Dirichlet Process Bayesian Density Estimation[a]

Longhai Li
Department of Statistics
University of Toronto
3 April 2004

[a] Presentation for STA4312 *Bayesian Linear Models*, taught by Professor Radu Craiu

Mixture Models

- Dataset often comprise the data points from different groups. The model for such data is called mixture models. For instance a dataset of the height may come from two groups, male and female.

- If the group label for each observation is known or we know the number of groups there have been lots of methods dealing with such data, such as **EM algorithm** and **Data Augmentation**. Such learning is usually referred to supervised learning. In many cases we don't have such information. In such cases we need to detect the groups underlying the dataset. Such learning is usually referred to unsupervised learning, or clustering analysis.

- Density estimation can be thought of as one sort of such analysis.

Kernel Density Estimation

- Suppose we observe a bunch of data $,y_1, y_2, \cdots, y_n$, we want to estimate the density function from which these data come.

- A widely used technique is kernel estimation. With a kernel function(a unit density function) $K(x)$, the density of Y_i is estimated by:

$$f(x) = \frac{1}{n\sqrt{h}} \sum_{i=1}^{n} K(\frac{x - Y_i}{\sqrt{h}})$$

- let's get familiar with it by taking $K(x) = \Phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$, the density of $N(0, 1)$. $\frac{1}{\sqrt{h}} K(\frac{x-Y_i}{\sqrt{h}})$ is the density of $N(Y_i, h)$. Kernel density estimation is just average of all the densities of Gaussian distribution centered at each data point Y_i and with variance h. Note that h is the same for all Y_i.

Bayesian Density Estimation

- The same idea can be generalized, i.e. by averaging over unit ordinary densities, such as of Gaussian distribution.

- The data point Y_i is viewed as coming from a mixture Gaussian models. I.e. the density of Y_i is given by

$$f_{Y_i}(x) = \sum_{i=1}^{p} \pi_i \Phi(x, \mu_i, v_i),$$

  where $\Phi(x, \mu_i, v_i)$ is the density of $N(\mu_i, v_i)$. Let $\theta_i = (\mu_i, v_i)$. Here π_i ,$\theta_i, i = 1, 2, \cdots p$ and p are all unknown. **Kernel technique takes** $\mu_i = Y_i$**,**$v_i = h$**,**$\pi_i = \frac{1}{n}$ **and** $p = n$.

- Bayesian methods require a prior for them. Here ϕ_i, $\theta_i = (\mu_i, v_i), i = 1, 2, \cdots p$ and p together form a discrete distribution. We need a prior over this distribution. Dirichlet process can be used to define this prior over distribution.

Dirichlet Distribution: Definition

- The density of a Dirichlet distribution $dir(\alpha_1, \alpha_2, \cdots, \alpha_p)$ is

$$f(\pi_1, \pi_2, \cdots, \pi_p | \alpha_1, \cdots, \alpha_p) = \frac{\Gamma(\alpha_1 + \alpha_2 + \cdots + \alpha_p)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\cdots\Gamma(\alpha_p)} \prod_{i=1}^{p} \pi_i^{\alpha_i}$$

where $\pi_i > 0$ and $\sum_{i=1}^{p} \pi_i = 1$. It is an extension of Beta distribution. It can be used to define a prior for the probability π_i's of a discrete distribution with p possible values,say $L_1, L_2, \cdots, L_p$, or of a multinomial distribution.

- Some properties on this distribution:

$$E(\pi_i) = \frac{\alpha_i}{\alpha_0}, Var(\pi_i) = \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0 + 1)}$$

where $\alpha_0 = \sum_{i=1}^{p} \pi_i$

Posterior Updating of Dirichlet Distribution

- Suppose $X_j \sim$ discrete $\begin{pmatrix} L_1 & L_2 & \cdots & L_p \\ \pi_1 & \pi_2 & \cdots & \pi_p \end{pmatrix}$, i.e. $P(X = L_j) = \pi_j$ and the π_i's are assigned prior $dir(\alpha_1, \alpha_2, \cdots, \alpha_p)$, then the posterior of π_i given $X = L_j$ is

$$P(\pi_1, \pi_2, \cdots, \pi_p | X = L_j) \propto \pi_j \prod_{i=1}^{p} \pi_i^{\alpha_i} = dir(\alpha_1, \cdots, \alpha_j+1, \cdots, \alpha_p)$$

- If $L_1, L_2, \cdots, L_n$ are unknown to us we have to extend Dirichlet distribution to Dirichlet process to define the prior we need for the components.

