Latent Dirichlet Allocation (LDA)

Mahdi Pakdaman
Intelligent systems Program
University of Pittsburgh

Outline

- Brief Review
- LDA
  - Dirichlet Distribution
  - The Model
  - Theoretical insights
  - Applications
  - Parameter Estimation
- Extensions to LDA
- Summary
How the story began!

- We Model the text corpora to:
  - similarity/relevance judgments, Classification, Summarization, ....
- Represent each document as a vector space
  - A word is an item from a vocabulary indexed by \( \{1, ..., V\} \). We represent words using unit-basis vectors. The \( \nu \)th word is represented by a \( V \)-vector \( w \) such that \( w^\nu = 1 \) and \( w^\nu = 0 \) for \( u \neq \nu \)
  \[ w = (w_1, w_2, \ldots, w_n) \]
  - A document is a sequence of \( N \) words denoted by where \( w_n \) is the \( n \)th word in the sequence.
  - A corpus is a collection of \( M \) documents denoted by
  \[ D = \{w_1, w_2, \ldots, w_M\} \]

The Problem with Vector space representation

- Three problems that arise using the vector space model:
  - The Vectors are very sparse
  - synonymy: many ways to refer to the same object, e.g. car and automobile
    - Will have small cosine but are related
    - leads to poor recall
  - polysemy: most words have more than one distinct meaning, e.g. model, python, chip
    - Will have large cosine but not truly related
    - leads to poor precision
Latent Semantic Space

- LSI maps terms and documents to a “latent semantic space”
- Comparing terms in this space should make synonymous terms look more similar

- SVD (in LSI)

\[
\begin{align*}
A & \xrightarrow{m \times n} U & \xrightarrow{m \times r} VT & \xrightarrow{r \leq \min(m,n)} A' & \xrightarrow{m \times n}
\end{align*}
\]

\[
\min \| A' - A \| \text{ for a given } k
\]

pLSI

- Latent Variable model for general co-occurrence data
  - Associate each observation \((w,d)\) with a class variable \(z \in Z\{z_1, ..., z_K\}\)

- Generative Model for document-term matrix \(D\)
  - Select a doc with probability \(P(d)\)
  - Pick a latent class \(z\) with probability \(P(z|d)\)
  - Generate a word \(w\) with probability \(p(w|z)\)
**pLSI Pitfalls**

- It is a generative model for the train data
- It can be easily overfitted to the train data

---

**LDA**

- Latent variable model
- Dirichlet Prior
- And that’s All

Father of Topic Modeling

So, LDA uses Latent variable model, and Dirichlet Distribution priors and that is All
Dirichlet Distributions

- Dirichlet distribution is the conjugate prior to the multinomial distribution. (This means that if our likelihood is multinomial with a Dirichlet prior, then the posterior is also Dirichlet!)
- The Dirichlet distribution is an exponential family distribution over the simplex, i.e., positive vectors that sum to one

\[
p(\theta | \alpha) = \frac{\Gamma \left( \sum_{i=1}^{k} \alpha_i \right)}{\prod_{i=1}^{k} \Gamma \left( \alpha_i \right)} \prod_{i=1}^{k} \theta_i^{\alpha_i - 1}
\]

\[
\alpha_0 = \sum_{i=1}^{k} \alpha_i
\]

\[
e[\theta_i] = \frac{\alpha_i}{\alpha_0} \quad \text{Var}[\theta_i] = \frac{\alpha_i (\alpha_0 - \alpha_i)}{\alpha_0^2 (\alpha_0 + 1)}
\]

- The Dirichlet parameter \( \alpha_i \) can be thought of as a prior count of the \( i \)th class.
- The parameter \( \alpha \) controls the mean shape and sparsity of \( \theta \).

\( \alpha = 1 \)
$\alpha = 10$

$\alpha = 100$
Each document is a mixture of corpus-wide topics

Each topic is a distribution over words

Each word is drawn from one of the topics
For each document,
- Choose $\theta$-Dirichlet($\alpha$)
- For each of the $N$ words $w_n$:
  - Choose a topic $z_n$ $\sim$ Multinomial($\theta$)
  - Choose a word $w_n$ from $p(w_n|z_n, \beta)$, a multinomial probability conditioned on the topic $z_n$. 

\[
[\beta]_{k \times V} \quad \beta_{ij} = p(w^j = 1|z^i = 1)
\]
LDA Model

- Introduces Dirichlet smoothing on $\beta$ to avoid the zero frequency word problem
- Fully Bayesian approach

