Case Study 5: Mixed Membership Modeling

Clustering Documents Revisited, Latent Dirichlet Allocation

Machine Learning for Big Data
CSE547/STAT548, University of Washington
Emily Fox
March 6th, 2014

Task 2: Cluster Documents

Then examined:
- Cluster documents based on topic
A Generative Model

- Documents: $x^1, \ldots, x^D$
- Associated topics: $z^1, \ldots, z^D$
- Parameters: $\theta = \{\pi, \beta\}$
- Generative model:

$$
\begin{align*}
  z^d & \sim \pi \quad \text{generate topic} \\
  w_i^d & \sim \beta_{z^d}^k \quad i=1, \ldots, N_d
\end{align*}
$$

Given topic $z^d=k$ for doc $d$, draw each word from $\beta_k^d$. 

Model In Pictures

- Mixture weights (on topics)
- Topic distributions (on words)

- For each document,

$$
\begin{align*}
  z^d & \sim \pi \\
  w_i^d & \sim \beta_{z^d}^k
\end{align*}
$$
Bayesian Document Model

- Model parameters \( \pi, \{\beta_k\} \) unknown
- Bayesian approach
  - place priors on parameters
- Need distribution on pmf's

\[
\sum_{k=1}^{K} \pi_k = 1 \quad \sum_{w} \beta_{kw} = 1
\]

\( \pi, \beta_k \) live on the simplex

1. What is the simplex?
2. What is a distribution on the simplex?

Dirichlet Distributions

- The Dirichlet distribution is defined on the simplex

\( \pi \sim \text{Dir}(\alpha_1, \ldots, \alpha_K) \)

\[
p(\pi | \alpha) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k \pi_k^{\alpha_k - 1}
\]

Moments:

\[
E_{\pi_k} = \frac{\alpha_k}{\alpha_0} \\
\text{Var}_{\pi_k} = \frac{K - 1}{K^2 (\alpha_0 + 1)}
\]
Model Summary

- Prior on model parameters
  - E.g., symmetric Dirichlet for $\pi$
  - Dirichlet prior for topic parameters
- Sample observations as
  $z^d \sim \pi$
  $w_i^d \mid z^d \sim \beta_{z^d}$

Posterior Inference via Sampling

- Iterate between sampling
  1. $\pi \sim p(\pi \mid \pi^d, \beta_k^d, \{w_i^d\})$
  2. For $k=1,...,K$
     $\beta_k \sim p(\beta_k \mid \pi, \{z_i^d\}, \{\beta_k^d, \{w_i^d\})$
  3. For $d=1,...,D$
     $z^d \sim p(z^d \mid \pi, \{z_i^d\}, \{\beta_k^d, \{w_i^d\})$

- What form do these complete conditionals take?
  - First a look at statements of conditional independence in directed graphical models
Conditional Independence in Bayes Nets

- Consider 4 different junction configurations

\[ \text{(a)} \quad \text{(b)} \quad \text{(c)} \quad \text{(d)} \]

- Conditional versus unconditional independence:
  \[
P(x, y, z) = p(x)p(y|x, z) \quad \Rightarrow \quad p(x, z) = p(x)p(z) \quad \Rightarrow \quad x \perp \!\!\!\!\!\!\!\!\!\!\perp z
  \]
  \[
p(x, z|y) \neq p(x, y, z) = p(x)p(y|x, z)
  \]

"Explaining away": \( x = \text{earthquake} \), \( z = \text{burglar} \), \( y = \text{car alarm} \)

\( x \perp \!\!\!\!\!\!\!\!\!\!\perp z \) \( \rightarrow \) ind. \( \rightarrow \) x \neq z | y

If \( x = \text{alarm} \) \( \rightarrow \) an increase in earthquake \( p(x|y) \), means \( p(z|y) \) lower

Bayes Ball Algorithm

- Consider 4 different junction configurations

\[ \text{(a)} \quad \text{(b)} \quad \text{(c)} \quad \text{(d)} \]

- Bayes ball algorithm

  - Start ball at one end or other.
  - If ball passes to a node (arrow) then, not cond/marg. ind.
  - If ball bounces back (walls + curved arrows) the nodes are cond/marg. ind.
Markov Blanket

- A node is conditionally independent of all other nodes in the graph given its Markov blanket.
  - Markov blanket: all parents
  - All children
  - All coparents

- Gibbs sampling iterates between full conditionals:
  \[ x_i \sim p(x_i | x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) \]
  \[ \rightarrow \text{simplify to} \]
  \[ x_i \sim p(x_i | MB(x_i)) \]

Unplated Document Model

- Recall that the plate notation is really indicating:
  - \( \alpha \)
  - \( \pi \)
  - \( \beta_k \)
  - \( \lambda \)
  - \( K \)
  - \( N_d \)
  - \( w_{i}^{d} \)
  - \( D \)
  - \( T \)
  - \( \theta \)
  - \( \phi \)
  - \( \lambda \)

