Bayesian Nonparametrics

A brief introduction

Will Grathwohl   Xuechen Li   Eleni Triantafillou

March 2, 2018

Outline

1. What is BNP? Why BNP?
2. Applications
   - Gaussian Processes
   - Dirichlet Processes
   - Indian Buffet Processes

Outline

1. What is BNP? Why BNP?

2. Applications
   - Gaussian Processes
   - Dirichlet Processes
   - Indian Buffet Processes

Statistical Inference

- In general, given some data X, we can assume that:
  data = underlying pattern + noise
- Can interpret $P(X|\theta)$ as $P(data|pattern)$
- The problem of statistical inference then is to figure out the underlying pattern
- Think of a model M as a set of probability measures on X according to some parameters θ. $M = \{P_\theta | \theta \in \mathbf{T}\}$ where $\mathbf{T}$ is the space in which θ takes values in.
- M is **parametric** if T has finite dimension, and **nonparametric** otherwise.

Example: Parametric vs Nonparametric Density Estimation

- Before discussing **Bayesian** nonparametrics, lets consider a simple example of a nonparametric model and compare it to a parametric alternative
- Assume we are given some observed data, shown below and want to perform density estimation

FIGURE 1.1. Density estimation with Gaussians: Maximum likelihood estimation (*left*) and kernel density estimation (*right*).

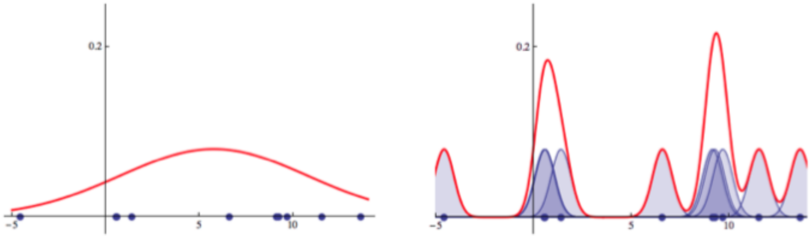


Figure from Lecture Notes on Bayesian Nonparametrics, Peter Orbanz

Example: Parametric vs Nonparametric Density Estimation

In the figure:

- Left: Fit 1 Gaussian to the data. In this case θ consists of a mean and standard deviation (regardless of the number of data points).
- Right: Kernel density estimation. Add a new Gaussian g for each data point x_i, centered at x_i. The density estimate is then $p(x) = \frac{1}{n}\sum_{i=1}^{n} g(x|x_i, \sigma)$
- The Gaussian model is parametric, with 2 degrees of freedom, while the Kernel density estimator is non-parametric, with the number of parameters growing as more data points are observed

Choosing the parameter space?

- How to decide on a parameter space to model data?
- For example, in the left figure below, a reasonable choice for the parameter is a line, so the parameter space $\mathbf{T} \in \mathcal{R}^2$ (slope and offset)
- If the data instead looks nonlinear like in the right subfigure, what is a reasonable parameter space? All possible (differentiable?) nonlinear functions?

FIGURE 1.2. Regression problems: Linear (*left*) and nonlinear (*right*). In either case, we regard the regression function (plotted in blue) as the model parameter.

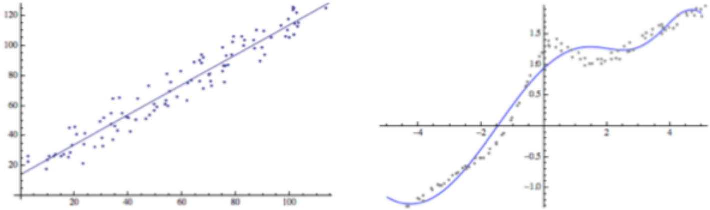


Figure from Lecture Notes on Bayesian Nonparametrics, Peter Orbanz

Bayesian Nonparametrics

- Bayesians treat uncertainty as randomness
- We do not know the parameter underlying the data - treat it as a random variable Θ taking values from $\mathbf{T}$.
- Make a modeling assumption, that $\Theta \sim Q$ for some distribution Q, referred to as the 'prior'.
- A Bayesian model consists of the prior Q and the observational model M as above
- Data is generated as $\Theta \sim Q$, $X1, X2, \ldots | \Theta \sim_{iid} P_\Theta$
- We are then interested in the posterior $Q(\Theta | X_1 = x_1, \ldots, X_n = x_n)$
- **Nonparametric Bayesian Model**: infinite parameter space $\mathbf{T}$. Therefore requires infinite-dimensional distributions for Q and M.