Smoothed LDA
The LDA equations

Joint Probability
\[ p(\theta, z, w | \alpha, \beta) = p(\theta | \alpha) \prod_{n=1}^{N} p(z_n | \theta) p(w_n | z_n, \beta) \]

Marginal Distribution of a document
\[ p(w | \alpha, \beta) = \int p(\theta | \alpha) \left( \prod_{n=1}^{N} \sum_{z_n} p(z_n | \theta) p(w_n | z_n, \beta) \right) d^k \theta \]

Probability of a corpus
\[ p(D | \alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d | \alpha) \left( \prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn} | \theta_d) p(w_{dn} | z_{dn}, \beta) \right) d^k \theta_d \]

More insights on LDA: Exchangeability

- A finite set of random variables \( \{x_1, \ldots, x_N\} \) is said to be exchangeable if the joint distribution is invariant to permutation. If \( \pi \) is a permutation of the integers from 1 to \( N \):
  \[ p(x_1, \ldots, x_N) = p(x_{\pi(1)}, \ldots, x_{\pi(N)}) \]

- An infinite sequence of random is infinitely exchangeable if every finite subsequence is exchangeable.
**bag-of-words Assumption**

- Word order is ignored
- "bag-of-words" – exchangeability, not i.i.d
- **Theorem (De Finetti, 1935)** – if \( (x_1, x_2, \ldots, x_N) \) are infinitely exchangeable, then the joint probability \( p(x_1, x_2, \ldots, x_N) \) has a representation as a mixture:

For some random variable \( \theta \)

\[
p(x_1, x_2, \ldots, x_N) = \int d\theta p(\theta) \prod_{i=1}^{N} p(x_i|\theta)
\]

**LDA and exchangeability**

- We assume that words are generated by topics and that those topics are infinitely exchangeable within a document.
- By de Finetti's theorem:

\[
p(w, z) = \int p(\theta) \left( \prod_{n=1}^{N} p(z_n|\theta)p(w_n|z_n) \right) d\theta
\]

- By marginalizing out the mixture component in eq 2, we get we will get the same distribution over observed and latent variables as above.
**Relationship with other latent variable models**

- **Unigram model**
  \[
  p(w) = \prod_{n=1}^{N} p(w_n)
  \]

- **Mixture of unigrams**
  - Each document is generated by first choosing a topic \( z \) and then generating \( N \) words independently from conditional multinomial.
  - \( k-1 \) parameters
  \[
  p(w) = \sum_z p(z) \prod_{n=1}^{N} p(w_n | z)
  \]

**Relationship with other latent variable models (cont.)**

- **Probabilistic latent semantic indexing**
  - Attempt to relax the simplifying assumption made in the mixture of unigrams models.
  - In a sense, it does capture the possibility that a document may contain multiple topics.
  - \( kv+kM \) parameters and linear growth in \( M \).

The \( k+kV \) parameters in a \( k \)-topic LDA model do not grow with the size of the training corpus.
Relationship with other latent variable models (cont.)

- The unigram model finds a single point on the word simplex and posits that all words in the corpus come from the corresponding distribution.
- The mixture of unigram models posits that for each document, one of the k points on the word simplex is chosen randomly and all the words of the document are drawn from the distribution.
- The pLSI model posits that each word of a training document comes from a randomly chosen topic. The topics are themselves drawn from a document-specific distribution over topics.
- LDA posits that each word of both the observed and unseen documents is generated by a randomly chosen topic which is drawn from a distribution with a randomly chosen parameter.

A geometric interpretation
A geometric interpretation
A geometric interpretation

Parameter Estimation

- Exact inference is not feasible
- Approximate methods
  - Gibbs Sampling
  - Variational inference
  - Collapsed Gibbs sampling
Gibbs Sampling

1. Initialize randomly the topic assignments
2. For each document “i” sample its topic mixture
   \[ p(\theta_i | .) = Dir(\{\alpha_k + \sum_l I(z_{il} = k)\}) \]
3. For each topic “k” sample from posterior of multinomial over vocabulary
   \[ p(\beta_k | .) = Dir(\{\gamma_v + \sum_i \sum_l I(w_{il} = v, z_{il} = k)\}) \]
4. For each document “i” and each word “l” sample its topic assignment
   \[ p(z_{il} = k | .) \propto \exp(\log \theta_{ik} + \log \beta_{kw_l}) \]

Parameter Estimation (Variational EM)

- Since we have latent variable model we need to use EM
  1. Find the expected value of the hidden variables (requires to run inference to compute \( p(\theta, z|w, \alpha, \beta) \))
  2. Use the expected counts to maximize the likelihood.
Inference and parameter estimation

