Notes on Gibbs Sampling for Dirichlet Process Mixtures

1. Dirichlet Process Mixtures for Density Estimation

Let $X_1, \ldots, X_n \sim F$ where F has density f and $X_i \in \mathbb{R}$. Our goal is to estimate f. The Dirichlet process is not a useful prior for this problem since it produces discrete distributions which do not even have densities. Instead, we use a modification of the Dirichlet process. But first, let us review the frequentist approach.

The most common frequentist estimator is the kernel estimator

$$\widehat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - X_i}{h}\right)$$

where K is a kernel and h is the bandwidth. A related method for estimating a density is to use a mixture model

$$f(x) = \sum_{j=1}^{k} w_j f(x; \theta_j).$$

For example, of $f(x;\theta)$ is Normal then $\theta = (\mu, \sigma)$. The kernel estimator can be thought of as a mixture with n components. In the Bayesian approach we would put a prior on $\theta_1, \ldots, \theta_k$, on $w_1, \ldots, w_k$ and a prior on k. We could be more ambitious and use an infinite mixture

$$f(x) = \sum_{j=1}^{\infty} w_j f(x; \theta_j).$$

As a prior for the parameters we could take $\theta_1, \theta_2, \ldots$ to be drawn from some F_0 and we could take $w_1, w_2, \ldots,$ to be drawn from the stick breaking prior; F_0 typically has parameters that require further priors.

This infinite mixture model is known as the Dirichlet process mixture model (Escobar and West, 1995). It is the same as the random distribution $F \sim \mathrm{DP}(\alpha, F_0)$ which had the form $F = \sum_{j=1}^{\infty} w_j \delta_{\theta_j}$ except that the point mass distributions δ_{θ_j} are replaced by smooth densities $f(x \,|\, \theta_j)$.

The model may be re-expressed as:

$$\begin{aligned} F &\sim \mathrm{DP}(\alpha, F_0) && (1)\\ \theta_1, \ldots, \theta_n \,|\, F &\sim F && (2)\\ X_i \,|\, \theta_i &\sim f(x \,|\, \theta_i), \quad i = 1, \ldots, n. && (3) \end{aligned}$$

Note that in the DPM, *the parameters θ_i of the mixture are sampled from a Dirichlet process, not the data X_i*. Because F is sampled from a Dirichlet process, it will be discrete. Hence there will be ties among the θ_i's; recall our earlier discussion of the Chinese Restaurant Process. The $k < n$ distinct values of θ_i can be thought of as defining clusters. The beauty of this model is that the discreteness of F automatically creates a clustering of the θ_j's. In other words, we have implicitly created a prior on k, the number of distinct θ_j's.

1.1. How to sample from the prior

Draw $\theta_1, \theta_2, \ldots, F_0$ and draw $w_1, w_2, \ldots,$ from the sick breaking process. Then, set

$$f(x) = \sum_{j=1}^{\infty} w_j f(x; \theta_j).$$

The density f is a random draw from the prior. We could repeat this process many times resulting in many randomly drawn densities from the prior. Plotting these densities could give some intuition about the structure of the prior.

1.2. How to sample from the prior marginal

If we want to draw a sample from the prior marginal $m(x_1, \ldots, x_n)$, we first draw F from a Dirichlet process with parameters α and F_0, and then generate θ_i independently from this realization. Then we sample $X_i \sim f(x \,|\, \theta_i)$. We can also use the Chinese restaurant process to draw the θ_j's sequentially. Given $\theta_1, \ldots, \theta_n$ we draw θ_{n+1} from

$$\frac{\alpha}{n+\alpha} F_0(\cdot) + \frac{1}{n+\alpha} \sum_{i=1}^{n} \delta_{\theta_i}(\cdot).$$

