A Bayesian approach for Non Parametric Regression

MTH673A Project

Arvind Singh Yadav

Department of Mathematics and Statistics
Indian Institute of Technology Kanpur

February 19, 2022

Overview

1. Setup
2. Mercer Theorem
3. An Example of Kernel
4. Non Parametric Regression using Gaussian Processes
5. Estimation
6. Prediction using Gaussian Process

Setup

- We are observing X_i, Y_i and have a model $\mathbb{M} = \{p(y|m) : m \in \Theta\}$ where $i = 1(1)n$. We put a prior $\pi(m)$ on the function m.
- Compute the posterior distribution using Bayes' rule
- $\pi(m|y) = \frac{L(m)\pi(m)}{m(Y)}$
- $Y = (Y_1, ...Y_n)$, $L(m) = \prod p(y_i|m)$ is the likelihood function
- m(y) is the marginal distribution for the data induced by the prior and the model.

Mercer Theorem

- The sample $S = x_1, ..., x_n$ includes n examples.
- For $S \in \mathcal{X} \subset R^d$ kernel function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$
- The Kernel matrix K is an n$\times$ n matrix such that $K_{i,j} = k(x_i, x_j)$ and K is symmetric.
- A symmetric function K is a mercer kernel iff for any finite sample S the kernel matrix for S is positive semi-definite.

An Example of Kernel

- Following is the structure of kernel

$$K(x) = \begin{bmatrix} k(x_1, x_1) & k(x_1, x_2) & \dots & k(x_1, x_n) \\ k(x_2, x_1) & k(x_2, x_2) & \dots & k(x_2, x_n) \\ \dots & \dots & \dots & \dots \\ k(x_n, x_1) & k(x_n, x_2) & \dots & k(x_n, x_n) \end{bmatrix}$$

- The most widely used co-variance function of this class is arguably the squared exponential function, given by:

$$k(x_i, x_j) = h^2 exp[-(\frac{x_i - x_j}{\lambda})^2] \tag{1}$$

where h and λ are hyperparameters.

Gaussian Process

- Consider the non parametric regression model:

$$Y_i = m(X_i) + \epsilon_i \tag{2}$$

where $E(\epsilon_i) = 0$, i=1(1)n

- A stochastic process $m(x)$ indexed by $x \in \mathcal{X} \subset R^d$ is a Gaussian process if for each $x_1, ..., x_n \in \mathcal{X}$ the vector $m(x) = (m(x_1), m(x_2), ..., m(x_n))$ is Normally distributed:

$$(m(x_1), m(x_2), ..., m(x_n)) \sim N_n(\mu(x), K(x)) \tag{3}$$

where

- $\mu(x) = \mathbb{E}(m(x))$ and $K(x)$ is a mercer kernel .The model is summarized as:

$$m \sim \pi \tag{4}$$

$$Y_1, ... Y_n | m \sim p(y|m) \tag{5}$$

Estimation

- Assume that $\mu = 0$. Then for given $x_1, ..., x_n$the density of the Gaussian process prior of $m = (m(x_1), ..., m(x_n))^T$ is given as:

$$\pi(m) = (2\pi)^{-n/2}|K|^{-1/2}exp(-\frac{1}{2}m^T K^{-1} m) \tag{6}$$

- Let $m = K\alpha$, then $\alpha \sim N_n(0, K^{-1})$ then density of alpha is given as :

$$\pi(\alpha) = (2\pi)^{-n/2}|K|^{-1/2}exp(-\frac{1}{2}\alpha^T K \alpha) \tag{7}$$

Estimation(continued...)

- Since $Y_i = m(X_i) + \epsilon_i$ and $\epsilon_i \sim N(0, \sigma^2)$ we can write the log likelihood as :

$$log(p(y|m)) = -\frac{1}{2\sigma^2}\sum(y_i - m(x_i))^2 + c_1 \tag{8}$$

- The log-posterior is given by:

$$log(p(y|m)) + log(\pi(m)) = -\frac{1}{2\sigma^2}||(y - K\alpha)||^2 - \frac{1}{2}\alpha^T K\alpha + c_2 \tag{9}$$

Estimation(continued...)

- In this Bayesian setup, MAP estimation corresponds to Mercer kernel regression is the posterior mean given as:

$$\mathbb{E}(\alpha|Y) = (K + \sigma^2 I)^{-1} Y \tag{10}$$

- Hence:

$$\hat{m} = \mathbb{E}(m|Y) = \mathbb{E}(K\alpha|Y) = K(K + \sigma^2 I)^{-1} Y \tag{11}$$

Prediction

- To compute the predictive distribution for a new point $Y_{n+1} = m(x_{n+1}) + \epsilon_{n+1}$, we note that $(Y_1, ..., Y_n) \sim N_n(0, (K + \sigma^2 I)$ also $(m(x_1), ..., m(x_n), m(x_{n+1}))$ will have following kernel :

$$K_{(x_1,\ldots,x_n,x_{n+1})} = \begin{bmatrix} k(x_1, x_1) & k(x_1, x_2) & \ldots & k(x_1, x_{n+1}) \\ k(x_2, x_1) & k(x_2, x_2) & \ldots & k(x_2, x_{n+1}) \\ \ldots & \ldots & \ldots & \ldots \\ k(x_{n+1}, x_1) & k(x_{n+1}, x_2) & \ldots & k(x_{n+1}, x_{n+1}) \end{bmatrix}$$

Prediction (Continued...)

- Let z be the vector such that $z = (k(x_1, x_{n+1}.......k(x_n, x_{n+1}))^T$ then $(Y_1, ..., Y_n, Y_{n+1})$ is jointly Gaussian with covariance

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

- so, conditional distribution of $_{Yn+1}$ is

$$Y_n + 1|Y \sim N(z^T(K + \sigma^2 I)^{-1}Y, k(x_{n+1}, x_{n+1}) + \sigma^2 - z^T(K + \sigma^2 I)^{-1}z)) \quad (12)$$

Conclusion

- $\hat{m} = K(K + \sigma^2 I)^{-1} Y$
- Comparing it with kernel regression it can be more complex because we have to choose the appropriate merecer kernel for every data.
- It is computationally expensive.