CS731 Spring 2011 Advanced Artificial Intelligence

Bayesian Nonparametrics

Lecturer: Xiaojin Zhu *jerryzhu@cs.wisc.edu*

Bayesian nonparametrics are Bayesian models where the underlying finite-dimensional random variable is replaced by a stochastic process. This replacement allows much richer nonparametric modeling (recall nonparametric regression methods such as Nadaraya-Watson, which is a "point estimate") but still in a Bayesian framework. Like nonparametric models, the model complexity generally is unlimited and grows with training data size, hence the name "infinite" models. But critically, the computation on any given training set is finite. A *stochastic process* is a (infinite) collection of random variables indexed by a set $\{\mathbf{x}\}$. The name originates from $\mathbf{x} \equiv t \in \mathbb{R}$ being "time," though in machine learning we generally have $\mathbf{x} \in \mathbb{R}^d$.

1 Gaussian Processes

A *Gaussian process* is a stochastic process where any finite number of random variables have a joint Gaussian distribution. Gaussian process is equivalent to *kriging* in some natural sciences. We will use $\mathbf{x} \in \mathbb{R}^d$ to denote the index, and $f(\mathbf{x}) \in \mathbb{R}$ the particular random variable indexed by $\mathbf{x}$. Thus f is the stochastic process. A Gaussian process is specified by a *mean function*

$$m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})] \tag{1}$$

and a *covariance function* (positive definite, a.k.a. kernel function)

$$k(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))]. \tag{2}$$

We write the Gaussian process as

$$f(\mathbf{x}) \sim GP(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')). \tag{3}$$

A draw from a GP is a function $f()$. A common choice of the mean function is $m(\mathbf{x}) = 0, \forall \mathbf{x}$ and we will follow it, though this is by no means necessary. A common choice of the covariance function is $k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{x} - \mathbf{x}'\|^2\right)$, though other covariance functions might be more appropriate for specific tasks.

By the definition of Gaussian process, for any finite set $\mathbf{x}_1, \ldots, \mathbf{x}_n$, we have

$$(f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n)) \sim N(\mu, \Sigma) \tag{4}$$

where

$$\mu = (m(\mathbf{x}_1), \ldots, m(\mathbf{x}_n)) \tag{5}$$

and

$$\Sigma = \begin{pmatrix} k(\mathbf{x}_1, \mathbf{x}_1) & \ldots & k(\mathbf{x}_1, \mathbf{x}_n) \\ & \ldots & \\ k(\mathbf{x}_n, \mathbf{x}_1) & \ldots & k(\mathbf{x}_n, \mathbf{x}_n) \end{pmatrix}. \tag{6}$$

That is, any finite subset of those random variables follow a Gaussian distribution with the mean vector and the covariance matrix determined pointwise by m and k, respectively. Therefore, GP can be thought of as an infinite-dimensional Gaussian distribution.

Bayesian nonparametric modeling with GP is basically the following realization: for the infinite set of random variables indexed by $\mathbf{x}$, the prior is a GP, and the posterior upon observing some finite subset of the random variables is another GP. We will start with regression, in which this relationship is the easiest to see.

1.1 Gaussian Process for Regression without Noise

Consider the standard regression setting where we are given a training set $\{(\mathbf{x}_i, y_i)\}_{i=1\dots n}$ where $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$. We want to predict y_* on a test set $\mathbf{x}_*$.

The key assumption we will make is that $y_{\mathbf{x}} = f(\mathbf{x}), \forall \mathbf{x}$ (noiseless), and $f \sim GP(0, k)$ for some covariance function k. This is the prior. This implies that the finite set of training and test outputs follow a (prior) Gaussian distribution:

$$\begin{pmatrix} f_{1:n} \\ f_* \end{pmatrix} \sim N\left(0, \begin{pmatrix} K_{nn} & K_{n*} \\ K_{*n} & K_{**} \end{pmatrix}\right). \tag{7}$$

Here, K_{nn} is the $n \times n$ covariance matrix defined on the training set as in (6), K_{n*} is the $n \times |\text{test}|$ covariance matrix, and so on.

