Bayesian Nonparametrics (Part 2)

Christof Seiler

Stanford University, Spring 2016, STATS 205

Overview

Last time:

- Bayesian estimating of CDF's and densities

Today:

- Example of Bayesian nonparametrics in practice
- Bayesian nonlinear regression

Nonlinear Regression

- Consider the nonparametric regression model

$$Y_i = r(x_i) + \epsilon_i, \quad i = 1, \ldots, n, \quad \epsilon \sim N(0, \sigma^2)$$

- The frequentist kernel estimator for r is

$$\widehat{r}(x) = \frac{\sum_{i=1}^{n} Y_i K\left(\frac{\|x - x_i\|}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{\|x - x_i\|}{h}\right)}$$

  with kernel K and bandwith h
- The Bayesian version requires a prior on π on the set of all regression functions $\mathcal{R}$
- A common choise is the **Gaussian process prior**

Nonlinear Regression

- A stochastic process $r(x)$ indexed by $x \in \mathcal{X} \subset \mathbb{R}^d$ is a Gaussian process if for each $x_1, \ldots, x_n \in \mathcal{X}$

$$r = \begin{bmatrix} r(x_i) \\ r(x_2) \\ \vdots \\ r(x_n) \end{bmatrix} \sim N(\mu(x), K(x))$$

Nonlinear Regression

- Assume that $\mu = 0$, then for $x_1, x_2, \ldots, x_n$, the Gaussian process prior is

$$\pi(r) = (2\pi)^{-n/2}|K|^{-1/2} \exp\left(-\frac{1}{2} r^T K^{-1} r\right)$$

- The log-likelihood is

$$-\log f(y|r) = \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - r(x_i))^2 + \text{const}$$

- The log-posterior is

$$-\log f(y|r) - \log \pi(r) = \frac{1}{2\sigma^2} \|y - r\|_2^2 + \frac{1}{2} r^T K^{-1} r + \text{const}$$

Nonlinear Regression

- What functions have high probability according to the Gaussian process prior?
- Consider the eigenvector v of K with eigenvalue λ

$$Kv = \lambda v$$

- Then

$$\frac{1}{\lambda} = v^T K^{-1} v$$

- The prior favors $r^T K^{-1} r$ being small
- Thus eigenfunctions with large eigenvalues are favored by the prior
- These corresponds to smooth functions
- The eigenfunctions that are very wiggly correspond to small eigenvalues

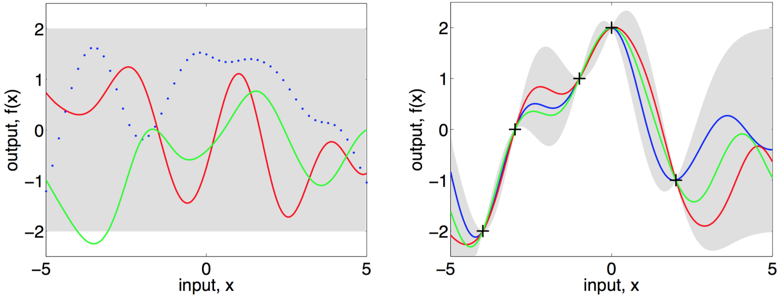
Nonlinear Regression

- The posterior mean is

$$\widehat{r} = \mathsf{E}(r|Y) = K(K + \sigma^2 I)^{-1} Y$$

- We see that this is a linear smoother

Nonlinear Regression



(a), prior (b), posterior

Source: Rasmussen and Williams (2006)

Nonlinear Regression

- To compute predictive distribution for a new point $Y_{n+1} = r(x_{n+1}) + \epsilon_{n+1}$
- The marginal distribution is $(Y_1, \ldots, Y_n)^T \sim N(0, (K + \sigma^2 I))$
- Let $\boldsymbol{k}$ be the vector $(K(x_1, x_{n+1}), \ldots, K(x_n, x_{n+1}))^T$
- Then $(Y_1, \ldots, Y_{n+1})^T$ is jointly Gaussian with covariance

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

- The conditional distribution of Y_{n+1} is

$$Y_{n+1} | Y_{1:n}, x_{1:n} \sim N(\mu_{n+1}, \sigma^2_{n+1})$$

with

$$\mu_{n+1} = \boldsymbol{k}^T (K + \sigma^2 I)^{-1} y$$

$$\sigma^2_{n+1} = K(x_{n+1}, x_{n+1}) + \sigma^2 - \boldsymbol{k}^T (K + \sigma^2 I)^{-1} \boldsymbol{k})$$

References

- Wasserman Lecture Notes
- Rasmussen and Williams (2006), Gaussian Processes for Machine Learning