Evaluation Methods for Topic Models

Hanna M. Wallach

University of Massachusetts Amherst
wallach@cs.umass.edu

April 13, 2009

Joint work with Iain Murray, Ruslan Salakhutdinov and David Mimno
Statistical Topic Models

- Useful for analyzing large, unstructured text collections

<table>
<thead>
<tr>
<th>bounds</th>
<th>units</th>
<th>policy</th>
<th>data</th>
<th>neurons</th>
</tr>
</thead>
<tbody>
<tr>
<td>bound</td>
<td>hidden</td>
<td>action</td>
<td>space</td>
<td>neuron</td>
</tr>
<tr>
<td>loss</td>
<td>network</td>
<td>reinforcement</td>
<td>clustering</td>
<td>spike</td>
</tr>
<tr>
<td>functions</td>
<td>layer</td>
<td>learning</td>
<td>points</td>
<td>synaptic</td>
</tr>
<tr>
<td>error</td>
<td>unit</td>
<td>actions</td>
<td>distance</td>
<td>firing</td>
</tr>
</tbody>
</table>

- Topic-based search interfaces ([http://rexa.info](http://rexa.info))
- Analysis of scientific trends (Blei & Lafferty, '07; Hall et al., '08)
- Information retrieval (Wei & Croft '06)
Latent Dirichlet Allocation (Blei et al., ’03)

- LDA generates a new document $\mathbf{w}$ by drawing:

\[
\begin{align*}
\theta & \sim \text{Dir}(\theta; \alpha \mathbf{m}) \quad \text{a document-specific topic dist.,} \\
z & \sim P(z \mid \theta) = \prod_n \theta_{z_n} \quad \text{a topic assignment for each token,} \\
\mathbf{w} & \sim P(\mathbf{w} \mid z, \Phi) = \prod_n \phi_{w_n \mid z_n} \quad \text{and finally the observed tokens.}
\end{align*}
\]

- The “topic” parameters $\Phi$, and $\alpha \mathbf{m}$, are shared by all documents
- For real-world data, only the tokens $\mathbf{w}$ are observed
Evaluating Topic Model Performance

- Unsupervised nature of topic models makes evaluation hard
- There may be extrinsic tasks for some applications...
- ... but we also want to estimate cross-task generalization
- Compute probability of held-out documents under the model
  - Classic way of evaluating generative models
  - Often used to evaluate topic models
- This talk: demonstrate that standard methods for evaluating topic models are inaccurate and propose two alternative methods
Evaluating LDA

Given training documents $\mathcal{W'}$ and held-out documents $\mathcal{W}$:

$$P(\mathcal{W} | \mathcal{W'}) = \int d\Phi d\alpha d\mathbf{m} P(\mathcal{W} | \Phi, \alpha \mathbf{m}) P(\Phi, \alpha \mathbf{m} | \mathcal{W'})$$

Approximate this integral by evaluating at a point estimate.

Variational or MCMC can be used to marginalize out topic assignments for training documents to infer $\Phi$ and $\alpha \mathbf{m}$.

The probability of interest is therefore:

$$P(\mathcal{W} | \Phi, \alpha \mathbf{m}) = \prod_d P(w^{(d)} | \Phi, \alpha \mathbf{m})$$
Computing $P(w | \Phi, \alpha m)$

- $P(w | \Phi, \alpha m)$ is the normalizing constant that relates the posterior distribution over $z$ to the joint distribution over $w$ and $z$:

  $$P(z | w, \Phi, \alpha m) = \frac{P(w, z | \Phi, \alpha m)}{P(w | \Phi, \alpha m)}$$

- Computing it involves marginalizing over latent variables:

  $$P(w | \Phi, \alpha m) = \sum_z \int d\theta 
  P(w, z, \theta | \Phi, \alpha m)$$
Methods for Computing Normalizing Constants

- Simple importance sampling methods:
  - e.g., MALLET’s “empirical likelihood”, “iterated pseudo-counts”
- The “harmonic mean” method (Newton & Raftery, '94):
  - Known to overestimate, yet used in topic modeling papers
- Annealed importance sampling (Neal, '01):
  - Accurate, but prohibitively slow for large data sets
- A Chib-style method (Murray & Salakhutdinov, '09)
- A “left-to-right” method (Wallach, '08)
Chib-Style Estimates