©Emily Fox 2014
Complete Conditional for $\pi$:

- Recall conjugate Dirichlet prior
  \[ \pi \sim \text{Dir}(\alpha_1, \ldots, \alpha_K) \]
  \[ p(\pi | \alpha) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k \pi_k^{\alpha_k - 1} \]

- Likelihood:
  \[ z^{d,k} \]
  \[ p(z^{d,k} | \pi) \]

- Dirichlet posterior
  - Count occurrences of $z^{d,k}$:
    \[ N_k = [\sum_d z^{d,k}] \]
  - Then,
    \[ p(\pi | z^{d,k}) \propto \prod_d p(z^{d,k} | \pi) p(\pi | \alpha) \]
    \[ \propto \prod_d \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k \pi_k^{\alpha_k - 1} \]
    \[ \cdot \prod_k \pi_k^{N_k + \alpha_k - 1} \]
    \[ = \text{Dir}(N_k + \alpha_k) \]

- Conjugacy: \textbf{Posterior} has same form as \textbf{prior}

Complete Conditional for $\beta_k$:

- Again, Dirichlet prior
  \[ \beta_k \sim \text{Dir}(\lambda, \ldots, \lambda) \]

- Consider docs $d$ such that $z^{d,k}$
  - For these observations, $w_{i}^{d} \sim \beta_k$
  - Do any other docs depend on $\beta_k$? \textbf{No}

- Then,
  \[ \text{Count } m_{k,v} = \sum_d w_{i}^{d} : w_{i}^{d} = v \forall d \text{ s.t. } z^{d,k} \text{ \& } \exists \text{ topic } \]
  \[ \beta_k \sim \text{Dir}(m_{k,v} + \lambda) \ldots, m_{k,v} + \lambda) \]

- Again, \textbf{posterior} has same form as \textbf{prior}
Complete Conditional for $z^d$

- We have $z^d \sim \pi$ “prior”
  
  $$w_i^d \mid z^d, \{\beta_k\} \sim \beta_{z^d}$$
  “likelihood”

- Calculate the posterior for each value of $z^d$
  (“responsibility” of each topic to the doc):

  $$r_{dk} = p(z^d = k \mid \{w_i^d\}, \pi, \beta) = \frac{\pi_k p(\{w_i^d\} \mid \beta_k)}{\sum_j \pi_j p(\{w_i^d\} \mid \beta_j)}$$

- Sample each cluster indicator as

  $$z^d \sim r_d$$

Collapsed Gibbs Sampler

- In conjugate models, can analytically marginalize some variables and only sample remaining

  $$z^d \sim p(z^d \mid z^1, \ldots, z^{d-1}, \{w_i^d\}, \lambda)$$

- Can improve efficiency if marginalized variables are high-dim
  - Reduced dimension of search space
  - But, often introduces dependences!
Collapsed Sampler Full Conditional

\[
p(z) = p(\pi | \alpha) \prod_{d=1}^{D} p(z_{d} | \pi) \left( \prod_{k=1}^{K} p(\beta_{k} | \lambda) \prod_{d=1}^{D} \prod_{i=1}^{N_{d}} p(w_{d}^{i} | z_{d}^{i}, \beta) \right)
\]

### Derivation

- \( p(z_{d} = k | z_{\backslash d}, \alpha, \lambda) \propto \int_{\pi} \int_{\beta_{1}} \cdots \int_{\beta_{K}} p(\cdot) \) full cond.

- \( \alpha p(z_{d} = k | z_{\backslash d}, \alpha) p(\{w_{d}^{i}\} | \{w_{c}^{i} : z_{c} = k, c \neq d\}) \) "prior"

- \( p\left(\frac{z_{d}^{i} = k | z_{\backslash d}, \alpha} \right) p(\{w_{d}^{i}\} | \{w_{c}^{i} : z_{c} = k, c \neq d\}) \) "likelihood"
Collapsed Sampler Intuition (MoG)

- Previously, \( p(z^i = k \mid x^i, \pi, \theta) \propto \pi_k p(x^i \mid \theta_k) \)

- If you’re not told \( \pi, \theta_k \)

Approx \( \pi \) by counts of occupancy of each cluster (plus prior)
Approx \( \Theta_k \) based on obs. already assigned to cluster \( k \)

Example – Uncollapsed Results

Figure 2.18. Learning a mixture of \( K = 4 \) Gaussians using the Gibbs sampler of Alg. 2.1. Columns show the current parameters after \( T=2 \) (top), \( T=10 \) (middle), and \( T=50 \) (bottom) iterations from two random initializations. Each plot is labeled by the current data log–likelihood.

Figure courtesy of Erik Sudderth.
Comparing Collapsed vs. Uncollapsed

Log Likelihood vs. Gibbs Iteration
(multiple chains)

Example – Collapsed Results

Comparing Collapsed vs. Uncollapsed

Log Likelihood vs. Gibbs Iteration
(multiple chains)
Task 3: Mixed Membership Models

**Now:** Document may belong to multiple clusters

---

**Latent Dirichlet Allocation (LDA)**

*COLD SPRING HARBOR, NEW YORK—* How many genes does an organism need to survive? Last week at the Genome Meeting here, two genome researchers and a relatively different approach presented complementary views of the basic genes needed for life.

