Case Study 5: Mixed Membership Modeling

LDA Collapsed Gibbs Sampler, Variational Inference

Machine Learning for Big Data
CSE547/STAT548, University of Washington
Emily Fox
May 28th, 2015

©Emily Fox 2015

Task 3: Mixed Membership Models

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

©Emily Fox 2015
Latent Dirichlet Allocation (LDA)

Top topics:
- Seeking Life’s Bare (Genetic) Necessities

Topic proportions and assignments:
- Topic proportions and assignments
- Every word is assigned to a topic
- Each doc has a doc-specific dist over topics
- Every word is labeled w/ a topic
- But we only observe the documents; the other structure is hidden.
- We compute the posterior $p$ topics, proportions, assignments |

All we see are words... for want: posterior $p$(topics | doc proportions, assign vars | words)
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:

\[
\begin{align*}
&\text{assign var. per word} \\
&\begin{array}{c}
\pi^d \sim \text{Dir}(\lambda_1, \ldots, \lambda_K) \\
\beta_k \sim \text{Dir}(\lambda_{1,1}, \ldots, \lambda_{K,K})
\end{array} \\
&w^d_i \sim \beta_{z^d_i} \quad i = 1, \ldots, N_d
\end{align*}
\]

Collapsed LDA Sampling

- Sample topic indicators for each word
  - Algorithm:

\[
p(z^d_i = k \mid z \_\setminus i, \{w^d_i\}, \alpha, \lambda) \\
\propto p(z^d_i = k \mid \{z^d_j, j \neq i\}, \alpha) p(w^d_i \mid \{w^c_j : z^c_j = k, (j, c) \neq (i, d)\}, \lambda)
\]
## Select a Document

<table>
<thead>
<tr>
<th>Etruscan</th>
<th>trade</th>
<th>price</th>
<th>temple</th>
<th>market</th>
</tr>
</thead>
</table>

All words in doc \(d\)

## Randomly Assign Topics

<table>
<thead>
<tr>
<th>(z^d)</th>
<th>3</th>
<th>2</th>
<th>1</th>
<th>3</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>(w^d)</td>
<td>Etruscan</td>
<td>trade</td>
<td>price</td>
<td>temple</td>
<td>market</td>
</tr>
</tbody>
</table>

To initialize sampler (one approach)
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>

Maintain Local Statistics

<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>

<table>
<thead>
<tr>
<th>Doc d</th>
<th>Topic 1</th>
<th>Topic 2</th>
<th>Topic 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>d</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

©Emily Fox 2015
### Maintain Global Statistics

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

<table>
<thead>
<tr>
<th>Topic 1</th>
<th>Topic 2</th>
<th>Topic 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Etruscan</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>market</td>
<td>50</td>
<td>0</td>
</tr>
<tr>
<td>price</td>
<td>42</td>
<td>1</td>
</tr>
<tr>
<td>temple</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>trade</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Resample Assignments

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

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

\[ p(z^d_A | \text{everything else}) \]

### Part I: How much does this document like each 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>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Topic 1</th>
<th>Topic 2</th>
<th>Topic 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doc d</td>
<td>2</td>
<td>0</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>3</td>
<td>trade</td>
<td>price</td>
<td>temple</td>
</tr>
<tr>
<td>1</td>
<td>trade</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[ z_{d,i} \quad w_{d,i} \]

\[ m_{i,d,k} = \lambda_{\text{trade}} \frac{m_{i,d,k} + \lambda}{\sum_k m_{i,d,k} + \lambda} \]
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?

\[
\frac{n_{id} - \alpha_k}{\sum_k \alpha_k \sum_{\gamma=1}^V m_{\gamma,di} + \lambda_{\gamma}}
\]

Sample a New Topic Indicator

\[
\frac{n_{id} - \alpha_k}{\sum_k \alpha_k \sum_{\gamma=1}^V m_{\gamma,di} + \lambda_{\gamma}}
\]
Update Counts

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

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

Geometrically...

inc. popularity of topic 1 in doc d

and prevalence of word "trade" in topic 1 (corpus wide)
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 include:
  – Large-scale LDA. For example,
  – Distributed LDA. For example,
  – And many, many more!

