Nonparametric Bayesian Modeling (Contd)

Piyush Rai

Topics in Probabilistic Modeling and Inference (CS698X)

March 27, 2019

Recap: A Nonparametric Bayesian Mixture Model

A brief sketch of a basic Gibbs sampler (samples $\mathbf{Z}$ and $\{\phi_k\}_{k=1}^K$) for this model with unbounded K (note: The mixing proportions π_k's were collapsed from the prior $p(\boldsymbol{z}_i|\pi)$)

Gibbs Sampler for NPBayes Mixture Model

- Set an initial K. Initialize $\mathbf{Z}^{(0)}$ and $\{\phi_k^{(0)}\}_{k=1}^K$
- For $t = 1, \ldots, T$
  - For each observation $i = 1, \ldots, N$, sample the cluster id $\boldsymbol{z}_i$

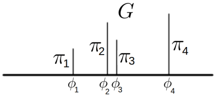
$$\begin{aligned}
p(\boldsymbol{z}_i = k|\mathbf{Z}_{-i}^{(t-1)}, \phi^{(t)}, \mathbf{X}) &\propto n_k^{(t-1)} \times p(\boldsymbol{x}_i|\phi_k^{(t-1)}) = \hat{\pi}_{ik} \quad (k = 1, \ldots, K) \\
p(\boldsymbol{z}_i = k_{new}|\mathbf{Z}_{-i}^{(t-1)}, \phi^{(t-1)}, \mathbf{X}) &\propto \alpha^{(t-1)} \times p(\boldsymbol{x}_i|G_0) = \hat{\pi}_{ik_{new}} \\
\boldsymbol{z}_i^{(t)} &\sim \text{multinoulli}(\hat{\pi}_{i1}, \hat{\pi}_{i2}, \ldots, \hat{\pi}_{ik_{new}}) \\
\text{set} \quad K &= K + 1 \quad (\text{if } \boldsymbol{x}_i \text{ assigned to a new cluster})
\end{aligned}$$

- Sample the mixture component parameters $\{\phi_k^{(t)}\}_{k=1}^K$ and $\alpha^{(t)}$ from the respective CPs

Note: "Markov Chain Sampling Methods for Dirichlet Process Mixture Models" (Neal, 2000) is an excellent reference for various MCMC sampling algorithms for nonparametric Bayesian mixture models (including collapsed versions that don't require sampling for $\{\phi_k\}_{k=1}^K$

Recap: An Alternate View of Mixture Models

- Can represent a mixture model as a discrete distribution as $G = \sum_{k=1}^{K} \pi_k \delta_{\phi_k}$



- Assume $\{\pi_k\}_{k=1}^{K}$ drawn from Dirichlet and parameters $\{\phi_k\}_{k=1}^{K}$ from some base distribution G_0

$$
\begin{aligned}
(\pi_1, \ldots, \pi_K) &\sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K) \\
\phi_k &\sim G_0 \qquad k = 1, \ldots, K
\end{aligned}
$$

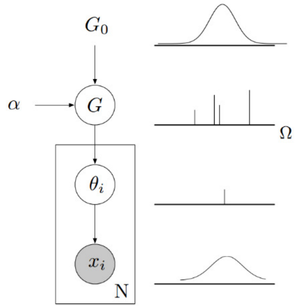
- The mixture model defined by G would generate observations $\boldsymbol{x}_i$ $(i = 1, \ldots, N)$ as follows

$$
\begin{aligned}
\theta_i &\sim G \\
\boldsymbol{x}_i &\sim p(\boldsymbol{x}|\theta_i)
\end{aligned}
$$

- Discrete G implies that many params θ_i's will be identical (leading to clustered observations)

Recap: An Alternate View of Mixture Models

$$
\begin{aligned}
\pi &\sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K) \\
\phi_k &\sim G_0 \\
G &= \sum_{i=1}^{K} \pi_k \delta_{\phi_k} \\
\theta_i &\sim G \\
x_i &\sim p(x|\theta_i)
\end{aligned}
$$



