Introduction to Nonparametric Bayes

Yuexin Wu
yuexinw@andrew.cmu.edu
Carnegie Mellon University

November 1, 2016
Outline

1 Introduction

2 Motivation
   - Why nonparametric methods
   - Roadmap

3 Application
   - Dirichlet Process
     - Sample from DP
     - Infer Posterior
     - Clustering and CDF Estimation
   - Gaussian Process
     - Sample from GP
     - Infer Posterior

4 Conclusion
Plan

1. Introduction

2. Motivation
   - Why nonparametric methods
   - Roadmap

3. Application
   - Dirichlet Process
     - Sample from DP
     - Infer Posterior
     - Clustering and CDF Estimation
   - Gaussian Process
     - Sample from GP
     - Infer Posterior

4. Conclusion
Nonparametric:
- Basically means models which have parameters of infinite dimensions
- # parameters usually grows with the size of data-points

Bayes:
- \( P[\theta|X] \propto P[X|\theta]P[\theta] \)
- Generative models: need specify prior \( P[\theta] \)
- know how to draw from it
Plan

1 Introduction

2 Motivation
   - Why nonparametric methods
   - Roadmap

3 Application
   - Dirichlet Process
     - Sample from DP
     - Infer Posterior
     - Clustering and CDF Estimation
   - Gaussian Process
     - Sample from GP
     - Infer Posterior

4 Conclusion
Motivation

- Bayesians vs. Frequentists
  - Choice of philosophy
  - Main advantage: obtain the posterior $P[\theta | X]$ instead of a point estimation
- Nonparametric
  - Natural choice in some cases
  - Data adaptive power
Why nonparametric methods

- Natural choice (infinite clusters)

Figure: Dendrogram (new species may come)
Why nonparametric methods

- Data adaptive

**Figure**: Parametric and Nonparametric Regression
Why nonparametric methods

- Data adaptive

**Figure**: Parametric and Nonparametric density estimation
Motivation

Roadmap

<table>
<thead>
<tr>
<th>Problem</th>
<th>Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clustering</td>
<td>Dirichlet process</td>
</tr>
<tr>
<td>Estimating cdf</td>
<td>Dirichlet process</td>
</tr>
<tr>
<td>Regression</td>
<td>Gaussian process</td>
</tr>
</tbody>
</table>

- Gaussian Process: a distribution on functions.
Plan

1. Introduction
2. Motivation
   - Why nonparametric methods
   - Roadmap
3. Application
   - Dirichlet Process
     - Sample from DP
     - Infer Posterior
     - Clustering and CDF Estimation
   - Gaussian Process
     - Sample from GP
     - Infer Posterior
4. Conclusion
Definition for DP

Definition: $F \sim DP(\alpha, F_0)$ if for any partition $(A_1, \ldots, A_k)$ of $\text{dom}(F_0)$, $(F(A_1), \ldots, F(A_k))$ follows a Dirichlet distribution:

$$\text{Dir}(\alpha F_0(A_1), \ldots, \alpha F_0(A_k)).$$

Different ways to derive:
- Stick-breaking Process
- Chinese Restaurant Process
Influence of $\alpha$ and $F_0$

$F_0 = N(0, \sigma^2)$

- Up to down: $\alpha = [1, 100]$
- Left to right: $\sigma^2 = [1, 10, 100]$

Figure: Sample of $F$ from Dirichlet Process
Draw from DP: Stick-breaking Process

- Draw $s_1, s_2, \ldots$ i.i.d. from $F_0$
- Draw $V_1, V_2, \ldots$ from $\text{Beta}(1, \alpha)$
- Let $w_1 = V_1$ and $w_j = V_j \prod_{i=1}^{j-1} (1 - V_i)$ for $j = 2, 3, \ldots$
- $F$ is the discrete distribution s.t. $F = \sum_{j=1}^{\infty} w_j \delta_{s_j}$
- Can be shown $\mathbb{E}(F) = F_0$

**Figure: Stick-breaking Process**
Sample from DP: Chinese Restaurant Process

- If only care about $X_i$, more computationally friendly
- Draw $X_1 \sim F_0$
- For $i = 2, \ldots$

\[
X_i | X_1, \ldots, X_{i-1} = \begin{cases} 
X \sim F_{i-1} & \text{with prob. } \frac{i-1}{i+\alpha-1} \\
X \sim F_0 & \text{with prob. } \frac{\alpha}{i+\alpha-1}
\end{cases}
\]

where $F_{i-1}$ is the empirical distribution of $X_1, \ldots, X_{i-1}$
Chinese Restaurant Process

- Empirical distribution $F_{i-1}$ are likely to have ties
- Let $X_1^*, X_2^*, \ldots$ denote unique values
- Demo

Figure: Chinese Restaurant Process
Chinese Restaurant Process

- Expected number of new tables at \( n \)-th customers is

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

**Proof Sketch**

Let \( Y_i \) denotes the event that customer \( i \) occupying a new table. Then

\[
\mathbb{P}[Y_i = 1] = \frac{\alpha}{i - 1 + \alpha}.
\]

