Bayesian Statistics

Debdeep Pati
Florida State University

October 4, 2016

Finite mixture model

- The finite mixture of normals can be equivalently expressed as

$$y_i \sim N(\mu_{S_i}; \tau_{S_i}^{-1}), S_i \sim \sum_{h=1}^{k} \pi_h \delta_h$$

δ_h = probability measure concentrated at the integer h, $S_i \in \{1, 2, \ldots, k\}$ indexes the mixture component for subject i, $i = 1, \ldots, n$.

- A prior on $\pi = (\pi_1, \pi_2, \ldots, \pi_k)$ and $(\mu_h, \tau_h), h = 1, \ldots, k$ is given by

$$\pi \sim \mathrm{Dir}(\alpha_1, \alpha_2, \ldots, \alpha_k)$$
$$(\mu_h, \tau_h) \sim \mathrm{N}(\mu_h; \mu_0, \kappa\tau_h^{-1})\mathrm{Ga}(\tau_h; a_\tau; b_\tau), h = 1, \ldots, k$$

Posterior Computation in finite mixture models

- Update S_i from its multinomial conditional posterior with

$$Pr(S_i = h|-) = \frac{\pi_h \mathrm{N}(y_i; \mu_h, \tau_h^{-1})}{\sum_{l=1}^{k} \pi_h \mathrm{N}(y_i; \mu_h, \tau_h^{-1})}, h = 1, \dots, k$$

- Let $n_h = \#\{S_i = h, i = 1, \dots, n\}$ and $\bar{y}_h = \frac{1}{n_h} \sum_{i:S_i=h} y_i$ and update (μ_h, τ_h^{-1}) from its conditional posterior

$$(\mu_h, \tau_h^{-1}|-) = \mathrm{N}(\mu_h, \hat{\mu}_h, \hat{\kappa}_h \tau_h^{-1}) \mathrm{Ga}(\tau_h, \hat{a}_{\tau_h}, \hat{b}_{\tau_h})$$

 where
$\hat{\kappa}_h = (\kappa^{-1} + n_h)^{-1}, \hat{\mu}_h = \hat{\kappa}(\kappa^{-1}\mu_0 + n_h \bar{y}_h), \hat{a}_{\tau_h} = a_\tau + \frac{n_h}{2}$,

$$\hat{b}_{\tau_h} = b_\tau + \frac{1}{2}\left\{ \sum_{i:S_i=h} (y_i - \bar{y}_h)^2 + \frac{n_h}{1 + \kappa n_h} (\bar{y}_h - \mu_0)^2 \right\}$$

- Update π as $(\pi|-) = \mathrm{Dir}(a_1 + n_1, \dots, a_k + n_k)$.

Some comments

- Gibbs sampler is trivial to implement
- Discarding a burn-in,monitor $f(y) = \sum_{h=1}^{k} \pi_h N(y; \mu_h, \tau_h^{-1})$ for a large number of iterations & a dense grid of y values
- Bayes estimate of $f(y)$ under squared error loss averages the samples
- Can also obtain 95% pointwise intervals for unknown density

Choosing the Dirichlet hyperparameters

- The choice of hyperparameters in the mixture model can have an important impact
- Focus initially on the choice of $(a_1, \ldots, a_k)'$ in the Dirichlet prior
- A common choice is $a_1 = \cdots = a_k = 1$,which seems "non-informative".
- However, this is actually a poor choice in many cases, as it favors assigning roughly equal weights to the different components
- Ideally, we could choose k as an upper bound and choose hyperparameters, which favor a small number of components with relatively large weights

Finite Approximation to Dirichlet Process

- Ishwaran and Zarepour (2002), "Dirichlet prior sieves in finite normal mixtures," Statistica Sinica, 12, 941-963) propose a finite approximation to the DP
- In particular, they propose letting
- $\pi \sim \text{Dir}(\alpha/k, \ldots, \alpha/k)$ iid
- Assuming also that $\theta_h = (\mu_h, \tau_h) \sim P_0$, they show that

$$\lim_{k\to\infty} \sum_{h=1}^{k} \pi_h \delta_{\theta_h} \to DP(\alpha P_0)$$

- In addition, the posterior for the density is L_1 consistent if $\log k/n \to 0$.

