Stochastic Divergence Minimization for Online Collapsed Variational Bayes Zero Inference of Latent Dirichlet Allocation

@KDD2015

Issei Sato & Hiroshi Nakagawa
The University of Tokyo
Latent Dirichlet allocation [Blei, 2003]

The annual ACM SIGKDD conference is the premier international forum for data mining researchers and practitioners from academia, industry, and government to share their ideas, research results and experiences.
Latent Dirichlet allocation [Blei, 2003]

The annual ACM SIGKDD conference is the premier international forum for data mining researchers and practitioners from academia, industry, and government to share their ideas, research results, and experiences.

Modeling co-occurrence:
Frequently co-occurring words are assigned to the same topic (color)
De Finetti theorem [De Finetti, 1930s]

A sequence of random variables \((x_1, x_2, \ldots)\) is infinitely exchangeable iff, for all \(n\)

\[
p(x_1, x_2, \ldots, x_n) = \int \prod_{i=1}^{n} p(x_i \mid \theta) p(\theta) d\theta
\]
Latent Dirichlet allocation \cite{Blei2003}

Topic distribution for each document
\[ \theta_d \sim \text{Dir}(\gamma) \quad (d = 1, \ldots, D) \]

Word distribution for each topic
\[ \phi_k \sim \text{Dir}(\beta) \quad (k = 1, \ldots, K) \]

For each words:
\[ z_{d,i} \sim \text{Multi}(\theta_d) \]
\[ w_{d,i} \sim \text{Multi}(\phi_{z_{d,i}}) \]
Priors Matter [Wallach+,2009]

Asymmetric Dirichlet prior

\[ \theta_d \sim \text{Dir}(\gamma_1, \gamma_2, \ldots, \gamma_K) \]

\[ \phi_k \sim \text{Dir}(\beta_1, \beta_2, \ldots, \beta_V) \]

Symmetric Dirichlet prior

For each words:

\[ z_{d,j,i} \sim \text{Dir}(\gamma, \gamma, \ldots, \gamma) \]

\[ w_{d,i} \sim \text{Dir}(\beta, \beta, \ldots, \beta) \]
Evaluation: Perplexity

Prediction of held-out words

$$\exp \left[ \frac{1}{N_{test}} \sum_{w^* \in W_{test}} \log p(w^* | W_{train}) \right]$$

$$\sum_{k=1}^{K} \frac{E[n_{k,w^*}^{train}]}{E[n_k^{train}]} + \beta \frac{E[n_{d,k}^{train}]}{E[n_d^{train}]} + \sum_k \gamma_k$$
Inference algorithms

- Variational Bayes (VB)  
  [Blei+, JMLR 2003]
- Collapsed Gibbs Sampling (CGS)  
  [Griffiths+, PNAS 2004]
- Collapsed Variational Bayes (CVB)  
  [Teh+, NIPS 2007]
- Collapsed Variational Bayes Zero (CVB0)  
  [Asuncion+, UAI 2009]

Marginalize out parameters
Why CVB0 works better?

CVB0 uses zero-order Taylor approximation for expectations in CVB

→ CVB0 is less accurate than CVB

CVB0 can be formulated as a local $\alpha$-divergence minimization

[Sato & Nakagawa, ICML2012]
### α-divergence minimization

<table>
<thead>
<tr>
<th>Inference</th>
<th>Marginalization</th>
<th>α-divergence</th>
</tr>
</thead>
<tbody>
<tr>
<td>VB [Blei+,03]</td>
<td></td>
<td>α→0</td>
</tr>
<tr>
<td>CVB [Teh+,07]</td>
<td>✓</td>
<td>α→0</td>
</tr>
<tr>
<td>CVB0 [Asuncion+, 09]</td>
<td>✓</td>
<td>α→1 (≒1)</td>
</tr>
<tr>
<td>EP [Minka+,02]</td>
<td></td>
<td>α→1</td>
</tr>
</tbody>
</table>

#### Zero forcing effect

The emphasis in the estimation is on high-frequency topics or low-frequency topics is forced to be zero.
Stochastic Optimization

Scaling up: Batch data $\rightarrow$ Sub-sampling

• Variational Bayes (VB) [Blei+, JMLR 2003]
  $\rightarrow$ Stochastic Variational Bayes (SVB)
  [Hoffman+, Sato+, NIPS 2010]

• Collapsed Variational Bayes Zero (CVB0)
  [Asuncion+, UAI 2009]
  $\rightarrow$ Stochastic Collapsed Variational Bayes Zero (SCVB0)
  [Foulds+, KDD 2013]
Framework of SCVB0

Problem
How to formulate SO of CVB0
- CVB0 integrates out parameters

Solution
When we manually adjust Dirichlet prior,
CVB0 update ~ MAP update.