- This representation doesn't have an explicit cluster id $\boldsymbol{z}_i$ for each observation $\boldsymbol{x}_i$. In this representation, clustering is implicit (non-uniqueness of θ_i's implies clustering of $\boldsymbol{x}_i$'s)
- G defines a prior on a K comp. mixture model with mix. prop. $\{\pi_k\}_{k=1}^K$ and params $\{\phi_k\}_{k=1}^K$

Nonparametric Bayesian Mixture Model (Contd)

- Also known as an "infinite mixture model"
- Can have an unbounded number of components (limited only by the data size)
- Can think of these models in two equivalent ways
  - An infinite mixture model can be obtaining using an infinite-dim Dirichlet on its mixing proportions
  - An infinite mixture model is as a discrete distribution G of the form

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$$

- Can view/define such infinite mixture models using various equivalent ways
  - Stick-breaking Process
  - Dirichlet Process
  - Chinese Restaurant Process
  - Pólya-Urn Scheme

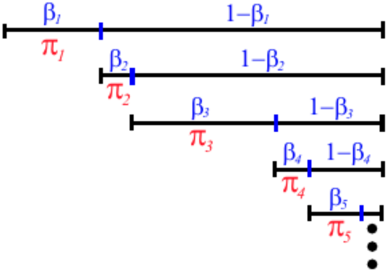
Stick-Breaking Process

- Sethuraman (1994) showed how to construct $G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$
- Sethuraman's stick-breaking construction provides a sequential way to generate π_k's
- We basically need to generate a sequence $\{\pi_k\}_{k=1}^{\infty}$ s.t. $\pi_k \in (0,1)$ and $\sum_{k=1}^{\infty} \pi_k = 1$
- Can be done using a stick-breaking construction for $\{\pi_k\}_{k=1}^{\infty}$ as follows

$$
\begin{aligned}
\beta_k &\sim \text{Beta}(1,\alpha) \qquad k = 1,\ldots,\infty \\
\pi_1 &= \beta_1 \\
\pi_k &= \beta_k \prod_{\ell=1}^{k-1} (1-\beta_{\ell-1}) \qquad k = 2,\ldots,\infty
\end{aligned}
$$

The Stick-Breaking Construction: Pictorial Illustration

- Assume a stick of length 1 to begin with. Now recursively break it as follows:
  - Choose a random location $\beta_k \in (0, 1)$ drawn from Beta$(1, \alpha)$ at which to break the stick
  - Record π_k as "β_k times the length of the remaining stick"



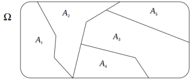
- It is also very popular in deriving inference algorithms for nonparametric Bayesian mixture models

Dirichlet Process (DP)

- A Dirichlet Process $\text{DP}(\alpha, G_0)$ defines a distribution over distributions
  - So $G \sim \text{DP}(\alpha, G_0)$ will give us a distribution
  - α : concentration param, G_0: base distribution (=mean of DP)
  - Large α means $G \to G_0$
- **Fact 1:** If $G \sim \text{DP}(\alpha, G_0)$ then any finite dim. marginal of G is Dirichlet distributed

$$[G(A_1), \ldots, G(A_K)] \sim \text{Dirichlet}(\alpha G_0(A_1), \ldots, \alpha G_0(A_K))$$

for any finite partition $A_1, \ldots, A_K$ of the space Ω (Ferguson, 1973)



- **Fact 2:** Any G drawn from $\text{DP}(\alpha, G_0)$ will be of the form $G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$ (Sethuraman, 1994)
- **Fact 3:** G is a discrete dist, i.e., only a few π_k's will be significant (an informal proof shown next)
- Intuitively, can think of DP as an infinite-dim generalization of a Dirichlet (hence the name)

Detour: Some Properties of Dirichlet Distribution

- Aggregation: If $(\pi_1, \pi_2, \ldots, \pi_K) \sim \text{Dirichlet}(\alpha_1, \alpha_2, \ldots, \alpha_K)$ then

$$\underbrace{(\pi_1 + \pi_2, \pi_3, \ldots, \pi_K)}_{K\text{-1 dim}} \sim \text{Dirichlet}(\alpha_1 + \alpha_2, \alpha_3, \ldots, \alpha_K)$$

- Expansion: If $(\pi_1, \pi_2, \ldots, \pi_K) \sim \text{Dirichlet}(\alpha_1, \alpha_2, \ldots, \alpha_K)$ and $\hat{\pi} \sim \text{Beta}(\alpha_1 b, \alpha_1(1-b))$ then

