Bayesian Nonparametrics

Christof Seiler

Stanford University, Spring 2016, STATS 205

Overview

Last week:

- Estimating functions frequentist style

Today:

- The Bayesian approach

Introduction

- In the **Bayesian** approach,
  - the true parameter θ is believed to be random variable
- In the **Frequentist** approach,
  - the true parameter θ is believed to be one value,
  - the randomness comes from sampling error

Parametric Bayesian

- In parametric Bayesian inference we have a model

$$M = \{f(y|\theta) : \theta \in \Theta\}$$

and data

$$Y_1, \dots, Y_n \sim f(y|\theta)$$

- We put a **prior** distribution $\pi(\theta)$ on the parameter θ and compute the **posterior** distribution using Bayes' rule

$$\pi(\theta|y) = \frac{\prod_{i=1}^n f(y_i|\theta)\pi(\theta)}{m(y)}$$

- with marginal distribution

$$m(y) = m(y_1, \dots, y_n) = \int f(y_1, \dots, y_n|\theta)\pi(\theta)d\theta$$

$$= \int \prod_{i=1}^n f(y_i|\theta)\pi(\theta)d\theta$$

Parametric Bayesian

- We can write a generative model as sampling parameters from the prior
$$\theta \sim \pi$$
- And then sampling data from the **likelihood** function
$$Y_1, \ldots, Y_n | \theta \sim f(y|\theta)$$
- We can use the posterior distribution to compute the posterior mean of θ
$$\bar{\theta} = \mathrm{E}(\theta|y) = \int \theta \pi(\theta|y) d\theta$$
- Also we can summarize the posterior by drawing a large sample
$$\theta_1, \ldots, \theta_N \sim \pi(\theta|y)$$
and plotting the samples

Parametric Bayesian

- Posterior distribution

$$\pi(\theta|y) = \frac{\prod_{i=1}^{n} f(y_i|\theta)\pi(\theta)}{m(y)}$$

- Even if we don't know $m(y)$, we can use tools like Markov chain Monte Carlo (MCMC) to **draw samples** from the posterior and plot them
- So even without being able to evaluate the posterior distribution, through sampling we **know** the posterior if we sample infinitly many times
- And approximately if we get a finite amount of samples

Nonparametric Bayesian

- We replace the finite dimensional model

$$\{f(y|\theta) : \theta \in \Theta\}$$

with an infinite dimensional model such as

$$\mathcal{F} = \left\{ f : \int (f''(y))^2 dy < \infty \right\}$$

- Surprisingly, sometimes neither the prior nor the posterior have a density function
- But the posterior is still defined

Nonparametric Bayesian

Some questions:

1. How do we construct a prior π on an infinite dimensional set $\mathcal{F}$?
2. How do we compute the posterior?
3. How do we draw random samples from the posterior?

Distributions on Infinite Dimensional Spaces

- We will need to put a prior π on an infinite dimensional space
- For example, suppose we observe

$$X_1, \ldots, X_n \sim F$$

  with unkown distribution F
- We put prior π on set of all distributions $\mathcal{F}$
- In many cases, we cannot explicitly write down a formula for π
- How can we describe a distribution π in another way than writing it down?
- If we know how to draw from π we can get many samples and then even without knowing the formula for π we can plot it

Distributions on Infinite Dimensional Spaces

- The idea: find an algorithm to sample from this model

$$F \sim \pi$$
$$X_1, \dots, X_n | F \sim F$$

- If we have such an algorithms then it is like being able to actually write down an explicit formula
- After we observe the data $X = (X_1, \dots, X_n)$, we are interested in the posterior distribution
- The same idea here, instead of writing down a forumla we describe an algorithm to sample for the posterior distribution

Estimating a Cumulative Distribution Function

- Suppose we observe $X_1, \ldots, X_n$ from an unkown distribution F ($X_i \in \mathbb{R}$)
- The usual frequentist estimate of F is the empirical distribution function

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x)$$

- To estimate F from a Bayesian perspective we put a prior on π on the set of all $\mathcal{F}$
- Compute the posterior on $\mathcal{F}$ given $X = \{X_1, \ldots, X_n\}$
- Such a prior was invented by Thomas Ferguson in 1973

Estimating a Cumulative Distribution Function

- The prior has two parameter: F_0 and α denoted by $\text{DP}(\alpha, F_0)$
- F_0 is a distribution function and should be thought of as a prior guess of F
- The number α controls how tightly concentrated the prior is around F_0
- The model is

$$F \sim \text{DP}(\alpha, F_0)$$
$$X_1, \dots, X_n | F \sim F$$

- But how to draw samples from this model?

