LDA Model and Sampling

Roni Khardon
Comp 136
Tufts University
Fall 2015

Dirichlet Distribution

\[ \mu = (\mu_1, \ldots, \mu_k)^T \]

Constraint \( \sum_{i=1}^{k} \mu_i = 1 \)

\[ \alpha = (\alpha_1, \ldots, \alpha_k)^T \]

\( \alpha_0 = \sum_{i=1}^{k} \alpha_i \)

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

Discrete Distribution

Discrete distribution over \( x \in \{1, \ldots, k\} \)

where value \( i \) is denoted in binary notation as \( i \)'th unit vector

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

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

Likelihood

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

Posterior

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

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

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

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

Predictive distribution

\[ P(x_\ell = 1 | D) = \int_{\mu} P(\mu | D) P(x_\ell = 1 | \mu) d\mu \]

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

The evidence function

\[ P(D | \alpha) = \int_{\mu} P(\mu | \alpha) P(D | \mu) d\mu \]

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

For each $d$ draw $\theta_{k|d} \sim \prod_d \text{Dir}(\theta_{d|\alpha})$
For each $k$ draw $\phi_{i|k} \sim \prod_d \text{Dir}(\phi_{k|\beta})$
For each $d$, for each location $j$ in $d$:
draw $z_{j,d} \sim \theta_{k|d}$
draw $w_{j,d} \sim \phi_{i|k=-z_{j,d}}$

Example Topics – Educational Text

Latent Semantics – Dim Reductions

Inference: Basic Ingredients

Prior: $[\prod_d \text{Dir}(\theta_{d|\alpha})][\prod_k \text{Dir}(\phi_{k|\beta})]$
Prior: \( \prod_d \text{Dir}(\theta_d | \alpha) \prod_k \text{Dir}(\phi_k | \beta) \)

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

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

Complete data Evidence function
\( \Pr(D, Z | \alpha, \beta) = \prod_d \frac{\Gamma(\alpha_d)}{\Gamma(\alpha_d + N_d)} \prod_k \frac{\Gamma(\beta + N_{i|k})}{\Gamma(\beta + N_{i|k} + N_{i|k})} \)

We have closed form expressions for everything!
Are we done?

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

What can we do?
Inference

- We have closed form expressions for everything!
- Are we done?
  
  No: we do not observe $Z$
  and do not have counts for $N_k, N_{k|d}, N_{i|k}$

- 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 $z_{i|d}$
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
  
  $Pr(V_i|V_1, \ldots, V_{i-1}, V_{i+1}, \ldots, V_N)$

Inference: Gibbs Sampling

$N^-_d, N^-_{k|d}, N^-_{i|k}$ are counts w/o $z_{i|d}, w_{i|d}$

$Pr(z_{k|d} = 1|D^-, Z^-, \text{word}^{i|d} = i) \propto$

$Pr(z_{k|d} = 1|D^-, Z^-) Pr(\text{word}^{i|d} = i|z_{k|d} = 1) =$

$
\frac{\alpha_k + N^-_{k|d}}{\alpha_t + N_{k|d}} \frac{\beta_{i} + N^-_{i|k}}{\beta_{o} + N_{i|k}}
$

Inference: Gibbs Sampling

- 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:
  
  Complete data Posterior
  
  $Pr(\theta_d|D, Z) = \text{Dir}(\theta_d|\alpha + N_{k|d})$
  
  $Pr(\phi_k|D, Z) = \text{Dir}(\phi_k|\beta + N_{i|k})$

Inference: Gibbs Sampling

- 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

[Graph showing stability of topics with KL distance]
Inference: Evidence Maximization

Complete data Evidence Function

\[ Pr(D, Z|\alpha, \beta) = \prod_d \frac{\Gamma(n_{d0})}{\Gamma(n_{d0} + N_{d1})} \prod_k \frac{\Gamma(n_{k0} + N_{k1})}{\Gamma(n_{k0})} \]

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 \( \sum_Z \log Pr(D, Z^1|\alpha, \beta) \)

---

Example Topics – Educational Text

<table>
<thead>
<tr>
<th>Topic 1</th>
<th>Topic 2</th>
<th>Topic 3</th>
<th>Topic 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARTS</td>
<td>MUSIC</td>
<td>BUDGETS</td>
<td>CHILDREN</td>
</tr>
<tr>
<td>SCHOOL</td>
<td>STUDENTS</td>
<td>PEOPLE</td>
<td>EDUCATION</td>
</tr>
<tr>
<td>BILLS</td>
<td>TIX</td>
<td>BILLS</td>
<td>BILLS</td>
</tr>
<tr>
<td>MUSICAL</td>
<td>BILLS</td>
<td>BILLS</td>
<td>BILLS</td>
</tr>
<tr>
<td>YORK</td>
<td>YORK</td>
<td>YORK</td>
<td>YORK</td>
</tr>
<tr>
<td>THEATRE</td>
<td>THEATRE</td>
<td>THEATRE</td>
<td>THEATRE</td>
</tr>
<tr>
<td>THEATRE</td>
<td>THEATRE</td>
<td>THEATRE</td>
<td>THEATRE</td>
</tr>
</tbody>
</table>

Figure 1. An illustration of four (out of 300) topics extracted from the TASA corpus.