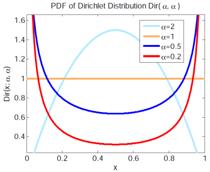
$$\underbrace{(\pi_1\hat{\pi}, \pi_1(1-\hat{\pi}), \pi_2, \ldots, \pi_K)}_{K+1 \text{ dim}} \sim \text{Dirichlet}(\alpha_1 b, \alpha_1(1-b), \alpha_2, \ldots, \alpha_K)$$

- Expansion: If $(\hat{\pi}_1, \ldots, \hat{\pi}_M) \sim \text{Dirichlet}(\alpha_1 b_1, \alpha_1 b_2, \ldots, \alpha_1 b_M)$ with $\sum_{m=1}^{M} b_m = 1$ then

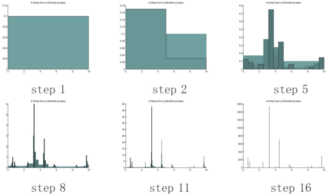
$$\underbrace{(\pi_1\hat{\pi}_1, \ldots, \pi_1\hat{\pi}_M, \pi_2, \ldots, \pi_K)}_{K+M-1 \text{ dim}} \sim \text{Dirichlet}(\alpha_1 b_1, \alpha_1 b_2, \ldots, \alpha_1 b_M, \alpha_2, \ldots, \alpha_K)$$

An Informal Proof: Discreteness of DP Draws

- Note: $(x, 1-x) \sim \text{Dirichlet}(\alpha, \alpha)$ is equivalent to $x \sim \text{Beta}(\alpha, 1)$



- If α is very small, x will be close to 0 or close to 1 (thus $(x, 1-x)$ will be skewed)
- Therefore, if we recursively keep expanding a Dirichlet, it will eventually become discrete



DP Posterior Distribution

- Assume $G \sim \text{DP}(\alpha, G_0)$. Note that G is discrete
- Assume N i.i.d. (non-unique) draws $\theta_1, \ldots, \theta_N$ from the discrete G
- What will be the posterior distribution of G?
- For the finite-dimensional marginal of G, due to Dirichlet-multinoulli conjugacy, we will have

$$[G(A_1), \ldots, G(A_K)] | \theta_1, \ldots, \theta_N \sim \text{Dirichlet}(\alpha G_0(A_1) + n_1, \ldots, \alpha G_0(A_K) + n_K)$$

.. where $n_k = \#\{i : \theta_i \in A_k\}$

- This implies that the posterior of G will also be a DP (a nice property!)

$$G | \theta_1, \ldots, \theta_N \sim \text{DP}(\alpha + N, \frac{\alpha}{\alpha + N} G_0 + \frac{1}{\alpha + N} \sum_{i=1}^{N} \delta_{\theta_i})$$

.. note also that $n_k = \sum_{i=1}^{N} \delta_{\theta_i}(A_k)$

- Note that the posterior's base dist. is a weighted avg. of prior base dist. and an empirical dist.

$$\frac{\alpha}{\alpha + N} G_0 + \frac{N}{\alpha + N} \frac{\sum_{i=1}^{N} \delta_{\theta_i}}{N}$$

DP Predictive Distribution

- We saw that the posterior of G is another DP

$$G|\theta_1,\ldots,\theta_N \sim \text{DP}\left(\alpha+N, \frac{\alpha}{\alpha+N}G_0 + \frac{1}{\alpha+N}\sum_{i=1}^N \delta_{\theta_i}\right)$$

  where $n_k = \sum_{i=1}^N \delta_i(A_k)$ or $n_k = \#\{i : \theta_i \in A_k\}$

- What will be the predictive posterior $p(\theta_{N+1}|\theta_1,\ldots,\theta_N)$?

$$p(\theta_{N+1}|\theta_1,\ldots,\theta_N) = \int p(\theta_{N+1}|G,\theta_1,\ldots,\theta_N)p(G|\theta_1,\ldots,\theta_N)dG = \int p(\theta_{N+1}|G)p(G|\theta_1,\ldots,\theta_N)dG$$