- For any “special” set of latent topic assignments $\mathbf{z}^*$:

$$P(\mathbf{w} | \Phi, \alpha \mathbf{m}) = \frac{P(\mathbf{w} | \mathbf{z}^*, \Phi) P(\mathbf{z}^* | \alpha \mathbf{m})}{P(\mathbf{z}^* | \mathbf{w}, \Phi, \alpha \mathbf{m})}$$

- Chib-style estimation:
  1. Pick some special set of latent topic assignments $\mathbf{z}^*$
  2. Compute $P(\mathbf{w} | \mathbf{z}^*, \Phi) P(\mathbf{z}^* | \alpha \mathbf{m})$
  3. Estimate $P(\mathbf{z}^* | \mathbf{w}, \Phi, \alpha \mathbf{m})$

- Can use a Markov chain to estimate $P(\mathbf{z}^* | \mathbf{w}, \Phi, \alpha \mathbf{m})$
Markov Chain Estimation

- Stationary condition for a Markov chain:

\[ P(z^* \mid w, \Phi, \alpha m) = \sum_z T(z^* \leftarrow z) \frac{P(z \mid w, \Phi, \alpha m)}{P(z^* \mid w, \Phi, \alpha m)} \]

- Estimate sum using a sequence of states \( Z = \{z^{(1)}, \ldots, z^{(S)}\} \) generated by a Markov chain that explores \( P(z \mid w, \Phi, \alpha m) \)
Overestimate of $P(w \mid \Phi, \alpha m)$

- $P(z^* \mid w, \Phi, \alpha m)$ is unbiased in expectation:

$$P(z^* \mid w, \Phi, \alpha m) = \mathbb{E} \left[ \frac{1}{S} \sum_{s=1}^{S} T(z^* \leftarrow z^{(s)}) \right]$$

- But, in expectation, $P(w \mid \Phi, \alpha m)$ will be overestimated (Jensen):

$$P(w \mid \Phi, \alpha m) = \frac{P(z^*, w \mid \Phi, \alpha m)}{\mathbb{E} \left[ \frac{1}{S} \sum_{s=1}^{S} T(z^* \leftarrow z^{(s)}) \right]} \leq \mathbb{E} \left[ \frac{P(z^*, w \mid \Phi, \alpha m)}{\frac{1}{S} \sum_{s=1}^{S} T(z^* \leftarrow z^{(s)})} \right]$$
Chib-Style Method \[\text{(Murray & Salakhutdinov, '09)}\]

- Draw $\mathcal{Z} = \{z^{(1)}, \ldots, z^{(S)}\}$ from a carefully designed distribution

$$P(w | \Phi, \alpha m) \simeq P(w, z^* | \Phi, \alpha m) \frac{1}{S} \sum_{s'=1}^{S} T(z^* \leftarrow z^{(s')})$$
Left-to-Right Method (Wallach, ’08)

- Can decompose $P(w \mid \Phi, \alpha m)$ as

$$
P(w \mid \Phi, \alpha m) = \prod_{n} P(w_n \mid w_{<n}, \Phi, \alpha m)
= \prod_{n} \sum_{z_{\leq n}} P(w_n, z_{\leq n} \mid w_{<n}, \Phi, \alpha m)
$$

- Approximate each sum over $z_{\leq n}$ using a MCMC algorithm
- “Left-to-right”: appropriate for language modeling applications
Left-to-Right Method (Wallach, ’08)

1: for each position \( n \) in \( w \) do
2:   for each particle \( r = 1 \) to \( R \) do
3:     for each position \( n' < n \) do
4:       resample \( z^{(r)}_{n'} \sim P(z^{(r)}_{n'} | w_{n'}, \{z^{(r)}_{< n'}\} \setminus n', \Phi, \alpha m) \)
5:     end for
6:   \( p^{(r)}_n := \sum_t P(w_n, z^{(r)}_n = t | z^{(r)}_{< n}, \Phi, \alpha m) \)
7:   sample a topic assignment: \( z^{(r)}_n \sim P(z^{(r)}_n | w_n, z^{(r)}_{< n}, \Phi, \alpha m) \)
8: end for
9: \( p_n := \sum_r p^{(r)}_n / R \)
10: \( l := l + \log p_n \)
11: end for
Relative Computational Costs

