COS 597C: Bayesian nonparametrics

Lecturer: David Blei
Scribes: Peter Frazier, Indraneel Mukherjee

Lecture # 1
September 21, 2007

In this first lecture, we begin by introducing the Chinese Restaurant Process. After a brief review of finite mixture models, we describe the Chinese Restaurant Process mixture, where the latent variables are distributed according to a Chinese Restaurant Process. We end by noting a connection between Dirchlet processes and the Chinese Restaurant Process mixture.

The Chinese Restaurant Process

We will define a distribution on the space of partitions of the positive integers, $\mathbb{N}$. This would induce a distribution on the partitions of the first n integers, for every $n \in \mathbb{N}$.

Imagine a restaurant with countably infinitely many tables, labelled $1, 2, \ldots$. Customers walk in and sit down at some table. The tables are chosen according to the following random process.

1. The first customer always chooses the first table.
2. The nth customer chooses the first unoccupied table with probability $\frac{\alpha}{n-1+\alpha}$, and an occupied table with probability $\frac{c}{n-1+\alpha}$, where c is the number of people sitting at that table.

In the above, α is a scalar parameter of the process. One might check that the above does define a probability distribution. Let us denote by k_n the number of different tables occupied after n customers have walked in. Then $1 \leq k_n \leq n$ and it follows from the above description that precisely tables $1, \ldots, k_n$ are occupied.

Example A possible arrangement of 10 customers is shown in Figure 1. Denote by z_i the table occupied by the customer i. The probability of this arrangement is

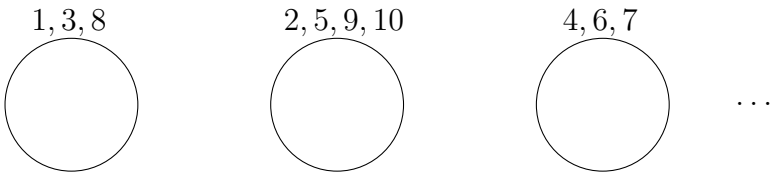

Figure 1: The Chinese restaurant process. Circles represent tables and the numbers around them are the customers sitting at that table.

$$\begin{aligned}\Pr(z_1,\ldots,z_{10}) &= \Pr(z_1)\Pr(z_2|z_1)\ldots\Pr(z_{10}|z_1,\ldots,z_9)\\ &= \frac{\alpha}{\alpha}\frac{\alpha}{1+\alpha}\frac{1}{2+\alpha}\frac{\alpha}{3+\alpha}\frac{1}{4+\alpha}\frac{1}{5+\alpha}\frac{2}{6+\alpha}\frac{2}{7+\alpha}\frac{2}{8+\alpha}\frac{3}{9+\alpha}\end{aligned}$$

We can make the following observations:

1. The probability of a seating is invariant under permutations. Permuting the customers permutes the numerators in the above computation, while the denominators remains the same. This property is known as *exchangeability*.

2. Any seating arrangement creates a partition. For example, the above seating arrangement partitions customers $1,\ldots,10$ into the following three groups $(1,3,8),(2,5,9,10),(4,6,7)$. Exchangeability now implies that two seating arrangements whose partitions consist of the same number of components with identical sizes will have the same probability. For instance, the probability of any seating arrangement of ten customers where three tables are occupied, with three customers each on two of the tables and the remaining four on the third table, will have the same probability as the seating in our example.

   Thus we could define a distribution on the space of all partitions of the integer n, where n is the total number of customers. The number of partitions is given by the partition function $p(n)$, which has no simple closed form. Asymptotically, $p(n) = \exp(O(\sqrt{n}))$.

3. The expected number of occupied tables for n customers grows logarithmically. In particular

$$\mathsf{E}[k_n|\alpha] = O(\alpha \log n)$$

This can be seen as follows: Let X_i be the IRV for the event that the i th customer starts a new table. The probability of this happening is $\Pr[X_i = 1] = \alpha/(i-1+\alpha)$. Since $k_n = \sum_i X_i$, by linearity of expectatin the summation is equal to $\sum_i \alpha/(\alpha+i-1)$ which is upper bounded by $O(\alpha H_n)$ where H_n is the nth harmonic sum.