- Intuitively, due to the discreteness of the DP posterior, this would simply be the mean of the DP posterior (= the posterior base distribution)

$$\theta_{N+1}|\theta_1,\ldots,\theta_N \sim \frac{\alpha}{\alpha+N}G_0 + \frac{1}{\alpha+N}\sum_{i=1}^N \delta_{\theta_i}$$

- Thus θ_{N+1} will be equal to a previous θ_i with probability proportional to $\sum_{j=1}^N \delta_{\theta_j=\theta_i}$, and will be a new value with probability proportional to α

A Sequential Generative Scheme

- The form of the DP predictive distribution

$$\theta_{N+1}|\theta_1,\ldots,\theta_N \sim \frac{\alpha}{\alpha+N}G_0 + \frac{1}{\alpha+N}\sum_{i=1}^{N}\delta_{\theta_i}$$

suggests the following scheme to generate a sequence of parameters $\theta_1,\ldots,\theta_N,\theta_{N+1},\ldots$

$$\begin{aligned}
\theta_1 &\sim G_0 \\
\theta_2|\theta_1 &\sim \frac{\alpha}{\alpha+1}G_0 + \frac{1}{\alpha+1}\delta_{\theta_1} \\
&\vdots \\
\theta_n|\theta_1,\ldots,\theta_{n-1} &\sim \frac{\alpha G_0 + \sum_{i=1}^{n-1}\delta_{\theta_i}}{\alpha+n-1}
\end{aligned}$$

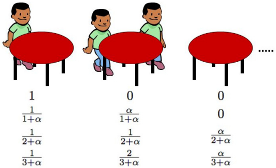
- Note that $\theta_1,\ldots,\theta_{n-1},\theta_n$ is an "exchangeable sequence" (joint probability invariant to ordering)

$$p(\theta_1,\theta_2,\ldots,\theta_n) = \prod_{i=1}^{n} p(\theta_i|\theta_1,\ldots,\theta_{i-1})$$

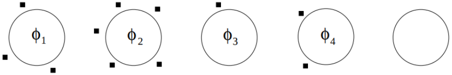
- Related to de-Finetti's theorem (next class)

Chinese Restaurant Process (CRP)

- The CRP is another (culinary) metaphor to describe the way $\theta_1, \ldots, \theta_n$ are sequentially generated
- Think of the θ_i's as customers who sequentially enter a restaurant (need not be Chinese!) and decide which table to sit at. All θ_i's sitting at the same table will be "colored"/labeled identical.



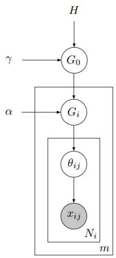
- Probability of sitting at an already occupied table $k \propto n_k$ (n_k: # of people sitting at table k)
- Probability of sitting at an unoccupied table $\propto \alpha$ (where α is a novelty hyperparameter)
- Imagine table k is associated with a unique ϕ_k. Then the arragement would look like..

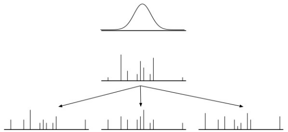


- The table assignment distribution is the same as the DP predictive distribution

Hierarchical Dirichlet Process (HDP)

- Defines a DP whose base distribution G_0 itself is drawn from another DP





$$
\begin{aligned}
G_0 &\sim \mathrm{DP}(\gamma, H)\\
G_i &\sim \mathrm{DP}(\alpha, G_0)\\
\theta_{ij}|G_i &\sim G_i\\
x_{ij}|\theta_{ij} &\sim p(x_{ij}|\theta_{ij})
\end{aligned}
$$

- Can be used if we would like to cluster m data sets, each using a DP mixture model
- The discreteness of the shared base distribution G_0 enables sharing information across the m clustering problems (reason: because the discreteness allows sharing clusters/atoms)
- Important: If G_0 were a continuous distribution, we won't be able to share atoms (probability of G_i and G_j sharing any atoms will be zero if G_0 is a continuous distribution)
- HDP used in nonparametric Bayesian version of LDA topic model

Hierarchical Dirichlet Processes (Teh et al, 2006)

Next Class

- Some other aspects of NPBayes mixture models
- Other examples of NPBayes models and wrap-up of discussion on NPBayes
- On to next topic: Deep Probabilistic Modeling