Bayesian Inference for Parameter Estimation

+ Topic Modeling

Matt Gormley
Lecture 20
Nov. 4, 2019
Reminders

• Homework 3: Structured SVM
  – Out: Fri, Oct. 24
  – Due: Wed, Nov. 6 at 11:59pm

• Homework 4: Topic Modeling
  – Out: Wed, Nov. 6
  – Due: Mon, Nov. 18 at 11:59pm
TOPIC MODELING
Topic Modeling

Motivation:
Suppose you’re given a massive corpora and asked to carry out the following tasks

• **Organize** the documents into **thematic categories**
• **Describe** the evolution of those categories **over time**
• Enable a domain expert to **analyze and understand** the content
• Find **relationships** between the categories
• Understand how **authorship** influences the content
Topic Modeling

Motivation:
Suppose you’re given a massive corpora and asked to carry out the following tasks
• **Organize** the documents into **thematic categories**
• **Describe** the evolution of those categories **over time**
• Enable a domain expert to **analyze and understand** the content
• Find **relationships** between the categories
• Understand how **authorship** influences the content

Topic Modeling:
A method of (usually unsupervised) discovery of latent or hidden structure in a corpus
• Applied primarily to text corpora, but **techniques are more general**
• Provides a **modeling toolbox**
• Has prompted the exploration of a variety of new **inference methods** to accommodate **large-scale datasets**
Topic Modeling

Dirichlet-multinomial regression (DMR) topic model on ICML (Mimno & McCallum, 2008)

Topic 0 [0.152]

Problem, optimization, problems, convex, convex optimization, linear, semidefinite programming, formulation, sets, constraints, proposed, margin, maximum margin, optimization problem, linear programming, programming, procedure, method, cutting plane, solutions.

Topic 54 [0.051]

Decision trees, trees, tree, decision tree, decision, tree ensemble, junction tree, decision tree learners, leaf nodes, arithmetic circuits, ensembles modts, skewing, ensembles, anytime induction decision trees, trees trees, random forests, objective decision trees, tree learners, trees grove, candidate split.

Topic 99 [0.066]

Inference, approximate inference, exact inference, markov chain, models, approximate, gibbs sampling, variational, bayesian, variational inference, variational bayesian, approximation, sampling, methods, exact, bayesian inference, dynamic bayesian, process, mcmc, efficient.

http://www.cs.umass.edu/~mimno/icml100.html
Topic Modeling

• Map of NIH Grants

(Tralley et al., 2011)

https://app.nihmaps.org/
Other Applications of Topic Models

• Spacial LDA

(Wang & Grimson, 2007)
Outline

• Applications of Topic Modeling
• Latent Dirichlet Allocation (LDA)
  1. Beta-Bernoulli
  2. Dirichlet-Multinomial
  3. Dirichlet-Multinomial Mixture Model
  4. LDA
• Bayesian Inference for Parameter Estimation
  – Exact inference
  – EM
  – Monte Carlo EM
  – Gibbs sampler
  – Collapsed Gibbs sampler
• Extensions of LDA
  – Correlated topic models
  – Dynamic topic models
  – Polylingual topic models
  – Supervised LDA
BAYESIAN INFERENCE FOR NAÏVE BAYES
Beta-Bernoulli Model

- Beta Distribution

\[
f(\phi | \alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1 - x)^{\beta-1}
\]
Beta-Bernoulli Model

• Generative Process

\[ \phi \sim \text{Beta}(\alpha, \beta) \]

For each word \( n \in \{1, \ldots, N\} \)

\[ x_n \sim \text{Bernoulli}(\phi) \]

[draw distribution over words]

[draw word]

• Example corpus (heads/tails)

<table>
<thead>
<tr>
<th>H</th>
<th>T</th>
<th>T</th>
<th>H</th>
<th>H</th>
<th>T</th>
<th>T</th>
<th>H</th>
<th>H</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>x_1</td>
<td>x_2</td>
<td>x_3</td>
<td>x_4</td>
<td>x_5</td>
<td>x_6</td>
<td>x_7</td>
<td>x_8</td>
<td>x_9</td>
<td>x_{10}</td>
</tr>
</tbody>
</table>
Dirichlet-Multinomial Model

- Dirichlet Distribution

\[
f(\phi|\alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1 - x)^{\beta-1}
\]
Dirichlet-Multinomial Model

- Dirichlet Distribution

\[
p(\phi | \alpha) = \frac{1}{B(\alpha)} \prod_{k=1}^{K} \phi_k^{\alpha_k - 1} \quad \text{where} \quad B(\alpha) = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{K} \alpha_k)}
\]
Dirichlet-Multinomial Model

