APMLA: Lecture 4

Caterina De Bacco and Isabel Valera

1 Introduction to Bayesian Nonparametrics and the Dirichlet Process

So far, we have assumed that the number of components or clusters (and thus the model complexity), is fixed and know. However, in real-world applications this is usually unknown. As a consequence, traditional parametric models using a fixed and finite number of parameters can suffer from data over- or under-fitting when there is a mismatch between the complexity of the model (often expressed in terms of the number of parameters) and the amount of available data. As a result, model selection, or the choice of a model with the right complexity, is often an important issue in parametric modeling. Unfortunately, model selection is an operation that is complicated and tedious, independently of the use cross validation or marginal probabilities as the basis for selection. The Bayesian nonparametric approach is an alternative to parametric modeling and selection. By using a model with an unbounded complexity, underfitting is mitigated, while the Bayesian approach of computing or approximating the full posterior over parameters mitigates overfitting.

Observation: Note that we refer as parametric model to a model with a finite number of parameters (e.g., a parametric MM is assumed to have a finite number of clusters, and thus parameters), a nonparametric model is such model with (a priori) an infinite number of parameters (e.g., a non-parametric MM is assumed to have an infinite number of clusters, and thus parameters).

A *Bayesian nonparametric* (BNP) model defines a probability distribution over an infinite-dimensional parameter space [3]. In practice, a Bayesian nonparametric model uses only a finite subset of the potentially infinite parameters to explain a finite sample of observations. The number of parameters used (i.e., model complexity), however, is adaptively chosen to match the effective complexity of the data, thus providing enough flexibility to deal with arbitrarily complex data [4]. As such, BNP priors are defined over latent variables as a way to perform joint Bayesian inference on both the model's latent variables and complexity.

The Dirichlet Process (DP). The Dirichlet process is currently one of the most popular BNP models, since it can be easily used to define infinite mixture models (a.k.a as DP-MMs), i.e., MMs with an infinite number of components/clusters. The DP is a stochastic process that defines a distribution over distributions, i.e., each draw from a Dirichlet process is itself a distribution. It is called a Dirichlet process because it has Dirichlet distributed finite dimensional marginal distributions, just as the Gaussian process, another popular stochastic process used for Bayesian nonparametric regression, has Gaussian distributed finite dimensional marginal distributions. Distributions drawn from a Dirichlet process are discrete, but cannot be described using a finite number of parameters, thus the classification as a nonparametric model.

More specifically, a DP is fully characterized by the probability distribution over the parameter space, a.k.a., *base measure* $H(\boldsymbol{\theta})$, and the *concentration parameter* $\alpha \geq 0$. And a sample G from a DP, i.e., $G \sim DP(\alpha, H(\boldsymbol{\theta}))$, is a random probability measure that has the same support as H and can be written

as

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\boldsymbol{\theta}_k}, \tag{1}$$

where the atoms weights satisfy $\sum_{k=1}^{\infty} \pi_k = 1$, and the atoms locations $\boldsymbol{\theta}_k \sim H(\boldsymbol{\theta})$. Figure 1 shows an example of a realization (sample) G_0 of a DP. For a formal definition (based on measure theory) refer to [3]. For our application purposes, it is not needed.

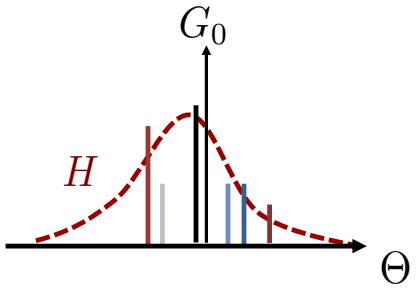

Figure 1: Visualization of a realization of a DP.

