Inference and Representation

David Sontag

New York University

Lecture 8, Nov. 3, 2015
Latent Dirichlet allocation (LDA)

- **Topic models** are powerful tools for exploring large data sets and for making inferences about the content of documents.

- Many applications in information retrieval, document summarization, and classification.

- LDA is one of the simplest and most widely used topic models.
1. Sample the document’s **topic distribution** $\theta$ (aka topic vector)

$$\theta \sim \text{Dirichlet}(\alpha_1:T)$$

where the $\{\alpha_t\}_{t=1}^T$ are fixed hyperparameters. Thus $\theta$ is a distribution over $T$ topics with mean $\theta_t = \alpha_t / \sum_{t'} \alpha_{t'}$.

2. For $i = 1$ to $N$, sample the **topic** $z_i$ of the $i$’th word

$$z_i | \theta \sim \theta$$

3. ... and then sample the actual **word** $w_i$ from the $z_i$’th topic

$$w_i | z_i \sim \beta_{z_i}$$

where $\{\beta_t\}_{t=1}^T$ are the **topics** (a fixed collection of distributions on words)
Sample the document’s **topic distribution** $\theta$ (aka topic vector)

$$\theta \sim \text{Dirichlet}(\alpha_1:T)$$

where the $\{\alpha_t\}_{t=1}^T$ are hyperparameters. The Dirichlet density, defined over

$$\Delta = \{\overrightarrow{\theta} \in \mathbb{R}^T : \forall t \theta_t \geq 0, \sum_{t=1}^T \theta_t = 1\},$$

is:

$$p(\theta_1, \ldots, \theta_T) \propto \prod_{t=1}^T \theta_t^{\alpha_t-1}$$

For example, for $T=3$ ($\theta_3 = 1 - \theta_1 - \theta_2$):

\[\begin{array}{c}
\alpha_1 = \alpha_2 = \alpha_3 = 1.9485 \\
\log \text{Pr}(\theta) \\
\theta_1 \in [0, 1], \theta_2 \in [0, 1]
\end{array}\]

\[\begin{array}{c}
\alpha_1 = \alpha_2 = \alpha_3 = .31717 \\
\log \text{Pr}(\theta) \\
\theta_1 \in [0, 1], \theta_2 \in [0, 1]
\end{array}\]
... and then sample the actual word \( w_i \) from the \( z_i \)'th topic

\[
w_i | z_i \sim \beta_{z_i}
\]

where \( \{\beta_t\}_{t=1}^{T} \) are the **topics** (a fixed collection of distributions on words)
Example of using LDA

**Topics**

- **gene** 0.04
- **dna** 0.02
- **genetic** 0.01
- ..., 

- **life** 0.02
- **evolve** 0.01
- **organism** 0.01
- ..., 

- **brain** 0.04
- **neuron** 0.02
- **nerve** 0.01
- ..., 

- **data** 0.02
- **number** 0.02
- **computer** 0.01
- ..., 

**Documents**

**Seeking Life’s Bare (Genetic) Necessities**

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here, two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today’s organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn’t be enough. Although the numbers don’t match precisely, those predictions are not all that far apart,” especially in comparison to the 75,000 genes in the human genome, notes Sir Andersson of the KTH University in Sweden, who arrived at the 300 number. But coming up with a consensus answer may be more than just a genetic numbers game; particularly as more and more genomes are completely mapped and sequenced. “It may be a way of organizing any newly sequenced genome,” explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing and

**Figure 1:** The intuitions behind latent Dirichlet allocation.

We assume that some number of “topics,” which are distributions over words, exist for the whole collection (far left). Each document is assumed to be generated as follows. First choose a distribution over the topics (the histogram at right); then, for each word, choose a topic assignment (the colored coins) and choose the word from the corresponding topic. The topics and topic assignments in this figure are illustrative—they are not fit from real data. See Figure 2 for topics fit from data.

The model assumes the documents arose. (The interpretation of LDA as a probabilistic model is fleshed out below in Section 2.1.)

We formally define a topic to be a distribution over a fixed vocabulary. For example, the genetics topic has words about genetics with high probability and the evolutionary biology topic has words about evolutionary biology with high probability. We assume that these topics are specified before any data has been generated.

