Latent Dirichlet Allocation

Spring 2022

Instructor: Shandian Zhe
zhe@cs.utah.edu
School of Computing
Outline

• Latent Dirichlet Allocation
• Variational inference
Outline

• Latent Dirichlet Allocation
• Variational inference
Latent Dirichlet Allocation

• A classical text mining model that extract topics from the text corpus

• Broadly used in all kinds of text mining and related tasks: information retrieval, text classification, advertisement keywords extraction, sentimental analysis, ....

• [https://medium.com/@fatmafatma/industrial-applications-of-topic-model-100e48a15ce4](https://medium.com/@fatmafatma/industrial-applications-of-topic-model-100e48a15ce4)

• A very good example to study Bayesian learning
Latent Dirichlet Allocation

<table>
<thead>
<tr>
<th>“Arts”</th>
<th>“Budgets”</th>
<th>“Children”</th>
<th>“Education”</th>
</tr>
</thead>
<tbody>
<tr>
<td>NEW</td>
<td>MILLION</td>
<td>CHILDREN</td>
<td>SCHOOL</td>
</tr>
<tr>
<td>FILM</td>
<td>TAX</td>
<td>WOMEN</td>
<td>STUDENTS</td>
</tr>
<tr>
<td>SHOW</td>
<td>PROGRAM</td>
<td>PEOPLE</td>
<td>SCHOOLS</td>
</tr>
<tr>
<td>MUSIC</td>
<td>BUDGET</td>
<td>CHILD</td>
<td>EDUCATION</td>
</tr>
<tr>
<td>MOVIE</td>
<td>BILLION</td>
<td>YEARS</td>
<td>TEACHERS</td>
</tr>
<tr>
<td>PLAY</td>
<td>FEDERAL</td>
<td>FAMILIES</td>
<td>HIGH</td>
</tr>
<tr>
<td>MUSICAL</td>
<td>YEAR</td>
<td>WORK</td>
<td>PUBLIC</td>
</tr>
<tr>
<td>BEST</td>
<td>SPENDING</td>
<td>PARENTS</td>
<td>TEACHER</td>
</tr>
<tr>
<td>ACTOR</td>
<td>NEW</td>
<td>SAYS</td>
<td>BENNETT</td>
</tr>
<tr>
<td>FIRST</td>
<td>STATE</td>
<td>FAMILY</td>
<td>MANIGAT</td>
</tr>
<tr>
<td>YORK</td>
<td>PLAN</td>
<td>WELFARE</td>
<td>NAMPHY</td>
</tr>
<tr>
<td>OPERA</td>
<td>MONEY</td>
<td>MEN</td>
<td>STATE</td>
</tr>
<tr>
<td>THEATER</td>
<td>PROGRAMS</td>
<td>PERCENT</td>
<td>PRESIDENT</td>
</tr>
<tr>
<td>ACTRESS</td>
<td>GOVERNMENT</td>
<td>CARE</td>
<td>ELEMENTARY</td>
</tr>
<tr>
<td>LOVE</td>
<td>CONGRESS</td>
<td>LIFE</td>
<td>HAITI</td>
</tr>
</tbody>
</table>

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. “Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services,” Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center’s share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.
Latent Dirichlet Allocation

• The original paper


LDA: sampling procedure

- Given $K$ topics
- First sample $K$ topics (word distributions)
- For each document in the corpus
  - Sample topic mixture distribution
  - For each word in the document
    - Sample the topic index, according to which to sample the word
LDA: graphical model

Suppose we have $V$ words, $M$ documents, document $n$ has $n_T$ words

- $\eta$ (K by 1 vector)
- $\alpha$ (K by 1 vector)
- $\theta_n = [\theta_{n1}, \ldots, \theta_{nK}]^T$ (K by 1 one-hot vector)
- $\beta_k = [\beta_{k1}, \ldots, \beta_{KV}]^T$ (Multinomial parameters)
- $\theta_n$ (Topic mixture)
- $\theta_n$ (Topic index)
- $z_{nj} = [z_{nj1}, \ldots, z_{njk}]^T$ (Only one element is 1)
- $n_T$ (M by 1)
- $w_{nj} = [w_{nj1}, \ldots, w_{njV}]^T$ (V by 1 one-hot vector)
- $w_{nj}$ (Only one element is 1)
- $\eta$ (K by 1 vector)
- $\alpha$ (K by 1 vector)
- $\theta_n = [\theta_{n1}, \ldots, \theta_{nK}]^T$ (K by 1 one-hot vector)
- $\beta_k = [\beta_{k1}, \ldots, \beta_{KV}]^T$ (Multinomial parameters)
- $\theta_n$ (Topic mixture)
- $\theta_n$ (Topic index)
- $z_{nj} = [z_{nj1}, \ldots, z_{njk}]^T$ (Only one element is 1)
- $n_T$ (M by 1)
- $w_{nj} = [w_{nj1}, \ldots, w_{njV}]^T$ (V by 1 one-hot vector)
- $w_{nj}$ (Only one element is 1)
LDA: sampling procedure