Now we observe the noiseless values $f_1 = y_1, \dots, f_n = y_n$, the posterior on f is another slightly degenerate GP. For the purpose of regression, however, it is important to realize that all we need to do is to compute the finite-dimensional conditional distribution $p(f_* \mid \mathbf{x}_*, \mathbf{x}_{1:n}, f_{1:n})$. This follows from the property of the Gaussian distribution:

$$p(f_* \mid \mathbf{x}_*, \mathbf{x}_{1:n}, f_{1:n}) = N(K_{*n} K_{nn}^{-1} f_{1:n}, K_{**} - K_{*n} K_{nn}^{-1} K_{n*}). \tag{8}$$

In particular, we have the Bayesian prediction for f_*, i.e., the mean in (8). We also have the uncertainty encoded in the covariance matrix above.

1.2 Gaussian Process for Regression with Noise

Most often, we assume that the observed output is noisy:

$$y_i = f_i + \epsilon, \tag{9}$$

where $\epsilon \sim N(0, \sigma_n^2)$. However, we still assume that the underlying f is a GP. In this case,

$$cov(y_i, y_j) = k(\mathbf{x}_i, \mathbf{x}_j) + \sigma_n^2 \delta_{ij}. \tag{10}$$

We now focus on the joint distribution between $y_{1:n}$ and f_*:

$$\begin{pmatrix} y_{1:n} \\ f_* \end{pmatrix} \sim N\left(0, \begin{pmatrix} K_{nn} + \sigma_n^2 I & K_{n*} \\ K_{*n} & K_{**} \end{pmatrix}\right). \tag{11}$$

A similar conditioning results in

$$p(f_* \mid \mathbf{x}_*, \mathbf{x}_{1:n}, y_{1:n}) = N(K_{*n}(K_{nn} + \sigma_n^2 I)^{-1} y_{1:n}, K_{**} - K_{*n}(K_{nn} + \sigma_n^2 I)^{-1} K_{n*}). \tag{12}$$

Note the above is the predictive distribution for f_*. The predictive distribution for y_* has the same mean but wider spread: its covariance can be obtained by adding $\sigma_n^2 I$ to the covariance in (12).

A quantity of interest is the *marginal likelihood* (or evidence)

$$p(y_{1:n} \mid \mathbf{x}_{1:n}) = \int p(y_{1:n} \mid f_{1:n}) p(f_{1:n} \mid \mathbf{x}_{1:n}) df_{1:n}, \tag{13}$$

where we integrate out (marginalize over) the function values $f_{1:n}$. Using the fact that $y_{1:n} \sim N(0, K_{nn} + \sigma_n^2 I)$, we have

$$\log p(y_{1:n} \mid \mathbf{x}_{1:n}) = -\frac{1}{2} y_{1:n}^\top (K_{nn} + \sigma_n^2 I)^{-1} y_{1:n} - \frac{1}{2} \log |K_{nn} + \sigma_n^2 I| - \frac{n}{2} \log 2\pi. \tag{14}$$

The marginal likelihood is used for model selection, e.g., tuning the kernel bandwidth σ in k.

1.3 Gaussian Process for Classification

For simplicity, let $y \in \{0, 1\}$. The idea is similar to regression with noise, where we assume an underlying GP for f. Instead of a Gaussian noise model, we have a sigmoid function which converts $f(\mathbf{x}_i) \in \mathbb{R}$ into

$$p(y_i = 1 \mid \mathbf{x}_i) = s(f(\mathbf{x}_i)). \tag{15}$$

A sigmoid function is a monotonically increasing function mapping from $\mathbb{R}$ to $[0, 1]$. Two common choices of the sigmoid function are the *logistic function*

$$s(z) = \frac{1}{1 + \exp(-z)} \tag{16}$$

and the *cumulative Gaussian*

$$s(z) = \Phi(z). \tag{17}$$

Inference is carried out in two steps. First, one computes the latent f_*:

$$p(f_* \mid \mathbf{x}_*, \mathbf{x}_{1:n}, y_{1:n}) = \int p(f_* \mid \mathbf{x}_*, \mathbf{x}_{1:n}, f_{1:n}) p(f_{1:n} \mid \mathbf{x}_{1:n}, y_{1:n}) df_{1:n}. \tag{18}$$

Second, the test labels are obtained by

$$p(y_* = 1 \mid \mathbf{x}_*, \mathbf{x}_{1:n}, y_{1:n}) = \int s(f_*) p(f_* \mid \mathbf{x}_*, \mathbf{x}_{1:n}, y_{1:n}) df_*. \tag{19}$$

The computation in (19) factors into individual test points. Unfortunately, unlike the regression case, the non-Gaussian likelihood in (18) and (19) makes the integrals analytically intractable. Much effort on GP classification has been on efficient approximations of these two integrals, see Rasmussen & Williams 3.4-3.6.