Let θ^*_j denote the unique values among the θ_i, with n_j denoting the number of elements in the cluster for parameter θ^*_i; that is, if $c_1, c_2, \ldots, c_n$ denote the cluster assignments $\theta_i = \theta^*_{c_i}$ then $n_j = |\{i \,:\, c_i = j\}|$. Then we can write

$$\theta_{n+1} = \begin{cases} \theta^*_j & \text{with probability } \frac{n_j}{n+\alpha} \\ \theta \sim F_0 & \text{with probability } \frac{\alpha}{n+\alpha}. \end{cases}$$

1.3. How to sample from the posterior

We sample from the posterior by Gibbs sampling. Our ultimate goal is to approximate the predictive distribuiton of a new observation x_{n+1}:

$$\widehat{f}(x_{n+1}) \equiv f(x_{n+1} \,|\, x_1, \ldots, x_n).$$

This density is our Bayesian density estimator.

Let c_{-i} denote the vector of the $n-1$ cluster assignments for all data points other than i. The Gibbs sampler cycles through indices i according to some schedule—for example randomly—and sets $c_i = k$ according to the conditional probability

$$p(c_i = k \,|\, x_{1:n}, c_{-i}).$$

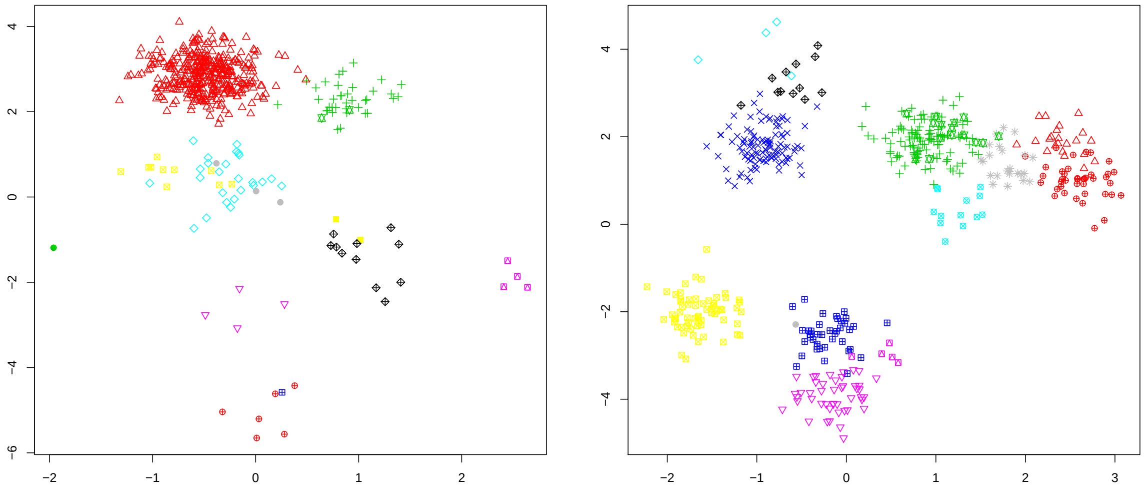

FIG 1. *Samples from a Dirichlet process mixture model with Gaussian generator,* $n = 500$.

This either assigns c_i to one of the existing clusters, or starts a new cluster. By Bayes' rule, this can be written as

$$p(c_i = k \,|\, x_{1:n}, c_{-i}) \propto p(c_i = k \,|\, c_{-i})\, p(x_i \,|\, x_{-i}, c_{-i}, c_i = k).$$

The cluster assignment probability $p(c_i = k \,|\, c_{-i})$ follows the Chinese restaurant process:

$$p(c_i = k \,|\, c_{-i}) = \begin{cases} \frac{n_{k,-i}}{n-1+\alpha} & \text{if } k \text{ is an existing cluster} \\ \frac{\alpha}{n-1+\alpha} & \text{if } k \text{ is a new cluster} \end{cases}$$

where $n_{k,-i} = \{i' : c_{i'} = k, i' \neq i\}$ is the number of data points other than x_i assigned to cluster k. The conditional probability of x_i is given by

$$p(x_i \,|\, x_{-i}, c_{-i}, c_i = k) = p\big(x_i \,|\, \text{other } x_j \text{ in cluster } k\big).$$

