Bayesian clustering

Estevão

November 5, 2021

1 Finite Mixture of Normals

Suppose we have a set of samples $X_1, X_2, ..., X_n$ can be modelled as

$$p(X_i, |\mu_{1:k}, \tau_{1:k}, q_{1:k}) = \sum_{j}^{k} q_j \mathrm{N}(\mu_j, \tau_j^{-1}), \tag{1}$$

where N(.) denotes the Normal distribution and q_j represents the weight of the j-th component, with $q_j > 0$ and $\sum_{j=1}^{k} q_j = 1$. In addition, let's introduce the latent variable $Z_{1:n}$ to induce the mixture. Thus, we have that $X_i|Z_i = j \sim \mathrm{N}(\mu_j, \tau_j)$. Given the introduction of the latent variable, we can rewrite the likelihood function as

$$p(X_{1:n}|Z_{1:n}, \mu_{1:k}, \tau_{1:k}, Z_{1:n}) = \prod_{j=1}^{k} \prod_{i:I_{(Z_i=j)}}^{n} \mathrm{N}(\mu_j, \tau_j), \tag{2}$$

where $I_{(Z_i=j)} = 1$ if $Z_i = j$ and $I_{(Z_i=j)} = 0$ otherwise. In addition, the latent variables $Z_{1:n} \sim$ Categorical$(1, q)$. That is,

$$p(Z_{1:n}|q_{1:k}) = \prod_{i=1} \prod_{j=1}^{k} q_j^{I_{(Z_i=j)}} = \prod_{j=1}^{k} q_j^{n_j}, \tag{3}$$

where $n_j = \sum_{i}^{n} I_{(Z_i=j)}$ represents the number of observations falling into component j. With that, we can define the joint distribution of $X_{1:n}$ and $Z_{1:n}$ as

$$p(X_{1:n}, Z_{1:n}|\mu_{1:k}, \tau_{1:k}, q_{1:k}) = p(X_{1:n}|Z_{1:n}, \mu_{1:k}, \tau_{1:k}) p(Z_{1:n}|q_{1:k}), \tag{4}$$

$$= \prod_{j=1}^{k} \left[\prod_{i:I_{(Z_i=j)}}^{n} \mathrm{N}(\mu_j, \tau_j) \right] q_j^{n_j}. \tag{5}$$

For notational convenience, let's denote $\theta_k = \{\mu_{1:k}, \tau_{1:k}, q_{1:k}\}$ and $\omega_{ij} = p(Z_i = j|\theta_k, X_{1:n}) / \sum_j p(Z_i = j|\theta_k, X_{1:n})$. Given the expressions above, we have that $Z_{1:n}$ conditioned on $X_{1:n}$ are independent with probability of classification given by

$$p(Z_i = j|\theta_k, X_{1:n}) \propto p(X_i|Z_i, \mu_j, \tau_j, Z_i) p(Z_i), \tag{6}$$

$$\propto \mathrm{N}(\mu_j, \tau_j) q_j. \tag{7}$$

In the end, we'll have that

$$p(Z_i|\theta_k, X_{1:n}) \sim \mathrm{Categorical}(1, \omega_{ij}). \tag{8}$$

To estimate the components of the finite mixture of Normals under the Bayesian paradigm, we consider the following priors:

$$\mu_j|\tau \sim \mathrm{N}(m_j, v_j/\tau_j), \tag{9}$$

$$\tau_j \sim \mathrm{G}(a_j, b_j), \tag{10}$$

$$q_{1:k} \sim \mathrm{Dirichlet}(r_1, r_2, ..., r_k). \tag{11}$$

To construct the MCMC structure, we need the full conditionals for μ_j, τ_j and q_j, which are given below.

$$p(\mu_j|-) \propto p(X_{1:n}|\theta_k, Z_{1:n})p(\mu_j),$$

$$\mu_j|- \sim \mathrm{N}(M_j, V_j), \tag{12}$$

where

$$M_j = (n_j + 1/v_j)^{-1}\left(\sum_{i:Z_i=j} x_i + m_j/v_j\right)$$

and

$$V_j = \frac{v_j}{(n_j v_j + 1)\tau_j}.$$

$$\tau|- \sim \mathrm{G}(A_j, B_j), \tag{13}$$

where

$$A_j = \frac{n_j}{2} + a_j,$$

$$B_j = b_j + \frac{m_j^2}{2v_j} + \frac{\sum_{i:Z_i=j} X_i^2}{2} - \frac{1}{2}\left(n_j + 1/v_j\right) M_j^2.$$

$$\begin{aligned} p(q_{1:k}|-) &\propto p(Z_{1:n}|q_{1:k})p(q_{1:k}), \\ &\propto \prod_{j=1}^{k} q_j^{n_j} p(q_{1:k}), \\ &\propto \mathrm{Multinomial}(n, q_{1:k}) \times \mathrm{Dirichlet}(r_{1:k}). \end{aligned}$$

$$q_{1:k}|- \sim \mathrm{Dirichlet}(r_{1:k} + n_{1:k}) \tag{14}$$

where $n = \sum_j n_j$.

