CSCI 2950-P Homework 4: Gibbs Samplers for Stochastic Block Models of Relational Data

Brown University, Spring 2013

Homework due at 11:59pm on April 26, 2013

In this problem set, we develop Markov chain Monte Carlo (MCMC) learning algorithms for a family of relational models, which are commonly used to model biological or social networks. Consider a set of N nodes, representing entities to be modeled. For node pairs $i \neq j$, we let $y_{ij} = 1$ if there is a relationship (e.g., friendship) from entity i to entity j, $i \neq j$, and $y_{ij} = 0$ otherwise. We focus on directed graphs, for which there is a distinct binary relationship y_{ji} from entity j to entity i. For some experiments, the y_{ij} variables will only be observed for a (known) subset of ordered entity pairs (i, j). All of your derivations and code should support such partial observations.

We model such relational data as being generated from K unknown latent communities, indexed by integers $k = 1, \ldots, K$. For each pair of communities k, ℓ, we define an interaction probability $W_{k\ell}$, and place a uniform prior distribution:

$$W_{k\ell} \sim \text{Beta}(1, 1), \qquad 1 \leq k, \ell \leq K.$$

We will then use Gibbs samplers to learn this non-symmetric $K \times K$ matrix of community interaction probabilities, as well as community memberships for individual entities.

The *stochastic block models* we consider are closely related to probabilistic mixture models and the latent Dirichlet allocation (LDA) topic model, and you may find it helpful to review Gibbs samplers for those models. Also remember that the Dirichlet distribution is conjugate to multinomial likelihoods, and the beta distribution to Bernoulli likelihoods.

Question 1: Stochastic Block Models

The basic stochastic block model assumes each entity i is a member of a single latent community z_i, sampled according to

$$z_i \sim \text{Cat}(\pi), \qquad i = 1, \ldots, N.$$

The K-dimensional distribution of community frequencies π has a symmetric Dirichlet prior:

$$\pi \sim \text{Dir}(\alpha, \ldots, \alpha).$$

For all experiments below, assume $\alpha = 1$. For any pair of nodes $i \neq j$, their relationship link variables are generated according to

$$p(y_{ij} = 1 \mid z_i = k, z_j = \ell, W) = W_{k\ell}, \quad p(y_{ji} = 1 \mid z_i = k, z_j = \ell, W) = W_{\ell k}.$$

Remember that we may only have partial observations of these link variables.

a) *Given fixed parameters π, W, and observations y for some subset of ordered entity pairs, derive a formula for the posterior distribution $p(z_i \mid z_{\setminus i}, y, \pi, W)$. Here, $z_{\setminus i}$ denotes the community assignments for all entities except node i.*

b) *Given fixed entity assignments z, derive formulas for the posterior distributions of the model parameters, $p(\pi \mid y, z, W)$ and $p(W \mid y, z, \pi)$. These should be members of some standard exponential family of distributions.*

c) *Implement a Gibbs sampler using the formulas from parts (a,b). Each iteration should resample each of the z, W, π variables once. Initialize by sampling the parameters π, W from their prior distributions, and z from its corresponding posterior. Also derive a formula for, and compute at each iteration, the log-probability $\log p(y, z, W, \pi)$.*

d) *To test your sampler, construct a synthetic dataset with $N = 30$ entities, 10 in each of $K = 3$ communities. Generate links y_{ij} by assuming a within-community link probability of $W_{kk} = 0.95$, and a between-community link probability of $W_{k\ell} = 0.10, \ell \neq k$. Then given only the observed links y, explore how accurately the sampler recovers the underlying W, z variables. Run your sampler for 500 iterations from each of 5 random initializations, and plot the resulting log-likelihood curves on a single set of axes. After each iteration, compute the Rand index (see provided script) between the true assignments and the sampled z, and plot these scores versus iteration.*

e) *Apply the stochastic block model to the Sampson monk data. Allow $K = 3$ communities, run the sampler for 500 iterations from each of 5 random initializations, and again plot log-likelihoods and Rand indexes (from the true faction labels) versus iteration.*

