February 11, 2020

Abstract

1

2 Introduction

This report presents the development of a C++ Bayesian Non parametric library containing Marcov Chain sampling methods for density estimation and clustering. In a Bayesian Non-parametric setting we focused on the Dirichlet process (DP) and its extensions, one of the most widely used priors due to its flexibility and computational ease.

3 Dirichlet Process

Formal definition : Let $M > 0$ and G_0 be a probability measure defined on S. A DP with parameters (M, G_0) is a random probability measure G defined on S which assigns probability $G(B)$ to every (measurable) set B such that for each (measurable) finite partition $B_1, ..., B_k$ of S, the joint distribution of the vector $(G(B_1), ..., G(B_k))$ is the Dirichlet distribution with parameters

$$(MG_0(B_1), ..., MG_0(B_k)). \tag{1}$$

The parameter M is called the precision or total mass parameter, G_0 is the centering measure, and the product MG_0 is referred to as the base measure of the DP.

The basic DP model has the form:

$$\begin{aligned} y_i|G &\overset{\text{iid}}{\sim} G, \quad i = 1, \dots, n \\ G &\sim DP(MG_0) \end{aligned}$$

A key property is that the DP is conjugate with respect to i.i.d sampling so that the posterior base distribution is a weighted average of the prior base distribution G_0 and the empirical distribution of the data, with the weighting controlled by M :

$$G|\mathbf{y} \sim DP(MG_0 + \sum_{i=1}^{n} \delta_{y_i}). \tag{2}$$

And the marginal distribution will be the result of the product of the conditionals:

$$p(y_i|y_1, ..., y_{i-1}) = \frac{1}{M+i-1} \sum_{h=1}^{n-1} \delta_{y_h}(y_i) + \frac{M}{M+i-1} G_0(y_i). \tag{3}$$

An important property of the DP is the discrete nature of G. As a discrete random probability measure we can always write G as a weighted sum of point masses. A useful property based on the discrete nature of the process is his stick-breaking representation, i.e. G can be written as:

$$G(\cdot) = \sum_{k=1}^{+\infty} w_k \delta_{m_k}(\cdot) \tag{4}$$

with $m_k \overset{\text{iid}}{\sim} G_0$ and the random weights constructed as $w_k = v_k \prod_{l<k}(1 - v_l)$ where v_k are independent Be(1,M)random variables.

In many applications in which we are interested in a continuous density estimation this discreteness can represents a limit. It's common choice to use a Dirichlet Process Mixture (DPM) model where the DP random measure is the mixing measure for the parameters of a parametric continuous kernel function.

4 Dirichlet Process Mixture Model

Extending the DP by convolving G with a kernel F, the model will have the form:

$$\begin{aligned} y_i|G \sim F_G(y) &= \int F(y, \vartheta)\, G(\mathrm{d}\vartheta), \quad i = 1, \dots, n \\ G &\sim DP(MG_0) \end{aligned}$$

An equivalent hierarchical model is:

$$\begin{aligned} y_i|\vartheta_i &\overset{\perp\!\!\!\perp}{\sim} F(\cdot, \vartheta_i), \quad i = 1, \dots, n \\ \vartheta_i|G &\overset{\text{iid}}{\sim} G, \quad i = 1, \dots, n \\ G &\sim DP(MG_0) \end{aligned}$$

where the *latent variables* ϑ_i are introduced, one per unit. Since G is discrete, we know that two independent draws ϑ_i and ϑ_j from G can be equal with positive probability. In this way the DPM model induces a probability model on clusters and an object of interest starting from this model is the partitioning induced by the clustering as well as the density estimation.

Considering n data units, each ϑ_i will have one of the k unique values ϕ_j. An estimation of the number of the unique values is $M \log(n) \ll n$. Calling c_i the *allocation* parameters to the clusters such that $c_i = j$ if $\vartheta_i = \phi_j$ the model can be thought as the limit as K goes to infinity of finite mixture model with K components:

$$\begin{aligned} (Y_i|\phi, c_i) &\sim F(\cdot, \phi_{c_i}) \\ (c_i|\mathbf{p}) &\sim \sum_{k=1}^{K} p_k \delta_k(\cdot) \\ \phi_c &\sim G_0 \\ \mathbf{p} &\sim \mathrm{Dir}(M/K, \dots, M/K) \end{aligned}$$

where $(p_1, ..., p_K)$ represent the mixing proportions for the classes and each theta is defined by the latent class c and the corresponding parameters ϕ_c.