2 Product Partition Models

Let $\mathbf{y} = (y_1, ..., y_n)$ be an n-dimensional vector of a variable we have interest in clustering. We define a partition ρ as a collection of clusters S_j, which are assumed to be non-empty and mutually exclusive. Following Quintana, Loschi, and Page [2], the parametric PPM is presented as

$$\begin{aligned} p(\mathbf{y}, \boldsymbol{\theta}, \rho) &= p(\mathbf{y}|\boldsymbol{\theta}, \rho)p(\boldsymbol{\theta})p(\rho), \\ &= \frac{1}{T}\prod_{j=1}^{k_n}\left[\left(\prod_{i\in S_j} p(y_i|\boldsymbol{\theta}_j)\right)p(\boldsymbol{\theta}_j)c(S_j)\right], \end{aligned}$$

where $c(S_j) = M \times (|S| - 1)!$ for some $M > 0$ is the cohesion function, $T = \sum_{\rho\in\mathcal{P}_n}\prod_{j=1}^{k_n(\rho)} c(S_j)$, and $\boldsymbol{\theta} = (\theta_1, ..., \theta_n)$ such that $\theta_i = \{\theta_j : i \in S_j\}$. For more detail, see Section 2 in Quintana, Loschi, and Page [2].

2.1 Example

Following Section 5 (2nd paragraph) in Quintana, Loschi, and Page [2], let's consider that

$$
\begin{aligned}
y_i|\mu_j, \sigma_j^2 &\sim \mathrm{N}(\mu_j, \sigma_j^2),\\
\mu_j|\mu_0, \sigma_0^2 &\sim \mathrm{N}(\mu_0, \sigma_0^2),\\
\sigma_j^2 &\sim \mathrm{U}(0,1),\\
\sigma_0^2 &\sim \mathrm{U}(0,2),\\
\mu_0 &\sim \mathrm{N}(0,100).
\end{aligned}
$$

Further, let's denote $n_j = |S_j|$ and k as the number of distinct clusters. Below, we present the full conditionals of the quantities/parameters of interest.

$$
\begin{aligned}
p(\mu_j|-) &\propto p(\mathbf{y}|\mu_j, \sigma_j^2)p(\mu_j),\\
&\propto \prod_{i\in S_j} \left[\mathrm{N}(y_i|\mu_j, \sigma_j^2)\right] p(\mu_j),\\
&\propto \exp\left(-\frac{1}{2\sigma_j^2}\sum_{i\in S_j}(y_i-\mu_j)^2\right)\exp\left(-\frac{1}{2\sigma_0^2}(\mu_j-\mu_0)^2\right),\\
&\propto \exp\left\{-\frac{1}{2}\left(\mu^2\left[\frac{n_j}{\sigma_j^2}+\frac{1}{\sigma_0^2}\right]-2\mu\left[\sum_{i\in S_j}\frac{y_i}{\sigma_j^2}+\frac{\mu_0}{\sigma_0^2}\right]\right)\right\},
\end{aligned}
$$

which is

$$
\mu_j|- \sim \mathrm{N}\left(\frac{\sigma_j^{-2}\sum_{i\in S_j} y_i + \mu_0/\sigma_0^2}{n_j/\sigma_j^2 + 1/\sigma_0^2}, \frac{1}{n_j/\sigma_j^2+1/\sigma_0^2}\right). \tag{15}
$$

$$
\begin{aligned}
p(\sigma_j^2|-) &\propto p(\mathbf{y}|\mu_j, \sigma_j^2)p(\sigma_j^2),\\
&\propto \prod_{i\in S_j} \left[\mathrm{N}(y_i|\mu_j, \sigma_j^2)\right] p(\sigma_j^2),\\
&\propto (\sigma_j^2)^{-n_j/2}\exp\left(-\frac{1}{2\sigma_j^2}\sum_{i\in S_j}(y_i-\mu_j)^2\right)\times 1,
\end{aligned}
$$

which is

$$
\sigma_j^2|- \sim \mathrm{IG}\left(\frac{n_j}{2}, \frac{\sum_{i\in S_j}(y_i-\mu_j)^2}{2}\right). \tag{16}
$$

