Case Study 2: Document Retrieval

Collapsed Gibbs and Variational Methods for LDA

Machine Learning/Statistics for Big Data
CSE599C1/STAT592, University of Washington
Emily Fox
February 7th, 2013

©Emily Fox 2013

Example – Collapsed MoG Sampling

\[ \pi \sim \text{Dir}(\alpha_1, \ldots, \alpha_K) \]
\[ \{\mu_k, \Sigma_k\} \sim F(\phi) \]
\[ x^i \mid z^i \sim N(x^i; \mu_{z^i}, \Sigma_{z^i}) \]

- Collapsed sampler

For \( i = 1, \ldots, N \)

\[ z^{(i)} \sim p(z^{(i)} \mid z^{(1)}, \ldots, z^{(i-1)}, z^{(i+1)}, \ldots, z^{(N)}, x_1:N, \phi) \]

©Emily Fox 2013
Example – Collapsed MoG Sampling

\[ \pi \sim \text{Dir}(\alpha_1, \ldots, \alpha_K) \]
\[ \{\mu_k, \Sigma_k\} \sim F(x) \]

- Derivation

\[ x^i \mid z^i \sim N(\mu_{z^i}, \Sigma_{z^i}) \]
\[ z^i \sim \pi \]

\[ p(z^i \mid z_{-i}, x_{-i}, \alpha) \propto p(z^i \mid z_{-i}, x_{-i}) p(x^i \mid z^i, z_{-i}, x_{-i}) \]
\[ p(z^i = k \mid z_{-i}, x_{-i}) = \int p(z^i = k \mid x_{-i}) p(x^i \mid z^i, x_{-i}) \, d\pi = \frac{n_k^x + \alpha_k}{N - 1 + \sum \alpha_k} \]

\[ p(x^i \mid z_{-i}, x_{-i}, \alpha) = \text{Student-t Dir post. Pred. likelihood} \]

- Important facts:

\[ p(z_{1:N} \mid \alpha) = \frac{\Gamma(\sum \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k \frac{\Gamma(n_k + \alpha_k)}{\Gamma(n_k + \alpha_k)} \frac{\Gamma(m + 1)}{\Gamma(m)} = m \]

Latent Dirichlet Allocation (LDA)

Each doc is a mixture of these corpus-wide topics

Every word is assigned to a topic

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

- Observations: \( w^d_1, \ldots, w^d_{N_d} \)
- Associated topics: \( z^d_1, \ldots, z^d_{N_d} \)
- Parameters: \( \theta = \{\{\pi^d\}, \{\beta_k\}\} \)
- Generative model:

\[
\begin{align*}
\pi^d & \sim \text{Dir}(\alpha_1, \ldots, \alpha_K) \quad d = 1, \ldots, D \\
\beta_k & \sim \text{Dir}(\lambda_1, \ldots, \lambda_V) \quad k = 1, \ldots, K \\
\end{align*}
\]

\[
p(\cdot) = \prod_{k=1}^{K} p(\beta_k \mid \lambda) \prod_{d=1}^{D} p(\pi^d \mid \alpha) \left( \prod_{i=1}^{N_d} p(z_i^d \mid \pi^d)p(w_i^d \mid z_i^d, \beta) \right)
\]
Collapsed LDA Sampling

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

- Sample topic indicators for each word

Derivation:

\[
p(z^d_{1:N_d} | \alpha) = \frac{\Gamma(\sum_k \alpha_k) \prod_k \Gamma(n^d_k + \alpha_k)}{\prod_k \Gamma(\alpha_k) \Gamma(\sum_k n^d_k + \alpha_k)}
\]

\[
p(w^d_i | z^d_i = k, \lambda) = \frac{\Gamma(\sum_k \lambda_k) \prod_k \Gamma(\nu^d_k + \lambda_k)}{\prod_k \Gamma(\lambda_k) \Gamma(\sum_k \nu^d_k + \lambda_k)}
\]

\[
p(z | \alpha) = \prod_{d=1}^D p(z^d_{1:N_d} | \alpha)
\]

\[
p(w | z, \lambda) = \prod_{k=1}^K p(w^d_i | z^d_i = k, \lambda)
\]
Sample Document

| Etruscan | trade | price | temple | market |

Randomly Assign Topics

<table>
<thead>
<tr>
<th>3</th>
<th>2</th>
<th>1</th>
<th>3</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Etruscan</td>
<td>trade</td>
<td>price</td>
<td>temple</td>
<td>market</td>
</tr>
</tbody>
</table>
### Randomly Assign Topics