1 Now for each document in the collection, we generate the words in a two-stage process.

1. Randomly choose a distribution over topics.
2. For each word in the document
   a. Randomly choose a topic from the distribution over topics in step #1.
   b. Randomly choose a word from the corresponding distribution over the vocabulary.

This statistical model reflects the intuition that documents exhibit multiple topics. Each document exhibits the topics with different proportion (step #1); each word in each document

1 Technically, the model assumes that the topics are generated first, before the documents.

3

\[
\theta_d \quad \beta_1 \quad \beta_T
\]

(\textit{Blei, Introduction to Probabilistic Topic Models, 2011})

David Sontag (NYU)
“Plate” notation for LDA model

Variables within a plate are replicated in a conditionally independent manner
Outline of lecture

- How to learn topic models?
  - Importance of hyperparameters
  - Choosing number of topics
  - Evaluating topic models
- Examples of extending LDA
  - Polylingual topic models
  - Author-topic model
Learning algorithm: Gibbs Sampling

By putting a prior distribution on the parameters, they become random variables which can be sampled within the Gibbs Sampling algorithm:

\[
\beta_0 \rightarrow \beta \\
\beta \rightarrow \alpha, \theta, z, w
\]

\[
\alpha \rightarrow \theta, z, w
\]

\[
w \rightarrow M
\]

\[
z \rightarrow M
\]

\[
\theta \rightarrow \alpha, z, w
\]

\[
\alpha \rightarrow \theta, z, w
\]

\[
\beta_0 \sim \text{Dirichlet}(\cdot; \beta_0)
\]

**Figure:** Putting a Bayesian prior on the parameters: \( \beta \sim \text{Dirichlet}(\cdot; \beta_0) \)
Learn using a \textit{collapsed} Gibbs sampler

After marginalizing out $\theta_d$ for all documents $d$ and $\beta$, we get:

$$P(z_i = t \mid z_{-i}, w) \propto \frac{n_{-i,t}^{(w_i)} + \beta_0}{n_{-i,.}^{(\cdot)} + W\beta_0} \frac{n_{-i,t}^{(d_i)} + \alpha}{n_{-i,.}^{(d_i)} + T\alpha}$$

\textit{n} derived from $z$, the assignments of words to topics (\textit{W} words, \textit{T} topics, and uniform hyperparameters $\alpha$ and $\beta_0$)

First ratio is probability of $w_i$ under topic $t$, second ratio is probability of topic $t$ in document $d_i$

Given a sample, can get an estimate for $\beta$ and $\theta_d$ by:

$$\hat{\beta}_{w,t} = \frac{n_t^{(w)} + \beta_0}{n_t^{(\cdot)} + W\beta_0}$$

$$\hat{\theta}_t^{(d)} = \frac{n_t^{(d)} + \alpha}{n_{.}^{(d)} + T\alpha}$$
Polylingual topic models (Mimno et al., EMNLP ’09)

- Goal: topic models that are aligned across languages
- Training data: corpora with multiple documents in each language
  - EuroParl corpus of parliamentary proceedings (11 western languages; exact translations)
  - Wikipedia articles (12 languages; not exact translations)
- How to do this?
Polylingual topic models (Mimno et al., EMNLP ’09)

3 Polylingual Topic Model

The polylingual topic model (PLTM) is an extension of latent Dirichlet allocation (LDA) (Blei et al., 2003) for modeling polylingual document tuples and the predictive distributions over words for articles in French, English and German. PLTM assumes that each tuple has its own document-specific distribution over topics. This is unsumed that the documents in a tuple share the same topics. Additionally, PLTM assumes that each language, e.g., corresponding Wikipedia articles, are loosely equivalent to each other, but written in different languages, e.g., corresponding Wikipedia articles in French, English and German. PLTM assigns a new document tuple \( w = (w_1, \ldots, w_L) \) can be inferred using Gibbs inference tasks of interest are: computing the probability of the test tuples given the training tuples—

\[ P(w_L | T) = \frac{1}{Z_T} \sum_{\theta} P(\theta) \prod_{l=1}^{L} P(w_l | \theta) \]

Given a corpus of training and test document tuples and inferring latent topic assignments for each tuple, a latent topic assignment is drawn for each token in that language: 

\[ z \sim \text{Dir}(\alpha) \]