Thus

\[
\mathbb{E}[k_n | \alpha] = \sum_i \mathbb{E}[Y_i | \alpha]
\]

\[
= \alpha \sum_i \frac{1}{\alpha + i - 1}
\]

\[
= O(\alpha H_n)
\]

\[
= O(\alpha \log n)
\]
Thereom for DP posterior

Let $X_1, \ldots, X_n \sim F$ and let $F$ have prior $\pi = DP(\alpha, F_0)$. Then the posterior of $F$ given $X_1, \ldots, X_n$ is $DP(\alpha + n, \bar{F}_n)$ where

$$\bar{F}_n = \frac{n}{n + \alpha} F_n + \frac{\alpha}{n + \alpha} F_0.$$ 

Thereom for sampling from DP posterior

For a new point $X$,

$$X | X_1, \ldots, X_n = \begin{cases} X \sim F_n \text{ with prob. } \frac{n}{n+\alpha} \\ X \sim F_0 \text{ with prob. } \frac{\alpha}{n+\alpha} \end{cases}$$
Infer Posterior

Figure: Sample from DP Posterior
Clustering and CDF Estimation

- **Clustering:**
  - Exactly the Chinese Restaurant Process
  - Cluster a new point using *theorem for sampling from DP posterior*

- **Density estimation:**
  - Mixture model for $X_1, \ldots, X_n$
  - $F \sim DP(\alpha, F_0)$
  - $\theta_1, \ldots, \theta_n \sim F$
  - $X_i|\theta_i \sim f(x|\theta_i), i = 1, \ldots, n$
  - Compute the cdf for a new point using Gibbs sampling (slow)
  - Refer to MacEachern for a more efficient algorithm.
Gaussian Process

Distribution over functions $\Theta : \mathbb{R}^d \rightarrow \mathbb{R}$

Definition for GP

A stochastic process indexed by $s \in S \subset \mathbb{R}^d$ is a Gaussian Process if for each $s_1, \ldots, s_n \in S$ vector, the drawn function $\Theta$ satisfies that

$$(\Theta(s_1), \ldots, \Theta(s_n))$$

is normally distributed:

$$(\Theta(s_1), \ldots, \Theta(s_n)) \sim \mathcal{N}(\mu(s), K(s))$$

where $K_{ij}(s) = K(s_i, s_j)$ is a Mercer kernel.
Sample from GP

- Take randomly sampled infinite sequence of \( s_1, s_2, \ldots \) from \( S \)
- Sample results using distribution \( \mathcal{N}(\mu(s), K(s)) \)

**Figure: GP sampling**

- Zero mean \( \mu(\cdot) \), with \( K(s_1, s_2) = \exp\left(-|s_1 - s_2|^2/(2\tau^2)\right) \)
- From left to right \( \tau^2 = [0.5, 2, 10] \).
Consider regression function $\Theta : S \rightarrow \mathbb{R}$ and noisy observations:

$$X_i = \Theta(s_i) + \epsilon_i$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$

Then for a new given point $s_{n+1}$, the covariance matrix of $(X_1, \ldots, X_{n+1})$ should be

$$
\begin{pmatrix}
K + \sigma^2 I & k \\
k^T & K(s_{n+1}, s_{n+1}) + \sigma^2
\end{pmatrix}
$$

Theorem for conditional distribution of $X_{n+1}$

The conditional distribution with zero mean prior $\mu(\cdot) = 0$ of value $X_{n+1}$ under new point $s_{n+1}$ is

$$X_{n+1}|X_{1:n}, s_{1:n+1} \sim \mathcal{N}(k^T (K + \sigma^2 I)^{-1} Y, k(x_{n+1}, x_{n+1}) + \sigma^2 - k^T (K + \sigma^2 I) k)$$
**Example for GP Posterior**

Zero mean \( \mu(\cdot) \), with

\[
K(s_1, s_2) = \exp\left(-\frac{|s_1 - s_2|^2}{2\tau^2}\right), \quad \tau = 0.1
\]

noise level \( \sigma = 1 \)

From left to right, training points take \([10, 20, 40]\)
Plan

1. Introduction

2. Motivation
   - Why nonparametric methods
   - Roadmap

3. Application
   - Dirichlet Process
     - Sample from DP
     - Infer Posterior
     - Clustering and CDF Estimation
   - Gaussian Process
     - Sample from GP
     - Infer Posterior

4. Conclusion
Take-home message

- Provides a posterior on parameters
- Data-adaptive estimation
- Might be the natural option when dealing with the infinite clustering problem
- Nevertheless, computationally expensive
Chuong B. Do: Gaussian Process, 
https://see.stanford.edu/materials/aimlcs229/cs229-gp.pdf

Larry Wasserman: Nonparametric Bayes, tutorials from 10-702
http://www.stat.cmu.edu/~larry/=sml/nonparbayes.pdf

Peter Orbanz: Lecture notes on bayesian nonparametrics,
http://ce.sharif.edu/courses/93-94/2/ce957-1/resources/root/References/porbanz_BNP.pdf

MacEachern: Estimating normal means with a conjugate style
Dirichlet process prior
Thanks!