4.1 Normal Normal-InverseGamma Model

A very common choice is the Gaussian Mixture Model, opting for a Normal kernel and as base measure G_0 the conjugate Normal-InverseGamma:

$$k(y|\vartheta) = N(y, \mu, \sigma^2)$$
$$G_0(\vartheta|\mu_0, \lambda_0, \alpha_0, \beta_0) = N\left(\mu|\mu_0, \frac{\sigma^2}{\lambda_0}\right) \text{Inv-Gamma}(\sigma^2|\alpha_0, \beta_0)$$

Thanks to the conjugacy the predictive distribution for a new observation $\widetilde{y}$ can be found analytically:

$$p(\widetilde{y}|\mu_0, \lambda_0, \alpha_0, \beta_0) = \int k(\widetilde{y}|\vartheta)p(\vartheta, G_0)d\vartheta = \frac{1}{\widetilde{\sigma}}\text{t-Student}\left(\frac{\widetilde{y} - \widetilde{\mu}}{\widetilde{\sigma}}, |\widetilde{v}\right) \tag{5}$$

where $\widetilde{v} = 2\alpha_0$, $\widetilde{\mu} = \mu_0$ and $\widetilde{\sigma} = \sqrt{\frac{\beta_0(\lambda_0+1)}{\alpha_0\lambda_0}}$

The posterior distribution is again a Normal-InvGamma:

$$p(\vartheta|y, \mu_0, \lambda_0, \alpha_0, \beta_0) = N\left(\mu|\mu_n, \frac{\sigma^2}{\lambda_0 + n}\right) \text{Inv-Gamma}(\sigma^2|\alpha_n, \beta_n) \tag{6}$$

with $\mu_n = \frac{\lambda_0\mu_0\bar{y}+n}{\lambda_0+n}$, $\alpha_n = \alpha_0 + \frac{n}{2}$ and $\beta_n = \beta_0 + \frac{1}{2}\sum_{i=1}^{n}(y_i - \bar{y})^2 + \frac{\lambda_0 n(\bar{y}-\mu_0)^2}{2(\lambda_0+n)}$

5 Methods

Starting from the hierarchical model (n) a direct approach is simply drawing values for each ϑ_i from its conditional given the data and the other ϑ_j, but as we discussed before there is an high probability for ties among them and this can result in a slow convergence.

5.1 Neal2

To solve and make it more efficient Neal proposed, starting from the discrete model (n), a Gibbs sampling method, integrating out the mixing proportion $\mathbf{p}$. Assuming the current state of Markov chain consists of $(c_1, ..., c_n)$ and the component parameters ϕ_c for all c, the Gibbs sampler consists of drawing values for each c_i given the conditional probabilities:

$$\mathbb{P}(c_i = c|\mathbf{c}_{-i}, y_i, \boldsymbol{\phi}) \propto \frac{n_{-i,c} + M/K}{n - 1 + M} F(y_i, \phi_c)$$

and consequently a new value for each ϕ_c given the data belonging to that class. The passage to the infinite case is done taking the limit of K to infinity in the conditional distribution of c_i that becomes:

$$\mathbb{P}(c_i = c|\mathbf{c}_{-i}, y_i, \boldsymbol{\phi}) \propto \frac{n_{-i,c}}{n - 1 + M} F(y_i, \phi_c)$$
$$\mathbb{P}(c_i \neq c_j \text{ for all } j|\mathbf{c}_{-i}, y_i, \boldsymbol{\phi}) \propto \frac{M}{n - 1 + M} \int F(y_i, \phi)\, G_0(\mathrm{d}\phi)$$

and considering only the ϕ_c associated with some observation, keeping feasible the sampling. At this point the algorithm works iteratively sampling c and ϕ. For each observation i c_i is updated according to its conditional distribution. It can be set either to one of the other components currently associated with some observation or to a new mixture component. If the new value of c_i is different from all the other c_j a value for ϕ_{ci} is drawn from the posterior distribution H_i, based on the prior G_0 and the single observation y_i. Then for all the classes the sample for ϕ_c is done considering the posterior distribution based on the prior and all the observations belonging to the specific class.

The probability of setting c_i to a new component involves the integral $\int F(y_i, \phi)\, G_0(\mathrm{d}\phi)$, which is difficult in the non-conjugate case, as well as the sample from H_i.

5.2 Neal8

To handle non-conjugate priors Neal proposed a second Markov chain sampling procedure where the state is extended by the addition of auxiliary parameters. This technique allows to update the c_i avoiding the integration with respect to G_0.

In this case the prior for c_i are:

$$\begin{aligned} \text{If } c = c_j \text{ for some } j\text{: } \mathbb{P}(c_i = c|\mathbf{c}_{-i}) &= \frac{n_{-i,c}}{n-1+M} \\ \mathbb{P}(c_i \neq c_j \text{ for all } j) &= \frac{M}{n-1+M} \end{aligned}$$

where the probability of selecting a new component is split among the m auxiliary components. Maintaining the same structure as the *Algorithm* 2, the *Algorithm* 8 is composed of two steps, where the components of the Markov Chain state (c, ϕ) are repeatedly sampled. The first step scans all the observations and evaluates each c_i. If this is equal to another c_j then all the auxiliary variables are drawn from G_0. If it is a singleton then it is linked to one of the auxiliary variable with the corresponding value of ϕ_c while the others are drawn as before from G_0. Then c_i is updated according to the conditional probabilities:

$$P(c_i = c|\mathbf{c}_{-i}, y_i, \phi_1, \dots, \phi_h) \propto \begin{cases} \frac{n_{-i,c}}{n-1+M} F(y_i, \phi_c), & \text{for } 1 \leq c \leq k^- \\ \frac{M/m}{n-1+M} F(y_i, \phi_c), & \text{for } k^- + 1 < c \leq h \end{cases}$$

Once all the ϕ_c not associated anymore with any observation are discarded, the algorithm proceeds with the sampling for ϕ_c for all the classes.

5.3 Blocked Gibbs