Case Study 5: Mixed Membership Modeling

Clustering Documents Revisited, Latent Dirichlet Allocation

Machine Learning for Big Data
CSE547/STAT548, University of Washington
Emily Fox
May 26\textsuperscript{th}, 2015

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:

  \[
  \begin{align*}
  \text{Sample topic:} & \quad z^d \sim \pi \\
  \text{Sample words:} & \quad w^d_i \mid z^d \sim \beta_{z^d_i} \\
  \text{Given topic } z^d = k \text{ for doc } d, \quad & \text{draw each word from } \beta_k
  \end{align*}
  \]

Bayesian Document Model

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

  \[
  \sum_{k=1}^{K} \pi_k = 1, \quad \sum_{v} \beta_{kv} = 1
  \]

  What is a distribution on the simplex?

  First, what is the simplex?
**Dirichlet Distributions**

- The Dirichlet distribution is defined on the simplex

\[ \alpha_k = 10 \forall k \]

\[
\begin{align*}
\text{Moments: } E_{\alpha} [\pi_k] &= \frac{\alpha_k}{\alpha_0} \\
\text{Var}_{\alpha} [\pi_k] &= \frac{K - 1}{K^2(\alpha_0 + 1)}
\end{align*}
\]

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

\[ \alpha_k = 0.1 \forall k \]

**Model Summary**

- Prior on model parameters
  - E.g., symmetric Dirichlet for \( \pi \)

\[ \pi \sim \text{Dir}(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K}) \]

- Dirichlet prior for topic parameters \( \beta_k \sim \text{Dir}(\frac{1}{V}, \ldots, \frac{1}{V}) \)

- Sample observations as

\[ z^d \sim \pi \quad d = 1, \ldots, D \]

\[ w_i^d \mid z^d \sim \beta_{z^d} \quad i = 1, \ldots, N_d \]
Posterior Inference via Sampling

• Iterate between sampling
  \[ \pi \sim p(\pi | z^d, \beta_k, \omega_k) \]
  For \( k = 1, \ldots, K \)
  \[ \beta_k \sim p(\beta_k | \pi, z^d, \omega_k) \]
  For \( d = 1, \ldots, D \)
  \[ z^d \sim p(z^d | \pi, \beta_k, \omega_k) \]

• What form do these complete conditionals take?
  – First a look at statements of conditional independence in directed graphical models

Markov Blanket

• A node is conditionally independent of all other nodes in the graph given its Markov blanket

Markov blanket = - all parents
  of \( x_i \)
- all children
  - all coparents

• Gibbs sampling iterates between full conditionals
  \[ x_i \sim p(x_i | x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_N) \]
  \[ \Rightarrow \text{simplify to} \]
  \[ x_i \sim p(x_i | \text{MB}(x_i)) \]
Unplated Document Model

- Recall that the plate notation is really indicating

\[ \begin{align*}
\alpha & \rightarrow \pi \\
\pi & \rightarrow z^d \\
w^d_i & \rightarrow N_d \\
\beta_k & \leftarrow \lambda
\end{align*} \]

Complete Conditional for \( \pi \)

- Recall conjugate Dirichlet prior

\( \pi \sim \text{Dir}(\alpha_1, \ldots, \alpha_K) \quad p(\pi | \alpha) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k \pi_k^{\alpha_k - 1} \)

- Likelihood:
- Dirichlet posterior
  - Count occurrences of
  - Then,

- Conjugacy: Posterior has same form as prior
Complete Conditional for $\beta_k$

- Again, Dirichlet prior
- Consider docs $d$ such that
  - For these observations,
  - Do any other docs depend on $\beta_k$?
- Then,

  - Again, posterior has same form as prior

Complete Conditional for $z^d$

- We have $z^d \sim \pi$
  
  \[ w_i^d \mid z^d, \{\beta_k\} \sim \beta_{z^d} \]
- 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
Collapsed Gibbs Sampler

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

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

Collapsed Sampler Full Conditional

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

- Derivation

\[
p(z^d = k | z_{\setminus d}, \{w^d_i\}, \alpha, \lambda) \propto \int_{\beta_1} \cdots \int_{\beta_K} p(\cdot)
\]
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 \( \pi, \theta_k \)

Example – Uncollapsed Results

Figure courtesy of Erik Sudderth
Given previous cluster assignments optima for many iterations (see right columns of Figs. 2.18 and 2.19). The Rao–Blackwellized sampler has superior typical performance, but occasionally remains trapped in low quantiles (thin dashed) of the resulting log–likelihood sequence for 20 different random initializations of each algorithm.

Figure 2.20. Efficient statistics of the parameters as defined in Fig. 2.16.

Algorithm 2.2.

