MTH673A: Robust Statistical Methods
A Bayesian approach for Non Parametric Regression

Arvind Singh Yadav (191026)

Feb 19, 2022

Abstract

As taught in lectures we have used kernel method to obtained the non parametric version of regression. Here we will see an bayesian approach to formulate that problem.

1 Bayesian Setup

We are observing X_i, Y_i and have a model $M = \{p(y|m) : m \in \Theta\}$ where $i = 1(1)n$.We put a prior $\pi(m)$ on the parameter m and compute the posterior distribution using Bayes' rule :

$$\pi(m|y) = \frac{L(m)\pi(m)}{m(Y)} \tag{1}$$

where $Y = (Y_1, ...Y_n)$, $L(m) = \prod p(y_i|m)$ is the likelihood function and m(y) is the marginal distribution for the data induced by the prior and the model.

2 Mercer Theorem

The sample $S = x_1, ..., x_n$ includes n examples. The Kernel matrix K is an n× n matrix such that $K_{i,j} = k(x_i, x_j)$ and K is symmetric.

A symmetric function K is a kernel iff for any finite sample S the kernel matrix for S is positive semi-definite.

3 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:

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

  where h and λ are hyperparameters.

4 Non Parametric Regression using Gaussian Processes

Consider the non parametric regression model:

$$Y_i = m(X_i) + \epsilon_i \tag{3}$$

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

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

$$(m(x_1), m(x_2), ..., m(x_n)) \sim N_n(\mu(x), K(x)) \tag{4}$$

where $\mu(x) = E(x)$ and $K(x)$ is a mercer kernel .The model is summarized as:

$$m \sim \pi \tag{5}$$

$$Y_1, ...Y_n|m \sim p(y|m) \tag{6}$$

5 Estimation

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

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

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

$$\pi(\alpha) = (2\pi)^{-n/2}|K|^{-1/2}exp(-\frac{1}{2}\alpha^T K\alpha) \tag{8}$$

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

$$log(p(y|m)) = -\frac{1}{2\sigma^2}\sum(y_i - m(x_i))^2 + c_1 \tag{9}$$

Now , and the log-posterior is given by:

$$log(p(y|m)) + log(\pi(m)) = -\frac{1}{2\sigma^2}||(y - K\alpha)||^2 - \frac{1}{2}\alpha^T K\alpha + c_2 \tag{10}$$

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

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

Hence:

$$\hat{m} = E(m|Y) = E(K\alpha|Y) = K(K + \sigma^2 I)^{-1}Y \tag{12}$$

6 Prediction using Gaussian Process

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 :

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

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 Y_{n+1} is

$$Y_{n+1}|Y \sim N(z^T (K + \sigma^2 I)^{-1} Y, k(x_{n+1}, x_{n+1}) + \sigma^2 - z^T (K + \sigma^2 I)^{-1} z) \tag{13}$$

7 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.

References

[1] https://www.stat.cmu.edu/ larry/=sml/nonparbayes