$$
\begin{aligned}
p(\mu_0|-) &\propto p(\mu_j|\mu_0, \sigma_0^2)p(\mu_0),\\
&\propto \prod_{j} \left[\mathrm{N}(\mu_j|\mu_0, \sigma_0^2)\right] p(\mu_0),\\
&\propto \exp\left(-\frac{1}{2\sigma_0^2}\sum_j(\mu_j-\mu_0)^2\right),
\end{aligned}
$$

which is

$$
\mu_0|- \sim \mathrm{N}\left(\frac{\sum_j \mu_j}{k}, \frac{\sigma_0^2}{k}\right). \tag{17}
$$

$$
\begin{aligned}
p(\sigma_0^2|-) &\propto p(\mu_j|\mu_0,\sigma_0^2)p(\sigma_0^2), \\
&\propto \prod_j \left[\mathrm{N}(\mu_j|\mu_0,\sigma_0^2)\right] p(\sigma_0^2), \\
&\propto (\sigma_0^2)^{-k/2} \exp\left(-\frac{1}{2\sigma_0^2}\sum_j(\mu_j-\mu_0)^2\right),
\end{aligned}
$$

$$
\sigma_0^2|- \sim \mathrm{IG}\left(\frac{k}{2}, \frac{\sum_j(\mu_j-\mu_0)^2}{2}\right). \tag{18}
$$

To simulate from the posterior distribution of the PPM, we use the algorithm 8 introduced by Neal [1]. This algorithm was proposed in the context of Dirichlet Process Mixture models, but it can be used for PPMs as well.

1. Let's denote the cluster labels as $c_i = \{j : i \in S_j\}$ with values in $\{1, ..., k\}$. For $i = 1, \cdots, n$, let $h = k + m$, where k is the number of distinct cluster labels c_j such that $j \neq i$ (i.e., the number of distinct clusters considering that observation i has been removed).

   If c_i is a singleton[1], i.e., $c_i \neq c_j$ for all $j \neq i$, let c_i have the label $k+1$, and draw values independently from the prior distribution for μ_j and σ_j^2 for those μ_j and σ_j^2 for which $k+1 < c \leq h$.

   If c_i is NOT a singleton, i.e., $c_i = c_j$ for some $j \neq i$, draw values independently from the prior distribution for μ_j and σ_j^2 for those μ_j and σ_j^2 for which $k < c \leq h$.

   For both cases, draw a new value for c_i from $\{1, \cdots, h\}$ using the following probabilities:

$$
p(c_i = c|c_{-i}, y_i, \{\mu_c\}, \{\sigma_c^2\}) = \begin{cases} b_i \frac{n_{-i,c}}{n-1+\alpha} p(y_i|\mu_c,\sigma_c^2) & \text{for } 1 \leq c \leq k, \\ \\ b_i \frac{\alpha/m}{n-1+\alpha} p(y_i|\mu_c,\sigma_c^2) & \text{for } k \leq c \leq h, \end{cases} \tag{19}
$$

   where $n_{-i,c}$ is the number of observations (excluding i) which have $c_j = c$, and α is the Dirichlet process concentration parameter. Change the state to contain only those μ_j and σ_j^2 that are now associated with one or more observations. Here, b_i is an appropriate normalising constant given by

$$
b_i^{-1} = \sum_{c=1}^{k} \frac{n_{-i,c}}{n-1+\alpha} p(y_i|\mu_c,\sigma_c^2) + \sum_{c=k}^{h} \frac{\alpha/m}{n-1+\alpha} p(y_i|\mu_c,\sigma_c^2). \tag{20}
$$

2. For all $c \in \{c_1, \cdots, c_n\}$: Draw new values from $\mu_j|-$, $\sigma_j^2|-$, $\mu_0|-$, and $\sigma_0^2|-$.

[1] A singleton is a cluster with only one observation. In contrast, any cluster with more than one observation is not a singleton.

Algorithm 1: PPM model

Set up $\mathbf{y}$ and assign all observation into a cluster.
Set values for α and m.
for *mcmc in 1:MCMCiter* **do**
    for *i in 1:n* **do**
        If c_i is NOT a singleton, call part 1 of Neal's algorithm accordingly.
        If c_i is a singleton, call part 1 of Neal's algorithm accordingly.
        If any cluster has been removed, adjust the labels so that there is no gap between them. Recall the labels should follow a sequence from 1 to k.
    end
    Update $\mu_j|-$.
    Update $\sigma_j^2|-$.
    Update $\mu_0|-$.
    Update $\sigma_0^2|-$.
end

References

[1] Radford M Neal. "Markov chain sampling methods for Dirichlet process mixture models". In: *Journal of computational and graphical statistics* 9.2 (2000), pp. 249–265.

[2] F. A. Quintana, R. H. Loschi, and G L Page. "Bayesian Product Partition Models". In: *Wiley StatsRef: Statistics Reference Online* 1.1 (2018), pp. 1–15.