2 Dirichlet Processes

A *Dirichlet process* is a stochastic process where any finite partitions follow a Dirichlet distribution. Let Θ be a probability space. Let H be a *base distribution* over Θ.

Example 1 *Let $\Theta = \mathbb{R}^d$, which is a probability space. An element $\theta \in \mathbb{R}^d$ serves as an index to the stochastic process we are going to define. $H = N(0, \Sigma)$ is a base distribution over Θ. Note, however, that $H(\theta) = N(\theta; 0, \Sigma)$ is* not *a random variable (it is a fixed value for a given θ). H does not define a stochastic process.*

Consider the following stick-breaking construction:

$$\begin{aligned}
\beta_k &\sim \text{Beta}(1, \alpha) && (20)\\
\pi_k &= \beta_k \prod_{i=1}^{k-1} (1 - \beta_i) && (21)\\
\theta_k^* &\sim H && (22)\\
G &= \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*}. && (23)
\end{aligned}$$

Here δ_z is the point mass function on z. Note $\pi_1, \pi_2, \ldots$ are the sequence of stick fragments, which tend to (but not necessarily) get smaller, and they sum to length 1 (the whole stick). For each fragment, we associate an index θ_k^* and this is where the base distribution H comes in. Now, G is a sample from a stochastic process. Here, the random variables are indexed by $\theta \in \Theta$. $G(\theta)$ is the value of the random variable indexed by θ: if we were "lucky" in that $\theta_k^* = \theta$ for a (let's assume) single k then $G(\theta) = \pi_k$ otherwise $G(\theta) = 0$. If we were to repeat the stick-breaking construction from scratch, $G(\theta)$ might take a different value.

★ *Contrast this with $H(\theta)$.*

The stochastic process that we are drawing G from is called the Dirichlet Process:

$$G \sim DP(\alpha, H). \tag{24}$$

Note the concentration parameter α appeared in the Beta distribution in stick-breaking. This sample G has the same status as a sample (a random function) f to a Gaussian Process. However, there are important differences:

- G is a probability measure on Θ. That is, it is naturally normalized. This makes G similar to H in the sense that both are probability measures on Θ. Therefore, one can draw samples $\theta \sim G$. In contrast, the sample $f \sim GP$ is a random function and not normalized.
- With probability one, G is a discrete measure. This is true even if H is a continuous measure (e.g., when H is Gaussian). This has the important consequence that θ's drawn from G have repeats. This is the basis for using DP to model (infinite) clustering. In contrast, $f \sim GP$ is a continuous function.

The Dirichlet Process has several interesting properties:

- For a random measure G to be distributed according to $DP(\alpha, H)$, its marginals have to be Dirichlet distributed in the following sense. Let $A_1, \ldots, A_r$ be any finite measurable partition of Θ, then

  $$(G(A_1), \ldots, G(A_r)) \sim \text{Dirichlet}(\alpha H(A_1), \ldots, \alpha H(A_r)). \tag{25}$$

  This is where Dirichlet Process gets its name.

- For any measurable $A \subseteq \Theta$,

  $$\mathbb{E}[G(A)] = H(A) \tag{26}$$

  $$\mathbb{V}[G(A)] = \frac{H(A)(1-H(A))}{1+\alpha}. \tag{27}$$

  As the concentration parameter α increases, the variance decreases. As $\alpha \to \infty$, $G(A) \to H(A)$ for any measurable A.

