Task 3: Mixed Membership Models

- **Now:** Document may belong to multiple clusters
Latent Dirichlet Allocation (LDA)

Topics
- gene 0.02
- race 0.02
- genetic 0.02
- life 0.02
- disease 0.02
- organ 0.01
- brain 0.02
- disease 0.02
- care 0.01
- data 0.02
- number 0.02
- computer 0.02

Documents

Previously, each doc had one topic
now, each is a mixture of topics

Topic proportions and assignments

Seeking Life’s Bare (Genetic) Necessities

We are all close to genes, especially in comparison to the US community. Just as before, we model the choice of words for each document as a function of the topic proportions and assignments.

Each topic is a distribution over words in vocabulary V:

\\[ p(v | \theta) = \prod_{k=1}^{K} \theta_{k,v} \]

We compute the posterior:

\\[ p(\theta, \phi, \beta | D) \]

where D is the observed documents.

Every word is assigned to a topic
Each doc has its own prevalence of topics in that doc

All we see are words \( \hat{w} \)

Want: posterior \[ p(\theta, \phi, \beta | \hat{w}) \]
assign vars. \( \theta, \phi, \beta \)

Want it to be fast!

©Emily Fox 2014
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:

\[
\begin{align*}
Z_{d} & \sim \Pi^d \quad d = 1, \ldots, D \\
W_{d}^i | Z_{d} & \sim \beta_{Z_{d}} \\
\end{align*}
\]

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

LDA Joint Probability

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

- Marginalize parameters
  - Document-specific topic weights
  - Corpus-wide topic-specific word distributions
  \[
p(z_i^d = k \mid z_{\backslash id}, \{w_i^d\}, \alpha, \lambda)
  \propto p(z_i^d = k \mid z_{\backslash id}, \alpha)p(w_i^d \mid z_i^d = k, z_{\backslash id}, w_{\backslash id}, \lambda)
  \]
- Unplate to see dependencies induced

Sample topic indicators for each word

Algorithm:

\[
p(z_i^d = k \mid z_{\backslash id}, \{w_i^d\}, \alpha, \lambda)
  \propto p(z_i^d = k \mid \{z_j^d, j \neq i\}, \alpha)p(w_i^d \mid \{w_j^c : z_j^c = d, (j, c) \neq (i, d)\}, \lambda)
\]
## Select a Document

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

Maintain Local Statistics
## Maintain Global Statistics

<table>
<thead>
<tr>
<th>Topic</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>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>
</tbody>
</table>

Total counts from all docs

## Resample Assignments

<table>
<thead>
<tr>
<th>Topic</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>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>
</tbody>
</table>

©Emily Fox 2014
What is the conditional distribution for this topic?

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

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

<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>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>

<table>
<thead>
<tr>
<th></th>
<th>Topic 1</th>
<th>Topic 2</th>
<th>Topic 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>trade</td>
<td>10</td>
<td>7</td>
<td>1</td>
</tr>
</tbody>
</table>

Etruscan trade price temple market
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>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>

Sample a New Topic Indicator

\[
\begin{align*}
&n_k^{-id} + \alpha_k v_{trade,k}^{-id} + \lambda_{trade} \\
&\frac{1}{N_d - 1 + \sum_k \alpha_k \sum_{\gamma=1}^V v_{\gamma,k}^{-id} + \lambda_{\gamma}}
\end{align*}
\]
Update Counts

<table>
<thead>
<tr>
<th>z^d</th>
<th>w^d</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>?</td>
</tr>
<tr>
<td>Etruscan</td>
<td>trade</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>10</td>
<td>7</td>
</tr>
</tbody>
</table>

... |

Topic 1  Topic 2  Topic 3
Doc d     2       0       2

Geometrically…

<table>
<thead>
<tr>
<th>z^d</th>
<th>w^d</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Etruscan</td>
<td>trade</td>
</tr>
</tbody>
</table>

Topic 1  Topic 2  Topic 3

©Emily Fox 2014
Issues with Generic LDA Sampling

- Slow mixing rates $\rightarrow$ Need many iterations
- Each iteration cycles through sampling topic assignments for all words in all documents
- Modern approaches include:
  - 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
Variational Methods Goal

- 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:

Variational Methods Cartoon

- Cartoon of goal:

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 \( q \)

- If \( p = q \) for all \( \theta \)

- Otherwise,

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

- Not symmetric
- \( p \) determines where the difference is important:
  - \( p(\theta)=0 \) and \( q(\theta)\neq0 \)
  - \( p(\theta)\neq0 \) and \( q(\theta)=0 \)