Outline

1. What is BNP? Why BNP?

2. Applications
   - Gaussian Processes
   - Dirichlet Processes
   - Indian Buffet Processes

Gaussian Processes Definition

- Let $\mathbf{T}$ be a space of functions from S to $\mathbb{R}$ where $S \subset \mathbb{R}^d$ (e.g. given d-dimensional points, predict a real-valued target for each one)
- Let Θ be a random element of $\mathbf{T}$. Then it is a random function.
- Let $s \in S$ be a (d-dimensional) point
- Then $\Theta(s)$ is a random variable in $\mathbb{R}$.
- Fixing n points then gives a random vector in $\mathbb{R}^n$: $(\Theta(s_1), \Theta(s_2), \ldots, \Theta(s_n))$
- Consider the quantity $\mu_{s_1,\ldots,s_n} = (\Theta(s_1), \Theta(s_2), \ldots, \Theta(s_n))$
- The distributions defined by μ are called 'finite-dimensional marginals' of μ

Gaussian Processes Definition

- μ is called a **Gaussian Process (GP)** on **T** if for any finite set $S_n = \{s_1, \ldots, s_n\}$, μ_{S_n} is an n-dimensional Gaussian.
- Define $m(s) = \mathbb{E}[\Theta(s)]$ and $k(s_1, s_2) = Cov[\Theta(s_1), \Theta(s_2)]$
- So, if μ is a GP, then each finite-dimensional marginal $\mu_{S_n} \sim \mathcal{N}(m(S_n), k(S_n))$ where

$$m(S_n) = \begin{bmatrix} m(s_1) \\ \ldots \\ m(s_n) \end{bmatrix} \quad \text{and} \quad k(S_n) = \begin{bmatrix} k(s_1, s_1) & \ldots & k(s_1, s_n) \\ \ldots & & \ldots \\ k(s_1, s_n) & \ldots & k(s_1, s_n) \end{bmatrix}$$

Gaussian Process Regression

- Assume we observe $\mathcal{D} = \{(\mathbf{x_i}, y_i)\}_{i=1}^{N} = (\mathbf{X}, \mathbf{y})$ where $\mathbf{x_i}$'s are observations in $\mathbb{R}^d$ and y_i's are targets in $\mathbb{R}$.
- The regression problem: find a function θ mapping observations to targets.
- One approach is to treat this function as a random variable Θ and infer a distribution over functions given data $p(\Theta|\mathbf{X}, \mathbf{y})$
- Since Θ is a random function, we can place a GP prior over it: $\Theta \sim GP(0, K)$.
- We can view the responses as random variables too: $Y_i = \Theta(x_i) + \epsilon_i$ where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ is some random independent noise

Gaussian Process Regression

- We are then looking for the posterior $p(\Theta | Y_1, \ldots, Y_N)$
- We can compute its finite dimensional marginals $p(\Theta(X_{1*}), \ldots, \Theta(X_{N*}) | Y_1, \ldots, Y_N)$ where $\{(X_{i*}, Y_{i*})\}_{i=1}^{N}$ denotes new data
- What is the distribution of the variables that we are conditioning on? Recall that each Y_i is the sum of 2 Gaussians.
- For convenience denote $Y_* = \{\Theta(X_{1*}), \ldots, \Theta(X_{N*}\}$ and $Y = \{Y_1, \ldots, Y_N\}$
- Let K be the covariance of the variables in Y

$$K = \begin{bmatrix} k(x_1, x_1) + \sigma^2 & \ldots & k(x_1, x_n) \\ \ldots & & \ldots \\ k(x_n, x_1) & \ldots & k(x_n, x_n) + \sigma^2 \end{bmatrix}$$

- Also let $K_* = k(Y_*, Y)$, and $K_{**} = k(Y_*, Y_*)$

Gaussian Process Regression

- The covariance of the joint $(\Theta(X_{1*}), \ldots, \Theta(X_{N*}), Y_1, \ldots, Y_N)$ is

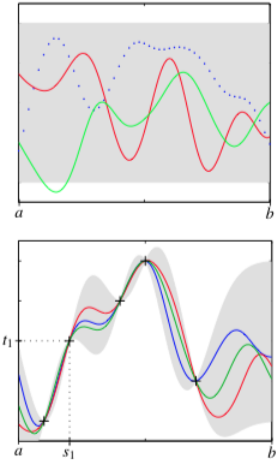
$$\begin{bmatrix} K & K* \\ K*^T & K** \end{bmatrix}$$

