Gaussian Process: A beautiful solution to Regression problems

Li Wang

April 14th, 2018
updated April 19th, 2018

1 Introduction

In this post, I will explain how to use Gaussian Process(GP) to solve linear regression problems, with the help of Gaussian distribution and proper choice of kernel function, it can be generalized to unlock non-linear problems. In the next post, I will talk more about applications of GP in classification scenarios.

1.1 Multivariate normal distribution

Multivariate Gaussian distribution is implemented for modelling the distribution of more than one variables. For a k-dimensional random vector $\boldsymbol{Y}$, the multivariate Gaussian distribution can be written as

$$\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{1}$$

$$p(\boldsymbol{Y}) = \frac{1}{(2\pi)^{k/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{ -\frac{1}{2}(\boldsymbol{Y}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{Y}-\boldsymbol{\mu}) \right\} \tag{2}$$

where

$$\begin{aligned} \boldsymbol{\mu} &= E[\boldsymbol{Y}] = \Big[E[Y_1], E[Y_2], \cdots, E[Y_k]\Big]^T \\ \boldsymbol{\Sigma} &= E[(\boldsymbol{Y}-\boldsymbol{\mu})^T(\boldsymbol{Y}-\boldsymbol{\mu})] \end{aligned}$$

we can also write eq(1) as joint distribution over k dimensional vectors

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \cdot \\ \cdot \\ \cdot \\ Y_k \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \\ \cdot \\ \cdot \\ \cdot \\ \mu_k \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} & \cdots & \Sigma_{1k} \\ \Sigma_{21} & \Sigma_{22} & \cdots & \Sigma_{2k} \\ \cdot & \cdot & & \\ \cdot & & \cdot & \\ \cdot & & & \cdot \\ \Sigma_{k1} & \Sigma_{k2} & \cdots & \Sigma_{kk} \end{bmatrix} \right) \tag{3}$$

For example, the joint distribution over two vectors $\boldsymbol{y_1}$ and $\boldsymbol{y_2}$ is

$$\begin{bmatrix} \boldsymbol{y_1} \\ \boldsymbol{y_2} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu_1} \\ \boldsymbol{\mu_2} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma_{11}} & \boldsymbol{\Sigma_{12}} \\ \boldsymbol{\Sigma_{21}} & \boldsymbol{\Sigma_{22}} \end{bmatrix}\right) \tag{4}$$

the condition distribution, is also a Gaussian

$$\boldsymbol{y_1}|\boldsymbol{y_2} \sim \mathcal{N}(\boldsymbol{\mu_{1|2}}, \boldsymbol{\Sigma_{1|2}}) \tag{5}$$

where

$$\begin{aligned} \boldsymbol{\mu_{1|2}} &= \boldsymbol{\mu_1} + \boldsymbol{\Sigma_{12}}\boldsymbol{\Sigma_{22}^{-1}}(\boldsymbol{y_2} - \boldsymbol{\mu_2}) \\ \boldsymbol{\Sigma_{1|2}} &= \boldsymbol{\Sigma_{11}} - \boldsymbol{\Sigma_{12}}\boldsymbol{\Sigma_{22}^{-1}}\boldsymbol{\Sigma_{21}} \end{aligned}$$

the product of two Gaussian distributions is also an Gaussian:

$$p_1(\boldsymbol{y})p_2(\boldsymbol{y}) \propto \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{6}$$

where

$$\begin{aligned} \boldsymbol{\mu} &= \boldsymbol{\Sigma}\Big(\boldsymbol{\Sigma_1^{-1}}\boldsymbol{\mu_1} + \boldsymbol{\Sigma_2^{-1}}\boldsymbol{\mu_2}\Big) \\ \boldsymbol{\Sigma} &= \Big(\boldsymbol{\Sigma_1^{-1}} + \boldsymbol{\Sigma_2^{-1}}\Big)^{\boldsymbol{-1}} \end{aligned}$$

2 Frequentist Linear regression: Maximum likelihood

Problems outline: assuming the relationship between observed variable y and its dependent vaiable $\boldsymbol{x}$ can be modeled as: $y = f(\boldsymbol{x}) + \epsilon$.

- systematic variation: function $f(\boldsymbol{x})$ can be accurately predicted if the underlying process is known.
- random variation: the inherent "unpredictability" of the system, assuming $\epsilon \in \mathcal{N}(0, \sigma_n^2)$,
- The goal is not to find the function, but to predict the distribution of $f(\boldsymbol{x}_{n+1})$ given data $D_{1:n}$.