---

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>BOOKS</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>ARTS</td>
<td>EXPEND</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>MARRIAT</td>
</tr>
<tr>
<td>YORK</td>
<td>PLAN</td>
<td>WELFARE</td>
<td>NAPLDRY</td>
</tr>
<tr>
<td>OPERA</td>
<td>MONEY</td>
<td>MIEN</td>
<td>STATE</td>
</tr>
<tr>
<td>THEATRE</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>

---

Example Topics in Text

Latent Semantics – Dim Reductions

\[ LSA \quad \text{documents} = \text{U} \cdot \text{D} \cdot \text{V}^T \]

\[ \text{TOPIC MODEL} \quad \text{documents} = \Phi \cdot \text{components} \cdot \text{weights} \]

Figure 6. The matrix factorization of the LSA model compared to the matrix factorization of the topic model.
Although these results need further substantiation, they suggest that the topic-based
in using the LDA-based features; indeed, in almost all cases the performance is improved with the

Figure 10: Classification results on two binary classification problems from the Reuters-21578
game

There

Document #213

r

owe

wanted

119

new

co

beid

0

It

268

was

hoped

268

for

hav

077

Bix

0.1

f i f t e e n

dataset for different proportions of training data. Graph (a) is

Proportion of data used for training

Figure 10 shows our results. We see that there is little reduction in classification performance

Figure 9: Three topics related to the word PLAY. [SG2007]

COMP 136

Roni Khardon, Tufts University

Multiple Senses

<table>
<thead>
<tr>
<th>Topic 77</th>
<th>Topic 82</th>
<th>Topic 106</th>
</tr>
</thead>
<tbody>
<tr>
<td>word</td>
<td>word</td>
<td>word</td>
</tr>
<tr>
<td>MUSI</td>
<td>030</td>
<td>DANCE</td>
</tr>
<tr>
<td>LITERA</td>
<td>013</td>
<td>POEM</td>
</tr>
<tr>
<td>DANCE</td>
<td>014</td>
<td>POET</td>
</tr>
<tr>
<td>SONG</td>
<td>033</td>
<td>POET</td>
</tr>
<tr>
<td>PLAY</td>
<td>039</td>
<td>PLAY</td>
</tr>
<tr>
<td>SING</td>
<td>026</td>
<td>POEMS</td>
</tr>
<tr>
<td>SING</td>
<td>026</td>
<td>POEMS</td>
</tr>
<tr>
<td>SING</td>
<td>026</td>
<td>POEMS</td>
</tr>
<tr>
<td>BAND</td>
<td>026</td>
<td>PLAY</td>
</tr>
<tr>
<td>PLAYS</td>
<td>023</td>
<td>LITERARY</td>
</tr>
<tr>
<td>SONG</td>
<td>022</td>
<td>WRITERS</td>
</tr>
<tr>
<td>SONG</td>
<td>022</td>
<td>WRITERS</td>
</tr>
<tr>
<td>SONG</td>
<td>022</td>
<td>WRITERS</td>
</tr>
<tr>
<td>PIANO</td>
<td>017</td>
<td>WRITERS</td>
</tr>
<tr>
<td>PLAYING</td>
<td>016</td>
<td>WRITERS</td>
</tr>
<tr>
<td>PLAY</td>
<td>016</td>
<td>WRITERS</td>
</tr>
<tr>
<td>RHYTHM</td>
<td>013</td>
<td>SHAKESPEARE</td>
</tr>
<tr>
<td>ALBERT</td>
<td>013</td>
<td>SCRIP</td>
</tr>
<tr>
<td>MUSI</td>
<td>013</td>
<td>SCRIP</td>
</tr>
</tbody>
</table>

Figure 9. Three topics related to the word PLAY. [SG2007]

COMP 136

Roni Khardon, Tufts University

Collaborative Filtering

User-> document    Movie-> word

Figure 11: Results for collaborative filtering on the EachMovie data. [BNJ2003]

COMP 136

Roni Khardon, Tufts University

Dim Reduction: Classification

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.

COMP 136

Roni Khardon, Tufts University