- Generative Process

\[
\phi \sim \text{Dir}(\beta) \quad \text{[draw distribution over words]}
\]

For each word \( n \in \{1, \ldots, N\} \)
\[
x_n \sim \text{Mult}(1, \phi) \quad \text{[draw word]}
\]

- Example corpus

<table>
<thead>
<tr>
<th>the</th>
<th>he</th>
<th>is</th>
<th>the</th>
<th>and</th>
<th>the</th>
<th>she</th>
<th>she</th>
<th>is</th>
<th>is</th>
</tr>
</thead>
<tbody>
<tr>
<td>(x_1)</td>
<td>(x_2)</td>
<td>(x_3)</td>
<td>(x_4)</td>
<td>(x_5)</td>
<td>(x_6)</td>
<td>(x_7)</td>
<td>(x_8)</td>
<td>(x_9)</td>
<td>(x_{10})</td>
</tr>
</tbody>
</table>
Dirichlet-Multinomial Mixture Model

• Generative Process

topics → documents \{ “mixture” \}

• Example corpus

<table>
<thead>
<tr>
<th>Document 1</th>
<th>Document 2</th>
<th>Document 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>he</td>
<td>is</td>
</tr>
<tr>
<td>$X_{11}$</td>
<td>$X_{12}$</td>
<td>$X_{13}$</td>
</tr>
</tbody>
</table>
Dirichlet-Multinomial Mixture Model

• Generative Process

For each topic $k \in \{1, \ldots, K\}$:
- $\phi_k \sim \text{Dir}(\beta)$
- $\theta \sim \text{Dir}(\alpha)$

For each document $m \in \{1, \ldots, M\}$
- $z_m \sim \text{Mult}(1, \theta)$

For each word $n \in \{1, \ldots, N_m\}$
- $x_{mn} \sim \text{Mult}(1, \phi_{z_m})$

• Example corpus

<table>
<thead>
<tr>
<th></th>
<th>the</th>
<th>he</th>
<th>is</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$x_{11}$</td>
<td>$x_{12}$</td>
<td>$x_{13}$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>the</th>
<th>and</th>
<th>the</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$x_{21}$</td>
<td>$x_{22}$</td>
<td>$x_{23}$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>she</th>
<th>she</th>
<th>is</th>
<th>is</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$x_{31}$</td>
<td>$x_{32}$</td>
<td>$x_{33}$</td>
<td>$x_{34}$</td>
</tr>
</tbody>
</table>
Bayesian Inference for Naïve Bayes

Whiteboard:

– Naïve Bayes is not Bayesian
– What if we observed both words and topics?
– Dirichlet-Multinomial in the fully observed setting is just Naïve Bayes
– Three ways of estimating parameters:
  1. MLE for Naïve Bayes
  2. MAP estimation for Naïve Bayes
  3. Bayesian parameter estimation for Naïve Bayes
Dirichlet-Multinomial Model

• The Dirichlet is conjugate to the Multinomial

\[
\phi \sim \text{Dir}(\beta) \quad \text{[draw distribution over words]}
\]
For each word \( n \in \{1, \ldots, N\} \)
\[
x_n \sim \text{Mult}(1, \phi) \quad \text{[draw word]}
\]

• The posterior of \( \phi \) is
\[
p(\phi|X) = \frac{p(X|\phi)p(\phi)}{P(X)}
\]

• Define the count vector \( n \) such that \( n_t \) denotes the number of times word \( t \) appeared

• Then the posterior is also a Dirichlet distribution:
\[
p(\phi|X) \sim \text{Dir}(\beta + n)
\]
LATENT DIRICHLET ALLOCATION (LDA)
Mixture vs. Admixture (LDA)

Diagrams from Wallach, JHU 2011, slides
Latent Dirichlet Allocation

• Generative Process

“admixture”

• Example corpus
Latent Dirichlet Allocation

• Generative Process

For each topic $k \in \{1, \ldots, K\}$:
\[
\phi_k \sim \text{Dir}(\beta) \quad \text{[draw distribution over words]}
\]

For each document $m \in \{1, \ldots, M\}$
\[
\theta_m \sim \text{Dir}(\alpha) \quad \text{[draw distribution over topics]}
\]

For each word $n \in \{1, \ldots, N_m\}$
\[
\tilde{z}_{mn} \sim \text{Mult}(1, \theta_m) \quad \text{[draw topic assignment]}
\]
\[
x_{mn} \sim \phi_{z_{mi}} \quad \text{[draw word]}
\]

