Hierarchical Dirichlet Processes

Yee Whye Teh, Michael I. Jordan
Matthew J. Beal, David M. Blei
Presented By: Qiang Fu
Outline

- Introduction
- Hierarchical Dirichlet Process (HDP)
- Representations of HDP
- Inference
- Experiments
- Hierarchical Dirichlet Process Hidden Markov Model (HDP-HMM)
Introduction

- **Problem Setting**
  - Groups of data
  - Observations within a group = Mixture Model
  - Mixture components are shared

- **Assumption:**
  - number of mixture components unknown
  - Exchangeability
HDP

- Consider a DP for each group
- One Simple Solution:
  \[ G_j \mid \alpha_o, G_o(\gamma) \sim DP(\alpha_o, G_o(\gamma)) \quad \text{for each } j \]
- But doesn’t work all the time
- Stick-Breaking Construction:
  \[ G \sim DP(\alpha_0, G_0) \quad G = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k} \]
HDP

HDP:

\[ G_0 \mid \gamma, H \sim \text{DP}(\gamma, H) \]
\[ G_j \mid \alpha_0, G_0 \sim \text{DP}(\alpha_0, G_0) \quad \text{for each } j, \]

Probability Model (Generative Process):

\[ \theta_{ji} \mid G_j \sim G_j \quad \text{for each } j \text{ and } i, \]
\[ x_{ji} \mid \theta_{ji} \sim F(\theta_{ji}) \quad \text{for each } j \text{ and } i, \]
Stick-Breaking Construction for DP

- Measures drawn from a Dirichlet process are discrete with probability one.

\[
\pi_k' \mid \alpha_0, G_0 \sim \text{Beta}(1, \alpha_0) \quad \phi_k \mid \alpha_0, G_0 \sim G_0
\]

\[
\pi_k = \pi_k' \prod_{l=1}^{k-1} (1 - \pi_l') 
\]

\[
G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k} 
\]

- Notation: \( \pi \sim \text{GEM}(\alpha_0) \)
Stick-Breaking Construction for HDP

- $G_o$ can be expressed as:
  \[ G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k} \]

- $G_j$ can be expressed similarly:
  \[ G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\phi_k} \]

- Let $(A_1, \ldots, A_r)$ be a measurable partition on $\Theta$.

- Define $K_l = \{ k : \phi_k \in A_l \}$ for $l = 1, \ldots, r$.

- $(K_1, \ldots, K_r)$ is a finite partitions of positive integers.
Stick-Breaking Construction for HDP

For each $j$, we have:

\[
(G_j(A_1), \ldots, G_j(A_r)) \sim \text{Dir}(\alpha_0 G_0(A_1), \ldots, \alpha_0 G_0(A_r))
\]

\[
\Rightarrow \left( \sum_{k \in K_1} \pi_{jk}, \ldots, \sum_{k \in K_r} \pi_{jk} \right) \sim \text{Dir} \left( \alpha_0 \sum_{k \in K_1} \beta_k, \ldots, \alpha_0 \sum_{k \in K_r} \beta_k \right)
\]
Stick-Breaking Construction for HDP

- Derive the explicit relationship
- For a partition \((\{1, \ldots, k-1\}, \{k\}, \{k+1, k+2, \ldots\})\):

\[
\left( \sum_{l=1}^{k-1} \pi_{jl}, \pi_{jk}, \sum_{l=k+1}^{\infty} \pi_{jl} \right) \sim \text{Dir} \left( \alpha_0 \sum_{l=1}^{k-1} \beta_l, \alpha_0 \beta_k, \alpha_0 \sum_{l=k+1}^{\infty} \beta_l \right)
\]

- Remove the first element:

\[
\frac{1}{1 - \sum_{l=1}^{k-1} \pi_{jl}} \left( \pi_{jk}, \sum_{l=k+1}^{\infty} \pi_{jl} \right) \sim \text{Dir} \left( \alpha_0 \beta_k, \alpha_0 \sum_{l=k+1}^{\infty} \beta_l \right)
\]
Stick-Breaking Construction for HDP

- Define: \[ \pi'_{jk} = \frac{\pi_{jk}}{1 - \sum_{l=1}^{k-1} \pi_{jl}} \]
- Observe that: \[ 1 - \sum_{l=1}^{k} \beta_l = \sum_{l=k+1}^{\infty} \beta_l \]
- We have:

\[
\pi'_{jk} \sim \text{Beta} \left( \alpha_0 \beta_k, \alpha_0 \left( 1 - \sum_{l=1}^{k} \beta_l \right) \right) \\
\pi_{jk} = \pi'_{jk} \prod_{l=1}^{k-1} (1 - \pi'_{jl})
\]
Chinese Restaurant Process

- Clustering effect of DP
- The metaphor
- After integrate out $G$, we have:

$$\theta_i | \theta_1, \ldots, \theta_{i-1}, \alpha_0, G_0 \sim \sum_{k=1}^{K} \frac{m_k}{i-1 + \alpha_0} \delta_{\phi_k} + \frac{\alpha_0}{i-1 + \alpha_0} G_0$$
Chinese Restaurant Franchise
Chinese Restaurant Franchise