2 Infinite mixture model (iMM)

We are now ready to build an infinite MM as:

$$\begin{aligned} G = \sum_{k=1}^{\infty} \pi_k \delta_{\boldsymbol{\theta}_k} &\sim DP(\alpha, H(\boldsymbol{\theta})), \\ z_n &\sim Cat(z|\boldsymbol{\pi}), \\ \mathbf{x}_n &\sim p(\mathbf{x}|z_n, \boldsymbol{\Theta}), \end{aligned} \tag{2}$$

where $\{\boldsymbol{\pi}_k\}_{k=1}^{\infty}$ are the mixing coefficients, $H(\boldsymbol{\theta})$ is the prior distribution over the cluster parameters $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_k\}_{k=1}^{\infty}$, the cluster assignments z_n take values in $\{1, \ldots, \infty\}$, and $p(\mathbf{x}|z_n, \boldsymbol{\Theta})$ is the likelihood model.

However while the model above defines an iMM, sampling from it seems quite difficult since it requires sampling from a DP (i.e., from an infinite dimensional distribution). Alternatively, we may decide to use a Dirichlet prior over the mixing coefficients of the form:

$$\boldsymbol{\pi} \sim Dirichlet(\boldsymbol{\pi}|(\alpha/K, \ldots, \alpha/K)),$$

where K is the number of clusters, α is a hyperparameter, and $(\alpha/K, \ldots, \alpha/K)$ is a K-dimensional vector. As shown in the previous lecture, conditioned on a sequence of cluster assignments $\{z_i\}_{i=1}^{n-1}$, we can compute the marginal probability of assigning a new sample z_n to cluster k as:

$$p\left(z_n = k|\{z_i\}_{i=1}^{n-1}\right) = \int p(z_n = k|\boldsymbol{\pi}) p\left(\boldsymbol{\pi}|\{z_i\}_{i=1}^{n-1}\right) d\boldsymbol{\pi} = \frac{\sum_{i=1}^{n-1}[z_i = k] + \alpha/K}{n-1+\sum_k \alpha/K} = \frac{\sum_{i=1}^{n-1}[z_i = k] + \alpha/K}{n-1+\alpha}.$$

If we take the limit $K \to \infty$ in the above expression, we obtain that

$$p\left(z_n = k|\{z_i\}_{i=1}^{n-1}\right) = \frac{m_k}{n-1+\alpha}, \tag{3}$$

where $m_k = \sum_{i=1}^{n-1}[z_i = k]$ is the number of samples already assigned to cluster k. Note that, since every sample z_i can only be assigned to a single cluster, given a finite number of samples $n-1$, we are only able to "see" a finite number of clusters $K^+ \leq n-1 << \infty$. Moreover, we also know that in order

for the $p\left(z_n|\{z_i\}_{i=1}^{n-1}\right)$ to be a valid pmf, we need to ensure that $\lim_{K\to\infty}\sum_{k=1}^{K} p\left(z_n=k|\{z_i\}_{i=1}^{n-1}\right)=1$. Combining these two observations, we can write: the probability of assigning z_n to a any of the "unseen" clusters $K^+ + 1$ as

$$p\left(z_n=k_{new}|\{z_i\}_{i=1}^{n-1}\right)=1-\sum_{k=1}^{K^+}\frac{m_k}{n-1+\alpha}=\frac{\alpha}{n-1+\alpha}. \tag{4}$$

We can now use this property to get samples from an infinite MM by using an algorithm known as *Chinese restaurant process* (CRP).

The Chinese Restaurant Process (CRP). The Chinese restaurant process (CRP) [1] provides an alternative representation of a DP, by exploiting the clustering property of the later via a restaurant metaphor. In this metaphor, customers enter a Chinese restaurant with an infinite number of tables. The first customer enters the restaurant and sits at the first table. The second customer enters and decides either to sit with the first customer, or by herself, by opening a new table. In general, the $n+1$-th customer can either sit at an already opened table k with probability proportional to the number of customers already sitting in the table, m_k, or open a new table with probability proportional to the concentration parameter α.