• Example corpus

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>he</td>
<td>is</td>
<td>the</td>
<td>and</td>
<td>the</td>
<td>she</td>
</tr>
<tr>
<td>$x_{11}$</td>
<td>$x_{12}$</td>
<td>$x_{13}$</td>
<td>$x_{21}$</td>
<td>$x_{22}$</td>
<td>$x_{23}$</td>
<td>$x_{31}$</td>
</tr>
</tbody>
</table>

Document 1  Document 2  Document 3
LDA for Topic Modeling

- The **generative story** begins with only a **Dirichlet prior** over the topics.
- Each **topic** is defined as a **Multinomial distribution** over the vocabulary, parameterized by $\phi_k$

(Blei, Ng, & Jordan, 2003)
LDA for Topic Modeling

- The *generative story* begins with only a **Dirichlet prior** over the topics.
- Each **topic** is defined as a **Multinomial distribution** over the vocabulary, parameterized by $\phi_k$

(Blei, Ng, & Jordan, 2003)
LDA for Topic Modeling

- A topic is visualized as its high probability words.

(Blei, Ng, & Jordan, 2003)

Dirichlet(\(\beta\))

\(\phi_1\) team, season, hockey, player, penguins, ice, canadiens, puck, montreal, stanley, cup

\(\phi_2\) puck, montreal, stanley, cup

\(\phi_3\) puck, montreal, stanley, cup

\(\phi_4\) puck, montreal, stanley, cup

\(\phi_5\)

\(\phi_6\)
LDA for Topic Modeling

A topic is visualized as its high probability words.

A pedagogical label is used to identify the topic.

(Blei, Ng, & Jordan, 2003)
LDA for Topic Modeling

- A topic is visualized as its high probability words.
- A pedagogical label is used to identify the topic.

(Blei, Ng, & Jordan, 2003)
LDA for Topic Modeling

(blei, ng, & jordan, 2003)

\[ \theta_1 = \text{Dirichlet}(\alpha) \]

\[ \phi_1 \{\text{Canadian gov.}\} \]
\[ \phi_2 \{\text{government}\} \]
\[ \phi_3 \{\text{hockey}\} \]
\[ \phi_4 \{\text{U.S. gov.}\} \]
\[ \phi_5 \{\text{baseball}\} \]
\[ \phi_6 \{\text{Japan}\} \]
The 54/40' boundary dispute is still unresolved, and Canadian and US
The 54/40' boundary dispute is still unresolved, and Canadian and US
The 54/40' boundary dispute is still unresolved, and Canadian and US Coast Guard
The 54/40' boundary dispute is still unresolved, and Canadian and US Coast Guard vessels regularly if infrequently detain each other's fish boats in the disputed waters off Dixon...
The 54/40' boundary dispute is still unresolved, and Canadian and US Coast Guard vessels regularly if infrequently detain each other's fish boats in the disputed waters off Dixon...

In the year before Lemieux came, Pittsburgh finished with 38 points. Following his arrival, the Pens finished...

The Orioles' pitching staff again is having a fine exhibition season. Four shutouts, low team ERA, (Well, I haven't gotten any baseball...
The 54/40' boundary dispute is still unresolved, and Canadian and US Coast Guard vessels regularly if infrequently detain each other's fish boats in the disputed waters off Dixon... In the year before Lemieux came, Pittsburgh finished with 38 points. Following his arrival, the Pens finished... The Orioles' pitching staff again is having a fine exhibition season. Four shutouts, low team ERA, (Well, I haven't gotten any baseball...
The 54/40' boundary dispute is still unresolved, and Canadian and US Coast Guard vessels regularly if infrequently detain each other's fish boats in the disputed waters off Dixon...

In the year before Lemieux came, Pittsburgh finished with 38 points. Following his arrival, the Pens finished...

The Orioles' pitching staff again is having a fine exhibition season. Four shutouts, low team ERA, (Well, I haven't gotten any baseball...
The 54/40' boundary dispute is still unresolved, and Canadian and US Coast Guard vessels regularly if infrequently detain each other's fish boats in the disputed waters off Dixon...

In the year before Lemieux came, Pittsburgh finished with 38 points. Following his arrival, the Pens finished...

The Orioles' itching staff again is having a fine exhibition season. Four shutouts, low team ERA, (Well, I haven't gotten any baseball...
Plate Diagrams

Whiteboard:

– Example #1: Plate diagram for Dirichlet-Multinomial model
– Example #2: Plate diagram for LDA
Latent Dirichlet Allocation

- Plate Diagram
Latent Dirichlet Allocation

- Plate Diagram
Latent Dirichlet Allocation

Questions:

• Is this a believable story for the generation of a corpus of documents?

