Lecture 10 - Learning Motif Models with Gibbs sampling

Colin Dewey
February 20, 2008
EM Theory

- Estimate parameters for models with latent (hidden) states
- Model: $X$ (observed), $Z$ (latent), $\Theta$ (params)
- Want to maximize $\log P(X|\Theta)$
- Much easier to maximize $\log P(X,Z|\Theta)$ but don’t know $Z$
- Instead, maximize expected value of $\log P(X,Z|\Theta)$
- Alternate expectation ($Z$) and maximization ($\Theta$) computations
- Theorem: this also maximizes (locally) $\log P(X|\Theta)$
Gibbs Sampling: An Alternative to EM

- a general procedure for sampling from the joint distribution of a set of random variables
- Iteratively sample from

\[ \Pr(X_1, \ldots, X_n) \]

\[ \Pr(X_j | X_1, \ldots, X_{j-1}, X_{j+1} \ldots X_n) \]

for each j

- application to motif finding: Lawrence et al. 1993

- can view it as a stochastic analog of EM for this task

- less susceptible to local minima than EM
Gibbs Sampling Approach

• in the EM approach we maintained a distribution $Z_i$ over the possible motif starting points for each sequence

• in the Gibbs sampling approach, we’ll maintain a specific starting point $a_i$ for each sequence but we’ll keep randomly resampling these
Gibbs Sampling Approach

given: length parameter $W$, training set of sequences

choose random positions for $a$

do

pick a sequence $X_i$

estimate $p$ given current motif positions $a$ \textbf{(update step)}

(using all sequences but $X_i$)

sample a new motif position $a_i$ for $X_i$ \textbf{(sampling step)}

until convergence

return: $p$, $a$
Sampling New Motif Positions

• for each possible starting position, \( a_i = j \), compute a weight

\[
A_j = \frac{\prod_{k=j}^{j+W-1} p_{c_k, k-j+1}}{\prod_{k=j}^{j+W-1} p_{c_k, 0}}
\]

• randomly select a new starting position according to these weights \( a_i \)
The Phase Shift Problem

• Gibbs sampler can get stuck in a local maxima that corresponds to the correct solution shifted by a few bases

• Solution: add a special step to shift the a values by the same amount for all sequences. Try different shift amounts and pick one in proportion to its probability score.
Convergence of Gibbs

![Graph showing the convergence of Gibbs process over iterations. The x-axis represents the number of iterations, and the y-axis represents information per parameter (bits). The graph highlights three distinct phases of convergence, designated as 1, 2, and 3.]
Markov Chain Monte Carlo

• Technique for sampling from probability distribution

• Construct Markov chain with stationary distribution equal to distribution of interest

• Transition probability: \( \tau(y|x) \quad x \rightarrow y \)

• Detailed balance: \( \mathbb{P}(x)\tau(y|x) = \mathbb{P}(y)\tau(x|y) \)

• If detailed balance, then: \( \frac{1}{N} \lim_{N \rightarrow \infty} C(y_i = x) = \mathbb{P}(x) \)
MCMC with Gibbs sampling

- Markov chain transitions by changing one variable at a time
- Transition probability is conditional distribution of the variable given all others
- Show that this obeys detailed balance

\[ \tau(X_{i}^{t+1} | X_{i}^{t}) = \mathbb{P}(X_{i}^{t+1} | X_{1}, \ldots, X_{i-1}, X_{i+1}, \ldots, X_{N}) \]
EM and Gibbs

- these methods are computing a *local, multiple* alignment
- both methods try to optimize the likelihood of the sequences
- EM converges to a local maximum
- Gibbs will converge to a global maximum, *in the limit*
- MEME can take advantage of background knowledge by
  - tying parameters
  - Dirichlet priors
Example: The Data

- Hidden motif of width 7 in 4 sequences of length 10
- Each motif occurrence differs from consensus (GATTACA) in two positions

ACCATGACAG
GAGTATAACCT
CATGCTTACT
CGGAATGCAT
Initialization

• Choose initial positions of motif at random

ACCATGACAG
GAGTATACCT
CATGCTTACT
CGGAATGCA
Predictive update step

- Update profile matrix based on motif and background frequencies and pseudocounts

```plaintext
<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>G</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>T</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>
```

---

Thursday, February 21, 2008
Predictive update step

- Calculate profile matrix from frequencies and pseudocounts

\[
p_{1,A} = \frac{c_{1,A} + b_A}{N - 1 + B} = \frac{0 + 0.5}{4 - 1 + 2} = 0.1
\]
\[
p_{1,C} = \frac{c_{1,C} + b_C}{N - 1 + B} = \frac{0 + 0.5}{4 - 1 + 2} = 0.1
\]
\[
p_{1,G} = \frac{c_{1,G} + b_G}{N - 1 + B} = \frac{2 + 0.5}{4 - 1 + 2} = 0.5
\]
\[
p_{1,T} = \frac{c_{1,T} + b_T}{N - 1 + B} = \frac{1 + 0.5}{4 - 1 + 2} = 0.3
\]
Sampling step

• For each possible motif start position, calculate ratio of likelihood of next W positions from motif vs. background

\[
A_1 = \frac{p_{1,A} \cdot p_{2,C} \cdot p_{3,C} \cdot p_{4,A} \cdot p_{5,T} \cdot p_{6,G} \cdot p_{7,A}}{p_{0,A} \cdot p_{0,C} \cdot p_{0,C} \cdot p_{0,A} \cdot p_{0,T} \cdot p_{0,G} \cdot p_{0,A}} \approx \frac{0.1 \cdot 0.3 \cdot 0.1 \cdot 0.5 \cdot 0.3 \cdot 0.3 \cdot 0.1}{0.31 \cdot 0.23 \cdot 0.23 \cdot 0.31 \cdot 0.23 \cdot 0.23 \cdot 0.31} \approx 0.16
\]

ACCATGA\[\text{CAG}\]
Sampling step

- Sample new position $i$ in chosen sequence based on $A_i$

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_i$</td>
<td>0.16</td>
<td>0.13</td>
<td>0.26</td>
<td>0.017</td>
</tr>
</tbody>
</table>

normalize

draw random sample from distribution

$a_3 = 2$

ACCATGACAG
Calculate likelihood

• Calculate likelihood (or some related value) after each iteration

• Iterate:
  • choose sequence
  • predictive update
  • sample new motif position in sequence

• After many iterations, choose motif positions and corresponding profile matrix
Inferring *cis* Regulatory Modules (CRMs)

Arrangement of these binding sites forms a *cis*-regulatory module.

A task of growing interest: infer models of CRMs that regulate certain sets of genes.
A Representation for CRMs
[Noto & Craven]

1. Multiple Binding Sites
   a collection of cooperative transcription factor binding sites

2. Multiple Motifs per Binding Site

3. Distance Constraints
   upper-bounds on the distance between binding sites

4. Strand Constraints

5. Order Constraints

6. Repressor Motifs
   binding of factors that deactivate a CRM