Bayesian nonparametrics

Presenter: Dr. Frank Wood

February 17, 2010

Bayesian nonparametrics (BNP)

Motivation

- Never know the full data generating mechanism
  - Want to make the most general assumptions.
  - Guard against possible gross model misspecification.

Bayesian nonparametrics

- Parameters can be described by functions or other infinite dimensional objects
  - Cumulative distribution function (CDF)
  - Density function.
  - Nonparametric regression function
  - Unknown link function in generalized linear model (GLM)

BNP

Challenges

- Construction of prior distribution involves specifying appropriate probability measure on function spaces.
- Knowledge of minute details of such distributions typically not available.

BNP

Challenges

- Construction of prior distribution involves specifying appropriate probability measure on function spaces.
- Knowledge of minute details of such distributions typically not available.

Standard Approach

- Prior typically chosen for computational practicality
- Key prior parameters might be chosen subjectively
- Important: prior should have large support
- Large support of the prior helps the posterior distribution to have good frequentist properties in large samples.

BNP Posterior consistency

Posterior consistency

- Basic frequentist validation of a Bayesian estimation procedure
- In limit of infinite data, does the posterior distribution converge to the true underlying parameter?
- Lack of consistency is *undesirable*.
- Rate of convergence can be used to distinguish different estimation procedures
  - How quickly can a ball around the true value shrink while retaining almost all of the posterior probability?

BNP Posterior consistency

Posterior consistency

- Basic frequentist validation of a Bayesian estimation procedure
- In limit of infinite data, does the posterior distribution converge to the true underlying parameter?
- Lack of consistency is *undesirable*.
- Rate of convergence can be used to distinguish different estimation procedures
  - How quickly can a ball around the true value shrink while retaining almost all of the posterior probability?

Posterior consistency of BNP models is an area of active research

BNP models

Examples

- Dirichlet process
  - Mixtures of Dirichlet processes [Antoniak, 1974, MacEachern and Muller, 1998]
  - Dirichlet process mixture [Escobar and West, 1995]
- Gaussian process
  - Gaussian processes for machine learning (book) [Rasmussen and Williams, 2006]
- Indian Buffet process, Chinese restaurant process, Beta process, Dependent Dirichlet process, Hierarchical Dirichlet Process (HDP), HDP-HMM, HDP-LDA, Sequence Memoizer, etc.

BNP models

To Start

- Look at role of Dirichlet process
- Discuss most important properties
- Informally talk about posterior convergence in such models

The Dirichlet Process

Motivation

Say we'd like to estimate a probability measure (or CDF) on the real line, with i.i.d. observations from it where the CDF is completely arbitrary.

The Dirichlet Process

Motivation

Say we'd like to estimate a probability measure (or CDF) on the real line, with i.i.d. observations from it where the CDF is completely arbitrary.

Classical approach

Build a nonparametric estimate of the CDF directly from the observations.

The Dirichlet Process

Motivation

Say we'd like to estimate a probability measure (or CDF) on the real line, with i.i.d. observations from it where the CDF is completely arbitrary.

Classical approach

Build a nonparametric estimate of the CDF directly from the observations.

Bayesian approach

Need a prior for the CDF (or for a random probability measure) and methods (algorthims, etc.) to estimate the posterior distribution.

Closest Parametric Analog

Multinomial / Dirichlet

- Multinomial distribution specifies an arbitrary probability distribution on the sample space of finitely many integers.
- Multinomial model can be derived from an arbitrary distribution by grouping the data in finitely man categories.
- Formalism
  - Let $(\pi_1, \ldots, \pi_k)$ be the probabilities of the categories with frequencies $n_1, \ldots, n_k$. The multinomial likelihood is proportional to $\pi_1^{n_1}, \ldots, \pi_k^{n_k}$.
  - The finite-dimensional Dirichlet prior has density proportional to $\pi_1^{c_1 - 1}, \ldots, \pi_k^{c_k - 1}$
  - The posterior has density proportional to $\pi_1^{n_1 + c_1 - 1}, \ldots, \pi_k^{n_k + c_k - 1}$ which is again Dirichlet.

Definition

The Dirichlet *process* is a probability distribution on the space of probability measures which induces finite-dimensional Dirichlet distributions when the data are grouped.

- For any measureable partition $\{B_1, \dots B_k\}$ of $\mathbb{R}$ the probability vector $(P(B_1), \dots, P(B_k))$ is a finite-dimensional Dirichlet distribution.
- This means that the parameters of the finite-dimensional Dirichlet dist. must be special.
- For instance, the joint distribution of r $(P(B_1), \dots, P(B_k))$ must agree with the joint distribution $(P(A_1), \dots, P(A_k))$ when $\{A_1, \dots A_k\}$ is finer than $\{B_1, \dots B_k\}$ since for any i, $P(B_i)$ would be the sum of some $P(A_j)$

Definition

A finite-dimensional Dirichlet distribution property is that summing the probabilities of different partitions gives rise to a new Dirichlet distribution whose parameters corresponding to the summed partitions are added. Let $\alpha(B)$ be the parameter corresponding to $P(B)$ in the specified Dirichlet joint distribution, it follows that $\alpha(\cdot)$ must be an additive set function.

Let α be a finite measure on a given Polish space $\mathfrak{X}$. A random measure P on $\mathfrak{X}$ is called a Dirichlet process if for every finite measureable partition $\{B_1, \ldots, B_k\}$ of $\mathfrak{X}$, the joint distribution of $(P(B_1), \ldots, P(B_k))$ is a k-dimensional Dirichlet distribution with parameters $\alpha(B_1), \ldots, \alpha(B_k)$

We call α the base measure of the Dirichlet process and denote the corresponding Dirichlet process $\mathcal{D}_\alpha$.

A Little Problem

Even when α is a measure, it still isn't clear that P is a probability measure, i.e. that it sums to one.

Several strategies could be taken towards demonstrating this, we will call them various constructions of the DP.

- Naive
- Countable generator
- Normalization

Can we use Kolmogorov's Existence Theorem ? (in short, no)

Bibliography

Antoniak, C. (1974).
Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems.
The Annals of Statistics, 2:1152–1174.

Escobar, M. D. and West, M. (1995).
Bayesian density estimation and inference using mixtures.
Journal of the American Statistical Association, 90:577–588.

MacEachern, S. and Muller, P. (1998).
Estimating mixture of Dirichlet process models.
Journal of Computational and Graphical Statistics, 7:223–238.

Rasmussen, C. and Williams, C. (2006).
Gaussian processes for machine learning.
Springer.