<table>
<thead>
<tr>
<th></th>
<th>3</th>
<th>2</th>
<th>1</th>
<th>3</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Etruscan</td>
<td>trade</td>
<td>price</td>
<td>temple</td>
<td>market</td>
<td></td>
</tr>
</tbody>
</table>

### Maintain Global Statistics

<table>
<thead>
<tr>
<th></th>
<th>3</th>
<th>2</th>
<th>1</th>
<th>3</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Etruscan</td>
<td>trade</td>
<td>price</td>
<td>temple</td>
<td>market</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Etruscan</td>
<td>1</td>
<td>0</td>
<td>35</td>
</tr>
<tr>
<td>market</td>
<td>50</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>price</td>
<td>42</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>temple</td>
<td>0</td>
<td>0</td>
<td>20</td>
</tr>
<tr>
<td>trade</td>
<td>10</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
What is the conditional distribution for this topic?

<table>
<thead>
<tr>
<th>3</th>
<th>?</th>
<th>1</th>
<th>3</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Etruscan</td>
<td>trade</td>
<td>price</td>
<td>temple</td>
<td>market</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>Etruscan</td>
<td>1</td>
<td>0</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td>market</td>
<td>50</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>price</td>
<td>42</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>temple</td>
<td>0</td>
<td>0</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>trade</td>
<td>10</td>
<td>8</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>
What is the conditional distribution for this topic?

- Part I: How much does this document like each topic?

<table>
<thead>
<tr>
<th></th>
<th>3</th>
<th>?</th>
<th>1</th>
<th>3</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Etruscan</td>
<td>trade</td>
<td>price</td>
<td>temple</td>
<td>market</td>
<td></td>
</tr>
</tbody>
</table>

Topic 1

Part II: How much does each topic like this word?

<table>
<thead>
<tr>
<th></th>
<th>3</th>
<th>?</th>
<th>1</th>
<th>3</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Etruscan</td>
<td>trade</td>
<td>price</td>
<td>temple</td>
<td>market</td>
<td></td>
</tr>
</tbody>
</table>

Topic 1

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>trade</td>
<td>10</td>
<td>7</td>
<td>1</td>
</tr>
</tbody>
</table>
What is the conditional distribution for this topic?

- Part I: How much does this document like each topic?
- Part II: How much does each topic like this word?

<table>
<thead>
<tr>
<th></th>
<th>Topic 1</th>
<th>Topic 2</th>
<th>Topic 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Etruscan</td>
<td><img src="image1.png" alt="Diagram" /></td>
<td><img src="image2.png" alt="Diagram" /></td>
<td><img src="image3.png" alt="Diagram" /></td>
</tr>
<tr>
<td>trade</td>
<td><img src="image4.png" alt="Diagram" /></td>
<td><img src="image5.png" alt="Diagram" /></td>
<td><img src="image6.png" alt="Diagram" /></td>
</tr>
<tr>
<td>price</td>
<td><img src="image7.png" alt="Diagram" /></td>
<td><img src="image8.png" alt="Diagram" /></td>
<td><img src="image9.png" alt="Diagram" /></td>
</tr>
<tr>
<td>temple</td>
<td><img src="image10.png" alt="Diagram" /></td>
<td><img src="image11.png" alt="Diagram" /></td>
<td><img src="image12.png" alt="Diagram" /></td>
</tr>
<tr>
<td>market</td>
<td><img src="image13.png" alt="Diagram" /></td>
<td><img src="image14.png" alt="Diagram" /></td>
<td><img src="image15.png" alt="Diagram" /></td>
</tr>
</tbody>
</table>

What is the conditional distribution for this topic?

