A Gibbs Sampler for Spatial Clustering with the Distance-dependent Chinese Restaurant Process

Richard Socher
Department of Computer Science
Stanford University
richard@socher.org

Christopher D. Manning
Department of Computer Science
Stanford University
manning@stanford.org

1 Introduction

The distance-dependent Chinese Restaurant Process (dd-CRP) is a flexible class of distributions over partitions which was recently introduced by [1, 2]. In their description and experiments Blei and Frazier focus on the sequential setting such as clustering over time. Their Gibbs sampler, while general in nature, does not explicitly handle the case of non-sequential (also called spatial) clustering. In this case further details are needed for a correct implementation. We introduce the Gibbs sampler for spatial clustering with the dd-CRP. For simplicity, we focus on infinite Gaussian mixture models (IGMM) [9].

The difference between the dd-CRP and the standard CRP is that in the dd-CRP customers sit down with other customers instead of directly at tables. Connected groups of customers sit together at a table only implicitly. Using a similar culinary metaphor, imagine a restaurant full of customers each seating at their own table. The ith customer choses to sit with some other customer j (denoted as $c_i = j$) with a probability proportional to a decreasing function of the distance between the two: $f(d_{ij})$, or by herself with a probability proportional to α. Hence, the larger your distance, the less likely you are to sit with somebody. This leads to the following multinomial over *customer assignments* conditioned on distances $D \in \mathbb{R}^{N\times N}$, where N is the number of customers and the decay function $f : \mathbb{R}^+ \to \mathbb{R}^+$ needs to be non-increasing and have $f(\infty) = 0$,

$$p(c_i = j|D, \alpha) \propto \begin{cases} f(d_{ij}) & \text{if } i \neq j \\ \alpha & \text{if } i = j. \end{cases} \tag{1}$$

The distance function is usually parameterized, e.g. for exponential decay, we have the parameter a: $f_a(d) = \exp(-d/a)$. Notice that this seating prior is not conditioned on $\mathbf{c}_{-i}$, the seating of other customers. Note also that customers may sit in cycles. Each connected component (of which cycles are a special case) forms its own table which may be joined by other customers who sit with a member of that component. Figure 1 illustrates a possible seating assignment.

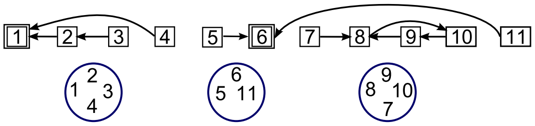

Figure 1: Illustration of the dd-CRP. Customers sit behind other customers. Each group of linked customers implicitly sits at a table. The distance to a customer determines seating, not the order. Hence, 5 may choose to sit with 6 (and hence at a table with 11). Cycles form tables. A customer that sits with somebody in a cycle joins that table, like 7 who joined the cycle of 8,9,10.

The dd-CRP provides another nonparametric alternative for learning infinite clustering models. Similar to the Dirichlet process mixture model, we can define a mixture model with a base distribution G_0 and use the dd-CRP as a prior over cluster assignments. The decay function f is applied to the pairwise distances in matrix D. Given the hyperparameter parameter α, we get the following generative process for observation $x_i \in \mathbb{R}^M$:

1. For each observation $i \in [1, N]$ draw seating assignment $c_i \sim$ dd-CRP(α, f, D).

2. For each cluster $k \in [1, K]$, induced by a connected component, draw parameters $\theta_k \sim G_0$
3. For each observation $i \in [1, N]$, draw $x_i \sim F(\theta_{k(i)})$,

where the function $k(i)$ returns the cluster number of the ith customer. $F(\theta_{k(i)})$ may for instance be a Gaussian distribution and $\theta = (\mu, \Sigma)$.

2 Posterior Inference via Gibbs Sampling

Blei and Frazier [1] provide a general Gibbs sampler for the dd-CRP together with the algorithmic details for language modeling and sequential clustering where customers can only sit with past customers. This results in a seating assignment in the form of a DAG. We introduce the details of a Gibbs sampler for the general case of mixture modeling where cycles are possible. While in principal it is the same sampler, one has to pay special attention to cycles which can start new tables even in cases where customers do not sit by themselves.