- After $G_j$ is integrated out:

$$
\theta_{j,i} \mid \theta_{j,1}, \ldots, \theta_{j,i-1}, \alpha_0, G_0 \sim \sum_{t=1}^{m_j} \frac{n_{jt}}{i - 1 + \alpha_0} \delta_{\psi_{jt}} + \frac{\alpha_0}{i - 1 + \alpha_0} G_0
$$

- After $G_0$ is integrated out:

$$
\psi_{jt} \mid \psi_{11}, \psi_{12}, \ldots, \psi_{21}, \ldots, \psi_{j,t-1}, \gamma, H \sim \sum_{k=1}^{K} \frac{m_k}{m_+ + \gamma} \delta_{\phi_k} + \frac{\gamma}{m_+ + \gamma} H
$$
Posterior Sampling in the CRF

- Sample $t$

- Integrate out the possible values of $k_{jt}^{new}$

\[
p(x_{ji} \mid t^{-ji}, t_{ji} = t^{new}, k) = \sum_{k=1}^{K} \frac{m_k}{m_+ + \gamma} f_k^{-x_{ji}}(x_{ji}) + \frac{\gamma}{m_+ + \gamma} f_{k^{new}}^{-x_{ji}}(x_{ji})
\]

- Then:

\[
p(k_{jt}^{new} = k \mid t, k^{-jt^{new}}) \propto \begin{cases} 
m_k f_k^{-x_{ji}}(x_{ji}) & \text{if } k \text{ previously used,} \\
\gamma f_{k^{new}}^{-x_{ji}}(x_{ji}) & \text{if } k = k^{new}.
\end{cases}
\]

\[
p(t_{ji} = t \mid t^{-ji}, k) \propto \begin{cases} 
n_{jt}^{-ji} f_{kjt}^{-x_{ji}}(x_{ji}) & \text{if } t \text{ previously used,} \\
\alpha_0 p(x_{ji} \mid t^{-ji}, t_{ji} = t^{new}, k) & \text{if } t = t^{new}.
\end{cases}
\]
Posterior Sampling in the CRF

Sample \( k \) will be similar:

\[
p(k_{jt} = k \mid t, k^{-jt}) \propto \begin{cases} 
    m^{-jt}_k f_k(x_{jt}) & \text{if } k \text{ is previously used}, \\
    \gamma f_{k_{new}}(x_{jt}) & \text{if } k = k_{new}.
\end{cases}
\]

\( \theta \) and \( \psi \) can be reconstructed from these index variables.
Posterior sampling with an augmented representation

- Based on the Dirichlet Posterior Distribution:

\[
G|\theta_1, \ldots, \theta_n \sim \text{DP} \left( \alpha + n, \frac{\alpha}{\alpha + n} H + \frac{n}{\alpha + n} \sum_{i=1}^{n} \delta \theta_i \right)
\]

- Rewrite it: \( G \) is distributed as

\[
\text{DP} \left( \gamma + m, \frac{\gamma H + \sum_{k=1}^{K} m_k \delta \phi_k}{\gamma + m} \right)
\]
Posterior sampling with an augmented representation

- Construct $G_o$:

$$
\beta = (\beta_1, \ldots, \beta_K, \beta_u) \sim \text{Dir}(m_1, \ldots, m_K, \gamma)
$$

$$
G_u \sim \text{DP}(\gamma, H) \quad G_0 = \sum_{k=1}^{K} \beta_k \delta_{\phi_k} + \beta_u G_u
$$

- Sampling for $t$ and $k$ will be similar to the previous algorithm
Posterior Sampling by Direct Assignment

- No Bookkeeping
- Sample $z$
  \[ p(z_{ji} = k \mid z^{\sim ji}, m, \beta) = \begin{cases} 
(n_{j,k}^{-ji} + \alpha_0 \beta_k) f_k^{-x_{ji}}(x_{ji}) & \text{if } k \text{ previously used,} \\
\alpha_0 \beta_u f_{k_{\text{new}}}^{-x_{ji}}(x_{ji}) & \text{if } k = k_{\text{new}}. 
\end{cases} \]
- Sample $m$
  \[ p(t_{ji} = t \mid k_{jt} = k, t^{-ji}, k, \beta) \propto n_{jt}^{-ji}. \]
  \[ p(t_{ji} = t_{\text{new}} \mid k_{jt_{\text{new}}} = k, t^{-ji}, k, \beta) \propto \alpha_0 \beta_k. \]
Experiment – Document Modeling

- HDP picks the number of topics for LDA
Experiment – Multiple Corpora

- Articles from the conference are divided into sections
- HDP is used to discover the shared topics among the articles within each section
- Want to exam relationships among the sections
Experiment – Multiple Corpora
Experiment – Multiple Corpora

Average perplexity over NIPS sections of 3 models

Generalization from LT, AA, AP to VS
Hidden Markov Models

- HMM is a dynamic variant of a mixture model: each row of the transition matrix is a set of mixing proportions for the choice of the next state.
HDP-HMM

- An HMM can be viewed as a set of mixture models: one mixture model for each value of the current state.
- When a new state arises, HDP shares this new state among the current states.
Experiments- Alice in Wonderland

Perplexity on test sentences of Alice

- ML
- MAP
- VB
- VBm

Number of hidden states

Perplexity