Implications

- We can implement a finite mixture model analysis with a carefully chosen prior & sufficiently large k to obtain an accurate approximation to the DP mixture (DPM) model:
- $f(y) = \int K(y;\theta)dP(\theta), P \sim DP(\alpha P_0)$
- Here, $K(y;\theta)$ is a kernel parameterized by θ - e.g., $K(y;\theta) = N(y;\mu,\tau^{-1})$ with $\theta = (\mu,\tau^{-1})$ for normal mixtures
- P is now an unknown mixing measure
- Hence, we no longer use the DP as a prior directly for the distribution of the data but instead use it for the mixture distribution

Dirichlet Process Mixtures - comments

- The discreteness of the Dirichlet process is not a problem when it is used for a mixture distribution instead of directly for the data distribution
- In fact, in this setting the discreteness is appealing in leading to a simple representation of the mixture distribution that leads to clustering of the observations as a side effect
- Focusing on the finite approximation, $P \sim DP_k(\alpha, P_0)$, let

$$f(y) = \int N(y; \mu, \tau) dP(\mu, \tau) = \sum_{h=1}^{k} \pi_h N(y; \mu_h, \tau_h^{-1})$$

- This induces a prior on f.

Dirichlet Process Mixtures

- For density estimation,consider the DP mixture (DPM)model

$$y_i \mid \mu_i, \tau_i \sim N(\mu_i, \tau_i^{-1}), \theta_i = (\mu_i, \tau_i) \sim P, P \sim \mathrm{DP}(\alpha P_0)(\cdot)$$

- Not immediate clear how to conduct posterior computation
- One strategy relies on marginalizing out P to obtain

$$(\theta_i \mid \theta_1, \ldots, \theta_{i-1}) \sim \left(\frac{\alpha}{\alpha + i - 1}\right) P_0 + \sum_{j=1}^{i-1} \frac{1}{\alpha + i - 1} \delta_{\theta_j}$$

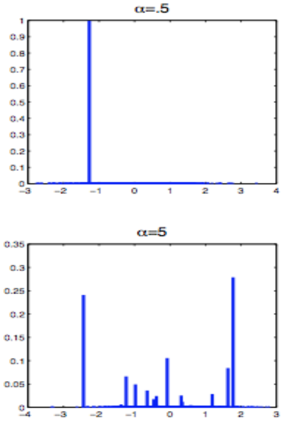
- DP prediction rule or Polya urn scheme (Blackwell & MacQueen, 73)

Avoiding Marginalization

- By marginalizing out the RPM P, we give up the ability to conduct inferences on P
- By having approaches that avoid marginalization, we open the door to generalizations of DPMs
- Stick-breaking representation (Sethuraman, 94),

$$\theta_i \sim P = \sum_{h=1}^{\infty} V_h \prod_{l<h} (1 - V_l) \delta_{\Theta_h}, V_h \overset{i.i.d}{\sim} \text{beta}(1, \alpha), \theta_h \overset{i.i.d}{\sim} P_0$$

Samples from Dirichlet process with precision α



Implications of Stick-Breaking

- For small α, most of the probability is allocated to the first few components, favoring few latent classes
- Expected number of occupied components $\propto \alpha \log n$
- Weights π_h decrease stochastically towards near zero rapidly in the index h
- Suggests truncation approximation (Muliere & Tardella, 98),

$$P = \sum_{h=1}^{N} V_h \prod_{l<h} (1 - V_l) \delta_{\Theta_h}$$

with $V_N = 1$ so that weights sum to one

Blocked Gibbs Sampler (Ishwaran & James, 01)

1. Update $S_i \in \{1, \ldots, N\}$ by multinomial sampling with

$$P(S_i = h \mid -) = \frac{\pi_h N(y_i; \Theta_h)}{\sum_{l=1}^{N} \pi_l N(y_i; \Theta_l)}, \quad h = 1, \ldots, N$$

2. Update stick-breaking weight $V_h, h = 1, \ldots, N-1$, from

$$\text{Beta}\left(1 + n_h, \alpha + \sum_{l=h+1}^{N} n_l\right).$$

3. Update $\Theta_h, h = 1, \ldots, N$, exactly as in the finite mixture model.

Comments on Blocked Gibbs

- N acts as an upper bound on the number of mixture components in the sample
- By choosing a large value, the approximation error should be small
- Possible to monitor this error during the MCMC
- Approximate inferences on functionals of P are possible
- Slice (Walker, 07) & retrospective sampling (Papaspiliopoulos & Roberts, 08) approaches avoid truncation - exact block Gibbs (Papaspiliopoulos, 08) combine these approaches

Choosing the DP precision parameter

- The DP precision parameter α plays a key role in controlling the prior on the number of clusters
- A number of strategies have been proposed in the literature -
  1. Fix α at a small number to favor allocation to few clusters relative to the sample size - a commonly used default value is $\alpha = 1$.
  2. Assign a hyperprior (typically gamma) to α - refer to technical report by West (92) & recent article by Dorazio (09, JSPI, 139, 3384-3390)
  3. Estimate α via empirical Bayes (Liu 96; McAulliffe et al. 06)