Estimating a Cumulative Distribution Function

- First to draw samples from the prior $\mathrm{DP}(\alpha, F_0)$, we follow four steps

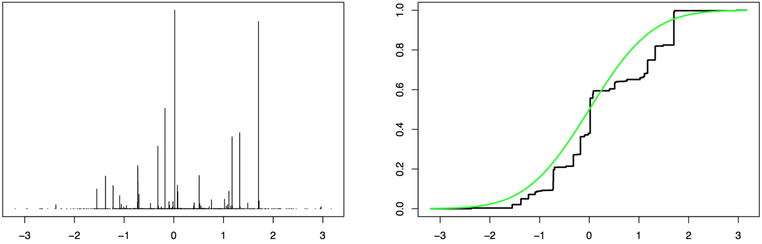
1. Draw $s_1, s_2, \ldots$ independently from F_0
2. Draw $V_1, V_2, \cdots \sim \mathrm{Beta}(1, \alpha)$
3. **Stick breaking process**: Let $w_1 = V_1$ and $w_j = V_j \prod_{i=1}^{j-1}(1 - V_i)$ for $j = 2, 3, \ldots$

Estimating a Cumulative Distribution Function

- **Stick breaking process:**
- Imagine a stick of unit length
- Then w_1 is obtained by breaking the stick at the random point V_1
- The stick has now length $1 - V_1$
- The second weight w_2 is obtained by breaking a proportion V_2 from the remaining stick
- The process continues and generates the whole sequence of weights $w_1, w_2, \ldots$

Estimating a Cumulative Distribution Function

4. Let F be the discrete distribution that puts mass w_j at s_j, that is, $F = \sum_{j=1}^{\infty} w_j \delta_{s_j}$ where δ_{s_j} is a point mass at s_j



Source: Wasserman (left: weights; right: F_0 and random draw)

- F is a discrete distribution
- The Dirichlet process is a generalization of the Dirichlet distribution

Estimating a Cumulative Distribution Function

- To sample from the posterior, we need the following theorem
- Let F_n be the empirical distribution
- **Theorem:** Let $X_1, \ldots, X_n \sim F$. Let F have prior $\pi = \mathrm{DP}(\alpha, F_0)$. Then the posterior π for F given $X_1, \ldots, X_n$ is $\mathrm{DP}(\alpha + n, \bar{F}_n)$ where

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

Estimating a Cumulative Distribution Function

- Since the posterior is again a Dirichlet process, we can sample from it as we did the prior
- We only replace α with $\alpha + n$ and we replace F_0 with $\bar{F}_n$
- Thus the posterior mean is $\bar{F}_n$ is a convex combination of the empirical distribution and the prior guess F_0
- To explore the posterior distribution, we could draw many random distribution functions form the posterior
- We could then numerically construct two functions L_n and U_n such that

$$\pi(L_n(x) \leq F(x) \leq U_n(x) \text{ for all } x | X_1, \ldots, X_n) = 1 - \alpha$$

- This is a Bayesian credible interval for F
- When n is large then $\bar{F}_n \approx F_n$

Density Estimation

- Let $X_1, \ldots, X_n \sim F$ where F has density f and $X_i \in \mathbb{R}$
- Our goal is to estimate f
- The Dirichlet process is not useful prior for this because it produces discrete distributions
- But we can make a modification
- The most popular frequentist estimator is the kernel estimator

$$\widehat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - X_i}{h}\right)$$

  with kernel K and bandwith h
- A related method is the mixture model

$$f(x) = \sum_{j=1}^{k} w_j f(x; \theta_j)$$

Density Estimation

- For example, if $f(x;\theta)$ is normal then $\theta = (\mu, \sigma^2)$
- The kernel estimator can be thought of as a mixture with k components
- In the Bayesian approach we would put a prior on $\theta_1, \dots, \theta_k$, on $w_1, \dots, w_k$, and on k
- Recently, it became more popular to use an infinite mixture model
$$f(x) = \sum_{j=1}^{\infty} w_j f(x;\theta_j)$$
- As prior for the paramters we could take $\theta_1, \theta_2, \dots$ to be drawn from some F_0 and
- We could take $w_1, w_2, \dots$ to be drawn from the stick breaking prior
- This is known as the **Dirichlet process mixture model**

Density Estimation

- This is the same as the random distribution $F \sim \mathrm{DP}(\alpha, F_0)$ which had the form $F = \sum_{j=1}^{\infty} w_j \delta_{\theta_j}$
- Except that the point mass distribution δ_{θ_j} are replaced by smooth densities $f(x|\theta_j)$
- The model is

$$\begin{aligned} F &\sim \mathrm{DP}(\alpha, F_0) \\ \theta_1, \ldots, \theta_n | F &\sim F \\ X_j|\theta_j &\sim f(x|\theta_j), \quad j = 1, \ldots, n \end{aligned}$$

- The beauty of this model is that the discreteness of F automatically creates a clustering of the θ_j's
- In other words, we have implicitly created a prior on k, the number of distinct θ_j's

References

- Wasserman Lecture Notes
- van der Vaart Lecture Notes
- Müller, Quintana, Jara, and Hanson (2015). Bayesian Nonparametric Data Analysis