1. Sample a random permutation
2. Set $z_i = 1$ for each of the $N$ observations
3. For each of the $K$ components, sequentially sample new assignments as follows:
   - Compute $p(z_i = k | x_i, z_{\neq i}, \pi, \theta)$ and $p(z_{\neq i} | z_i = k, x_{\neq i}, \pi, \theta)$
   - Sample $z_i \sim$ Multinomial($\pi_k$, $\tau_k$)
   - Optionally, mixture parameters may be sampled via steps 2–3 of Alg. 2.2.

This likelihood can be computed from cached sufficient statistics via Prop. 2.14.

Comparison of standard (Alg. 2.1) and Rao–Blackwellized (Alg. 2.2) samplers for a mixture of 4 two–dimensional Gaussians. We compare data log–likelihoods at $T=2$ (top), $T=10$ (middle), $T=50$ (bottom) iterations from two random initializations. Each plot is labeled by the current data log–likelihood.

Log–likelihood vs. Gibbs iteration (multiple chains)

Figure courtesy of Erik Sudderth ©Emily Fox 2015
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 with radically different approaches presented complementary views of the basic genes needed for life.

One research team, using computer analyses to compare known genomes, concluded that today’s organisms 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 species and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 would be enough.

Although the numbers don’t match precisely, those predictions

---


“are not all that far apart,” especially in comparison to the 75,000 genes in the human genome, notes Sir Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more organisms are completely mapped and sequenced. “It may be a way of organizing any newly sequenced genome,” explains Aradhya Murthy, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an
Latent Dirichlet Allocation (LDA)

 Topics

 Documents

 Topic proportions and assignments

 Seeking Life's Bare (Genetic) Necessities

 "Life's bare necessities..." (i.e., survival need, most critical for survival of the species, not just one individual) provide an excellent model for our approach to protein function prediction. In this paper, we describe our approach to predicting protein function from the protein's DNA sequence. Our approach is based on the idea that proteins are related at the DNA level, and that this relationship can be used to predict protein function. We show that our approach is able to accurately predict protein function, and that it can be used to identify new protein functions.

 But we only observe the documents; the other structure is hidden.

 We compute the posterior $p$.

 Topics

 Documents

 Topic proportions and assignments

 Seeking Life's Bare (Genetic) Necessities

 "Life's bare necessities..." (i.e., survival need, most critical for survival of the species, not just one individual) provide an excellent model for our approach to protein function prediction. In this paper, we describe our approach to predicting protein function from the protein's DNA sequence. Our approach is based on the idea that proteins are related at the DNA level, and that this relationship can be used to predict protein function. We show that our approach is able to accurately predict protein function, and that it can be used to identify new protein functions.

 But we only observe the documents; the other structure is hidden.

 We compute the posterior $p$.

 ©Emily Fox 2015
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:

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

---

**Seeking Life's Bare (Genetic) Necessities**

Cold Spring Harbor, New York—How many genes does a sperm need to fertilize an egg? Researchers have an answer. As stated last year, from the bad genes needed for life, one can think of a computer analogy. Is it possible to imagine a computer that requires no bad genes, not even a 1% chance of bad genes? The German researchers have shown that it’s possible. They have shown that the genetic code of 120 would be enough. Although the numbers don’t match perfectly, these predictions are encouraging.

---

Example Inference – Topic Words

- **Topics:**
  - human
  - genome
  - dna
  - genetic
  - genes
  - sequence
  - gene
  - molecular
  - sequencing
  - map
  - information
  - genetics
  - mapping
  - project
  - sequences
  - evolution
  - evolutionary
  - species
  - organisms
  - life
  - origin
  - biology
  - groups
  - phylogenetic
  - living
  - diversity
  - group
  - new
  - two
  - common
  - disease
  - host
  - bacteria
  - organisms
  - resistance
  - control
  - infectious
  - malaria
  - parasite
  - united
  - tuberculosis
  - computer
  - models
  - information
  - data
  - system
  - network
  - method
  - networks
  - software
  - simulations
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

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

Maintain Local Statistics

©Emily Fox 2015
Maintain Global Statistics

\[
\begin{array}{cccc}
\text{Etruscan} & \text{trade} & \text{price} & \text{temple} & \text{market} \\
3 & 2 & 1 & 3 & 1 \\
\end{array}
\]

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

Total counts from all docs

Resample Assignments

\[
\begin{array}{cccc}
\text{Etruscan} & \text{trade} & \text{price} & \text{temple} & \text{market} \\
3 & 2 & 1 & 3 & 1 \\
\end{array}
\]

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

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

©Emily Fox 2015
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>3</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>trade</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>price</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>temple</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>market</td>
<td></td>
<td></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>
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?

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

Sample a New Topic Indicator

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

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

©Emily Fox 2015
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:
  - And many, many more!

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

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