LDA Model and Sampling

Dirichlet Distribution

\[ \mu = (\mu_1, \ldots, \mu_k)^T \text{ Constraint } \sum_{i=1}^{k} \mu_i = 1 \]

\[ \alpha = (\alpha_1, \ldots, \alpha_k)^T \quad \alpha_0 = \sum_{i=1}^{k} \alpha_i \]

\[ Pr(\mu|\alpha) = \text{Dir}(\mu|\alpha) = \frac{\Gamma(\alpha_0)}{\prod \Gamma(\alpha_i)} \prod \mu_i^{\alpha_i-1} \]

Discrete Distribution

Discrete distribution over \( x \in \{1, \ldots, k\} \)
where value \( i \) is denoted in binary notation as \( i \)'th unit vector

\[ Pr(x|\mu) = \prod_{i=1}^{k} \mu_i^x_i \]

We observe \( N \) values \( D = x_1, \ldots, x_N \) and \( x_n \in \{1, \ldots, k\} \) is denoted in binary notation

Likelihood

\[ \mathcal{L} = Pr(D|\mu) = \prod_{n=1}^{N} \prod_{i=1}^{k} \mu_i^{x_{n,i}} = \prod_{i=1}^{k} \mu_i^{m_i} \]

Posterior

\[ Pr(\mu|D) \propto \prod_{i=1}^{k} \mu_i^{m_i+\alpha_i-1} \]

\[ Pr(\mu|D) = \text{Dir}(\mu|\alpha + m) \]

\[ m = (m_1, \ldots, m_k)^T \]

\[ \sum \alpha_i + m_i = \alpha_0 + N \]

Predictive distribution

\[ Pr(x_\ell = 1|D) = \int \! Pr(\mu|D)Pr(x_\ell = 1|\mu)d\mu \]

\[ = \ldots \]

\[ = \frac{\alpha_\ell + m_\ell}{\alpha_0 + N} \]

The evidence function

\[ Pr(D|\alpha) = \int \! Pr(\mu|\alpha)Pr(D|\mu)d\mu \]

\[ = \ldots \]

\[ = \frac{\Gamma(\alpha_0)}{\prod \Gamma(\alpha_i)} \prod \frac{\Gamma(\alpha_i + m_i)}{\Gamma(\alpha_0 + N)} \]
LDA Model

\[ \theta_k \sim \prod_d \text{Dir}(\theta_d | \alpha) \]

\[ \phi_{i|k} \sim \prod_k \text{Dir}(\phi_k | \beta) \]

For each \( d \), draw \( z_{i|d} \sim \theta_k | d \)

For each \( k \), sample \( \phi_{i|k} \) from \( \text{Dir}(\phi_k | \beta) \)

For each \( d \), for each location \( j \) in \( d \):

draw \( w_{j|d} \sim \phi_{i|k} \)

\[ \text{Topic 1} \]

\[ \text{Topic 2} \]

\[ \text{Topic 3} \]

LDA Example

- Vocabulary of 5 words: A,B,C,D,E
- 3 topics: t1, t2, t3
- 4 documents of varying lengths

LDA Example: V=5, T=3, D=4

\[ \phi_k \]

\[ \theta_d \]

LDA Generative Model

Example Topics – Educational Text

Figure 1. An illustration of four (out of 100) topics extracted from the TASA corpus.
Latent Semantics – Dim Reductions

Inference: Basic Ingredients

Prior: \( \prod_d \text{Dir}(\theta_d|\alpha) \prod_k \text{Dir}(\phi_k|\beta) \)

Complete data:
all topic assignments \( Z \) and all words \( D \)

- \( N \) words total: \( N = \sum N_d \)
- \( N_d \) number of words in \( d \)
- \( N_k \) number of times topic \( k \) is used
- \( N_{k|d} \) number of times topic \( k \) is used in \( d \)
- \( N_{i|k} \) number of times word \( i \) is used in \( k \)

Complete data Likelihood
\( \mathcal{L} = P_r(D, Z|\phi, \theta) = \prod_d \prod_k \phi_{k|d}^{N_{k|d}} \prod_k \prod_i \phi_{i|k}^{N_{i|k}} \)

Complete data Posterior
\( \prod_d \text{Dir}(\theta_d|\alpha + N_k) \prod_k \text{Dir}(\phi_k|\beta + N_{i|k}) \)

Inference

Prior: \( \prod_d \text{Dir}(\theta_d|\alpha) \prod_k \text{Dir}(\phi_k|\beta) \)

Complete data Likelihood
\( \mathcal{L} = P_r(D, Z|\phi, \theta) = \prod_d \prod_k \phi_{k|d}^{N_{k|d}} \prod_k \prod_i \phi_{i|k}^{N_{i|k}} \)

Complete data Posterior
\( \prod_d \text{Dir}(\theta_d|\alpha + N_k) \prod_k \text{Dir}(\phi_k|\beta + N_{i|k}) \)
### Inference

Complete data Evidence function

\[ Pr(D, Z|\alpha, \beta) = \prod_d \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_z)} \prod_i \frac{\Gamma(\alpha_i + N_{k|d})}{\Gamma(\alpha_i + N_{i|k})} \prod_k \frac{\Gamma(\beta_0)}{\Gamma(\beta_z)} \prod_i \frac{\Gamma(\beta_i + N_{i|k})}{\Gamma(\beta_i + N_{k|d})} \]

### We have closed form expressions for everything!

- Are we done?