Dirichlet Process:Definition

Definition: Let Θ be a set, and $\mathcal{A}$ a σ-field of subsets of Θ. Let ν be a finite, non-null,non-negative, finitely additive measure on $(\Theta, \mathcal{A})$. We say a random probability measure P on $(\Theta, \mathcal{A})$ is a Dirichlet Process on $(\Theta, \mathcal{A})$ with BASE measure ν, denoted $P \in DP(\nu)$, if for every $k = 1, 2, \cdots$, and measurable partition $\mathcal{B}_1, \mathcal{B}_2, \cdots, \mathcal{B}_k$ of Θ, the joint distribution of the random probabilities is

$$(P(\mathcal{B}_1), P(\mathcal{B}_2), \cdots, P(\mathcal{B}_k)) \sim Dir(\nu(\mathcal{B}_1), \nu(\mathcal{B}_2), \cdots, \nu(\mathcal{B}_k))$$

Properties of Dirichlet Process

- If $P \sim DP(\nu)$ and $A \in \mathcal{A}$, then

$$E(P(A)) = \frac{\nu(A)}{\nu(\Theta)}$$

$$Var(P(A)) = \frac{\nu(A)(\nu(\Theta) - \nu(A))}{\nu(\Theta)^2(\nu(\Theta) + 1)}$$

- If $\nu = \alpha G_0$ where G_0 is a probability measure,

$$E(P(A)) = G_0(A)$$

$$Var(P(A)) = \frac{G_0(A)(1 - G_0(A))}{\alpha + 1}$$

So G_0 is called BASE distribution(probability measure) and α is called precision parameter. Higher precision leads to smaller variance of $P(A)$ and the effect of G_0 is stronger. This suggests that Dirichlet Process is easy to adjust.

Dirichlet Process Mixture Models

Dirichlet Process Mixture Models is defined as

$$
\begin{aligned}
y_i|\theta_i &\sim F(\theta_i) && (1)\\
\theta_i|G &\sim G && (2)\\
G &\sim DP(G_0, \alpha) && (3)
\end{aligned}
$$

Equivalent model can be obtained by taking $K \rightarrow \infty$ of the following model:

$$
\begin{aligned}
y_i|c, \phi &\sim F(\phi_{c_i}) && (4)\\
\phi_c &\sim G_0 && (5)\\
c_i|\pi &\sim \text{Discrete}(\pi_1, \pi_2, \cdots, \pi_K) && (6)\\
\pi_1, \cdots, \pi_K &\sim \text{Dir}(\alpha/K, \cdots, \alpha/K) && (7)
\end{aligned}
$$

Posterior Updating

- (2) and (3) induce the following posterior distribution of the following form:

$$\theta_i|\theta_1, \cdots, \theta_{i-1} \sim \frac{1}{i-1+\alpha} \sum_{j=1}^{i-1} \delta(\theta_j) + \frac{\alpha}{i-1+\alpha} G_0 \tag{8}$$

- (6) and (7) induce the following posterior updating:

$$P(c_i = c|c_1, \cdots, c_{i-1}) = \frac{n_i^c + \alpha/K}{i-1+\alpha} \tag{9}$$

where $n_i^c = \sum_{j=1}^{i-1} I(c_j = c)$, c is one possible value in the discrete distribution (6), say one of $L_1, \cdots, L_K$.

Posterior Updating: Cont.

(9) implies that when $K \to \infty$

$$
\begin{aligned}
P(c_i = c) &\to \frac{n_i^c}{i-1+\alpha}, c \in (c_1, \cdots, c_{i-1}) \qquad (10) \\
P(c_i \notin (c_1, \cdots, c_{i-1})) &\to \frac{\alpha}{i-1+\alpha} \qquad (11)
\end{aligned}
$$

Let $\theta_i = \phi_{c_i}$, we can see that the limiting distribution (10) and (11) are equivalent to (8).

Gibbs Sampling I

For the DP model defined by (1)-(3), we have a value of θ_i for each observation y_i, the following posterior distribution is directly implied by equation (8):

$$\theta_i|\theta_{-i} \sim \frac{1}{n-1+\alpha}\sum_{j\neq i}\delta(\theta_j) + \frac{\alpha}{n-1+\alpha}G_0 \tag{12}$$

Combining with the data distribution (1), we have

$$\theta_i|\theta_{-i}, y_i \sim \sum_{j\neq i} q_{i,j}\delta(\theta_j) + r_i H_i \tag{13}$$

where,

$$q_{i,j} = bF(y_i, \theta_j)$$

$$r_i = b\alpha \int F(y_i, \theta)dG_0(\theta)$$

Gibbs Sampling II

For the model defined by (4)-(7), by the parallel derivation, we can get that

$$P(c_i = c|c_{-i}, y_i, \phi) = b\frac{n^c_{-i}}{n-1+\alpha}F(y_i, \phi_c), c \in c_{-i} \tag{14}$$

$$P(c_i \notin c_{-i}|c_{-i}, y_i, \phi) = b\frac{\alpha}{n-1+\alpha}\int F(y_i, \phi)dG_0(\phi) \tag{15}$$

