Announcements

- Tutorial: derivation of the Gibbs sampler
- Meet in pairs to discuss
- Quiz #2 will cover the main ideas of recent lectures and the tutorial
- Coming up: LDA project

Objectives

- Pose questions about how to use the samples
- Understand the behavior of the Gibbs sampler for clustering using the Mixture of Multinomials model
- Review the experimental results

Acknowledgments

- Dan Walker’s Ph.D. dissertation: ch. 2

What Now?

- \( C \leftarrow \text{GibbsSampler}(x, a, b, \text{burn}, \text{length}) \)
- Now we have a matrix \( C \) of samples
  - \( c_{d,j} \) is the \( j^{th} \) sample for the \( d^{th} \) document
- How can we best use \( C \) to assign documents to clusters?
- Four questions:
  - How many samples should we discard (burn)?
  - How many samples do we need to take (length)?
  - How should we summarize the samples?
    - MAP sample?
    - Random sample?
    - Marginal posterior?
    - Is it worth it?

Convergence and # Samples

- Time-series of likelihood \( p(x, c | a, b) \)
  - gives evidence of MCMC chain convergence (in distribution) very early

Del.icio.us      Enron      20 Newsgroups
Single-Sample Summarization

- **MAP Sample**
  - Choose the sample with the highest likelihood
  \[ \text{label}_d \leftarrow \arg \max_{c_d \in \mathcal{C}_d} p(c_d | x, \alpha, \beta) \]
  - A good candidate since quality is high according to model

- **Random Sample**
  - Chose an arbitrary sample after burn
  - We can just use the first sample after burn, and not take any more samples
  - Cheap!

Multi-Sample Summarization

- **Marginal posterior**
  - Or just "marginal"
  - Choose the mode of the document’s (post-burn) sample vector
  \[ \text{label}_d \leftarrow \arg \max_c \sum_i \delta(c_{d,i}, c) \]

Non-Identifiability

- Cluster labels have no real meaning
  - If we label all of the 1s as 2s and the 2s as 1s, nothing has changed
  - This can make it impossible to use summarizations that use more than 1 sample
  - If labels switch mid-chain, then the resulting distributions may turn out uniform

- To diagnose label-switching,
  - \( p_{\text{burn}+1}(w|c) \) was calculated for a reference sample (after burn-in)
  - \( p_i(w|c) \) also calculated for subsequent samples \( \text{burn} + 1 < i < \text{length} \)
  - For all pairs of classes, compute KL-divergence
  \[ KL(p_{\text{burn}+1} || p_i) \]

What’s going on?

- Stuck in one region of the state space
- Is that bad?

20 Newsgroups

<table>
<thead>
<tr>
<th>Metric</th>
<th>Marginal</th>
<th>MAP</th>
<th>Random</th>
</tr>
</thead>
<tbody>
<tr>
<td>F-Measure</td>
<td>0.42174</td>
<td>0.42017</td>
<td>0.42118</td>
</tr>
<tr>
<td>VI</td>
<td>2.14728</td>
<td>2.16088</td>
<td>2.15931</td>
</tr>
<tr>
<td>ARI</td>
<td>0.27530</td>
<td>0.27507</td>
<td>0.27692</td>
</tr>
<tr>
<td>V-Measure</td>
<td>0.55637</td>
<td>0.55330</td>
<td>0.55405</td>
</tr>
<tr>
<td>Q_2</td>
<td>0.73058</td>
<td>0.72968</td>
<td>0.72969</td>
</tr>
</tbody>
</table>

20 Newsgroups: 140 Chains, 750 Samples each
**Enron**

<table>
<thead>
<tr>
<th>Metric</th>
<th>Marginal</th>
<th>MAP</th>
<th>Random</th>
</tr>
</thead>
<tbody>
<tr>
<td>F-Measure</td>
<td>0.35023</td>
<td>0.35100</td>
<td>0.34989</td>
</tr>
<tr>
<td>VI</td>
<td>3.50433</td>
<td>3.49994</td>
<td>3.50624</td>
</tr>
<tr>
<td>ARI</td>
<td>0.13499</td>
<td>0.13564</td>
<td>0.13461</td>
</tr>
<tr>
<td>V-Measure</td>
<td>0.32366</td>
<td>0.32348</td>
<td>0.32322</td>
</tr>
<tr>
<td>Q_2</td>
<td>0.71328</td>
<td>0.71315</td>
<td>0.71309</td>
</tr>
</tbody>
</table>

Enron: 100 Chains, 7300 Samples each

**Del.icio.us**

<table>
<thead>
<tr>
<th>Metric</th>
<th>Marginal</th>
<th>MAP</th>
<th>Rand</th>
</tr>
</thead>
<tbody>
<tr>
<td>F-Measure</td>
<td>0.38591</td>
<td>0.38571</td>
<td>0.38586</td>
</tr>
<tr>
<td>VI</td>
<td>3.74822</td>
<td>3.76847</td>
<td>3.74871</td>
</tr>
<tr>
<td>ARI</td>
<td>0.22893</td>
<td>0.22860</td>
<td>0.22840</td>
</tr>
<tr>
<td>V-Measure</td>
<td>0.47867</td>
<td>0.47850</td>
<td>0.47859</td>
</tr>
<tr>
<td>Q_2</td>
<td>0.72949</td>
<td>0.72938</td>
<td>0.72945</td>
</tr>
</tbody>
</table>

Del.icio.us: 100 Chains, 200 Samples each

### Summarization Conclusions

- The Marginal summarization frequently outperforms the other two but ...
- It turns out that the choice of summarizer doesn’t effect cluster quality (according to metrics) too much
- This suggests that the cheapest summary should be used (Random)
  - Huge time savings
  - Small loss of quality

### “Random” Summaries

- Choose the n<sup>th</sup> sample in each chain as the single-sample summary, sweep n (del.icio.us dataset averaged over 100 chains):

### Qualitative Analysis

- **Datasets:**
  - ~400 documents each
  - 1 data set very non-uniform, to test effects of uniform priors

<table>
<thead>
<tr>
<th>Data set</th>
<th>Natural Classes</th>
<th>Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>del.icio.us</td>
<td>8</td>
<td>Uniform</td>
</tr>
<tr>
<td>Newsgroups</td>
<td>4</td>
<td>Uniform</td>
</tr>
<tr>
<td>Reuters</td>
<td>5</td>
<td>26, 12, 185, 0.8, 34</td>
</tr>
<tr>
<td>Enron</td>
<td>1</td>
<td>Uniform</td>
</tr>
</tbody>
</table>

Table 2.7: Example contingency table showing the relationship of natural classes to clusters produced by the collapsed Gibbs sampler, using the 100<sup>th</sup> sample as the summary for the 20 Newsgroups data.
Qualitative Analysis - Words

Hyper-parameters: Gibbs

Hyper-parameters: EM

Del.icio.us

20 Newsgroups

Enron
**Conclusions**

- Gibbs consistently achieves better quality clusters according to all external metrics.