One research team, using computer analyses to compare known genomes, concluded that today’s bacteria can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a single parasite and estimated that for this organism, 882 genes are necessary to do the job—but that anything short of 100 would not be enough.

Although the numbers don’t match precisely, those predictions

---

Latent Dirichlet Allocation (LDA)

We compute the posterior $p$. topics, proportions, assignments.

Every word is assigned to a topic.

Each doc has its own prevalence of topics in that doc.
LDA Generative Model

- Observations: $w_1^d, \ldots, w_{N_d}^d$
- Associated topics: $z_1^d, \ldots, z_{N_d}^d$
- Parameters: $\theta = \{\{\pi^d\}, \{\beta_k\}\}$
- Generative model:

\[
\begin{align*}
  z_i^d & \sim \pi^d, \quad d = 1, \ldots, D, \quad i = 1, \ldots, N_d \\
  w_i^d | z_i^d & \sim \beta_{z_i^d}
\end{align*}
\]

Priors:
\[
\pi^d \sim \text{Dir}(\alpha_1, \ldots, \alpha_K), \quad k = 1, \ldots, D \\
\beta_k \sim \text{Dir}(\lambda_1, \ldots, \lambda_V), \quad k = 1, \ldots, K
\]

LDA Joint Probability

\[
p(\cdot) = \prod_{k=1}^{K} p(\beta_k | \lambda) \prod_{d=1}^{D} p(\pi^d | \alpha) \left( \prod_{i=1}^{N_d} p(z_i^d | \pi^d) p(w_i^d | z_i^d, \beta) \right)
\]
Example Inference – Topic Weights

- **Data:** The OCR’ed collection of *Science* from 1990-2000
  - 17K documents
  - 11M words
  - 20K unique terms (stop words and rare words removed)

- **Model:** 100-topic LDA model

**Example Inference – Topic Words**

<table>
<thead>
<tr>
<th>topic1</th>
<th>topic2</th>
</tr>
</thead>
<tbody>
<tr>
<td>human</td>
<td>disease</td>
</tr>
<tr>
<td>genome</td>
<td>evolution</td>
</tr>
<tr>
<td>dna</td>
<td>genetic</td>
</tr>
<tr>
<td>species</td>
<td>genes</td>
</tr>
<tr>
<td>bacteria</td>
<td>sequence</td>
</tr>
<tr>
<td>life</td>
<td>origin</td>
</tr>
<tr>
<td>resistance</td>
<td>gene</td>
</tr>
<tr>
<td>data</td>
<td>new</td>
</tr>
<tr>
<td>computers</td>
<td>molecular</td>
</tr>
<tr>
<td>system</td>
<td>networks</td>
</tr>
<tr>
<td>network</td>
<td>methods</td>
</tr>
<tr>
<td>models</td>
<td>parasites</td>
</tr>
<tr>
<td>information</td>
<td>map</td>
</tr>
<tr>
<td>data</td>
<td>phylogenetic</td>
</tr>
<tr>
<td>control</td>
<td>information</td>
</tr>
<tr>
<td>model</td>
<td>diversity</td>
</tr>
<tr>
<td>parallel</td>
<td>group</td>
</tr>
<tr>
<td>networks</td>
<td>living</td>
</tr>
<tr>
<td>parasite</td>
<td>new</td>
</tr>
<tr>
<td>software</td>
<td>sequences</td>
</tr>
<tr>
<td>new</td>
<td>common</td>
</tr>
<tr>
<td>simulations</td>
<td>tuberculosis</td>
</tr>
</tbody>
</table>

©Emily Fox 2014
Collapsed LDA Sampling

- Marginalize parameters
  - Document-specific topic weights
  - Corpus-wide topic-specific word distributions

\[
p(z_i^d = k \mid z_{\setminus id}, \{w_i^d\}, \alpha, \lambda) \\
\propto p(z_i^d = k \mid z_{\setminus id}, \alpha) p(w_i^d \mid z_i^d = k, z_{\setminus id}, w_{\setminus id}, \lambda)
\]

- Unplate to see dependencies induced

What you need to know...

- Bayesian specification of document clustering model
- Rules of conditional and unconditional independence in directed graphical models (Bayes nets)
  - Bayes’ ball
  - Markov blanket
- Gibbs sampling for Bayesian document model
- Latent Dirichlet allocation (LDA)
  - Motivation and generative model specification
  - Collapsed Gibbs sampler
Reading

- **Mixed Membership Models: KM Sec. 27.3**
  - Basic LDA:
  - Introduction:
  - Sampling:

Acknowledgements

- Thanks to Dave Blei for some material in this lecture relating to LDA