- The key inferential problem is that of computing the posteriori distribution of the hidden variable given a document

\[
p(\theta, z | w, \alpha, \beta) = \frac{p(\theta, z, w | \alpha, \beta)}{p(w | \alpha, \beta)}
\]

\[
p(w | \alpha, \beta) = \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \int \left(\prod_{i=1}^{k} \theta_i^{\alpha_i - 1}\right) \left(\prod_{n=1}^{N} \sum_{i=1}^{k} \sum_{j=1}^{V} (\theta_i \beta_{ij})^{n_{wi}}\right) d\theta
\]

It is intractable to compute in general, due to the coupling between \(\theta\) and \(\beta\) in the summation over latent topics.

Variational Inference

- The basic idea of convexity-based variational inference is to make use of Jensen’s inequality to obtain an adjustable lower bound on the log likelihood.

- Essentially, one considers a family of lower bounds, indexed by a set of variational parameters.

- A simple way to obtain a tractable family of lower bound is to consider simple modifications of the original graph model in which some of the edges and nodes are removed.
In variational inference, we consider a simplified graphical model with variational parameters $\gamma$, $\phi$ and minimize the KL Divergence between the variational and posterior distributions.

$$p(\theta, z | w, \alpha, \beta) = \frac{p(\theta, z, w | \alpha, \beta)}{p(w | \alpha, \beta)} \quad q(\theta, z | \gamma, \phi) = q(\theta | \gamma) \prod_{n=1}^{N} q(z_n | \phi_n)$$

$$(\gamma^*, \phi^*) = \arg \min_{(\gamma, \phi)} KL(q(\theta, z | \gamma, \phi) || p(\theta, z | \alpha, \beta))$$
Application: Document modeling

- Unlabeled data – our goal is density estimation.
- Compute the perplexity of a held-out test to evaluate the models – lower perplexity score indicates better generalization.

\[
\text{perplexity}(D_{\text{test}}) = \exp \left\{ - \frac{\sum_{d=1}^{M} \log p(w_d)}{\sum_{d=1}^{M} N_d} \right\}
\]
Both pLSI and mixture suffer from overfitting.

pLSI – overfitting due to dimensionality of the \( p(z|d) \) parameter.

<table>
<thead>
<tr>
<th>Num. topics (k)</th>
<th>Perplexity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mult. Mixt.</td>
</tr>
<tr>
<td>2</td>
<td>22,266</td>
</tr>
<tr>
<td>5</td>
<td>( 2.20 \times 10^8 )</td>
</tr>
<tr>
<td>10</td>
<td>( 1.93 \times 10^{17} )</td>
</tr>
<tr>
<td>20</td>
<td>( 1.20 \times 10^{22} )</td>
</tr>
<tr>
<td>50</td>
<td>( 4.19 \times 10^{100} )</td>
</tr>
<tr>
<td>100</td>
<td>( 2.39 \times 10^{150} )</td>
</tr>
<tr>
<td>200</td>
<td>( 3.51 \times 10^{554} )</td>
</tr>
</tbody>
</table>
Application: Classification (Dimensionality Reduction)

(a) Accuracy vs. Proportion of data used for training

(b) Accuracy vs. Proportion of data used for training

Application: Recommendation Systems

LDA, Fold in pLSI, Smoothed Mixt. Unigrams

Predictive Perplexity vs. Number of Topics
Number of Topics K

- Cross validation, using log likelihood on the test set
- Use variational lower bound as a proxy for log p(D|K)
- Use non-parametric Bayesian Methods (The et al. 2006)
- Use annealed importance sampling to approximate the evidence (Wallach et al. 2009)

LDA Extensions

- Correlated Topic Model (Blei and Lafferty 2007)
- Supervised LDA (Blei and McAlliffe 2010)
- Dynamic Topic Model (Blei and Lafferty 2006)
- LDA-HMM (Griffiths et al. 2004)
- ......
Summary

- LDA is a flexible generative probabilistic model for collection of discrete data.
- Arguably, could be considered as the best possible model based on the Bag of Word assumption
- Can be viewed as a dimensionality reduction technique
- Exact inference is intractable, however it is possible to use approximate inference instead
- Can be used in other collection, e.g. images, collaborative filtering, ...
- There are lots of extensions and applications to LDA

References

- Latent Dirichlet allocation presentation Slides, David M. Blei, Princeton University
- Latent Dirichlet allocation presentation Slides, Ido Abramovich, Hebrew University
- LSI, PLSI, LDA presentation slides, Alexander Yates et al., Temple University