Notes on the Discrete Hierarchical Dirichlet Process

David Pfau

12 Movember 2009

We consider the case of a Hierarchical Dirichlet Process (HDP) mixture with an arbitrary tree depth, and show that inference is considerably simpler in the case of discrete data. Rather than data coming from j separate groups, as in the basic HDP mixture model, consider the different groups of data as coming from the leaves of a tree, where the root of this tree is a DP over the base distribution H, and each node of the tree is a DP with base distribution given by the DP of the parent node. Denote a node in the tree by an index vector $\mathbf{u}$ whose length corresponds to the distance from the root of that node. Leaf nodes have index vectors of length N. To avoid confusion, denote indices at the leaf node $\mathbf{v}$. Our generative model is as follows:

$$\begin{aligned} x_{\mathbf{v}i}|\theta_{\mathbf{v}i} &\sim F(\theta_{\mathbf{v}i}) \\ \theta_{\mathbf{v}i} &\sim G_{\mathbf{v}} \\ G_{\mathbf{u}} &\sim DP(\alpha_{|\mathbf{u}|}, G_{\pi(\mathbf{u})}) \\ G_{[]} &\sim DP(\gamma, H) \end{aligned} \tag{1}$$

For pedagogical clarity, rather than computational efficiency, we describe inference when all parameters are instantiated (this is equivalent to the hierarchical version of algorithm 1 in Neal (1998)). Inference in this scheme will involve some complicated bookkeeping. At each node in our tree we have a set of observations, which correspond to the active clusters in the node below, except in the leaf nodes, where they correspond directly to data. To use the parlance of the Chinese Restaurant Franchise, each customer in a restaurant above the leaf level corresponds to a table in some restaurant beneath it. Thus in each restaurant $\mathbf{u}$ we have two sets of parameters to keep track of: the $\theta_{\mathbf{u}j}$ which are customers and the $\psi_{\mathbf{u}k}$ which are the dishes being served at the tables at which we seat the customers, i.e., the set of active parameters to which the lower-level parameters may be assigned. There is a direct mapping $g(\mathbf{u}, k) = j$ that maps the index of a table in $\mathbf{u}$ to a customer in its parent restaurant $\pi(\mathbf{u})$, thus $\psi_{\mathbf{u}k} = \theta_{\pi(\mathbf{u})g(\mathbf{u},k)}$. The assignments of one θ to one ψ is constantly resampled during inference, but the assignment of ψ to θ in the restaurant above it is constant (save creation/deletion when tables are added or removed) - it's simply two different ways of indexing the same thing.

Just as we have two sets of parameters to keep track of in each restaurant, we have two counts to keep track of: the number of customers and number of tables in a particular restaurant. Let $c_{\mathbf{u}k}$ denote the number of customers at table k while $c_{\mathbf{u}}$ gives the total number of customers in a restaurant. Similarly let $t_{\mathbf{u}}$ be the total number of tables in restaurant $\mathbf{u}$. Then $c_{\mathbf{u}} = \sum_{\mathbf{u}' \, s.t. \, \pi(\mathbf{u}')=\mathbf{u}} t_{\mathbf{u}'}$. To keep everyone from pulling their hair out, we leave out the additional index for the value of the parameter at table k, instead using an indicator variable $\delta_{\psi_{\mathbf{u}k}}$ if we are interested in the particular value of a parameter at some table. If we wish to refer to the counts with a particular element removed, we add a superscript with the index of that element.

To construct a posterior Gibbs sampler, note that the posterior over a single $\theta_{\mathbf{u}j}$ factors as

$$\begin{aligned} p(\theta_{\mathbf{u}j}|\theta_{\mathbf{u}-j}, \alpha_{|\mathbf{u}|}, G_{\pi(\mathbf{u})}, \mathbf{x}) &\propto p(x_{\mathbf{v}i} \, s.t. \, x_{\mathbf{v}i} \to \theta_{\mathbf{u}j}|\theta_{\mathbf{u}j}) p(\theta_{\mathbf{u}j}|\theta_{\mathbf{u}-j}, \alpha_{|\mathbf{u}|}, G_{\pi(\mathbf{u})}) \\ &= \prod_{\mathbf{v}i \, s.t. \, x_{\mathbf{v}i} \to \theta_{\mathbf{u}j}} F(x_{\mathbf{v}i}|\theta_{\mathbf{u}j}) p(\theta_{\mathbf{u}j}|\theta_{\mathbf{u}-j}, \alpha_{|\mathbf{u}|}, G_{\pi(\mathbf{u})}) \end{aligned} \tag{2}$$