Question 2: Mixed Membership Stochastic Block Models

The mixed membership stochastic block model generalizes basic block models by allowing each entity to participate in multiple communities. We begin by sampling a K-dimensional community membership distribution for each entity:

$$\pi_i \sim \text{Dir}(\alpha, \ldots, \alpha), \qquad i = 1, \ldots, N.$$

For all experiments below, assume $\alpha = 1$. Observation y_{ij} is then determined by link-specific source and receiver community assignments, s_{ij} and r_{ij}, sampled as follows:

$$p(y_{ij} = 1 \mid r_{ij} = k, s_{ij} = \ell, W) = W_{k\ell}, \quad r_{ij} \sim \text{Cat}(\pi_i), \quad s_{ij} \sim \text{Cat}(\pi_j).$$

Intuitively, s_{ij}, r_{ij} are the communities which “explain” the interactions of node i with node j.

a) *Given fixed parameters π, W, and observations y for some subset of ordered entity pairs, derive a formula for the posterior distributions $p(r \mid s, y, \pi, W)$ and $p(s \mid r, y, \pi, W)$. Are the indicators r_{ij}, s_{ij} for different links conditionally independent?*

b) *Given fixed link assignments r, s, derive formulas for the posterior distributions of the model parameters, $p(\pi \mid y, r, s, W)$ and $p(W \mid y, r, s, \pi)$. These should be members of some standard exponential family of distributions.*

c) *Implement a Gibbs sampler using the formulas from parts (a,b). Each iteration should resample each of the r, s, W, π variables once. Initialize by sampling the parameters π, W from their prior distributions. Also derive a formula for, and compute at each iteration, the log-probability $\log p(y, r, s, W, \pi)$.*

d) *Apply your sampler to the synthetic data from part 1(d), assuming $K = 3$, running for 500 iterations from each of 5 initializations, and plotting log-likelihood versus iteration. Looking at the result from the highest-likelihood Markov chain, does the mixed membership model provide a reasonable interpretation of this data?*

e) *Apply your sampler to the Sampson monk data, assuming $K = 3$, running for 500 iterations from each of 5 initializations, and plotting log-likelihood versus iteration. Looking at the result from the highest-likelihood Markov chain, how many monks have significant membership in more than one community?*

f) *The basic sampler outlined above may mix slowly, due to correlations between the source and receiver link variables. To address this, derive a formula for the joint posterior distribution $p(r, s \mid y, \pi, W)$. Are the paired indicators (r_{ij}, s_{ij}) for different links conditionally independent?* Hint: *You should be able to sample from this conditional distribution by considering appropriate K^2-dimensional categorical distributions.*

g) *Repeat the experiments from parts (d,e) using the blocked sampler, including creation of log-likelihood plots. Are there substantial performance differences between the samplers?*

Question 3: Link Prediction

In this question, we use our mixed membership models to predict the presence of likely links in a partially sampled social network. We focus on a network describing advice relationships among $N = 71$ attorneys in a New England law firm.

a) *Write code which randomly subsamples half of the y_{ij} link variables to use for training, and reserves the other half for testing. Note that the training data is half of the directed node pairs (potential links),* not *half of the actually present links.*

b) *Using a single train-test split of the attorney network, apply your Gibbs sampler from problem 1, as well as the blocked Gibbs sampler from problem 2. For both models assume $K = 6$, run the sampler for 500 iterations from each of 3 random initializations, and reserve the final sample from the most probable Markov chain.*

c) *For the standard block model, compute the posterior distribution of each of the test link variables y_{ij}, given the parameters z, π, W learned during training. Use the held-out test labels to create an ROC curve summarizing classification performance.*

d) *For the mixed membership block model, compute the posterior distribution of each of the test link variables y_{ij}, given the parameters π, W learned during training and marginalizing r_{ij}, s_{ij}. Use the held-out test labels to create an ROC curve summarizing classification performance, and compare to part (c).*

e) *With more computational effort, is there a more sophisticated way you could predict test link variables based on the output of your Gibbs sampling algorithms?*