Memory Bounded Inference in Topic Models

Ryan Gomes*, Max Welling**, and Pietro Perona*

*Caltech

**UC Irvine
Motivation

What type of algorithms support unsupervised learning from very large datasets over long stretches of time?
Motivation

What type of algorithms support unsupervised learning from very large datasets over long stretches of time?

- Model complexity (number of categories/topic) should adapt as new data is collected.
Motivation

What type of algorithms support unsupervised learning from very large datasets over long stretches of time?

- Model complexity (number of categories/topic) should adapt as new data is collected.
- Space and time required to update model must scale gradually with amount of data.
Online Approaches

Maintain multiple parallel hypotheses with differing number of clusters/topics.

Very expensive: combinatorial explosion in the number of possible assignments.

- “Particle filters for mixture models with an unknown number of components.” (P. Fearnhead 2004).
- “Online Model Selection Based on the Variational Bayes.” (M. Sato 2001).
Overview

Estimate Model

Document 1

Document 2
Overview

Estimate Model  →  Compression

Document 1

Document 2

Document 1

Document 2
Overview

Estimate Model → Compression → Get more data, estimate model

Document 1

Document 2

Document 3

Document 4
Overview

Estimate Model → Compression → Get more data, estimate model → Compression

Document 1

Document 2

Document 3

Document 4
Topic Model

\[
p(x, z, \eta, \pi, \alpha) = \prod_{i,j} p(x_{ij} | z_{ij}; \eta) \pi_{j,z_{ij}} \prod_k p(\eta_k | \beta) G(\alpha_k; a, b) \prod_j \mathcal{D}(\pi_j; \alpha)
\]

\[x_{ij}: \text{ word i in document j}\]
\[z_{ij}: \text{ topic assignment variable for word i in document j}\]
\[\eta_k: \text{ parameter for topic k}\]
\[\pi_j: \text{ mixture of topics for document j (with Dirichlet prior)}\]
\[\alpha: \text{ topic mixture prior parameter (with Gamma priors)}\]
\[\beta: \text{ topic prior hyperparameter}\]
\[a, b: \text{ Gamma prior hyperparameters}\]
Variational Approximation

\[ q(\eta) = \prod_k q(\eta_k; \xi_k) \]
\[ q(\pi) = \prod_j D(\pi_j; \zeta_j) \]
\[ q(z) = \prod_{ij} q(z_{ij}) \]
\[ q(\alpha_k) = \delta(\alpha_k - \hat{\alpha}_k) \]

\[ F(x; q) \leq \log p(x; \beta, a, b) \]

- Topic mixture prior is point estimated to avoid conjugacy issues. Efficient update rules based on (Minka, 2000).
- Optimizing the number of topics: Truncate q(z) at K topics. Use free energy cost function to compare solutions with different K.
- Topic split and merge (Ueda et al., 1999)
- Other approaches exist, e.g. (Teh et al., 2007).
“Clumps”

if $x_{ij}$ and $x_{i'j'}$ are in clump $c$:

$q(z_{ij}) = q(z_{i'j'}) = q(z_c)$

Key assumption: $p(x_{ij} | z_{ij})$ in exponential family with conjugate prior $p(\eta_k | \beta)$
Document Groups

if document \( j \) and \( j' \) are in group \( s \):

\[
q(\pi_j) = q(\pi_{j'}) = q(\pi_s)
\]

- Update rules for \( q(\pi_s) \) depend on average topic counts
Compression

Clumps

- Recursively split groups of data points
- Compression is irreversible: must account for future data with modified free energy cost function.
- Halt when memory cost to store clumps (MC) exceeds predefined bound:

\[ MC = \left( \frac{d^2 + 3d}{2} \right) |N_c > 1| + |S||N_c > 1| + d|N_c = 1| \]

Document Groups

\[ DM_{s,s'} = \sum_k \frac{E[\pi_{sk}]E[\pi_{s'k}]}{|E[\pi_s]|||E[\pi_s']||} \]
Joint Segmentation

- Aligned Faces
- Segments of same color belong to same topic.
- \(~4.5\) times faster than batch baseline

Free Energy Ratio vs. Batch

<table>
<thead>
<tr>
<th># of groups/total images processed</th>
<th>0.2</th>
<th>0.4</th>
<th>0.6</th>
<th>0.8</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Free Energy Ratio vs. Batch</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.975</td>
<td>0.98</td>
<td>0.985</td>
<td>0.99</td>
<td>0.995</td>
</tr>
</tbody>
</table>

Free Energy Ratio vs. Batch

<table>
<thead>
<tr>
<th>Number of clumps</th>
<th>0</th>
<th>100</th>
<th>200</th>
<th>300</th>
</tr>
</thead>
<tbody>
<tr>
<td>Free Energy Ratio vs. Batch</td>
<td>0.95</td>
<td>0.96</td>
<td>0.97</td>
<td>0.98</td>
</tr>
</tbody>
</table>
Joint Segmentation

- 435 documents, 13.9 million words
- ~38 fold memory savings relative to batch baseline

![Graphs showing the number of model components and minutes per learning round over the number of images processed and learning rounds.]

Clumps
Model Segments (topics)
Object Recognition

Caltech 101 Object Categories Dataset

- 3000 training images, 1000 test images
- 500 128-dimension SIFT features per image
- Topic model defines similarity between images.
- Similarity enables image retrieval and nearest neighbor classification.
Object Recognition

1-NN Accuracy vs Memory Bound

Free Energy vs Memory Bound

1-NN Accuracy vs # of groups/total images processed

Free Energy vs # of groups/total images processed
Future Work

- Generalize to other models
- More principled insight into compression phase
- Bounded primary memory (RAM) but unlimited secondary memory