The function sitBehind(i), returns the set of all customers that sit behind i, including i itself.[1] This function is recursive. Examples based on the seating of Fig. 1 are: $\text{sitBehind}(1) = \{1, 2, 3, 4\}$, $\text{sitBehind}(2) = \{2, 3\}$, $\text{sitBehind}(3) = \{3\}$. Note that in a cycle, everybody sits behind each other: $\text{sitBehind}(10) = \text{sitBehind}(8) = \{7, 8, 9, 10\}$.

During sampling, we want to compute the probability of each customer i to sit with any customer j: $p(c_i = j | \mathbf{c}_{-i}, \mathbf{x}_{1:N}, S, \alpha, \Lambda_0, \Theta_{1:K})$, where $\Theta_{1:K} = (\mu_k, \Sigma_k)_{k=1:K}$ are the table/cluster parameters, $\mathbf{x}_{1:N}$ are the observations and $\mathbf{c}_{-i}$ are other customers' seating assignments. The two hyperparameters are α (see Eq. 1) and Λ_0, the prior on the covariance of the Gaussian components. In general, we have that

$$p(c_i = j | \mathbf{c}_{-i}, \mathbf{x}_{1:N}, \cdot) \propto p(c_i = j | \cdot) p(\mathbf{x}_{1:N} | c_i = j, \mathbf{c}_{-i}) \tag{2}$$

The prior $p(c_i = j | \alpha, f, D)$ is defined in Eq. 1. It is left to show the different likelihoods that can arise from the seating choice c_i. There are two main cases: either customer i implicitly creates a new table with her seating choice, or she connects to an existing table. New tables can be created in two ways: Either the customer sits by herself or she sits with somebody behind her (and has not previously done so). The latter case creates a cycle. This is captured by the boolean predicate newTable:

$$\text{newTable}(c_i) \text{ is true iff } \left(c_i \in \text{sitBehind}(i) \wedge c_i^{(old)} \notin \text{sitBehind}(i)\right). \tag{3}$$

The likelihood computation has some resemblance to Gibbs sampling for Dirichlet process mixture models [7]. The difference is that we compute the likelihood for all the customers that sit behind i, denoted $X_i = \mathbf{x}_{\text{sitBehind}(i)}$. This can be seen as a sort of blocked sample. Note that we can ignore all other customers as their likelihood is not affected. In the case of a new table, we integrate over the normal-inverse Wishart base distribution $G_0 = \mathcal{NW}$:

$$\text{If newTable}(c_i), p(\mathbf{x}_{1:N} | c_i = j, \mathbf{c}_{-i}) \propto \int \mathcal{N}(X_i | \mu, \Sigma) \mathcal{NW}(\mu, \Sigma | \nu_0, \mu_0, \Lambda_0) d(\mu, \Sigma). \tag{4}$$

Since, we use a conjugate prior, the above integral has a closed form solution in a form of a multivariate Student-t distribution. We approximate this distribution by a moment-matched Gaussian. We sample a new cluster covariance matrix from the inverse-Wishart prior (with hyperparameters Λ_0 which is fixed and $\nu_0 = M$) and a mean which depends on $\mu_0 = 0$ and this covariance matrix as described in [10],

$$\Sigma_{K+1} \sim \mathcal{W}(\nu_0, \Lambda_0), \qquad \mu_{K+1} \sim \mathcal{N}(\mu_0, \Sigma_{K+1}) \tag{5}$$

and then computing the likelihood for all $l \in \text{sitBehind}(i)$ given this Gaussian distribution.

In the case of i sitting with customer j and at its table $k(j)$, we compute the likelihood of all the customers behind i, given that table's parameters.

$$\text{If } \neg\text{newTable}(c_i), p(\mathbf{x}_{1:N} | c_i = j, \mathbf{c}_{-i}) \propto \mathcal{N}(X_i | \mu_{k(j)}, \Sigma_{k(j)}) \tag{6}$$

[1] For notational convenience that will become apparent soon, each customer sits *behind* herself. Intuitively, this is the set of customers that point to i, including i but excluding c_i (unless i and c_i are in a cycle).

