Gibbs Sampler

Jialei Wang Peter I. Frazier

July 8, 2013

1 Gibbs Sampler

In this section, we consider classification problems in which nature only gives us positive labels. This occurs in bioinformatics because our data comes from phage display, a combinatorial chemistry experiment using split pool, or from organisms evolved via natural selection. It occurs in other applications as well, e.g., in the arxiv, where we only observe the papers that a user looked at, and not the larger set of papers in which a user is interested.

To address this problem, we propose a probabilistic model, and then propose a Gibbs sampler to do inference within the context of this probabilistic model.

1.1 the Model

Sample a peptide pool $W_1, \ldots, W_M$ i.i.d. from the peptide library with some known distribution F_W, and M is fixed. Then we observe peptides $X_1, \ldots, X_N$. Because $Y(X_n, 1) = 1 \, \forall n \in \{1, \ldots, N\}$, we can treat $X_1, \ldots, X_N$ as being drawn independently with replacement from $\{W_m : Y(W_m, 1) = 1\}$. $Y(W_m, 1)$ is drawn from Bernoulli distribution with parameter $f(\theta, W_m)$, where $\theta \sim F_\theta$ and f is some known function.

1.2 Gibbs Sampler

First we define some variables:

- $Y_m = Y(W_m, 1)$
- $(X_n)_{n=1}^N$ are tested peptides and clearly $Y(X_n, 1) = 1$.
- Z_n is index of W_m that was chosen to make X_n.

The conditional distributions for this model are

1. $\theta | (Y_m)_{m=1}^M, (W_m)_{m=1}^M, (X_n)_{n=1}^N, M, (Z_n)_{n=1}^N$ is sampling from the posterior distribution estimated in (**??**).
2. $$Y_m | \theta, (W_m)_{m=1}^M, (X_n)_{n=1}^N, M, (Z_n)_{n=1}^N = \begin{cases} 1 & \text{if } Z_n = m \ , \\ \text{sample from (??) given } \theta, W_m & \text{o/w.} \end{cases}$$

3.

$$W_m|(Y_m)_{m=1}^M, \theta, (X_n)_{n=1}^N, M, (Z_n)_{n=1}^N = \begin{cases} X_n & \text{if } Z_n = m \ , \\ \text{sample from (??) or (??) given } \theta, Y_m & \text{o/w.} \end{cases}$$

4. $Z_n|(Y_m)_{m=1}^M, \theta, (W_m)_{m=1}^M, (X_n)_{n=1}^N, M$ is chosen uniformly among $\{m : W_m = X_n\}$.

Initialization: we fix M and set $Z_n = n$. Then

$$W_m = \begin{cases} X_n & \text{if } Z_n = m \ , \\ \text{sample from } F_W & \text{o/w.} \end{cases}$$

Then draw θ from prior distribution and set $(Y_m)_{m=1}^M$ according to step 2 above.

1.3 Algorithm

Require: inputs M, $(X_n)_{n=1}^N$ and prior parameter of Dirichlet distribution $\boldsymbol{\alpha} = (\alpha(1), \ldots, \alpha(6))$

for $n = 1$ to N **do**
  $Z_n \leftarrow n$
end for
for $m = 1$ to M **do**
  if $Z_n = m$ **then**
    $W_m \leftarrow X_n$
    $Y_m \leftarrow 1$
  else
    $W_m \leftarrow$ sample from F_W
    $Y_m \leftarrow$ sample from (??) given θ, W_m
  end if
end for
loop
  $\theta \leftarrow$ sample from the posterior distribution in (??) given $(Y_m)_{m=1}^M, (W_m)_{m=1}^M, (X_n)_{n=1}^N, M, (Z_n)_{n=1}^N$
  for $m = 1$ to M **do**
    if $m = (Z_n)_{n=1}^N$ **then**
      $Y_m \leftarrow 1$
    else
      $Y_m \leftarrow$ sample from posterior distribution in (??) given $\theta, (W_m)_{m=1}^M, (X_n)_{n=1}^N, M, (Z_n)_{n=1}^N$
    end if
  end for
  for $m = 1$ to M **do**
    if $m = (Z_n)_{n=1}^N$ **then**
      $W_m \leftarrow X_n$
    else
      $W_m \leftarrow$ sample from (??) or (??) given θ, Y_m
    end if
  end for

29: **for** $n = 1$ to N **do**
30: $Z_n \leftarrow$ sample from $\{m : W_m = X_n\}$ uniformly
31: **end for**
32: **end loop**

1.4 Proposed Data to test the algorithm

To test the algorithm, let's use simulated data to start. Then, we can use some standard data sets (the UCI machine learning repository). Then, in addition to testing on chemistry applications, we may be able to do this in the context of the arxiv:

- Consider the set of users whose number of papers viewed exceeds a threshold.
- Take the set of papers at which that the user looked (obtained via the mysql database on whale) as our list of hits. Take the (much larger) set of papers that were available for viewing as our list of items in the pool from which the hits were drawn. Use features obtained from the Lucene index (Xiaoting can show us how).
- As a test, we can do an offline experiment where we take papers from a time period not in our training set, and predict for each paper whether the user will be interested or not. Then we can look at the number of times we correctly predicted a hit. This will underestimate the number of true hits. To do this test more precisely, we can take the list of papers on which we do our testing from the "new" or "recent" pages that we know that the user visited. This information is also available in the mysql database.
- In the offline experiment, we could compare to the method that counts every item not viewed by the user as a negative, and uses standard naive bayes classification. We could also try to improve this baseline, e.g., by using a different prior. We need to look in the literature to find what other techniques people might use.

We may be able to do a more sophisticated model where we incorporate the probability that a paper was seen.