CS281B/Stat241B: Advanced Topics in Learning & Decision Making

Dirichlet Process Mixture Models (cont'd)

Lecturer: Michael I. Jordan *Scribes: Kofi A. Boakye*

1 DP Mixture Models

1.1 Infinite Mixture Model

Given data $x_i, \ldots, x_n$ which are regarded as exchangeable, we model the distribution from which the x_i are drawn as a mixture of distributions of the form $F(\theta)$. We represent the mixing distribution over θ as G and set the prior for this mixing distribution to be a Dirichlet process with concentration parameter α and base distribution G_0. This is compactly represented in the graphical model below:

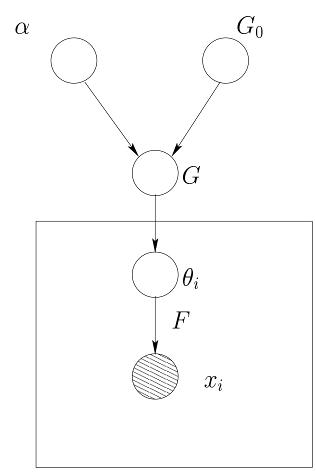

Figure 1: Graphical Model for DP Mixture Model.

From the model we obtain the following relations:

$$\begin{aligned} x_i \mid \theta_i &\sim F(\theta_i) \\ \theta_i \mid G &\sim G \\ G &\sim DP(\alpha, G_0) \end{aligned} \tag{1}$$

As realizations of the DP are discrete with probability one, one observes clustering of the data as in figure 2. In general for N data points there will be $logN$ ϕ_i's, where $\{\phi\}$ is the set of distinct θ's. Another consequence of the discrete nature of the DP is that the mixture model can be viewed as countably infinite mixtures. We can also see this by integrating over G in the model given in figure 3: From this we get:

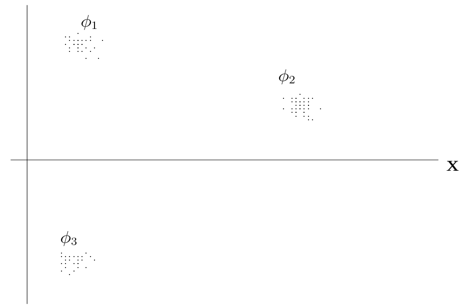

Figure 2: Realization of DP mixture.

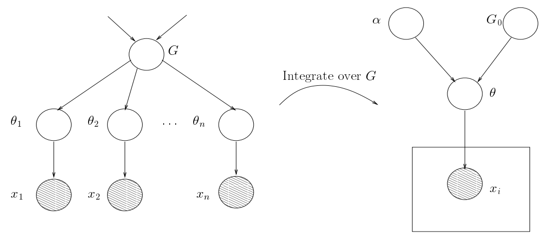

Figure 3: Integration over latent variable G.

$$\theta_n \mid \theta_1, \ldots, \theta_{n-1} \;\sim\; \frac{1}{n-1+\alpha} \sum_{j<n} \delta(\theta_j) + \frac{\alpha}{n-1+\alpha} G_0 \tag{2}$$

where $\delta(\theta)$ is the distribution concentrated at a single point θ.

1.2 Finite Mixture Model

An alternative method of arriving at the DP mixture is by taking the limit of finite mixtures of the following form:

$$\begin{aligned} x_i \mid c, \phi \;&\sim\; F(\phi_{c_i}) \\ c_i \mid p \;&\sim\; Mult(p_1, \ldots, p_k) \\ p \;&\sim\; Dir(\alpha/K, \ldots, \alpha/K) \\ \phi_c \;&\sim\; G_0 \end{aligned} \tag{3}$$

The graphical model is shown in figure 4. c_i indicates which class is associated with a given observation x_i. The parameters ϕ_c determine the observation distribution for the class and p_c are the mixing proportions, which are given a symmetric Dirichlet prior with concentration parameter α/K so as to have it approach zero as K goes to infinity.

If we integrate over the mixing proportions, we get the following for the prior for c_i:

$$P(c_i = c \mid c_1, \ldots, c_{i-1}) \;=\; \frac{n_{i,c} + \alpha/K}{i-1+\alpha} \tag{4}$$

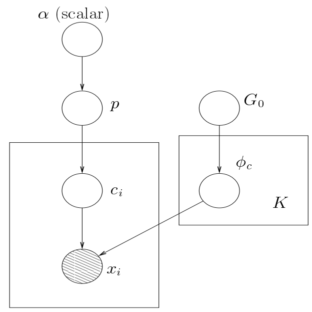

Figure 4: Finite mixture model of DP mixture.

where $n_{i,c}$ is the number of c_j for $j < i$ that are equal to c. To get from this finite mixture model to the DP we now let $K \to \infty$:

$$\begin{aligned} P(c_i \;=\; c \mid c_1, \ldots, c_{i-1}) \;&\to\; \frac{n_{i,c}}{i-1+\alpha} \\ P(c_i \;\neq\; c_j \;\forall\; j < i \mid c_1, \ldots, c_{i-1}) \;&\to\; \frac{\alpha}{i-1+\alpha} \end{aligned} \tag{5}$$

Note the correspondence between the conditional probabilities for θ_i in (2) and those implied by (5).

2 Gibbs sampling for the case of conjugate priors

Algorithm 1
State $(\theta_, \ldots, \theta_n)$

- For $i = 1, \ldots, n$
  Draw $\theta_i \mid \theta_{-i}, x_i \;\sim\; \sum_{j \neq i} q_{ij} \delta(\theta_j) + r_i H_i$
  where
  $q_{ij} \;=\; bF(x_i, \theta_j)$
  $r_i \;=\; b\alpha \int F(x_i, \theta) \; dG_0(\theta)$
  with b such that $\sum_{j \neq i} q_{ij} + r_i \;=\; 1$ and H_i being the posterior distribution for θ based on the prior G_0 and observation x_i

For this algorithm convergence to the posterior distribution may be slow. The reason is that the algorithm cannot change the θ for more than one observation at a time, though there are often groups of observations associated with the same θ; it would be preferable to move data points *en masse*. This issue is avoided in the following algorithm, which does Gibbs sampling on the model shown below:

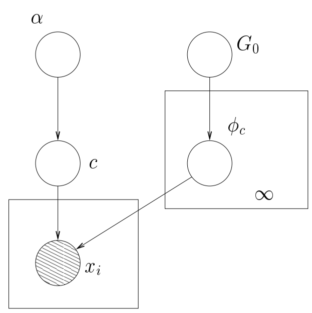

Figure 5: Model for Gibbs sampling for Algorithm 2.

Algorithm 2
State$(c_1, \ldots, c_n, \phi_1, \ldots)$

- For $i = 1, \ldots, n$
  If $c = c_j$ for some $j \neq i$
  $p(c_i = c \mid c_{-i}, x_i, \phi) = b\frac{n_{-i,c}}{n-1+\alpha} \int F(x_i, \phi) dH_{-i,c}(\phi)$
  $p(c_i \neq c_j \; \forall \; j \neq i \mid c_{-i}, x_i) = b\frac{\alpha}{n-1+\alpha} \int F(x_i, \phi) dG_0(\phi)$
  where $H_{-i,c}$ is the posterior distribution of ϕ based on the prior G_0 and all observations for which $j \neq i$ and $c_j = c$