>Description</th>
<th>PSNR</th>
<th>SSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noisy</td>
<td></td>
<td>8.14</td>
<td>0.033</td>
</tr>
<tr>
<td>HDP-HMT (Emp. Bayes)</td>
<td></td>
<td>24.24</td>
<td>0.519</td>
</tr>
<tr>
<td>HDP-HMT (Transfer)</td>
<td></td>
<td>26.50</td>
<td>0.794</td>
</tr>
<tr>
<td>Original</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BLS-GSM</td>
<td></td>
<td>25.59</td>
<td>0.726</td>
</tr>
<tr>
<td>BM3D</td>
<td></td>
<td>25.74</td>
<td>0.751</td>
</tr>
</tbody>
</table>
Natural Scene Categorization

Goals:

- Visually recognize natural scene categories
- Accurately model the statistics of natural scene categories
HDP-HMT Scene Model

- Hidden states $\tilde{z}_{ti}$ generate vectors of clean wavelet coefficients $\tilde{x}_{ti}$ at multiple orientations, or dense multiscale SIFT descriptors
... versus baseline HDP-BOF

HDP-HMT

HDP-BOF

Nonparametric Bayesian extension of LDA scene models (Fei-Fei & Perona, 2005) which ignore spatial locations of locally extracted image features
Number of States

Wavelet (sp5)

SIFT
Samples given MAP states

Input Image

HDP Hidden Markov Tree

HDP Bag of Features
### Categorizing Natural Scenes

#### Wavelet (sfp7)

<table>
<thead>
<tr>
<th>Category</th>
<th>HDP-BOF [75.3 %]</th>
<th>HDP-HMT [80.7 %]</th>
</tr>
</thead>
<tbody>
<tr>
<td>coast</td>
<td></td>
<td></td>
</tr>
<tr>
<td>forest</td>
<td></td>
<td></td>
</tr>
<tr>
<td>highway</td>
<td></td>
<td></td>
</tr>
<tr>
<td>inside city</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mountain</td>
<td></td>
<td></td>
</tr>
<tr>
<td>open country</td>
<td></td>
<td></td>
</tr>
<tr>
<td>street</td>
<td></td>
<td></td>
</tr>
<tr>
<td>tall building</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### SIFT

<table>
<thead>
<tr>
<th>Category</th>
<th>HDP-BOF [82.4 %]</th>
<th>HDP-HMT [86.5 %]</th>
</tr>
</thead>
<tbody>
<tr>
<td>coast</td>
<td></td>
<td></td>
</tr>
<tr>
<td>forest</td>
<td></td>
<td></td>
</tr>
<tr>
<td>highway</td>
<td></td>
<td></td>
</tr>
<tr>
<td>inside city</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mountain</td>
<td></td>
<td></td>
</tr>
<tr>
<td>open country</td>
<td></td>
<td></td>
</tr>
<tr>
<td>street</td>
<td></td>
<td></td>
</tr>
<tr>
<td>tall building</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>