By taking the limit as n goes to infinity, we could perhaps define a distribution on the space of all natural numbers. However, technical difficulties might arise while dealing with distributions over infinite sequences, and appropriate sigma algebras have to be chosen. Chapters 1-4 of [5], and the lecture notes from ORFE 551 (sent to course mailing list) are recommended for reading up basic measure theory.

Review: Finite mixture models

Finite mixture models are latent variable models. To model data via finite mixtures of distributions, the basic steps are

1. Posit hidden random variables
2. Construct a joint distribution of observed and hidden random variables
3. Compute the posterior distribution of the hidden variables given the observed variables.

Examples include Gaussian mixtures, Kalman filter, Factor analysis, Hidden Markov models, etc. We review the Gaussian mixture model.

A Gaussian mixture with K components, for fixed K, can be described by the following generative process:

1. Choose cluster proportions $\pi \sim \mathsf{Dir}(\alpha)$, where $\mathsf{Dir}(\alpha)$ is a dirichlet prior, with parameter α, over distributions over k points.
2. Choose K means $\mu_k \sim N(0, \sigma_\mu^2)$

3. For each data point:

   (a) Choose cluster assignment $z \sim \mathsf{Mult}(\pi)$

   (b) Choose $x \sim N(\mu_z, \sigma_X^2)$.

The above gives a model for the joint distribution $\Pr(\pi, \mu_{1:K}, z_{1:N} x_{1:N} | \alpha, \sigma_\mu^2, \sigma_X^2)$. We will drop $\alpha, \sigma_\mu^2, \sigma_X^2$ from now on, and assume they are known and fixed. The posterior distribution, given data $x_{1:N}$ is $\Pr(\pi, \mu_{1:K}, z_{1:N} | x_{1:N})$ and has the followin:

1. This decomposed the data over the latent space, thus revealing underlying structure when present.

2. The posterior helps predict a new data point via the *predictive distribution* $\Pr(x|x_{1:N})$. For simplicity, we show how to calculate this quantity when the cluster proportions π is fixed.

$$\begin{aligned}
\Pr(x|x_{1:N}, \pi) &= \sum_z \int_{\mu_z} \Pr(x, z, \mu_z | x_{1:N}, \pi) d\mu_z \\
&= \sum_z \int_{\mu_z} \pi_z \Pr(x|\mu_z) \Pr(\mu_z | x_{1:N}) d\mu_z \\
&= \sum_z \pi_z \int_{\mu_z} \Pr(x|\mu_z) \Pr(\mu_z | x_{1:N}) d\mu_z
\end{aligned}$$

   We could compute $\Pr(\mu_z | x_{1:N}) d\mu_z$ from the posterior $\Pr(\mu_{1:K}, z_{1:N} | x_{1:N}, \pi)$ by marginalizing out the $z_{1:N}$ and the μ_k for $k \neq z$. This would enable us to complete the above calculation to obtain the predictive distribution.

Chinese Restaurant Process Mixture

A Chinese restaurant process mixture is constructed by the following procedure:

1. Endow each table with a mean, $\mu_k^* \sim N(0, \sigma_\mu^2)$, $k = 1, 2, \ldots$.

2a. Customer n sits down at table $z_n \sim \mathrm{CRP}(\alpha; z_1, \ldots, z_{n-1})$.

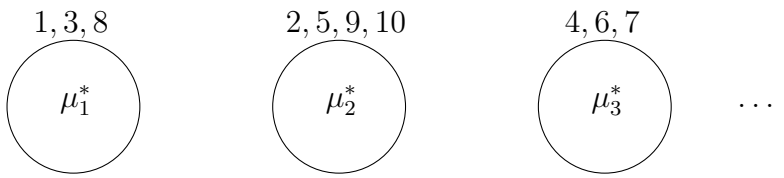

Figure 2: The Chinese restaurant process mixture. Each table k has a mean μ_k^* and customers sitting at it are distributed according to that mean. Tables 1 through 3 and customers 1 through 10 are pictured.