- Part I: How much does this document like each topic?
- Part II: How much does each topic like this word?

$$
\sum_{k=1}^{K} \left( \frac{n^d_k + \alpha_k}{\sum_{j=1}^{K} n^d_j + \alpha_j} \frac{v^k_{\text{trade}}}{\sum_{j=1}^{V} v^k_j + \lambda_j} \right)
$$
Sample a New Topic Indicator

<table>
<thead>
<tr>
<th></th>
<th>?</th>
<th>1</th>
<th>3</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Etruscan</td>
<td>trade</td>
<td>price</td>
<td>temple</td>
<td>market</td>
</tr>
</tbody>
</table>

Topic 1  Topic 2  Topic 3

Update Counts

<table>
<thead>
<tr>
<th></th>
<th>?</th>
<th>1</th>
<th>3</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Etruscan</td>
<td>trade</td>
<td>price</td>
<td>temple</td>
<td>market</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Etruscan</td>
<td>1</td>
<td>0</td>
<td>35</td>
</tr>
<tr>
<td>market</td>
<td>50</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>price</td>
<td>42</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>temple</td>
<td>0</td>
<td>0</td>
<td>20</td>
</tr>
<tr>
<td>trade</td>
<td>10</td>
<td>7</td>
<td>1</td>
</tr>
</tbody>
</table>

...
Geometrically…

<table>
<thead>
<tr>
<th>3</th>
<th>1</th>
<th>1</th>
<th>3</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Etruscan</td>
<td>trade</td>
<td>price</td>
<td>temple</td>
<td>market</td>
</tr>
</tbody>
</table>

Issues with Generic LDA Sampling

- Slow mixing rates → Need many iterations
- Each iteration cycles through sampling topic assignments for all words in all documents
- Modern approaches:
- Alternative: Variational methods instead of sampling
  - Approximate posterior with an optimized variational distribution
Variational Methods

- Recall task: Characterize the posterior
- Turn posterior inference into an optimization task
- Introduce a “tractable” family of distributions over parameters and latent variables
  - Family is indexed by a set of “free parameters”
  - Find member of the family closest to:

Questions:
- How do we measure “closeness”?  
- If the posterior is intractable, how can we approximate something we do not have to begin with?

A Measure of Closeness

- Kullback-Leibler (KL) divergence
  - Measures “distance” between two distributions $p$ and $q$

- Not symmetric
- $p$ determines where the difference is important:
  - $p(x)=0$ and $q(x)\neq 0$
  - $p(x)\neq 0$ and $q(x)=0$

- Want

- Just as hard as the original problem!
Reverse Divergence

- Divergence $D(q \parallel p)$
  - true distribution $p$ defines support of diff.
  - the "correct" direction
  - will be intractable to compute
- Reverse divergence $D(q \parallel p)$
  - approximate distribution defines support
  - tends to give overconfident results
  - will be tractable

Interpretations of Minimizing Reverse KL

- Similarity measure:

- Evidence lower bound (ELBO)

- Therefore, minimizing KL is equivalent to maximizing a lower bound on the marginal likelihood:
  - $\max \mathcal{L}(q) = \min D(q||p) = \max \text{lower bound of } \log p(x)$
Mean Field

- How do we choose a $Q$ such that the following is tractable?