SCVB0 ~ Stochastic Approx. of MAP infer.
Question and Problem on SCVB0

• Why MAP works better than VB?
• We cannot use Asym. Dirichlet prior

Our contribution

• Formulation of SCVB0
  → Stochastic divergence minimization (SDM)

• Estimation of Dirichlet prior
  → Reformulate DM of [Sato+, ICML2012]
Main Idea

[Sato & Nakagawa, ICML2012]

Infer \( q(Z) = \prod_{d,i} q(z_{d,i}) \) by DM

This work

Infer

\[
q(Z, W \mid \gamma, \beta) = \prod_{d,i} q(z_{d,i} \mid w_{d,i}) q(w_{d,i} \mid \gamma, \beta)
\]

=CVB0 update

by DM

Stochastic Approx.
Our contributions

\[ q(w_{d,i} \mid \gamma, \beta) = \sum_{k=1}^{K} \frac{E[n_{k,w_{d,i}}]}{E[n_{k}^{\text{loo}}] + V\beta} + \frac{E[n_{d,k}]}{E[n_{d}^{\text{loo}}] + \sum_{k} \gamma_k} \]

Leave-One-Out-Perplexity

\[ \exp \left[ -\frac{1}{N} \sum_{d,i} \log q(w_{d,i} \mid \gamma, \beta) \right] \]

This work is CVB0 update

Infer

\[ q(Z,W \mid \gamma, \beta) = \prod_{d,i} q(z_{d,i} \mid w_{d,i}) \]

\[ q(w_{d,i} \mid \gamma, \beta) \]

= CVB0 update by DM

Stochastic Approx.
Testset perplexity $\times$ LOO perplexity

NY times $\times$ CVB0 with Sym.Dir

Correlation coefficient : 0.9913
Testset perplexity $\times$ LOO perplexity

NY times $\times$ CVB0 with Sym.Dir

Graph showing the change in perplexity over the number of iterations.
Empirical Bayes

\[(\gamma^*, \beta^*) = \arg \max_{\gamma, \beta} \log p(D | \gamma, \beta)\]

Variational EM

\[(\gamma^*, \beta^*) = \arg \max_{\gamma, \beta} L(D | \gamma, \beta)\]

\[\log p(D | \gamma, \beta) \geq L(D | \gamma, \beta)\]

This work

Stochastic Approx.

\[(\gamma^*, \beta^*) = \arg \max_{\gamma, \beta} \log q(D | \gamma, \beta)\]

\[\prod_{d, i} q(w_{d, i} | \gamma, \beta)\]

\[\iff \min \text{ Leave-One-Out Perplexity}\]
<table>
<thead>
<tr>
<th></th>
<th>VB ‘03</th>
<th>SVB ‘10</th>
<th>CVB0 ‘07</th>
<th>SCVB0 ‘13</th>
<th>This work</th>
</tr>
</thead>
<tbody>
<tr>
<td>Memory</td>
<td>O(VK)</td>
<td>O(VK)</td>
<td>O(NK)</td>
<td>O(VK)</td>
<td>O(VK)</td>
</tr>
<tr>
<td>Update/mini-batch</td>
<td>-</td>
<td>O(VK)</td>
<td>-</td>
<td>O(VK)</td>
<td>O(VsK)</td>
</tr>
<tr>
<td>HDP(Asym. Dir)</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>-</td>
<td>✔</td>
</tr>
</tbody>
</table>

V: Vocabulary size  
Vs: Vocab. size in sub-samples  
K: # of topics  
N: Total # of words

Ignore!
Experimental settings

4 datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th># of Doc.</th>
<th>Vocab. size</th>
</tr>
</thead>
<tbody>
<tr>
<td>DBLP</td>
<td>0.6M</td>
<td>19K</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>1M</td>
<td>130K</td>
</tr>
<tr>
<td>Pubmed1M</td>
<td>1M</td>
<td>50K</td>
</tr>
<tr>
<td>Pubmed5M</td>
<td>5M</td>
<td>122K</td>
</tr>
</tbody>
</table>

Evaluation: Testset Perplexity

Algorithms: SVB, SCVB0, SDM (This work)

# of Topics (Truncation level): 1000
Bayesian optimization for tuning hyper-parameters

Dataset: Pubmed1M

![Graph showing the performance of different methods over the number of BO trials. The graph compares SCVB0+BO, SVB+BO, and SCVB0 [Foulds+13] on minimizing perplexity. The y-axis represents the minimum perplexity, and the x-axis represents the number of BO trials. The graph shows how each method performs over time, with SCVB0+BO and SVB+BO generally improving over the trials.]
Experimental result

Dataset: Pubmed5M
反省点:査読者との戦いを経て

- スコープが狭い
  - （建前）LDAは引用数1万を超える論文なので、LDAのアルゴリズムの改良自体は重要
  - （本音）汎用性大事。少なくとも汎用性がりそうな書き方を心がけるべき

- 実験が少ない。Twitterの解析とかもやったら？
  - （建前）[Foulds+, KDD2013]と同等の実験
  - （本音）実験の種類はやはり多いほうが良い
    少なくとも[Sato+, KDD2012]のときは、Perplexity(4datasets), リンク予測(2datasets), 文書分類(2datasets)としたら褒められた

- 外部リンク先に証明のある定理は貢献に入れるべきではない
  - （建前）
    事前にCo-Chairsに可能か確認済なので、メタ査読者に確認を
    以前KDDの査読者に外部リンクに置くように言われたことがある
    そのようにしているKDD論文も過去にある
  - （本音）Supplemental materialの無い会議では、やはりペーパー内に収めるように書きべき