Finally, the probability of x_i conditioned on the event that it starts a new cluster is

$$p(x_i \,|\, F_0) = \int p(x_i \,|\, \theta)\, dF_0(\theta).$$

The algorithm iteratively updates the cluster assignments in this manner. After appropriate convergence has been determined, the approximation procedure is to collect a set of partitions $c^{(b)}$, for $b = 1, \ldots, B$. The predictive distribution is then approximated as

$$p(x_{n+1} \,|\, x_{1:n}) \approx \frac{1}{B} \sum_{b=1}^{B} p(x_{n+1} \,|\, c_{1:n}^{(b)}, x_{1:n})$$

where the probabilities are computed just as in the Gibbs sampling procedure; that is,

$$p(x \,|\, c_{1:n}^{(b)}, x_{1:n}) = \sum_j \frac{n_j^{(b)}}{n+\alpha} p\big(x \,|\, x_i \text{ in cluster } c_j^{(b)}\big) + \frac{\alpha}{n+\alpha}\, p(x \,|\, F_0).$$

The calculations are simplest if F_0 is conjugate. Otherwise, MCMC is significantly more complicated; see Neal (2000) for a discussion of MCMC algorithms for this case.

Each partition $c^{(b)}$ has a random number of clusters $k^{(b)} \leq n$. The posterior sampling scheme described above therefore gives an approximation of the posterior distribution over the number components of the mixture model. For example, an estimate of the posterior mean of the number of clusters is

$$\widehat{k} = \frac{1}{B} \sum_b k^{(b)}.$$

A histogram of the number of clusters can also be plotted.

2. Gaussian calculations

Let $X \sim N(\theta, \sigma^2)$ and $\mathcal{D}_n = \{x_1, \ldots, x_n\}$ be the observed data. For simplicity, let us assume that σ is known and we want to estimate $\theta \in \mathbb{R}$. Suppose we take as a prior $\theta \sim N(\mu_0, \tau_0^2)$. Let $\overline{x}_n = \sum_{i=1}^n x_i/n$ be the sample mean. It can be shown that the posterior for θ is

$$\theta \,|\, \mathcal{D}_n \sim N(\overline{\theta}_n, \tau_n^2)$$

where

$$\begin{aligned}\overline{\theta}_n &= w_n \overline{x}_n + (1 - w_n)\mu_0 \\ w_n &= \frac{1}{1 + \frac{\sigma^2/n}{\tau_0^2}} \\ \tau_n^2 &= \frac{\sigma^2/n}{1 + \frac{\sigma^2/n}{\tau_0^2}}\end{aligned}$$

This is another example of a conjugate prior. Note that $w_n \to 1$ and $\tau_n / \frac{\sigma}{\sqrt{n}} \to 1$ as $n \to \infty$. So, for large n, the posterior is approximately $N(\overline{x}_n, se^{\,2})$, and the frequentist and Bayesian inferences agree. The same is true if n is fixed but $\tau_0 \to \infty$, which corresponds to letting the prior become very flat.

The predictive distribution under this Bayesian model is

$$X_{n+1} \,|\, x_1, \ldots, x_n \sim N(\overline{\theta}_n, \tau_n^2 + \sigma^2).$$

To see this, write $X_{n+1} = (X_{n+1} - \theta) + \theta$ and note that $X_{n+1} - \theta$ and θ are uncorrelated given $x_1, \ldots, x_n$, with variances

$$\begin{aligned}\text{Var}(X_{n+1} - \theta \,|\, \theta) &= \sigma^2 \\ \text{Var}(\theta \,|\, x_1, \ldots, x_n) &= \tau_n^2.\end{aligned}$$

References

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430):577–588.

Neal, R. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265.