2b. A datapoint is drawn $x_n \sim N(\mu_{z_n}^*, \sigma_x^2)$.

We will consider the posterior and predictive distributions. The hidden variables are the infinite collection of means, $\mu_1^*, \mu_2^*, \ldots$, and the cluster assignments $z_1, z_2, \ldots$. Consider the posterior on these hidden variables given the first N datapoints,

$$p(\mu_{1:N}^*, z_{1:N} \mid x_{1:N}, \theta),$$

where we define $\theta = (\sigma_\mu^2, \sigma_x^2, \alpha)$ to contain the fixed parameters of the model. Note that we only need to care about the means of the first N tables because the rest will have posterior distribution $N(0, \sigma_\mu^2)$ unchanged from the prior. Similarly, we only need to care about the cluser assignments of the first N customers because the rest will have posterior equal to the prior.

The predictive distribution given the data and some additional hidden variables is

$$p(x \mid x_{1:N}, \mu_{1:N+1}^*, z_{1:N}, \theta) = \sum_{z=1}^{1+K_N} p(z \mid z_{1:N}) p(x \mid z, \mu_z^*).$$

These hidden variables may be integrated out with the posterior on the hidden variables to give the predictive distribution conditioned only on the data. Note that, whereas in the finite mixture model the cluster proportions were modeled explicitly, here the cluster proportions are within the z variables. Also note that permuting the $x_{1:N}$ results in the same predictive distribution, so we have exchangeability here as we did earlier.

Why is exchangeability important? Having exchangeability is as though we drew a parameter from a prior and then drew data independently and

identically from that prior. Thus, the data are independent conditioned on the parameter, but are not independent in general. This is weaker than assuming independence.

Specifically, DeFinetti's exchangeability theorem [1] states that the exchangeability of a random sequence $x_1, x_2, \ldots$ is equivalent to having a parameter θ drawn from a distribution $F(\cdot)$ and then choosing x_n iid from the distribution implied by θ. That is,

$$\begin{aligned} \theta &\sim F(\cdot) \\ x_n &\overset{iid}{\sim} \theta. \end{aligned}$$

We may apply this idea to the Chinese restaurant process mixture, which is really a distribution on distributions, or a distribution on the $(\mu_{z_1}, \mu_{z_2}, \ldots)$. The random means $\mu_{z_1}, \mu_{z_2}, \ldots$ are exchangeable, so this implies that their distribution may be expressed in the form given by DeFinetti's exchangeability theorem. Their distribution is given by

$$\begin{aligned} G &\sim \text{DirichletProcess}(\alpha G_0) \\ \mu_{z_i} &\overset{iid}{\sim} G. \end{aligned}$$

Here G_0 is the distribution of the μ^* on the reals, e.g., $N(0, \sigma_u^2)$. Note that we get repeated values when sampling the z_i whenever a customer sits at an already occupied table. Then customer i has μ_{z_i} identical to the μ_{z_j} of all customers j previously seated at the table. In fact, G is an almost surely discrete distribution on the reals with a countably infinite number of atoms.

The reading for next week, [2, 4], is on how to compute the posterior of the hidden parameters given the data, and also includes two new perspectives on the Chinese restaurant process. One perspective is the one just described, of the Chinese restaurant process as a Dirichlet process, and the other is as an infinite limit of finite mixture models. In the reading, focus on [4]. In addition, a good general reference on Bayesian statistics that may be helpful in the course is [3].

References

[1] Jose-M. Bernardo and Adrian F. M. Smith. *Bayesian theory*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Ltd., Chichester, 1994.

[2] Michael D. Escobar and Mike West. Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.*, 90(430):577–588, 1995.

[3] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian data analysis*. Texts in Statistical Science Series. Chapman & Hall/CRC, Boca Raton, FL, second edition, 2004.

[4] Radford M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Statist.*, 9(2):249–265, 2000.

[5] David Williams. *Probability with martingales*. Cambridge Mathematical Textbooks. Cambridge University Press, Cambridge, 1991.