- For each topic
  \[ p(\beta_k | \eta) = \text{Dir}(\beta_k | \eta) = \frac{\Gamma(V \eta)}{\prod_{v=1}^{V} \Gamma(\eta_v)} \prod_{v=1}^{V} \beta_{kv}^{\eta_v - 1} \]

- For each document
  \[ p(\theta_n | \alpha) = \text{Dir}(\theta_n | \alpha) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_{nk}^{\alpha_k - 1} \]

- For each word
  \[ p(z_{nj} | \theta_n) = \text{Mul}(z_{nj} | \theta_n) = \prod_{k=1}^{K} \theta_{nk}^{z_{njk}} \]
  \[ p(w_{nj} | z_{njk} = 1, \beta) = \text{Mul}(w_{nj} | \beta_k) = \prod_{v=1}^{V} \beta_{kv}^{w_{njv}} \]
  \[ p(w_{nj} | z_{nj}, \beta) = \prod_{k=1}^{K} \prod_{v=1}^{V} \beta_{kv}^{z_{njk} w_{njv}} \]
LDA: joint probability

\[ p(\beta, \theta, Z, W | \eta, \alpha) \]

\[
= \prod_{k=1}^{K} \text{Dir}(\beta_k | \eta) \prod_{n=1}^{M} \text{Dir}(\theta_n | \alpha) \left\{ \prod_{j=1}^{n_T} \text{Mul}(z_{nj} | \theta_n) \left[ \prod_{k=1}^{K} \prod_{v=1}^{V} \beta_{kv}^{\tilde{z}_{nj} w_{njv}} \right] \right\} 
\]
Outline

• Latent Dirichlet Allocation
• Variational inference
LDA: inference

How can we compute the posterior of

\[ \beta = \{\beta_1, \ldots, \beta_K\} \quad \text{Topic words: critical for numerous tasks} \]

\[ \theta = \{\theta_1, \ldots, \theta_M\} \quad \text{Topic mixture: low-rank representation of docs} \]

True posterior is intractable to compute
LDA: variational inference

- We use variational EM algorithm (empirical Bayes)

  - E step: mean-field update

    \[ q(\beta, \theta, Z) = \prod_{k=1}^{K} q(\beta_k) \prod_{n=1}^{M} \left[ q(\theta_n) \prod_{j=1}^{n_T} q(z_{nj}) \right] \]

  - M step

    Maximize variational lower bound of the model evidence w.r.t \( \alpha, \eta \)
LDA: variational E update

\[ p(\beta, \theta, Z, W | \eta, \alpha) \]
\[ = \prod_{k=1}^{K} \text{Dir}(\beta_k | \eta) \prod_{n=1}^{M} \text{Dir}(\theta_n | \alpha) \left\{ \prod_{j=1}^{n_T} \text{Mul}(z_{nj} | \theta_n) \left[ \prod_{k=1}^{K} \prod_{v=1}^{V} \beta_{kv}^{z_{njk} w_{njv}} \right] \right\} \]

Update each \( q(z_{nj}) \)

\[ q(z_{nj}) \propto \exp \left( \mathbb{E}_q \log \left[ \text{Mul}(z_{nj} | \theta_n) \prod_{k=1}^{K} \prod_{v=1}^{V} \beta_{kv}^{z_{njk} w_{njv}} \right] \right) \]
\[ \exp \left( \sum_{k=1}^{K} z_{njk} \left[ \mathbb{E}_q[\log \theta_{nk}] + \sum_{v=1}^{V} w_{njv} \mathbb{E}_q[\log \beta_{kv}] \right] \right) \]

\[ q(z_{nj}) = \text{Mul}(z_{nj} | \phi_{nj}) \]

\[ \phi_{njk} \propto \exp \left( \mathbb{E}_q[\log \theta_{nk}] + \sum_{v=1}^{V} w_{njv} \mathbb{E}_q[\log \beta_{kv}] \right) \]
LDA: variational E update

\[ p(\beta, \theta, Z, W | \eta, \alpha) \]
\[ = \prod_{k=1}^{K} \text{Dir}(\beta_k | \eta) \prod_{n=1}^{M} \text{Dir}(\theta_n | \alpha) \left\{ \prod_{j=1}^{n_T} \text{Mul}(z_{nj} | \theta_n) \left[ \prod_{k=1}^{K} \prod_{v=1}^{V} \beta_{kv}^{z_{nk} w_{njv}} \right] \right\} \]