In such a metaphor tables can be seen as clusters of customers and are characterized by the dish that they serve, i.e., the cluster parameters $\boldsymbol{\theta}_k$, which are sampled from the distribution (formally called base measure $H(\boldsymbol{\theta})$). Indeed, after n customers have sat down the tables define a partition of the n customers. Note that one does not need to keep track of which customer sits at each table, but only of the total number of customers sitting at a table. As a result, the CRP leads to partitions that are exchangeable in the customer order [4]. Remarkably, the CRP defines a distribution over random partitions (customer-table assignment), whose realizations together with (dish) samples from the base measure $H(\boldsymbol{\theta})$ yield to random samples of a DP with base measure $H(\boldsymbol{\theta})$ and concentration parameter α. Moreover, this generative construction of the DP is particularly suited by Monte Carlo based inference approaches, such as Gibbs sampling.

Generative model. We may therefore sample from an infinite mixture model as:

1. Assigne $z_1=1$ and initialize $K^+=1$
2. Sample likelihood parameters for the first cluster $\boldsymbol{\theta}_1\sim H(\boldsymbol{\theta})$.
   For example, in the case of a GMM, if we assume the covariance matrices to be equal and known for all clusters, i.e., $\boldsymbol{\Sigma}_k=\boldsymbol{\Sigma}_x$, the likelihood parameters correspond only to the mean vectors. Then, we can sample $\boldsymbol{\mu}_1\sim\mathcal{N}(\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)$.
3. For n=2,..., N
   (a) Sample component/cluster indicator as $z_n\sim Cat(\frac{m_1}{n-1+\alpha},\ldots,\frac{m_{K^+}}{n-1+\alpha},\frac{\alpha}{n-1+\alpha})$.
   (b) If $z_n=K^++1$ (i.e., we sample a new cluster):
   - Sample the likelihood parameters for the new cluster as $\boldsymbol{\theta}_{K^++1}\sim H(\boldsymbol{\theta})$
   - Increase the number of seen clusters $K^+=K^++1$
   (c) Sample the observation as $\mathbf{x}_n\sim p(\mathbf{x}|z_n,\boldsymbol{\Theta})$. In the case of the GMM, $p(\mathbf{x}|z_n,\boldsymbol{\Theta})=\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{z_n},\boldsymbol{\Sigma}_x)$

Inference. Similarly as for the parametric Bayesian MM studied in the previous lecture we can rely on Gibbs sampling to infer the posterior distribution of the latent variables and parameters in the iMM (i.e., $p\left(\boldsymbol{\Theta},\{z_n\}_{n=1}^N|\{\mathbf{x}_n\}_{n=1}^N\right)$ using Gibbs sampling.

Observation: Since in the case of the iMM the component coefficient vector $\boldsymbol{\pi}$ is an infinite-dimensional vector and we do not know how to directly sample from an infinite dimensional Dirichlet distribution (or alternatively from a DP), Algorithm 1 from previous lecture is not suitable in this case. In contrast we can easily apply Algorithms 2 and 3, since they work on the posterior distribution after marginalizing out $\boldsymbol{\pi}$ (refer to [2] for a more complete discussion on MCMC methods for iMM). Note that in order

to do so, we should keep in mind that the posterior distribution for the cluster assignment is given by

$$p\left(z_n = k|\mathbf{x}_n, \{z_i\}_{i\neq n}\right) \propto \frac{m_k}{N-1+\alpha} p(\mathbf{x}_n|z_n = k, \boldsymbol{\theta}_k), \tag{5}$$

$m_k = \sum_{i\neq n}[z_i = k]$, for the already "seen" clusters, i.e., $k \in \{1,\dots,K^+\}$. We can also compute the probability of assigning the n-th observation to a new "unseen" cluster as

$$p\left(z_n = k_{new}|\mathbf{x}_n, \{z_i\}_{i\neq n}\right) \propto \frac{\alpha}{N-1+\alpha} \int p(\mathbf{x}_n|z_n = k_{new}, \boldsymbol{\theta}) H(\boldsymbol{\theta}) d\boldsymbol{\theta}, \tag{6}$$

where the second term in the integral accounts for the likelihood after marginalizing the likelihood parameter. Roughly speaking, this terms takes into account that the new cluster can be located in any value of $\boldsymbol{\theta}$ with density given by the prior $H(\boldsymbol{\theta})$. As a result, we can re-write the Gibbs algorithms 2 and 3 for the iMM as follows:

Algorithm 2. We can approximate the distribution $p\left(\boldsymbol{\Theta}, \{z_n\}_{n=1}^N | \{\mathbf{x}_n\}_{n=1}^N\right)$ using Gibbs sampling, as follows:

1. Initialize the initial number of "seen" clusters K^+ (e.g., to one).
2. Initialize the cluster assignments $\{z_n^{(0)}\}_{n=1}^N$ and the model parameters $\{\boldsymbol{\theta}_k^{(0)}\}_{k=1}^{K^+}$.
3. For $\tau = 1,\dots,N_{it}$
   (a) For n=1,..., N
   - If necessary, remove empty cluster and update the total number of "seen" clusters K^+ accordingly.
   - Sample $z_n^{(\tau)} \sim p\left(z_n = k|\mathbf{x}_n, \{z_i^{(\tau)}\}_{i<n}, \{z_i^{(\tau-1)}\}_{i>n}, \{\boldsymbol{\theta}_k^{(\tau-1)}\}_{k=1}^{K^+}\right)$ according to Equations 5 and 6.
   - If $z_n^{(\tau)} = K^+ + 1$, then sample the new cluster parameters $\boldsymbol{\theta}_{K^+ +1}$ from its posterior distribution (i.e., conditioned on $\mathbf{x}_n$) and increase the number of "seen" clusters to $K^+ = K^+ + 1$.
   (b) Sample likelihood model parameters for the already seen clusters ($k \in \{1,\dots,K^+\}$) from their posterior distribution $\boldsymbol{\theta}_k^{(\tau)} \sim p(\boldsymbol{\theta}|\{z_n^{(\tau)}\}_{n=1}^N, \{\mathbf{x}_n\}_{n=1}^N)$.

Algorithm 3. We can approximate the distribution $p\left(\{z_n\}_{n=1}^N | \{\mathbf{x}_n\}_{n=1}^N\right)$ using Gibbs sampling, as follows:

1. Initialize the initial number of "seen" clusters K^+ (e.g., to one).
2. Initialize the cluster assignments $\{z_n^{(0)}\}_{n=1}^N$.
3. For $\tau = 1,\dots,N_{it}$
   (a) For n=1,..., N
   - If necessary, remove empty cluster and update the total number of "seen" clusters K^+ accordingly.
   - Sample $z_n^{(\tau)} \sim p\left(z_n = k|\{\boldsymbol{x}_i\}_{i=1}^N, \{z_i^{(\tau)}\}_{i<n}, \{z_i^{(\tau-1)}\}_{i>n}\right)$ using Eq. 6 for a new cluster and the posterior predictive for the already seen clusters, which is given by

$$p\left(z_n = k|\mathbf{x}_n, \{z_i\}_{i\neq n}\right) \propto \frac{m_k}{N-1+\alpha} \int p(\mathbf{x}_n|z_n = k, \boldsymbol{\theta}_k) p(\boldsymbol{\theta}_k|\{\boldsymbol{x}_i, z_i\}_{i\neq n}) d\boldsymbol{\theta}_k, \tag{7}$$

   - If $z_n^{(\tau)} = K^+ + 1$, then increase the number of "seen" clusters to $K^+ = K^+ + 1$.

References

[1] D. J. Aldous. Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII – 1983*, pages 1–198. Springer, 1985.

[2] R. M. Neal. Markov chain sampling methods for dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265, 2000.

[3] P. Orbanz and Y. W. Teh. Bayesian nonparametric models. In *Encyclopedia of Machine Learning*, pages 81–89. Springer, 2011.

[4] Y. W. Teh. Dirichlet process. In *Encyclopedia of machine learning*, pages 280–287. Springer, 2011.