- Gibbs sampling dominates cost for most methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Parameters</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iterated pseudo-counts</td>
<td># itns. I, # samples S</td>
<td>((I + S) N)</td>
</tr>
<tr>
<td>Empirical likelihood</td>
<td># samples S</td>
<td>(SN)</td>
</tr>
<tr>
<td>Harmonic mean</td>
<td>burn-in B, # samples S</td>
<td>(N (B + S))</td>
</tr>
<tr>
<td>AIS</td>
<td># temperatures S</td>
<td>(SN)</td>
</tr>
<tr>
<td>Chib-style</td>
<td>chain length S</td>
<td>(2SN)</td>
</tr>
<tr>
<td>Left-to-right</td>
<td># particles R</td>
<td>(RN (N - 1) / 2)</td>
</tr>
</tbody>
</table>

- Costs are in terms of # Gibbs site updates required (or equivalent)
### Data Sets

- Two synthetic data sets, three real data sets:

<table>
<thead>
<tr>
<th>Data set</th>
<th>$V$</th>
<th>$\bar{N}$</th>
<th>St. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synthetic, 3 topics</td>
<td>9242</td>
<td>500</td>
<td>0</td>
</tr>
<tr>
<td>Synthetic, 50 topics</td>
<td>9242</td>
<td>200</td>
<td>0</td>
</tr>
<tr>
<td>20 Newsgroups</td>
<td>22695</td>
<td>120.4</td>
<td>296.2</td>
</tr>
<tr>
<td>PubMed Central abstracts</td>
<td>30262</td>
<td>101.8</td>
<td>49.2</td>
</tr>
<tr>
<td>New York Times articles</td>
<td>50412</td>
<td>230.6</td>
<td>250.5</td>
</tr>
</tbody>
</table>

- $V$ is the vocabulary size, $\bar{N}$ is the mean document length, “St. Dev.” is the estimated standard deviation in document length.
Average Log Prob. Per Held-Out Document (20 Newsgroups)

- **AIS**: Annealed importance sampling. **HM**: Harmonic mean. **LR**: Left-to-right. **CS**: Chib-style. **IS-EL**: Importance sampling (empirical likelihood). **IS-IP**: Importance sampling (iterated pseudocounts)
Conclusions

- Empirically determined that the evaluation methods currently used in the topic modeling community are inaccurate:
  - Harmonic mean method often significantly overestimates
  - Simple IS methods tend to underestimate (but not by as much)

- Proposed two, more accurate, alternatives
  - A Chib-style method (Murray & Salakhutdinov, ’09)
  - A left-to-right method (Wallach, ’08)
Questions?

wallach@cs.umass.edu

http://www.cs.umass.edu/~wallach/
Average Log Prob. Per Held-Out Document (Synth., 3 Topics)

- **AIS**: Annealed importance sampling.  
- **HM**: Harmonic mean.  
- **LR**: Left-to-right.  
- **CS**: Chib-style.  
- **IS-EL**: Importance sampling (empirical likelihood).  
- **IS-IP**: Importance sampling (iterated pseudocounts)
Average Log Prob. Per Held-Out Document (PubMed Central)

- **AIS**: Annealed importance sampling.
- **HM**: Harmonic mean.
- **LR**: Left-to-right.
- **CS**: Chib-style.
- **IS-EL**: Importance sampling (empirical likelihood).
- **IS-IP**: Importance sampling (iterated pseudocounts)

- **AIS**: Annealed importance sampling. **HM**: Harmonic mean. **LR**: Left-to-right. **CS**: Chib-style. **IS-EL**: Importance sampling (empirical likelihood). **IS-IP**: Importance sampling (iterated pseudocounts)
Choosing a “Special” State $z^*$

- Run regular Gibbs sampling for a few iterations
- Iteratively maximize the following quantity:

$$P(z_n = t \mid w, z_{\backslash n}, \Phi, \alpha \mathbf{m})$$

$$\propto P(w_n \mid z_n = t, \Phi) \, P(z_n = t \mid z_{\backslash n}, \alpha \mathbf{m})$$

$$\propto \phi_{w_n \mid t} \, \frac{\{N_t\}_{\backslash n} + \alpha \mathbf{m}_t}{N - 1 + \alpha},$$

- $\{N_t\}_{\backslash n}$ is # times topic $t$ occurs in $z$ excluding position $n$