No: we do not observe \( Z \) and do not have counts for \( N_{k|d}, N_{i|k} \)

- What can we do?
  - ... EM ... or ... Sampling

### Inference: Gibbs Sampling

We marginalize \( \phi, \theta \) and sample \( Z \) directly.

For Gibbs sampling we want to resample each \( z_{j,d} \) conditioned on the rest of the complete data.

Gibbs sampling from distribution over \( V_1, \ldots, V_n \):

- Repeat
  - Pick \( i \in \{1, \ldots, N\} \) uniformly
  - Draw new value for \( V_i \) from distribution

\[ Pr(V_i|V_1, \ldots, V_{i-1}, V_{i+1}, \ldots, V_N) \]

### From \( Z \)'s to estimates:

- After long walk in Markov chain we have a random sample for \( Z \) from the posterior
- Get multiple samples using independent runs or skip-X-steps in same chain
- From \( Z \)'s to estimates:

\[ Pr(\theta_d|D, Z) = \text{Dir}(\theta_d|\alpha + N_{k|d}) \]
\[ Pr(\phi_k|D, Z) = \text{Dir}(\phi_k|\beta + N_{i|k}) \]

### From \( Z \)'s to estimates:

Complete data Posterior

\[ Pr(\theta_d|D, Z) = \text{Dir}(\theta_d|\alpha + N_{k|d}) \]
\[ Pr(\phi_k|D, Z) = \text{Dir}(\phi_k|\beta + N_{i|k}) \]

- Because of exchangeability cannot use multiple samples for these estimates.
- Instead each \( Z \) can make its own model or estimate
- Use for prediction & err estimates
Stability of Topics

![Stability of Topics](image)

LDA Model

![LDA Model](image)

Example Topics in Text

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 official said. Lincoln Center’s plans will be $20 million 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, whose music and performing arts are taught, will get $290,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its total award $300,000 donation, too.

Example Topics in Text

![Example Topics in Text](image)

Inference: Evidence Maximization

Complete data Evidence function

$$Pr(D, Z|\alpha, \beta) = \prod_{D} \frac{\Gamma(\alpha_{0})}{\Gamma(\alpha_{0} + N_{D})} \prod_{i} \frac{\Gamma(\alpha_{i} + N_{i}(d))}{\Gamma(\alpha_{i} + N_{i})} \prod_{k} \frac{\Gamma(\beta_{k} + N_{k}(z))}{\Gamma(\beta_{k} + N_{k})}$$

Evidence hard to eval directly

$$Pr(D|\alpha, \beta) = \sum_{Z} Pr(D, Z|\alpha, \beta)$$

Evaluate using samples $Z^{1}, Z^{2}, \ldots$
pick $\alpha, \beta$ to max $\sum_{Z}$, log $Pr(D, Z^{i}|\alpha, \beta)$

Example Topics

<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>MARRIAGE</td>
</tr>
<tr>
<td>YORK</td>
<td>PLAN</td>
<td>WELFARE</td>
<td>NAACP</td>
</tr>
<tr>
<td>OPERA</td>
<td>MONEY</td>
<td>MEN</td>
<td>STATE</td>
</tr>
<tr>
<td>THEATER</td>
<td>PROGRAMS</td>
<td>PENDENT</td>
<td>PRESENT</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>

Multiple Senses

<table>
<thead>
<tr>
<th>Topic 1</th>
<th>Topic 2</th>
<th>Topic 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>MUSIC</td>
<td>LITERATURE</td>
<td>PLAY</td>
</tr>
<tr>
<td>DANCE</td>
<td>POETRY</td>
<td>DANCE</td>
</tr>
<tr>
<td>SONG</td>
<td>STAGE</td>
<td>SONG</td>
</tr>
<tr>
<td>PLAY</td>
<td>PLAY</td>
<td>PLAY</td>
</tr>
<tr>
<td>SMOKING</td>
<td>SMOKING</td>
<td>SMOKING</td>
</tr>
<tr>
<td>BAND</td>
<td>BLUES</td>
<td>BAND</td>
</tr>
<tr>
<td>PLAYING</td>
<td>WRITING</td>
<td>PLAYING</td>
</tr>
<tr>
<td>SONGS</td>
<td>DANCE</td>
<td>DANCE</td>
</tr>
<tr>
<td>DANCING</td>
<td>SONG</td>
<td>DANCING</td>
</tr>
<tr>
<td>CHORE</td>
<td>CHORE</td>
<td>CHORE</td>
</tr>
<tr>
<td>PLAYING</td>
<td>PLAY</td>
<td>PLAYING</td>
</tr>
<tr>
<td>RHYTHM</td>
<td>RHYTHM</td>
<td>RHYTHM</td>
</tr>
<tr>
<td>ALBINO</td>
<td>ALBINO</td>
<td>ALBINO</td>
</tr>
<tr>
<td>MUSICAL</td>
<td>MUSICAL</td>
<td>MUSICAL</td>
</tr>
</tbody>
</table>

![Multiple Senses](image)
Figure 10: Classification results on two binary classification problems from the Reuters-21578 dataset for different proportions of training data. Graph (a) is EARN vs. NOT EARN. Graph (b) is GRAIN vs. NOT GRAIN.

Figure 11: Results for collaborative filtering on the EachMovie data.