- Let $G \sim DP(\alpha, H)$. Recall G is itself a probability measure. Suppose we observe $\theta_1, \ldots, \theta_n \sim G$. These observations should help us reduce some uncertainty in G. That is, we are interested in the posterior distribution of G given $\theta_1, \ldots, \theta_n$ (the original DP is the prior). For any finite measurable partition $A_1 \ldots A_r$ of Θ, let $n_k = \sum_{i=1}^n 1_{\theta_i \in A_k}$. Then

  $$(G(A_1), \ldots, G(A_r)) \mid \theta_1, \ldots, \theta_n \sim \text{Dirichlet}(\alpha H(A_1) + n_1, \ldots, \alpha H(A_r) + n_r). \tag{28}$$

  The posterior is

  $$G \mid \theta_1, \ldots, \theta_n \sim DP\left(\alpha + n, \frac{\alpha}{\alpha+n} H + \frac{1}{\alpha+n} \sum_{i=1}^n \delta_{\theta_i}\right). \tag{29}$$

- Let $G \sim DP(\alpha, H)$ and we observe $\theta_1, \ldots, \theta_n \sim G$. What is the predictive distribution of θ_{n+1}? As standard in Bayesian methods, we need to integrate out G and arrive at

  $$\theta_{n+1} \sim \frac{\alpha}{\alpha+n} H + \frac{1}{\alpha+n} \sum_{i=1}^n \delta_{\theta_i}. \tag{30}$$

  Note this is also the base distribution in the DP posterior.

  Following (30), there is a chance that $\theta_{n+1} = \theta_i$ for some $i \leq n$. For simplicity assume H is smooth so that any repeated values come from the δ point mass functions. In fact, let $\theta_1^* \ldots \theta_m^*$ be the *unique* values in $\theta_1 \ldots \theta_n$, and $n_k = \sum_{i=1}^n 1_{\theta_i = \theta_k^*}$ for $k = 1 \ldots m$. Then θ_{n+1} can be generated with the following procedure:

1. With probability $\alpha/(\alpha+n)$, draw a new value from H and assign it to θ_{n+1};
2. Otherwise, reuse value θ_k^* with probability n_k/n.
3. We add θ_{n+1} to the samples, and repeat this process.

This process, as defined by (30) is known as the Blackwell-MacQueen urn scheme.

- The equality relationship in $\theta_1 \ldots \theta_n$ defines a *partition* of n items. Think of the samples as customers to a Chinese restaurant with infinite tables. Initially the restaurant is empty, and the first customer sits at the first table. Let n_k be the number of customers at table k after n customers have arrived. Then with probability $\alpha/(\alpha+n)$ the $(n+1)$-th customer sits at a new table; otherwise he joins an existing table with probability proportional to the number of people already sitting there. This is known as the *Chinese Restaurant Process* (CRP), which defines a distribution over partitions of items.

  The CRP can be thought of as "half of DP." If, whenever a new table is used, a dish is drawn from the base distribution $\theta \sim H$, and all future customers sitting on this table will eat this dish, it would be equivalent to DP.

★ *The facts that there are infinite tables and the customers can never leave make the restaurant a rather eerie place. Maybe we should rename it the Hotel California Process.*

2.1 Dirichlet Process Mixture Models (DPMMs)

The most common application of DP is in "infinite mixture models," where the number of clusters is unknown *a priori* and is unbounded. The idea is simple:

$$G \sim DP(\alpha, H) \tag{31}$$

$$\theta_i \sim G \tag{32}$$

$$\mathbf{x}_i \sim F(\theta), \tag{33}$$

where $F(\theta)$ is an appropriate distribution parametrized by θ.

Example 2 *Let $\theta \in \mathbb{R}^d \times S_d$, $H =$ Normal-Inverse-Wishart$(\mu_0, \kappa_0, \Lambda_0, \nu_0)$, then $\theta_i = (\mu_i, \Sigma_i)$ is the mean vector and covariance matrix of a d-dim Gaussian. Let $F(\theta) = N(\mu_i, \Sigma_i)$, then we draw $\mathbf{x}_i \in \mathbb{R}^d$ from this Gaussian.*

Each observation $\mathbf{x}_i$ has its own parameter θ_i. However, many of the θ_i's are identical, naturally inducing a clustering structure over $\mathbf{x}$. Also note that, even though G is discrete with probability 1, the marginal distribution over $\mathbf{x}$ is smooth if F is a smooth distribution.

Given observations $\mathbf{x}_1 \ldots \mathbf{x}_n, \alpha, H, F$, it is possible to sample using MCMC $\theta_1 \ldots \theta_n$ according to the posterior; This gives a distribution over the clusterings of the data points, from which one can also obtain the most likely number of clusters.