- Want

- Just as hard as the original problem!
Reverse Divergence

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

Interpretations of Minimizing Reverse KL

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

- Similarity measure:
  - Evidence lower bound (ELBO)
Interpretations of Minimizing Reverse KL

- Evidence lower bound (ELBO)
\[
\log p(x) = D(q(z, \theta) \parallel p(z, \theta|x)) + \mathcal{L}(\theta) \geq \mathcal{L}(\theta)
\]

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

Mean Field
\[
\mathcal{L}(\theta) = E_{\theta} \left[ \log p(z, \theta, x) \right] - E_{\theta} \left[ \log q(z, \theta) \right]
\]

- 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

Original graph

\begin{center}
\begin{tikzpicture}
\node at (0,0) (a) {}; \node at (0.866,0.5) (b) {}; \node at (1.732,0) (c) {}; \node at (2.6,0.5) (d) {}; \node at (0,1.732) (e) {}; \node at (0.866,2.293) (f) {}; \node at (1.732,1.732) (g) {}; \node at (2.6,2.293) (h) {};
\end{tikzpicture}
\end{center}

Naïve mean field

\begin{center}
\begin{tikzpicture}
\node at (0,0) (i) {}; \node at (0.866,0.5) (j) {}; \node at (1.732,0) (k) {}; \node at (2.6,0.5) (l) {}; \node at (0,1.732) (m) {}; \node at (0.866,2.293) (n) {}; \node at (1.732,1.732) (o) {}; \node at (2.6,2.293) (p) {};
\end{tikzpicture}
\end{center}
Mean Field

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

- 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

  \[ E_q[\log q(\theta | z, x)] = E_q[\log q(\theta | z, x)] - E_q[\log p(z, x)] \]

- Can (always) rewrite joint term as

  \[ E_q[\log p(\theta, z, x)] = E_q[\log p(\theta | z, x)] + E_q[\log p(z, x)] \]

  \[ E_q[\log p(\theta, z, x)] = E_q[\log p(z^i | z_{\setminus i}, \theta, x)] + E_q[\log p(z_{\setminus i}, \theta, x)] \]

Mean Field – Optimize \( \gamma \)

- Examine one free parameter, e.g., \( \gamma \)

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

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

  \[ \mathcal{L}^\gamma = \]
Mean Field – Optimize $\phi^i$

- Examine another free parameter, e.g., $\phi^i$

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

- Look at terms of ELBO just depending on $\phi^i$

\[
L^{\phi^i} = 
\]

- 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 $q_{(0)}$

- **E-Step:** Given parameters, find posterior of hidden data

\[
q_z^{(t)} = \text{arg max}_{q_z} L(q_z, q_{(t-1)}^{(t)})
\]

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

\[
q_{(t)} = \text{arg max}_{q_{}} L(q_{z}^{(t)}, q_{})
\]

- **Iteration:** Alternate E-step & M-step until convergence
Mean Field for LDA

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

$\quad z = \{z^d_i\}$

The variational distribution factorizes as

- 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

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

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

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

- Let's look at some of these terms

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

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

- Other terms follow similarly

Optimize via Coordinate Ascent

- Algorithm:
Optimize via Coordinate Ascent

Algorithm:

Generalizing...

\[ \log p(x) \geq \int q_z(z)q_{\theta}(\theta) \log \frac{p(x, z, \theta)}{q_z(z)q_{\theta}} dz d\theta \]

- Condition 1: Complete data likelihood is in exponential family
- Condition 2: Parameter prior is conjugate to complete data likelihood

<table>
<thead>
<tr>
<th>EM for MAP estimation</th>
<th>Variational Bayesian EM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Goal: maximise ( p(\theta</td>
<td>x) ) w.r.t. ( \theta )</td>
</tr>
<tr>
<td>E Step: compute ( q_z^{(t+1)}(z) = p(z</td>
<td>x, \theta^{(t)}) )</td>
</tr>
<tr>
<td>M Step: ( \theta^{(t+1)} = \arg \max_{\theta} \int q_z^{(t+1)}(z) \ln p(z, x, \theta) , dz )</td>
<td>VB-M Step: ( q_{\theta}^{(t+1)}(\theta) \propto \exp \left[ \int q_z^{(t+1)}(z) \ln p(z, x, \theta) , dz \right] )</td>
</tr>
</tbody>
</table>
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
  - Mean field for LDA

Reading

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

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