Update each \( q(\theta_n) \)

\[ q(\theta_n) \propto \exp \left( \mathbb{E}_q \left[ \log[\text{Dir}(\theta_n | \alpha)] + \sum_{j=1}^{n_T} \log[\text{Mul}(z_{nj} | \theta_n)] \right] \right) \]
\[ = \sum_{k=1}^{K} (\alpha_k + \sum_{j=1}^{n_T} \mathbb{E}_q[z_{nkj}] - 1) \log \theta_{nk} \]

\[ q(\theta_n) = \text{Dir}(\theta_n | \gamma_n) \]

\[ \gamma_{nk} = \alpha_k + \sum_{j=1}^{n_T} \mathbb{E}_q[z_{nkj}] \]
LDA: variational E update

\[ p(\beta, \theta, Z, W | \eta, \alpha) \]

\[ = \prod_{k=1}^{K} \text{Dir}(\beta_k | \eta) \prod_{n=1}^{M} \text{Dir}(\theta_n | \alpha) \left\{ \prod_{j=1}^{n_T} \text{Mul}(z_{nj} | \theta_n) \right\} \]

\[ \left\{ \prod_{k=1}^{K} \prod_{v=1}^{V} \beta_{kv}^{z_{njk}w_{njv}} \right\} \]

Update each \( q(\beta_k) \)

\[ q(\beta_k) \propto \exp \left( \mathbb{E}_q \left[ \log[\text{Dir}(\beta_k | \eta)] + \sum_{n=1}^{M} \sum_{j=1}^{n_T} \sum_{v=1}^{V} z_{njk}w_{njv} \beta_{kv} \right] \right) \]

\[ \sum_{v=1}^{V} \log[\beta_{kv}](\eta + \sum_{n=1}^{M} \sum_{j=1}^{n_T} \mathbb{E}_q[z_{njk}]w_{njv} - 1) \]

\[ q(\beta_k) = \text{Dir}(\beta_k | \psi_k) \]

\[ \psi_{kv} = \eta + \sum_{n=1}^{M} \sum_{j=1}^{n_T} \mathbb{E}_q[z_{njk}]w_{njv} \]
LDA: variational E update

• How to compute the required moments in the update?

\[ q(z_{nj}) = \text{Mul}(z_{nj} | \phi_{nj}) \]

\[ \mathbb{E}_q[z_{njk}] \]

\[ q(\theta_n) = \text{Dir}(\theta_n | \gamma_n) \]

\[ \mathbb{E}_q \log[\theta_{nk}] \]

\[ q(\beta_k) = \text{Dir}(\beta_k | \psi_k) \]

\[ \mathbb{E}_q \log \beta_{kv} \]

Leave it as your review!
LDA: variational M step

\[ p(\beta, \theta, Z, W | \eta, \alpha) \]

\[ = \prod_{k=1}^{K} \text{Dir}(\beta_k | \eta) \prod_{n=1}^{M} \text{Dir}(\theta_n | \alpha) \left\{ \prod_{j=1}^{n_T} \text{Mul}(z_{nj} | \theta_n) \left[ \prod_{k=1}^{K} \prod_{v=1}^{V} \beta_{kv}^{z_{nj}w_{nv}} \right] \right\} \]

\[ \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k} \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_{nk}^{\alpha_k - 1} \]

Variational lower bound

\[ \mathcal{L}(\alpha) = \sum_{n=1}^{M} \log \Gamma(\sum_{k=1}^{K} \alpha_k) - \sum_{n=1}^{M} \sum_{k=1}^{K} \log \Gamma(\alpha_k) \]

\[ + \sum_{n=1}^{M} \sum_{k=1}^{K} (\alpha_k - 1) \mathbb{E}_q[\log \theta_{nk}] + \text{const} \]
LDA: variational M step

Variational lower bound

\[ \mathcal{L}(\alpha) = \sum_{n=1}^{M} \log \Gamma\left(\sum_{k=1}^{K} \alpha_k\right) - \sum_{n=1}^{M} \sum_{k=1}^{K} \log \Gamma(\alpha_k) \]

\[ + \sum_{n=1}^{M} \sum_{k=1}^{K} (\alpha_k - 1) \mathbb{E}_q[\log \theta_{nk}] + \text{const} \]

\[ \frac{\partial \mathcal{L}}{\partial \alpha_k} = M \Psi\left(\sum_{k=1}^{K} \alpha_k\right) - M \Psi(\alpha_k) + \sum_{n=1}^{M} \mathbb{E}_q[\log \theta_{nk}] \]

Digamma function
LDA: variational M step

- Derive $\frac{\partial \mathcal{L}}{\partial \eta}$

Note that $\eta$ is a scalar

Leave it as your exercise

Use any gradient based algorithm with constraints

$\alpha > 0, \eta > 0$ e.g., LBFGS-B
What you need to do

• Write down LDA sampling procedure and joint probability
• Derive the variational E updates and gradients for M step
• Implement an algorithm for LDA inference and test it on real-world data (see homework assignments).