Latent Dirichlet Allocation

David Blei, Andrew Ng, Michael Jordan

27 April, 2010

presented by Zhaoyin Jia, Ainur Yessenalina

Intuition behind LDA

(from David Blei)

Simple intuition: Documents exhibit multiple topics.

Problabilistic model

(from David Blei)

- Each document is a random mixture of corpus-wide topics
- Each word is drawn from one of those topics

Problabilistic model (2)

(from David Blei)

- We only observe the documents
- Our goal is to infer the underlying topic structure

Problabilistic model (2)

- The observations are generated from a generative probabilistic process that includes hidden variables
- Infer the hidden structure using posterior inference. What are the topics that describe this collection?
- Situate new data into the estimated model.
  - How does this query or new document fit into the estimated topic structure?

Notation

1. word: \(1..V\)
2. document: \(\mathbf{w} = (w_1, w_2, ..., w_N)\) sequence of \(N\) words
3. corpus: \(D = \{\mathbf{w}_1, ..., \mathbf{w}_M\}\) collection of \(M\) documents
Graphical models notation

- Nodes are random variables
- Edges denote possible dependence
- Observed variables are shaded
- Plates denote replicated structure

Other models of the discrete data.

Latent Dirichlet allocation

LDA assumes the following generative process:
1. Choose $N \sim \text{Poisson}(\xi)$
2. Choose $\theta \sim \text{Dir}(\alpha)$
3. For each of $N$ words $w_n$:
   a. Choose topic $z_n \sim \text{Multinomial}(\theta)$
   b. Choose word $w_n \sim P(w_n|z_n, \beta)$

Recap on distributions: Poisson

Recap on distributions: Dirichlet example

$\text{Dir}(\alpha); \alpha = (3, 2, 1)$
Cut strings (each of initial length 1.0) into $K$ pieces with different lengths
(from Wikipedia)
Recap on distributions: Dirichlet example (2)

Dirichlet distribution, $K=3$ for various parameter vectors $\alpha$
Clockwise from top left:
$\alpha = (6, 2, 2), (3, 7, 5), (6, 2, 6), (2, 3, 4)$.
(from Wikipedia)

The Dirichlet distribution

- The Dirichlet distribution is an exponential family distribution over
the simplex, i.e., positive vectors that sum to one
$$p(\theta | \alpha) = \frac{\Gamma \left( \sum_{k} \alpha_k \right)}{\prod \Gamma(\alpha_k)} \prod_k \theta_k^{\alpha_k - 1}.$$  
- The Dirichlet is conjugate to the multinomial. Given a multinomial
observation, the posterior distribution of $\theta$ is a Dirichlet.
- The parameter $\alpha$ controls the mean shape and sparsity of $\theta$.
- The topic proportions are a $K$ dimensional Dirichlet. The
topics are a $V$ dimensional Dirichlet.

Geometric intuition

From a collection of documents, infer
- Per-word topic assignment $z_{d,n}$
- Per-document topic proportions $\theta_d$
- Per-corpus topic distributions $\beta_k$

Inference

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 to 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.
Inference

- Given corpus \((w \text{ is observed})\), parameters \((\alpha, \beta)\), calculate \(p(\theta, z | \alpha, \beta, w)\)
- Intractable
  - Gibbs sampling
  - Variational inference

Variational Inference

- \(\alpha\) controls proportion distribution of topics in one document.
- \(\beta\) is the probability matrix of topics and words

Parameter Estimation

- Try to estimate parameters \((\alpha, \beta)\), given corpus \(\{w\}\).
- EM algorithm:
  - E step: find the optimizing value of \(\gamma, \phi\)
  - M step: maximize log likelihood w.r.t \(\alpha\) and \(\beta\).

Smoothing for unseen words

- For unseen word, MLE of \(\beta\) will assign zero probability during inference.
- Take \(\beta\) as Dirichlet distribution parameterized by \(\eta\).
Parameter Estimation Example

- 16,000 documents of TREC AP corpus
- 100-topic LDA model

<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>MANICAT</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>

Top words of $p(w|z)$

Inference example

<table>
<thead>
<tr>
<th>“Arts”</th>
<th>“Budgets”</th>
<th>“Children”</th>
<th>“Education”</th>
</tr>
</thead>
</table>

$q(z|w)>0.9$

Bag-of-words assumption

Application/Empirical Results

- Mixture of unigrams model:
  - pLSI:
    - Heuristic Inference:
      $$p(W) = \sum_d \prod_{n=1}^N p(w_n|z)p(z|d)p(d).$$
    - Fold-in pLSI: refit $p(z|d)$

Overfitting discussion

- Never seen in art topic, $p(d)$ decreases a lot, Perplexity explodes

Document classification

- Topic distribution

Application in Vision

LDA is modular, general, useful

LDA can be embedded in a more complicated model, embodying further intuition of structure of text

Summary

• Better graphic model
  – Compared to unigram, mixture of unigram, PLSI

• Approximate inference/Parameter estimation

• Applications:
  – generalizing documents/Images
  – Feature reduction
  – Other extensions

Thanks

Questions?