- Finally there is a lemma that given a partition (A, B) with $X = (X_A, X_B)$ Gaussian in $\mathbb{R}^d = \mathbb{R}^A x \mathbb{R}^B$, computes the conditional distribution $X_A|(X_B = x_B)$
- Using this lemma we find that the posterior of a $GP(0, K)$ under the observations $Y_i = \Theta(x_i) + \epsilon_i$ is again Gaussian. Its finite-dimensional marginal distributions at any finite set $\{X_{*1}, \ldots, X_{*N}\}$ is the Gaussian with mean and covariance defined below

$$\mathbb{E}[Y_*|Y] = K_*(K + \sigma^2\mathbf{I})^{-1}Y$$
$$Cov[Y_*|Y] = K_{**} - K_*^T(K + \sigma^2\mathbf{I})^{-1}K_*$$

Posterior GP

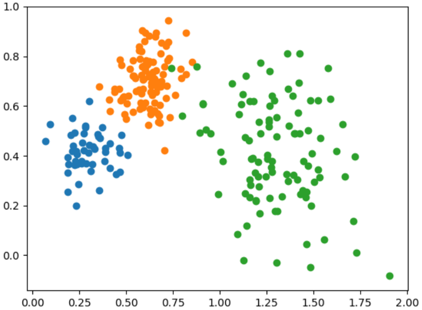
- So we've seen that the posterior $p(\Theta|data)$ is also a Gaussian process (distribution over functions).
- This can be thought of as quantifying prediction uncertainty.



Dirichlet Processes Motivation

- Consider the task of clustering with a *finite* mixture model.
- Let $\theta_1, ..., \theta_k$ be parameters associated with each cluster.
- Let $c_1, ..., c_k$ be cluster weightings, i.e. $\sum_i c_i = 1$ and $\forall i, c_i \geq 0$.
- Assuming continuous data, the mixture density is:

$$p(x) = \sum_i c_i p(x|\theta_i)$$



Dirichlet Processes Motivation

- A **Bayesian Mixture** treats c_i and θ_i as random variables.
- A simple way to instantiate c_i and θ_i is to sample them i.i.d. from fixed distributions $p(c)$ and $p(\theta)$
- To ensure the cluster weightings c_i are valid ($\sum_i c_i = 1$ and $\forall i, c_i \geq 0$), we need apply normalization.
- However, naive normalization schemes (e.g. divide by sum, softmax) fail when there are infinitely many positive i.i.d. variables.
- The Dirichlet Process (DP) solves this problem and extends Bayesian mixtures to infinite components.

Dirichlet Processes Stick-Breaking Construction

- An intuitive construction of the DP is via **stick-breaking**.
- Consider a stick of unit length, we break it into infinite pieces.
- The length of each piece would be the weighting for each cluster.
- To do this, we sample *ratio* v_i from a distribution on $[0, 1]$ each time.
- We take v_i of the stick and leave the rest $1 - v_i$ for next iteration.
- The stick lengths (cluster weightings) are $c_i = (1 - \sum_{j=1}^{i-1} c_j)v_i$.

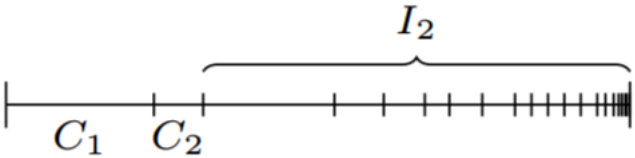


Figure from Lecture Notes on Bayesian Nonparametrics, Peter Orbanz

Dirichlet Processes Stick-Breaking Construction

Definition

If $\alpha > 0$ and G_0 is a probability measure on the parameter space Ω_θ, the random discrete probability measure Θ generate by:

$$V_1, V_2, \ldots \sim_{iid} Beta(1, \alpha)$$

$$C_k := V_k \prod_{j=1}^{k-1} (1 - V_j)$$

$$\Theta_1, \Theta_2, \ldots \sim_{iid} G_0$$

is called a Dirichlet Process (DP), with base measure G_0 and concentration parameter α, denoted by $DP(\alpha, G_0)$.