Each tuple is a set of documents that are topically similar document tuples (where there is a high probability of held-out documents originating from the same topics). The probability of held-out documents originating from the same topics is approximated by 

\[ P(w_L | T) = \frac{1}{Z_T} \sum_{\theta} P(\theta) \prod_{l=1}^{L} P(w_l | \theta) \]

We take a simpler approach that is more suitable for topically similar document tuples (where there is a high probability of held-out documents originating from the same topics) and use the MAP estimate for the latent topic assignments. We take the limiting property of the test tuples given the training tuples 

\[ P(w_L | T) = \frac{1}{Z_T} \sum_{\theta} P(\theta) \prod_{l=1}^{L} P(w_l | \theta) \]

Finally, the observed tokens are themselves drawn from a language-specific symmetric Dirichlet prior with concentration parameter \( \alpha \), a latent topic assignment is drawn for each token in that language: 

\[ z \sim \text{Dir}(\alpha) \]

In other words, rather than using a single set of word-to-word alignments as part of their inference procedures, which would become exponentially more complex if additional languages were added. Word-to-word alignments have been studied previously using the HM-bitam model (Zhao and Xing, 2007). Tam, Lane and Schultz (Tam et al., 2007) also show improvements in machine translation using bilingual topic models. Both translation-focused topic models infer translation using bilingual topic models. Both of these translation-focused topic models infer translation using bilingual topic models.
### Learned topics

<table>
<thead>
<tr>
<th>Language</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>DA</td>
<td>centralbank europæiske ecb s lån centralbanks</td>
</tr>
<tr>
<td>DE</td>
<td>zentralbank ezb bank europäischen investitionsbank darlehen</td>
</tr>
<tr>
<td>EL</td>
<td>τράπεζα τράπεζας κεντρική εκτ κεντρικής τράπεζας</td>
</tr>
<tr>
<td>EN</td>
<td><strong>bank central ecb banks european monetary</strong></td>
</tr>
<tr>
<td>ES</td>
<td>banco central europeo bce bancos centrales</td>
</tr>
<tr>
<td>FI</td>
<td>keskuspankin ekp n euroopan keskuspankki eip</td>
</tr>
<tr>
<td>FR</td>
<td>banque centrale bce européenne banques monétaire</td>
</tr>
<tr>
<td>IT</td>
<td>banca centrale bce europea banche prestiti</td>
</tr>
<tr>
<td>NL</td>
<td>bank centrale ecb europese banken leningen</td>
</tr>
<tr>
<td>PT</td>
<td>banco central europeu bce bancos empréstimos</td>
</tr>
<tr>
<td>SV</td>
<td>centralbanken europeiska ecb centralbankens s lån</td>
</tr>
</tbody>
</table>
Learned topics

<table>
<thead>
<tr>
<th>Language</th>
<th>Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>DA</td>
<td>børn familie udnyttelse børns børnene seksuel</td>
</tr>
<tr>
<td>DE</td>
<td>kinder kindern familie ausbeutung familien eltern</td>
</tr>
<tr>
<td>EL</td>
<td>παιδιά παιδιών οικογένεια οικογένειας γονείς παιδικής</td>
</tr>
<tr>
<td>EN</td>
<td>children family child sexual families exploitation</td>
</tr>
<tr>
<td>ES</td>
<td>niños familia hijos sexual infantil menores</td>
</tr>
<tr>
<td>FI</td>
<td>lasten lapsia lapset perheen lapsen lapsiin</td>
</tr>
<tr>
<td>FR</td>
<td>enfants famille enfant parents exploitation familles</td>
</tr>
<tr>
<td>IT</td>
<td>bambini famiglia figli minori sessuale sfruttamento</td>
</tr>
<tr>
<td>NL</td>
<td>kinderen kind gezin seksuele ouders familie</td>
</tr>
<tr>
<td>PT</td>
<td>crianças família filhos sexual criança infantil</td>
</tr>
<tr>
<td>SV</td>
<td>barn barnen familjen sexuellt familj utnyttjande</td>
</tr>
</tbody>
</table>
Discussion

- How would you use this?
- How could you extend this?
Goal: topic models that take into consideration author interests
Training data: corpora with label for who wrote each document
- Papers from NIPS conference from 1987 to 1999
- Twitter posts from US politicians

Why do this?