- Simplest case = mean field approximation
  - Assume each parameter and latent variable is conditionally independent given the set of free parameters

- Then, entropy term decomposes as

---

Mean Field

- Examine one free parameter, e.g., $\gamma$
  - Can rewrite joint as
    $$E_q[\log p(\theta, z, x)] = E_q[\log p(\theta | z, x)] + E_q[\log p(z, x)]$$
  - Look at terms of ELBO just depending on $\gamma$
    $$L^\gamma =$$

- Likewise,
  $$L^{\phi_{\gamma}} =$$

- This motivates using a coordinate ascent algorithm for optimization
  - Iteratively optimize each free parameter holding all others fixed
Mean Field for LDA

- In LDA, our parameters are \( \theta = \{\pi^d\}, \{\beta_k\} \)
  \( z = \{z_i^d\} \)

- The variational distribution factorizes as

\[
q(\pi, \beta, z) = \prod_{k=1}^{K} q(\beta_k | \eta_k) \prod_{d=1}^{D} q(\pi^d | \gamma^d) \prod_{i=1}^{N_d} q(z_i^d | \phi_i^d)
\]

- The joint distribution factorizes as

\[
p(\pi, \beta, z, w) = \prod_{k=1}^{K} p(\beta_k | \lambda) \prod_{d=1}^{D} p(\pi^d | \alpha) \prod_{i=1}^{N_d} p(z_i^d | \pi^d)p(w_i^d | z_i^d, \beta)
\]

- Examine the ELBO

\[
\mathcal{L}(q) = \sum_{k=1}^{K} E_q[\log p(\beta_k | \lambda)] + \sum_{d=1}^{D} E_q[\log p(\pi^d | \alpha)]
\]
\[
+ \sum_{d=1}^{D} \sum_{i=1}^{N_d} E_q[\log p(z_i^d | \pi^d)] + E_q[\log p(w_i^d | z_i^d, \beta)]
\]
\[
- \sum_{k=1}^{K} E_q[\log q(\beta_k | \eta_k)] - \sum_{d=1}^{D} E_q[\log q(\pi^d | \gamma^d)] - \sum_{i=1}^{N_d} E_q[\log q(z_i^d | \phi_i^d)]
\]
Mean Field for LDA

Let’s look at some of these terms

\[ E_q[\log p(z_i^d | \pi^d)] \]

\[ E_q[\log q(z_i^d | \phi_i^d)] \]

- Other terms follow similarly

Optimize via Coordinate Ascent

Algorithm:
Optimize via Coordinate Ascent

Algorithm:

Alternative Optimization Schemes

- Inefficient:
  - Start from randomly initialized $\eta_k$ (topics)
  - Analyze whole corpus before updating $\eta_k$ again
  - If streaming data scenario, can’t compute even one iteration!

- Didn’t have to do coord. ascent. Could have used gradient ascent.
Alternative Optimization Schemes

Recall stochastic gradient ascent:
- Assume $M = 1$
- Unbiased, but noisy

Here,
\[
L = E_q[\log p(\beta)] - E_q[\log q(\beta)] + \sum_{d=1}^{D} \left[ E_q[\log p(\pi^d)] - E_q[\log q(\pi^d)] \right]
+ \sum_{d=1}^{D} E_q[\log p(z^d, x^d | \pi^d, \beta)] - E_q[\log q(z^d)]
\]

\[
L_t = E_q[\log p(\beta)] - E_q[\log q(\beta)] + D \left( E_q[\log p(\pi^t)] - E[\log q(\pi^t)] \right)
+ D \left( E_q[\log p(z^t, x^t | \pi^t, \beta)] - E_q[\log q(z^t)] \right)
\]

Stochastic Variational Inference for LDA

- Initialize $\eta^{(0)}$ randomly.
- Repeat (indefinitely):
  - Sample a document $d$ uniformly from the data set.
  - For all $k$, initialize $\gamma_k^{(0)} = 1$
  - Repeat until converged
    - For $i=1,\ldots,N_d$
      \[
      \phi_{ik}^{(t)} \propto \exp\{E[\log \pi_k^t] + E[\log \beta_{k,w_i}^t]\}
      \]
    - Set $\gamma^d = \alpha + \sum_{i=1}^{N_d} \phi_{ik}^{(t)}$
  - Take a stochastic gradient step $\eta^{(t)} = \eta^{(t-1)} + \rho_t \nabla_q L_d$
Acknowledgements

- Thanks to Dave Blei, David Mimno, and Jordan Boyd-Graber for some material in this lecture relating to LDA