As we noted above, the dd-CRP needs to take into account the current and all connected customers since it is not marginally invariant. While this results in a higher computational cost for computing each step, it allows for larger steps through the state space[2] and therefore faster convergence than Gibbs sampling in a CRP mixture. Note also that unless seating assignments actually change, we can use cached likelihood computations of previous iterations.

After each iteration of sampling the seating assignments, we need to update the table parameters given the new table members. Since our $\mathcal{NW}$-prior is conjugate, we can sample the new table parameters from the posterior density in the same family [4]. Let $x_1, \ldots, x_n$ be the customers at a table, $\bar{x}$ the sample mean and we define $Q = \sum_{i=1}^{n}(x_i - \overline{x})(x_i - \overline{x})^T$; then the posterior is $\mathcal{NW}(\mu_n, \kappa_n, \nu_n, \Lambda_n)$ with the following parameters ($\kappa_n = \kappa_0 + n, \nu_n = \nu_0 + n$):

$$\mu_n = \frac{\kappa_0}{\kappa_0 + n}\mu_0 + \frac{n}{\kappa_0 + n}\bar{x}, \;\; \Lambda_n = \Lambda_0 + Q + \frac{\kappa_0 n}{\kappa_0 + n}(\bar{x} - \mu_0)(\bar{x} - \mu_0)^T.$$

Typically, in such conjugate models, one could simply integrate out the parameters and only sample the assignments of observations to clusters. In our experiments this worked very well on synthetic data that we sampled from true Gaussian mixture distributions. However, in all experiments on real data we achieved higher performance when explicitly sampling the cluster parameters.

3 Experiments

We compare the dd-CRP to model based clustering (MBC) [3], the CRP and k-means in preliminary experiments on handwritten digits [6]. The dataset consists of 10 different digits of which we use a subset of digits 1-4. We use the unsupervised method of [5] to extract features from them and then apply spectral dimensionality reduction based on similarities [8] and compare clustering algorithms in this reduced dimensional space. The covariance prior is set to $\Lambda_0 = 0.005 \cdot \text{diag}(1)$ for the CRP and dd-CRP. We set $\alpha = 10^{-6}$ for all experiments. The dd-CRP uses the exponential decay function with $a = 0.01$: $f(d) = \exp(-d/a)$. The table to the right shows results. The dd-CRP outperforms other methods on the mutual information criterion (mutI). k-means, which was given the true number of clusters, outperforms other methods on the rand Index criterion. Results are averaged from samples over 5 independent runs.

Method	mutI	randI	K
Oracle	1.38	1	(4)
$\mathcal{N}$-Oracle	0.98	0.90	(4)
k-means	0.93	0.88	(4)
MBC	0.96	0.86	5
CRP	0.72	0.82	4.2
dd-CRP	0.98	0.86	6.0

Figure 2: Clustering Digits 1-4. MBC stands for model based clustering. (See text for more details.)

References

[1] D. M. Blei and P. I. Frazier. Distance dependent chinese restaurant processes, 2009.

[2] D. M. Blei and P. I. Frazier. Distance dependent chinese restaurant processes. In *ICML*, 2010.

[3] C. Fraley and A. E. Raftery. How many clusters? which clustering method? answers via model-based cluster analysis. *The Computer Journal*, 41:578–588, 1998.

[4] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis, Second Edition*. Chapman & Hall/CRC, 2 edition, July 2003.

[5] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, July 2006.

[6] Y. Lecun and C. Cortes. The mnist database of handwritten digits.

[7] R. M. Neal. Markov chain sampling methods for dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9:249–265, 2000.

[8] A. Ng, M. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *NIPS 14*. MIT Press, 2001.

[9] C. E. Rasmussen. The infinite gaussian mixture model. In *NIPS 12*, volume 12, pages 554–560, 2000.

[10] E. B. Sudderth. *Graphical Models for Visual Object Recognition and Tracking*. Ph.D. thesis, MIT, Cambridge, MA, 2006.

[2] Larger steps through the state space are a result of the seating assignments. When customer i moves to a different table, all customers who recursively sit behind her also move to that table. As a special case, if one customer of a cycle moves, all of them move.