Where $x_{\mathbf{v}i} \to \theta_{\mathbf{u}j}$ denotes all the data points that are assigned to $\theta_{\mathbf{u}j}$. The prior given all other $\theta_{\mathbf{u}-j}$ is given by the familiar Blackwell-MacQueen distribution for the posterior predictive with $G_{\mathbf{u}}$ integrated out:

$$\theta_{\mathbf{u}j}|\theta_{\mathbf{u}-j}, \alpha_{|\mathbf{u}|}, G_{\pi(\mathbf{u})} \sim \sum_{k=1}^{t_{\mathbf{u}}} \frac{c_{\mathbf{u}k}^{-j}}{c_{\mathbf{u}}^{-j} + \alpha_{|\mathbf{u}|}} \delta_{\psi_{\mathbf{u}k}} + \frac{\alpha_{|\mathbf{u}|}}{c_{\mathbf{u}}^{-j} + \alpha_{|\mathbf{u}|}} G_{\pi(\mathbf{u})} \tag{3}$$

To sample from this, we have to integrate out $G_{\pi(\mathbf{u})}$ as well, and so on up the tree, such that sampling $\theta_{\mathbf{u}j}$ depends on all $\theta_{\mathbf{u}'j'}$ in nodes above it. We can write this out in simple recursive form. Our base case is the root of the HDP, $\mathbf{u} = []$. Thus the posterior of $\theta_{[]j}$ has the form

$$p(\theta_{[]j} = \psi|\theta_{-[]j}, x_{\mathbf{v}i} \to \theta_{[]j}) \propto \begin{cases} c_{[]k}^{-j} \prod_{\mathbf{v}i} F(x_{\mathbf{v}i}|\psi_{[]k}) & \text{if } \psi = \psi_{[]k} \\ \gamma p(\theta_{[]j}|H, x_{\mathbf{v}i}) & \text{if } \psi \text{ is new} \end{cases} \tag{4}$$

Note that this is valid given *all* θ other than $\theta_{[]j}$, not only those in the same "restaurant". For an arbitrary node in the tree, the posterior is given by

$$p(\theta_{\mathbf{u}j} = \psi|\theta_{-\mathbf{u}j}, x_{\mathbf{v}i} \to \theta_{\mathbf{u}j}) \propto \begin{cases} c_{\mathbf{u}k}^{-j} \prod_{\mathbf{v}i} F(x_{\mathbf{v}i}|\psi_{\mathbf{u}k}) & \text{if } \psi = \psi_{\mathbf{u}k} \\ \alpha_{|\mathbf{u}|} p(\theta_{\pi(\mathbf{u})c_{\pi(\mathbf{u})}+1} = \psi|\theta_{-\mathbf{u}j}, x_{\mathbf{v}i}) & \text{if } \psi \text{ is new} \end{cases} \tag{5}$$

In other words, the probability of assigning $\theta_{\mathbf{u}j}$ to a new value ψ not represented in restaurant $\mathbf{u}$ is proportional to $\alpha_{|\mathbf{u}|}$ times the probability of assigning a new customer in restaurant $\pi(\mathbf{u})$ the value ψ if it is associated with the same data as $\theta_{\mathbf{u}j}$. In the case of

discrete data, there is only one possible assignment of parameters to a data point: the value of that data point itself. Thus the likelihood becomes trivial: $F(x_{\mathbf{v}i}|\theta_{\mathbf{u}j}) = \delta_{\theta_{\mathbf{u}j}}(x_{\mathbf{v}i})$ and the products above reduce to indicator variables. Thus as long as we pass the value of our data point up the tree at each step of inference, inference at one level becomes completely independent of the levels of the tree beneath it, and our posterior update becomes

$$p(\theta_{\mathbf{u}j} = \psi|\theta_{-\mathbf{u}j}) \propto \begin{cases} c_{\mathbf{u}k}^{-j}\delta_{\psi_{\mathbf{u}k}}(\theta_{\mathbf{u}j}) & \text{if } \psi = \psi_{\mathbf{u}k} \\ \alpha_{|\mathbf{u}|}p(\theta_{\pi(\mathbf{u})c_{\pi(\mathbf{u})}+1} = \psi|\theta_{-\mathbf{u}j}) & \text{if } \psi \text{ is new} \end{cases} \tag{6}$$

With the obvious extension to the base case.