• Alternative: Variational methods instead of sampling
  – Approximate posterior with an optimized variational distribution

Case Study 5: Mixed Membership Modeling

Variational Methods

Machine Learning for Big Data
CSE547/STAT548, University of Washington
Emily Fox
May 28th, 2015
Variational Methods Goal

- Recall task: Characterize the posterior

\[ p(\theta, z | x) \]

- 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:

\[ p(\theta, z | x) \]

Call family \( Q \) and want \( q \in Q \) that is closest to \( p(\theta, z | x) \)

Variational Methods Cartoon

- Cartoon of goal:

- Questions:
  1. How do we measure “closeness”? 
  2. 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 \( q \)

\[
KL(p \| q) \triangleq D(p \| q) = \mathbb{E}_p \left[ \log \frac{p(\theta)}{q(\theta)} \right] = \int_{\Theta} p(\theta) \log \frac{p(\theta)}{q(\theta)} d\theta
\]

• If \( p = q \) for all \( \theta \)

\[
D(p \| q) = \int p(\theta) \log 1 d\theta = 0
\]

• Otherwise, \( D(p \| q) > 0 \)

---

A Measure of Closeness

\[
KL(p \| q) \triangleq D(p \| q) = \int_\Theta p(\theta) \log \frac{p(\theta)}{q(\theta)} d\theta
\]

• Not symmetric \( D(p \| q) \neq D(q \| p) \) ... not a true distance metric

• \( p \) determines where the difference is important:
  \( \exists \theta \)
  - \( p(\theta) = 0 \) and \( q(\theta) > 0 \) \( \Rightarrow \log 0 \text{ undefined} \)
  - \( p(\theta) > 0 \) and \( q(\theta) = 0 \) \( \Rightarrow \log \frac{0}{0} = \infty \)

If \( D(p \| q) \) finite, \( \text{supp}(q) \supseteq \text{supp}(p) \)

• Want \( \hat{q} = \arg \min_{q \in Q} D(p \| q) \)

• Just as hard as the original problem!

©Emily Fox 2015
Reverse Divergence

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

Interpretations of Minimizing Reverse KL

$$D(q\|p) = E_q \left[ \log \frac{q}{p} \right]$$

- Similarity measure:
  $$D(q(\theta, z)\| p(\theta, z | x)) = E_q[\log q(\theta, z)] - E_q[\log p(\theta, z | x)]$$
  $$= E_q[\log q(\theta, z)] - E_q[\log p(\theta, z | x)] + \log p(x)$$

- Evidence lower bound (ELBO)
  $$\log p(x) = D(q(\theta, z)\| p(\theta, z | x)) + L(q) \geq L(q)$$
  “ELBO”
Interpretations of Minimizing Reverse KL

- Evidence lower bound (ELBO)

\[
\log p(x) = D(q(z, \theta) \| p(z, \theta | x)) + \mathcal{L}(q) \geq \mathcal{L}(q)
\]

- Therefore,
  - ELBO provides a lower bound on marginal likelihood
  - Maximizing ELBO is equivalent to minimizing KL

\[
\max \mathcal{L}(q) = \min D(q \| p) = \max \text{lower bound on } \log p(x)
\]

Mean Field

\[
\mathcal{L}(q) = E_q[\log p(z, \theta, x)] - E_q[\log q(z, \theta)]
\]

- How do we choose a \( Q \) such that the following is tractable?

\[
\hat{q} = \arg \max_{q \in Q} \mathcal{L}(q) \leftarrow \text{new objective}
\]

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

\[
q(\theta | z) = q(\theta | y) \prod_i q(z_i | \phi_i)
\]

\( \theta, z = z_1, \ldots, z_N \) are free parameters

\( z_1 \) et al. are also structured approx.
Mean Field
\[ L(q) = E_q[\log p(z, \theta, x)] - E_q[\log q(z, \theta)] \]

- Naïve mean field decomposition:
  \[ q(z, \theta) = q(\theta | \gamma) \prod_{i=1}^{N} q(z^i | \phi^i) \]

- Under this approximation, entropy term decomposes as
  \[ -E_q[\log q(\theta, z)] = E_q[\log q(\theta | z)] - \sum_{i} E_q[\log q(z^i | \phi^i)] \]

- Can (always) rewrite joint term as
  \[ E_q[\log p(\theta, z, x)] = E_q[\log p(\theta | z, x)] + E_q[\log p(z, x)] \]

Mean Field – Optimize \( \gamma \)

- Examine one free parameter, e.g., \( \gamma \)
  \[ L(q) = E_q[\log p(\theta | z, x)] + E_q[\log p(z, x)] - E_q[\log q(\theta | \gamma)] - \sum E_q[\log q(z^i | \phi^i)] \]

- Look at terms of ELBO just depending on \( \gamma \)
  \[ L^\gamma = E_q[\log p(\theta | z, x)] - E_q[\log q(\theta | \gamma)] + \text{const} \]

really just \( q_0 = q(\theta | \gamma) \) needed here
Mean Field – Optimize

- Examine another free parameter, e.g., \( \phi_i \)

\[
\mathcal{L}(q) = E_q[\log p(z^i | z, \theta, x)] - E_q[\log p(z, \theta, x)] - E_q[\log q(\theta | z^i, \phi)] - \sum_i E_q[\log q(z^i | \phi^i)]
\]