Repeatedly sample as follows:

- For $i = 1, 2, \cdots, n$, draw a new value for c_i by (14) and (15), If a new value for c_i other than c_{-i} is drew, we draw a new value from H_i for ϕ_{c_i}, where H_i is the posterior distribution of ϕ given y_i defined by $F(y_i, \phi)$ and G_0.
- For all $c \in (c_1, \cdots, c_n)$, draw a new value from $\phi_c|y_i$ $s.t.$ $c_i = c$

Return to Our Problem

From (13) and (14) we can see a particular property that θ_i and c_i has a tendency to get closer to each other. So the Gibbs sampling of θ_i and c_i is possible to converge to a few dominating points. Our problem is modeled using the limiting model defined by (4)-(7):

$$
\begin{aligned}
y_i | c, \phi &\sim F(y_i, \phi_{c_i}) = \Phi(y_i, \mu_{c_i}, v_{c_i}), v_{c_i} = 1/\tau_{c_i} && (16) \\
\phi_c &\sim G_0(\mu_c, \tau_c) = \Phi(\mu_c, 0, \sigma_g^2) \times \gamma(\tau_c, a, b) && (17) \\
c_i | \pi &\sim \text{Discrete}(\pi_1, \pi_2, \cdots, \pi_K) && (18) \\
\pi_1, \cdots, \pi_K &\sim \text{Dir}(\alpha/K, \cdots, \alpha/K) && (19)
\end{aligned}
$$

where $\gamma(\cdot, a, b)$ is the density of Gamma distribution.

More detailed Implementation

- For $i = 1, 2, \cdots, n$, update c_i by (14) and (15), where

$$\Phi(y_i, \mu_{c_i}, v_{c_i}) = \tau_c^{1/2} e^{-1/2\tau_c(y_i - \phi_c)^2}$$

$$\int F(y_i, \phi) dG_0(\phi) = \frac{1}{\sqrt{v_c + \sigma_g^2}} e^{-1/2 \frac{y_i^2}{v_c + \sigma_g^2}}$$

$$H(\mu, \tau | y_i) \propto F(y_i, \mu, \tau) \times \Phi(\mu, 0, \sigma_g^2) \times \gamma(\tau, a, b)$$

- For all $c \in (c_1, \cdots, c_n)$, update ϕ_c and τ_c:

$$\phi_c | c, y, \tau \sim N\left(\frac{\sum y_i I(c_i = c)}{m_c + \frac{\tau_g}{\tau_c}}, \frac{1}{m_c \tau_c + \tau_g} \right)$$

$$\tau_c | \phi, y, c \sim \Gamma(a + m_c/2, b + 1/2 \sum_{i=1}^{n} (y_i - \theta_i)^2 I(c_i = c))$$

Density Estimation

After running the MCMC we got a bunch of data of $\theta = (\mu, \tau)$, θ might be only part of the final samples, let $n = length(\theta)$,then the density is estimated as,

$$f(x) = \frac{1}{n} \sum_{i=1}^{n} \Phi(x, \mu_i, 1/\tau_i)$$

Deconvolution Problem

Sometimes we can not directly observe y_i's themselves, instead we got a bunch of data $z_i = y_i + N(y_i, \sigma_e^2)$,$\sigma_e^2$ is known to us. We want to estimate the density of y_i. Such problem is called deconvolution problem.It is not hard to use above models to solve it.
Indeed, we need only add one more layer to our models. i.e. $z_i|y_i \sim N(0, \sigma_e^2)$. Then we could learn about y_i by z_i then use y_i to learn about θ_i. Alternatively, we see that adding one more layer to our model is equivalent to $y_i|\theta_i \sim N(\mu_i, \sigma_e^2 + \sigma_{c_i}^2)$, σ_e^2 is a fixed We can also assign Gamma distribution to σ_c^2, the only difficulty is that when we are updating tau using $H(\phi, \tau|y_i)$ and updating τ_c in the final step the Gamma prior for tau_c is not conjugate any more. This problem can be fixed by using some non-standard sampling for it, such as importance sampling, rejection sampling or Markov chain sampling.

Reference

- Antoniak,C.E.(1974),Mixtures of Dirichlet processes with application to Bayesian Nonparametric Problems, Annals of Statistics, vol. 2 pp. 1152-1174
- Escobar,M.D. and West,M. (1995) Bayesian density estimation and inference using mixtures, JASA, vol. 90,pp577-588
- Neal,Radford M.(1998), Markov Chain Samping Methods for Dirichlet Process Mixture Models, Technical Report No.9815 Department of Statistics, Univeristy of Toronto