CSC 2541, Small exercise #3, due in class February 7, worth 5% of the mark

In this exercise you are to implement Gibbs sampling for a Bayesian mixture model with conjugate priors. Due to conjugacy, you can integrate out all the model parameters, so MCMC is done on a state space of just the indicators of which mixture component each observation comes from.

The data (available from the web page) consists of $n = 100$ independent observations of $p = 10$ binary variables, with each observation having the same mixture distribution with K components, in which the 10 binary variables are independent given the mixture component.

Writing $y_{i,j}$ for the value of the jth variable in the ith observation, the model is

$$\begin{aligned} y_{i,j} \mid c_i,\, \phi &\sim \text{Bernoulli}(\phi_{c_i,j}) \\ c_i \mid \rho_1, \ldots, \rho_K &\sim \text{Discrete}\,(\rho_1, \ldots, \rho_K) \\ \phi_{c,j} &\sim \text{Beta}(\beta/2, \beta/2) \\ \rho_1, \ldots, \rho_K &\sim \text{Dirichlet}\,(\alpha/K, \ldots, \alpha/K) \end{aligned}$$

Here, α and β are known constants. The Bernoulli(p) distribution is that of a binary variable with probability p for the value 1.

You should be able to integrate out ϕ and ρ, and hence get simple forms for the conditional distribution of c_i given the other $c_{i'}$, and for the conditional distributon of $y_{i,j}$ given c_i and all the other $c_{i'}$ and $y_{i',j}$. These should allow you to do Gibbs sampling on the space of just the component indicators, c_i for $i = 1, \ldots, n$.

You should try three variations of this model, with $K = 5$ and $\alpha = 1$, with $K = 20$ and $\alpha = 1$, and with $K = 20$ and $\alpha = 10$. You should fix $\beta = 1$ for all of these.

For each model variation, you should run your Gibbs sampler for 500 iterations (each iteration being a scan of all 100 component indicators), starting from the state where all observations are associated with component 1.

To see what each of these Gibbs sampling runs did, you should produce a plot with iteration number (1 to 500) on the horizontal axis, and with the vertical axis being the fraction of observations that are associated with the k most common mixture components. You should plot a line for each k from 1 to K. (In other words, you plot a line showing how the fraction of observations associated with the most common mixture component varies over the 500 MCMC iterations, plus one showing the fraction of observations associated with the two most common mixture components, etc.) Note that the most common mixture component may change from one iteration to the next.

You should hand in your program code and these three plots, and comment briefly on how rapidly the Gibbs sampler converges, and how well it samples after convergence, and on the differences you see among the three model variations.