Lecture 9: More Bayesian nonparametrics

04-01-2016
Scribe: Benjamin Bloem-Reddy

Plan for lecture:

- Basic ideas behind CRMs
- HDP, hierarchical models
- Discuss Zhou & Carin (2015)

Summary of the reader responses to Zhou & Carin (2015):

- Difficult
- Satisfying
- A lot of material

Construction of the DP through CRMs

We spent much of last class talking about the Dirichlet Process (DP) and its connection to the Chinese Restaurant Process (CRP). The DP is a special case of a *completely random measure* (CRM). CRMs underlie most Bayesian nonparametric models that are currently in use.

Recall from last week that a DP is a random discrete probability measure and can be written as

$$G_{DP}(\bullet) := \sum_{i=1}^{\infty} p_i \delta_{\theta_i}(\bullet) \quad \text{with } \theta \in \Omega \text{ and } \sum_{i=1}^{\infty} p_i = 1.$$

Let's relax constraint that G is a *probability* measure (i.e., we no longer require that $\sum_{i=1}^{\infty} p_i = 1$). G is now a random measure. Another (useful) characterization is as a stochastic process.

Let Ω' denote the space on which G is defined, with measurable sets $A \subseteq \Omega'$. Then G is said to be **completely random** if $G(A_i)$ is independent from $G(A_j)$ for all $A_i, A_j \subset \Omega'$ such that $A_i \cap A_j = \emptyset$. Roughly speaking, for disjoint sets A_i, A_j of

Ω', $G(A_i)$ and $G(A_j)$ are independent random variables (the randomness comes from the fact that G is random). (Note to the Measure Theoretic Police: G is a stochastic process indexed by the σ-algebra $\mathcal{A}$ when this is true for all sets in $\mathcal{A}$.)

A property of any CRM is that it is a nonhomogeneous *Poisson process* (PP) with Levy measure $\mu(d\omega)$. As a reminder, the number of points in any set $A \subseteq \Omega$ is $N(A) \sim \text{Poisson}(\mu(A))$. Technicalities aside, we can think of $\mu(\cdot)$ as the intensity of the PP. (A great reference is Kingman's monograph *Poisson Processes*.)

In the BNP world, CRMs serve almost exclusively as infinite-dimensional priors on Ω. Their PP construction lives on a product space $\Omega' := \Omega \times \mathbb{R}_+$, which factors as follows:

- Ω is the space where the latent parameters live. For example, in a DP mixture of two-dimensional normals with fixed covariance, Ω is where the component means live, so $\Omega = \mathbb{R}^2$.
- $\mathbb{R}_+$ is the "weight space" of the CRM.

The PP has Levy measure $\mu(d\omega\ dp)$, and a realization of the PP is a set of weighted atoms (points) $\{\omega_i, p_i\}_{i=1}^{N(\Omega')}$, where the total number of atoms $N(\Omega')$ may be infinite. We can use these points to construct a random measure:

$G(\bullet) := \sum_{i=1}^{\infty} p_i \delta_{w_i}(\bullet)$

(See the illustration in Jordan's paper from last class.)

Note that G is itself a realization of a draw from a distribution defined by the Levy measure $\mu(\cdot)$.

How are these useful in BNP modeling? Typically, it boils down to specifying $\mu(d\omega\ dp)$, which is most often assumed to be *homogeneous*: $\mu(d\omega\ dp) = G_0(d\omega)\nu(dp)$. G_0 is often a probability distribution on Ω (and we will make that assumption for the rest of lecture), which will yield draws of the parameters in a hierarchical model, and $\nu(\cdot)$ is the Levy measure for a PP on $\mathbb{R}_+$.

Most of the different BNP models are differentiated by these two ingredients. G_0 is chosen to be appropriate as a prior for the parameters of the data likelihood; ν is chosen according to whether the data are being modeled with latent clusters, latent features, etc.

Scribe's note: One of the things that has made BNP models so useful/popular for a wide range of applications is that these ingredients are largely modular in that models can often be constructed as a set of blocks; the important properties that make the individual blocks appealing from an interpretability and tractability perspective often carry through with little or no modification needed.