- Look at terms of ELBO just depending on \( \phi_i \)

\[
\mathcal{L}^{\phi_i} = E_q[\log p(z^i | z, \theta, x)] - E_q[\log q(z^i | \phi^i)]
\]

- This motivates using a coordinate ascent algorithm for optimization

  - Iteratively optimize each free parameter holding all others fixed

Algorithm Outline

- **Initialization:** Randomly select starting distribution \( q^{(0)}_{\theta} \)
- **E-Step:** Given parameters, find posterior of hidden data

  \[
  q^{(t)}_z = \arg \max_{q_z} \mathcal{L}(q_z, q^{(t-1)}_{\theta})
  \]

- **M-Step:** Given posterior distributions, find likely parameters

  \[
  q^{(t)}_{\theta} = \arg \max_{q_{\theta}} \mathcal{L}(q^{(t)}_z, q^{(t)}_{\theta})
  \]

- **Iteration:** Alternate E-step & M-step until convergence
Case Study 5: Mixed Membership Modeling

Variational Inference for LDA

Mean Field for LDA

- In LDA, our parameters are $\theta = \{\pi^d\}, \{\beta_k\}$
  $z = \{z^d_i\}$

- The variational distribution factorizes as

$$q(\pi, \beta, z) = \prod_{k=1}^{K} q(\beta_k | \eta_k) \prod_{d=1}^{D} \left[ q(z^d_i | \phi^d_i) \prod_{i=1}^{N_d} q(z^d_i | \phi^d_i) \right]$$

- The joint distribution factorizes as

$$p(\pi, \beta, z, w) = \prod_{k=1}^{K} p(\beta_k | \lambda) \prod_{d=1}^{D} p(\pi^d | \alpha) \prod_{i=1}^{N_d} p(z^d_i | \pi^d) p(w^d_i | z^d_i, \beta)$$
Mean Field for LDA

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

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

- Examine the ELBO

\[ \mathcal{L}(q) = \sum_{k=1}^{K} E_q[\log p(\beta_k | \lambda)] + \sum_{d=1}^{D} E_q[\log p(\pi^d | \alpha)] \]

\[ + \sum_{d=1}^{D} \sum_{i=1}^{N_d} E_q[\log p(z^d_i | \pi^d)] + E_q[\log p(w^d_i | z^d_i, \beta)] \]

\[- \sum_{k=1}^{K} E_q[\log q(\beta_k | \eta_k)] - \sum_{d=1}^{D} E_q[\log q(\pi^d | \gamma^d)] - \sum_{d=1}^{D} \sum_{i=1}^{N_d} E_q[\log q(z^d_i | \phi^d_i)] \]

Let's look at some of these terms

\[ E_q[\log p(z^d_i | \pi^d)] = E_q[\log \prod_{k=1}^{K} \phi^d_{ik}] = E_q[\log \prod_{k=1}^{K} \psi(\beta_k) \phi^d_{ik}] \]

\[ = \sum_{k=1}^{K} E_q[I(\beta_k = 1) \log \phi^d_{ik}] = \sum_{k=1}^{K} E_q[I(\beta_k = 1)] E_q[\log \phi^d_{ik}] \]

\[ = \sum_{k=1}^{K} E_q[I(\beta_k = 1)] \log \phi^d_{ik} \]

\[ E_q[\log q(z^d_i | \phi^d_i)] \]

- Other terms follow similarly
Optimize via Coordinate Ascent

- **Algorithm:**

  \[
  \begin{align*}
  &\text{for } d = 1, \ldots, D \\
  &\frac{\partial L}{\partial \gamma^d} = 0 \quad \Rightarrow \quad \gamma^{d(t+1)} = \alpha + \sum_{i=1}^{K} \phi_i^d
  \\
  &\text{for } i = 1, \ldots, Nd \\
  &\frac{\partial L}{\partial \phi_i^d} = 0 \quad \Rightarrow \quad \phi_i^d \propto \exp \left( \psi \left( y_i^{d(t+1)} \right) \right) + \psi \left( \bar{y}_{1:k} \right) - \psi \left( \sum_{i=1}^{K} \bar{y}_{1:k} \right)
  \\
  &\text{use Lagrange multiplier to enforce } \sum \phi_i^d = 1
  \\
  &\text{DATA PARALLEL across } \phi_i^d
  \end{align*}
  \]

©Emily Fox 2015
What you need to know...

- Latent Dirichlet allocation (LDA)
  - Motivation and generative model specification
  - Collapsed Gibbs sampler

- Variational methods
  - Overall goal
  - Interpretation in terms of minimizing (reverse) KL
  - Mean field approximation

Acknowledgements

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