• Why might it work well anyway?
Latent Dirichlet Allocation

Why does LDA “work”? 

• LDA trades off two goals. 
  ① For each document, allocate its words to as few topics as possible. 
  ② For each topic, assign high probability to as few terms as possible. 

• These goals are at odds. 
  • Putting a document in a single topic makes #2 hard: 
    All of its words must have probability under that topic. 
  • Putting very few words in each topic makes #1 hard: 
    To cover a document’s words, it must assign many topics to it. 

• Trading off these goals finds groups of tightly co-occurring words.

Slide from David Blei, MLSS 2012
Latent Dirichlet Allocation

How does this relate to my other favorite model for capturing low-dimensional representations of a corpus?

• Builds on latent semantic analysis (Deerwester et al., 1990; Hofmann, 1999)
• It is a mixed-membership model (Erosheva, 2004).
• It relates to PCA and matrix factorization (Jakulin and Buntine, 2002)
• Was independently invented for genetics (Pritchard et al., 2000)

Slide from David Blei, MLSS 2012
Outline

• Applications of Topic Modeling
• Latent Dirichlet Allocation (LDA)
  1. Beta-Bernoulli
  2. Dirichlet-Multinomial
  3. Dirichlet-Multinomial Mixture Model
  4. LDA
• Bayesian Inference for Parameter Estimation
  – Exact inference
  – EM
  – Monte Carlo EM
  – Gibbs sampler
  – Collapsed Gibbs sampler
• Extensions of LDA
  – Correlated topic models
  – Dynamic topic models
  – Polylingual topic models
  – Supervised LDA
BAYESIAN INFERENCE FOR PARAMETER ESTIMATION
LDA Inference

- Fully Observed MLE

Learning like this would be easy, but in practice we do not observe the topic assignments $z_{mn}$. 

\[ 	heta_m \] 

Document-specific topic distribution

\[ z_{mn} \] 
Topic assignment

\[ x_{mn} \] 
Observed word

\[ N_m \] 
Optimized

\[ M \] 

\[ \phi_k \] 

\[ K \] 

For each document

For each topic

For each word

\[ \text{Product of Experts} \]

\[ \text{LDA Inference} \]

\[ \text{Generative process} \] 

\[ \text{IBP} \]

\[ \text{PoE} \] 

\[ \text{Shared Components Topic Models} \]
LDA Inference

• Full Observed MAP Estimation

Learning like this would be easy, but in practice we do not observe the topic assignments $z_{mn}$
Unsupervised Learning

Three learning paradigms:

1. Maximum likelihood estimation (MLE)
   \[ \arg \max_{\theta} p(X|\theta) \]

2. Maximum a posteriori (MAP) estimation
   \[ \arg \max_{\theta} p(\theta|X) \propto p(X|\theta)p(\theta) \]

3. Bayesian approach
   Estimate the posterior:
   \[ p(\theta|X) = \ldots \]
LDA Inference

- Standard EM (MLE)
LDA Inference

- Standard EM (MAP Estimation)
LDA Inference

- Monte Carlo EM (MAP Estimation)
LDA Inference

• Bayesian Approach
Bayesian Inference

**Whiteboard:**

- Posteriors over parameters
- Bayesian inference for parameter estimation
LDA Inference

- Bayesian Approach

Diagram:
- \( \alpha \) (Dirichlet parameter)
- \( \theta_m \) (Document-specific topic distribution)
- \( z_{mn} \) (Topic assignment)
- \( x_{mn} \) (Observed word)
- \( \phi_k \) (Topic distribution)
- \( \beta \) (Dirichlet parameter)

Intractable

- Document-specific topic distribution
- Observed word

Exact Inference?
Exact Inference in LDA

- Exactly computing the posterior is intractable in LDA
  - Junction tree algorithm: exact inference in general graphical models
    1. “moralization” converts directed to undirected
    2. “triangulation” breaks 4-cycles by adding edges
    3. Cliques arranged into a junction tree
  - Time complexity is exponential in size of cliques
  - LDA cliques will be large (at least $O(\# \text{ topics})$), so complexity is $O(2^{\# \text{ topics}})$

- Exact MAP inference in LDA is NP-hard for a large number of topics (Sontag & Roy, 2011)
LDA Inference

- Explicit Gibbs Sampler
LDA Inference

- Collapsed Gibbs Sampler