For $\mu(\cdot)$ to be useful in a latent variable model, we need it to satisfy:

- $\int_{A \times \mathbb{R}_+} \mu(d\omega\ dp) = \infty$ (infinite number of points on any set $A \subseteq \Omega$)
- $\int_{\Omega \times \mathbb{R}_+} \min\{p, 1\}\ \mu(d\omega\ dp) < \infty$

When $\mu(d\omega\ dp) = G_0(d\omega)\nu(dp)$ and G_0 is a probability distribution, these requirements simplify to be requirements on $\nu(\cdot)$.

(Note to the MTP: These are stronger requirements than what is needed to ensure that $\mu(\cdot)$ is well-defined, but if we want it to be useful, e.g. we can almost surely normalize the CRM for a latent cluster model, or an observation expresses an almost surely finite number of features in a latent feature model, the finite first moment property is necessary. See Gonzalo's note on Piazza for why the second condition emplies that $\sum_{i=1}^{\infty} p_i < \infty$.)

Some examples of CRMs that have appeared in the BNP literature:

- **Beta process:** $\mu(d\omega\ dp) := cp^{-1}(1-p)^{c-1}dp\ G_0(d\omega)$ where $\Omega' = \Omega \times [0, 1]$. This process underlies the Indian Buffet Process and most BNP latent feature models, in which Ω is the space of features, and an observation in our data expresses a feature ω_i with probability p_i. Note that $\mu(\cdot)$ satisfies both of the conditions above, so it will produce an infinite number of atoms with a finite total weight, which ensures that each observation expresses a finite number of features. References are Hjort (1990) and Thibeaux & Jordan (2007).
- (Pause for input from Gonzalo: connection to pure jump processes on $\mathbb{R}$, in particular the Poisson process and the Gamma process. The beta process on $\mathbb{R}$ has been used as a prior for survival analysis and was the subject of Hjort (1990).)
- **Gamma process:** $\nu(d\omega\ dp) := cp^{-1}e^{-cp}dp\ G_0(d\omega)$, where $\Omega' = \Omega \times \mathbb{R}_+$. This process underlies many BNP latent cluster models. Note that the second condition above is satisfied and implies that the total mass of a Gamma process

is finite, which is crucial for construction of the DP.

- **Dirichlet process:** Draw a Gamma process $G = \sum_{i=1}^{\infty} p_i \delta_{w_i}$ and normalize by the total mass $T := \sum_{i=1}^{\infty} p_i$. The result is a Dirichlet process $\tilde{G} = \frac{G}{T} = \sum_{i=1}^{\infty} \tilde{p}_i \delta_{w_i}$, where $\sum_{i=1}^{\infty} \tilde{p}_i = 1$. The DP is characterized by the fact that it is the **only** normalized CRM that is conjugate to itself, which makes posterior inference easy and enables useful hierarchical constructions such as the HDP.

Hierarchical Dirichlet Processes (HDP)

Suppose we have grouped data that we would like to model with its own distribution over latent objects, e.g. topics, but we would like the groups to *share* the set of latent objects. One way to do that is to pass a DP in as the base measure to another DP.

Written out:

- $G_0 \sim DP(\gamma\ H_0)$
- $G_j | G_0 \sim DP(\alpha G_0)$, for $j = 1, 2, ...$
- $\theta_{ij} | G_j \sim G_j$, for $i = 1, 2, ..., n_j$
- $x_{ij} | \theta_{ij} \sim f(\cdot | \theta_{ij})$

Parameters:

- H_0: base distribution
- γ, α: scaling parameters

As a mixed membership model:

- draw a distribution, G_0, over mixture components (e.g. topics, communities, etc.).
- for each $j = 1, 2, ...$, draw a j-specific distribution G_j that has the same atom locations as G_0, but assigns different weight to each atom
- for each observation x_{ij} in group j, draw a parameter θ_{ij} from the discrete distribution with probability p_{ij}
- feed that parameter into the likelihood, $f(x | \theta_{ij})$ to draw an observation x_{ij}.

The key property that makes this useful as a hierarchical model is that G_0 is discrete, so each G_j has the same set of atoms, which ensures that each group j

has a distribution over the same set of latent objects. If G_0 is a continuous distribution, then each G_j will have different atoms and we couldn't interpret any of the clusters to be shared across groups.

Effects of the scaling parameters: γ controls how many occupied clusters are shared across groups; if it is large, not many clusters will be occupied in multiple groups. α controls, in each group, how many clusters are occupied.

(Dave drew a picture of the Chinese Restaurant Franchise.)

Zhou & Carin (2015)

Next week.