How to do this?
Author-topic model (Rosen-Zvi et al., UAI ’04)

Figure 1: Generative models for documents. (a) Latent Dirichlet Allocation (LDA; Blei et al., 2003), a topic model. (b) An author model. (c) The author-topic model.
### Most likely author for a topic

<table>
<thead>
<tr>
<th>Topic 31</th>
<th>Topic 61</th>
<th>Topic 71</th>
<th>Topic 100</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>WORD</strong></td>
<td><strong>PROB.</strong></td>
<td><strong>WORD</strong></td>
<td><strong>PROB.</strong></td>
</tr>
<tr>
<td>SPEECH</td>
<td>0.0823</td>
<td>BAYESIAN</td>
<td>0.0450</td>
</tr>
<tr>
<td>RECOGNITION</td>
<td>0.0497</td>
<td>GAUSSIAN</td>
<td>0.0364</td>
</tr>
<tr>
<td>HMM</td>
<td>0.0234</td>
<td>POSTERIOR</td>
<td>0.0355</td>
</tr>
<tr>
<td>SPEAKER</td>
<td>0.0226</td>
<td>PRIOR</td>
<td>0.0345</td>
</tr>
<tr>
<td>CONTEXT</td>
<td>0.0224</td>
<td>DISTRIBUTION</td>
<td>0.0259</td>
</tr>
<tr>
<td>WORD</td>
<td>0.0166</td>
<td>PARAMETERS</td>
<td>0.0199</td>
</tr>
<tr>
<td>SYSTEM</td>
<td>0.0151</td>
<td>EVIDENCE</td>
<td>0.0127</td>
</tr>
<tr>
<td>ACOUSTIC</td>
<td>0.0134</td>
<td>SAMPLING</td>
<td>0.0117</td>
</tr>
<tr>
<td>PHONEME</td>
<td>0.0131</td>
<td>COVARIANCE</td>
<td>0.0117</td>
</tr>
<tr>
<td>CONTINUOUS</td>
<td>0.0129</td>
<td>LOG</td>
<td>0.0112</td>
</tr>
<tr>
<td><strong>AUTHOR</strong></td>
<td><strong>PROB.</strong></td>
<td><strong>AUTHOR</strong></td>
<td><strong>PROB.</strong></td>
</tr>
<tr>
<td>Waibel_A</td>
<td>0.0936</td>
<td>Bishop_C</td>
<td>0.0563</td>
</tr>
<tr>
<td>Makhoul_J</td>
<td>0.0238</td>
<td>Williams_C</td>
<td>0.0497</td>
</tr>
<tr>
<td>De-Mori_R</td>
<td>0.0225</td>
<td>Barber_D</td>
<td>0.0368</td>
</tr>
<tr>
<td>Bourlard_H</td>
<td>0.0216</td>
<td>MacKay_D</td>
<td>0.0323</td>
</tr>
<tr>
<td>Cole_R</td>
<td>0.0200</td>
<td>Tipping_M</td>
<td>0.0216</td>
</tr>
<tr>
<td>Rigoll_G</td>
<td>0.0191</td>
<td>Rasmussen_C</td>
<td>0.0215</td>
</tr>
<tr>
<td>Hochberg_M</td>
<td>0.0176</td>
<td>Opper_M</td>
<td>0.0204</td>
</tr>
<tr>
<td>Franco_H</td>
<td>0.0163</td>
<td>Attias_H</td>
<td>0.0155</td>
</tr>
<tr>
<td>Abrash_V</td>
<td>0.0157</td>
<td>Sollich_P</td>
<td>0.0143</td>
</tr>
<tr>
<td>Movellan_J</td>
<td>0.0149</td>
<td>Schottky_B</td>
<td>0.0128</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Lecture 8, Nov. 3, 2015
Perplexity as a function of number of observed words

\[
\text{perplexity}(\mathbf{w}_{\text{test},d} \mid \mathbf{w}_{\text{train},d}, \mathbf{a}_{d}) = \exp \left[ -\frac{\ln p(\mathbf{w}_{\text{test},d} \mid \mathbf{w}_{\text{train},d}, \mathbf{a}_{d})}{N_{\text{test},d}} \right]
\]