Dirichlet Processes Posterior

- Assume true data generating process first generates a discrete measure from DP, i.e. $G \sim DP(\alpha, G_0)$.
- Assume observations are generated from G i.i.d., i.e. $\theta_1, ..., \theta_n \sim_{iid} G$.
- It is shown (by Ferguson) that the posterior over G is also a DP:

$$p(G|\theta_1, ..., \theta_n) = DP(\alpha + n, \frac{\alpha G_0 + \sum_{i=1}^{n} \delta_{\theta_i}}{\alpha + n})$$

.

- δ_θ denotes the dirac delta (point mass) at θ.
- Conjugacy makes posterior inference easy for DP.

- Chinese Restaurant Process (CRP) is another interpretation of DP.
- Recall DP deals with the task of clustering.
- In clustering, if we abstract away the details of each cluster and only care about the cluster indices, we end up defining a partition.
- For instance, the clustering $(\{X_1, X_2, X_5\}, \{X_3\}, \{X_4\})$ defines the partition $(\{1, 2, 5\}, \{3\}, \{4\})$.
- The partition can also be extended to (countably) infinite sets.

Dirichlet Processes and Chinese Restaurant Processes

- CRP defines distribution on partitions of the naturals.
- More formally, $CRP(\alpha)$ defines a generative process:
- For $n = 1, 2, 3, ...$
  - insert n into an existing block Ψ_k with probability $\frac{|\Psi_k|}{\alpha+(n-1)}$
  - create a new block with only n with probability $\frac{\alpha}{\alpha+(n-1)}$
- CRP does not have a base measure parameter G_0 because we abstract away the "location" of clusters.
- One intuition is that each time a person indexed by n comes into a restaurant and decides to sit at a random table with probability proportional to the number of people seated or α if no one is seated.

Dirichlet Process Mixture Models

- We can add a further hierarchy to DPs to create an infinite mixture model.
- Such models are called Dirichlet Process Mixtures (DPM).
- Assume the true data generating process is:

$$
\begin{aligned}
G &\sim DP(\alpha, G_0) \\
\theta_i &\sim_{iid} G \\
x_i &\sim_{iid} p(x|\theta_i)
\end{aligned}
$$

- In this case, θ_i is a local latent variable of the observed x_i.

Indian Buffet Processes

Dirichlet Processes give us a distribution over potentially infinite partitions $\{\{e_1, e_4, e_2\}, \{e_3, e_5\}, \{e_6\}, \ldots\}$ where each element e_i belongs to exactly 1 partition.
What if elements could belong to multiple groups? Enter the Indian Buffet Process.

$$\text{Partition} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix} \quad \text{Multiple Groups} = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

Simple when number of groups is fixed, but what if number of groups is infinite?

Indian Buffet Processes

Indian restaurant interpretation. Dishes = Groups. Assume an infinite number of dishes ordered arbitrarily. Has 1 parameter α.

- Customer 1 takes first Poisson(α) dishes
- Customer i:
  - takes dish k with probability $= \frac{\#\text{ times } k \text{ previously chosen}}{i}$
  - takes Poisson$\left(\frac{\alpha}{i}\right)$ new dishes

Like the Chinese Restaurant Process, this process is exchangeable in the ordering of the customers. Also in the dishes!

Alternate Generative Process: $X_{ij} = \mathrm{I}[\text{customer i takes dish j}]$.

- $w_j \sim \text{Beta}(1, \alpha/j)$
- $X_{ij} \sim \text{Bernoulli}(w_j)$

Applications of the IBP: Latent Feature Models

Assumes datapoint X_i is dependent on a finite number of unobserved attributes z_j where there are an infinite number of potential z_j. X_i could be the set of movies that user i has viewed and each z_j could be a type of movie. So X_i is determined by which types of movies user i likes.

Definitions:

- $X_{ij} = \mathrm{I}[\text{user i has watched movie j}], i \in [1, N], j \in [1, D]$
- $Z_{ij} = \mathrm{I}[\text{user i likes movie type } j], i \in [1, N], j \in [1, \infty]$
- $\phi_{ij} =$ movie i's relation to type j
- $X_{ij} = \sum_{k=1}^{\infty} Z_{ik}\phi_{jk} + \epsilon_{ij}, \epsilon_{ij} \sim p(\epsilon_{ij})$

Inference performed via MCMC or with variational inference and truncated IBP posterior with maximum T features.