Shane T. Jensen
May 20, 2009

Correction to Gibbs Sampler with Uniform Process Prior?

I will try to stick to with the notation of Zhu et.al. where they use $\boldsymbol{d} = (d_1, \ldots, d_n)$ to denote the observed data (e.g. documents) and $\boldsymbol{t} = (t_1, \ldots, t_n)$ to denote the ordering/time of these documents. We also have a set of true clusters $\boldsymbol{s} = (s_1, \ldots, s_n)$ for each data point, and we have a specified distribution $p(d_i|s_i)$ that models the observed data as a function of these unobserved clusters. I'm not sure what form this takes for your topic model, so I'll keep it as a general density $p(d_i|s_i)$. Our goal is to estimate the posterior distribution of the clusters,

$$p(\boldsymbol{s}|\boldsymbol{d}) \quad \propto \quad p(\boldsymbol{s}) \times \prod_{i=1}^{n} p(d_i|s_i)$$

Obviously, our focus is on the prior distribution of the cluster memberships $p(\boldsymbol{s})$. Zhu et.al. consider a general constructive framework,

$$p(s_i = j|s_1, \ldots, s_{i-1}) = \begin{cases} \frac{w(t_i,j)}{\sum_{j' \neq j} w(t_i,j') + \alpha} & \text{for current cluster } j \\ \frac{\alpha}{\sum_{j' \neq j} w(t_i,j') + \alpha} & \text{for new cluster } j \end{cases}$$

The weight $w(t_i, j)$ is a function of the cluster memberships of previous observations assigned to cluster j:

$$w(t_i, j) = \sum_{k:t_k<t_i} k(t_i - t_k) \cdot \text{Ind}(s_k = j)$$

where $\text{Ind}(s_k = j)$ is an indicator function and $k(\cdot)$ can be an arbitrary function of the previous observations. Of course, we are really only considering two particular choices of $w(t_i, j)$. For the Dirichlet process, we have

$$w(t_i, j|\text{DP}) = \sum_{k:t_k<t_i} \text{Ind}(s_k = j)$$

For the Uniform process, we have

$$w(t_i, j|\text{UN}) = 1$$

The Gibbs sampling implementation of these models involves iteratively removing each observation i from its current cluster, and re-assigning it to a cluster based on the probability

$$p(s_i = j|\boldsymbol{s}_{-i}, \boldsymbol{d}) \quad \propto \quad p(s_i = j|\boldsymbol{s}_{-i}) \times \prod_{i=1}^{n} p(d_i|\boldsymbol{d}^j_{-i})$$

where $\boldsymbol{d}^j_{-i}$ is the set of observations in cluster j excluding observation i. The crucial component of this calculation is $p(s_i = j|\boldsymbol{s}_{-i})$,

$$p(s_i = j|\boldsymbol{s}_{-i}) \quad \propto \quad p(s_i = j, \boldsymbol{s}_{-i})$$

which is given in the Zhu et.al. paper as

$$\begin{aligned} p(s_i = j|\boldsymbol{s}_{-i}) \quad &\propto \quad \prod_{m=1}^{i-1} p(s_m|s_1, \ldots, s_{m-1}) \cdot p(s_i = j|s_1, \ldots, s_{i-1}) \cdot \prod_{m=i+1}^{n} p(s_m|s_1, \ldots, s_{m-1}) \\ &\propto \quad p(s_i = j|s_1, \ldots, s_{i-1}) \cdot \prod_{m=i+1}^{n} p(s_m|s_1, \ldots, s_{m-1}) \end{aligned}$$

However, the Zhu et.al. paper has a different model than our situation in the sense that they have a real time-ordering that implies that s_i follows $s_1, \ldots, s_{i-1}$ and precedes $s_{i+1}, \ldots, s_n$. However, since the uniform process is not exchangable across orderings, we will also condition on a particular ordering $s_1, \ldots, s_n$ for our cluster memberships as well. So, we must factor in the subsequent observations in our probability calculations,

$$p(s_i = j|\boldsymbol{s}_{-i}) \quad \propto \quad p(s_i = j|s_1, \ldots, s_{i-1}) \cdot \prod_{m=i+1}^{n} p(s_m|s_1, \ldots, s_{m-1})$$

Consider the Gibbs move for s_i. First, let $K^\star$ be the total number of clusters formed by the first $i-1$ indicators $(s_1, \ldots, s_{i-1})$. Then,

$$p(s_i = j|s_1, \ldots, s_{i-1}) = \begin{cases} \frac{1}{K^\star+\alpha} & \text{for current cluster } j \\ \frac{\alpha}{K^\star+\alpha} & \text{for new cluster } j \end{cases}$$

Now, we focus on the set of indicators $(s_{i+1}, \ldots, s_n)$ that follow s_i. Let $n^\star$ be the number of decisions among the indicators $(s_{i+1}, \ldots, s_n)$ to form a new cluster. Treating $(s_{i+1}, \ldots, s_n)$ as a sequence, let q_1 be the number of indicators starting from s_{i+1} up until the first $s_m = new$ decision to form a new cluster. Let q_2 equal the number of indicators after that $s_m = new$ decision until the next decision $s'_m = new$ decision to form a new cluster, etc. By the time we get to the last indicator s_n, we have a set of waiting times $q_1, q_2, \ldots, q_{n^\star}$ between decisions to form a new cluster. With this notation, we are finally ready to calculate

$$\prod_{m=i+1}^{n} p(s_m|s_1, \ldots, s_{m-1}) = \begin{cases} \prod_{k=1}^{n^\star} \left(\frac{\alpha}{K^\star+\alpha+k}\right)^{q_k} & \text{if } s_i \text{ joins current cluster } j \\ \prod_{k=1}^{n^\star} \left(\frac{\alpha}{K^\star+\alpha+k+1}\right)^{q_k} & \text{if } s_i \text{ forms new cluster } j \end{cases}$$

which gives us finally that

$$p(s_i = j|\boldsymbol{s}_{-i}) = \begin{cases} \frac{1}{K^\star+\alpha} \times \prod_{k=1}^{n^\star} \left(\frac{\alpha}{K^\star+\alpha+k}\right)^{q_k} & \text{if } s_i \text{ joins current cluster } j \\ \frac{\alpha}{K^\star+\alpha} \times \prod_{k=1}^{n^\star} \left(\frac{\alpha}{K^\star+\alpha+k+1}\right)^{q_k} & \text{if } s_i \text{ forms new cluster } j \end{cases}$$