Suppose $f(\boldsymbol{x}) = \boldsymbol{\omega^T x}$, it is easy to write the likelihood function as

$$p(y_1, y_2, \cdots, y_n|\boldsymbol{x_1}, \boldsymbol{x_1}, \cdots, \boldsymbol{x_n}, \boldsymbol{\omega}) = \prod_{i=1}^{n} p(y_i|\boldsymbol{x}_i, \boldsymbol{\omega}) = \mathcal{N}(\boldsymbol{y}; \boldsymbol{X^T}\boldsymbol{\omega}, \sigma_n^2\boldsymbol{I}) \tag{7}$$

where

$$\boldsymbol{X} = [\boldsymbol{x_1}, \boldsymbol{x_2} \cdots \boldsymbol{x_n}] = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{k1} & x_{k2} & \cdots & x_{kn} \end{bmatrix}, \quad \boldsymbol{\omega} = \begin{bmatrix} \omega_1 \\ \omega_2 \\ \vdots \\ \omega_k \end{bmatrix} \text{ and } \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \tag{8}$$

$\boldsymbol{X}$ can be considered as a $k \times n$ matrix, where each x has k dimensions, and we have n samples in total. Using maximum likelihood $\min_{\boldsymbol{\omega}} \left\{ \sum_{i=1}^{n} (y_i - \boldsymbol{\omega}^T \boldsymbol{x}_i)^2 \right\}$, we obtain

$$\hat{\boldsymbol{\omega}} = (\boldsymbol{X}\boldsymbol{X}^{\boldsymbol{T}})^{-1}\boldsymbol{X}\boldsymbol{y} \tag{9}$$

3 Bayesian Linear regression: Maximum posterior

The posterior distribution of the weights is

$$p(\boldsymbol{\omega}|\boldsymbol{y}, \boldsymbol{X}) = \frac{p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\omega})p(\boldsymbol{\omega})}{p(\boldsymbol{y}|\boldsymbol{X})} \tag{10}$$

Assuming the prior $p(\boldsymbol{\omega})$ and likelihood $p(\boldsymbol{y}|\boldsymbol{X}, \omega)$ are

$$p(\boldsymbol{\omega}) \sim \mathcal{N}(\boldsymbol{0}, \sigma_{\boldsymbol{\omega}}^2 \boldsymbol{I}) \tag{11}$$

$$p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\omega}) \sim \mathcal{N}(\boldsymbol{y}; \boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{\omega}, \sigma_n^2 \boldsymbol{I}) \tag{12}$$

From eq(10), we obtain

$$\begin{aligned} p(\boldsymbol{\omega}|\boldsymbol{y}, \boldsymbol{X}) &\propto p(\boldsymbol{\omega})p(\boldsymbol{y}|\boldsymbol{X}, \omega) \\ \log(p(\boldsymbol{\omega}|\boldsymbol{y}, \boldsymbol{X})) &\propto -\frac{1}{2\sigma_n^2}(\boldsymbol{y} - \boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{\omega})^T(\boldsymbol{y} - \boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{\omega}) - \frac{1}{2\sigma_{\boldsymbol{\omega}}^2}\boldsymbol{\omega}^{\boldsymbol{T}}\boldsymbol{\omega} \end{aligned}$$

the resulting distribution is also a Gaussian

$$p(\boldsymbol{\omega}|\boldsymbol{y}, \boldsymbol{X}) \sim \mathcal{N}(\boldsymbol{\omega}, \boldsymbol{\mu}_{\boldsymbol{\omega}}, \boldsymbol{\Sigma}_{\boldsymbol{\omega}}) \tag{13}$$

where

$$\boldsymbol{\mu}_{\boldsymbol{\omega}} = \frac{1}{\sigma_n^2}\left(\frac{1}{\sigma_n^2}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X} + \frac{1}{\sigma_{\boldsymbol{\omega}}^2}\boldsymbol{I}\right)^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y} \tag{14}$$

$$= \left(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X} + \frac{\sigma_n^2}{\sigma_{\boldsymbol{\omega}}^2}\boldsymbol{I}\right)^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y} \tag{15}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{\omega}} = \left(\frac{1}{\sigma_n^2}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X} + \frac{1}{\sigma_{\boldsymbol{\omega}}^2}\boldsymbol{I}\right)^{-1} \tag{16}$$

$$\tag{17}$$

3.1 Linear regression Prediction

Frequentist prediction

$$\begin{aligned} p(y_{n+1}|x, \boldsymbol{x}, \boldsymbol{X}, \boldsymbol{y}) &= \int p(y_{n+1}|\boldsymbol{x}, \boldsymbol{\omega})p(\boldsymbol{\omega}|\boldsymbol{X}, \boldsymbol{y})\, \mathrm{d}\boldsymbol{\omega} \\ &= \int \mathcal{N}(y; \boldsymbol{\omega}^{\boldsymbol{T}}\boldsymbol{x}_{n+1}, \sigma_n^2)\delta(\boldsymbol{\omega} - \hat{\boldsymbol{\omega}})\, \mathrm{d}\boldsymbol{\omega} \\ &= \mathcal{N}(y; \hat{\boldsymbol{\omega}}^{\boldsymbol{T}}\boldsymbol{x}_{n+1}, \sigma_n^2 \boldsymbol{I}) \end{aligned}$$

from above equation, the deterministic part is, $f(\boldsymbol{x}_{n+1}) = \hat{\boldsymbol{\omega}}^{\boldsymbol{T}} \boldsymbol{x}_{n+1}$

Bayesian prediction In bayesian paradigm, $\boldsymbol{\omega}$ is a probability distribution, thus the deterministic part is a marginal distribution over $\boldsymbol{\omega}$

$$\begin{aligned} p(f(\boldsymbol{x}_{n+1})|\boldsymbol{x}, \boldsymbol{X}, \boldsymbol{y}) &= \int p(f(\boldsymbol{x}_{n+1})|\boldsymbol{x}, \boldsymbol{\omega}) p(\boldsymbol{\omega}|\boldsymbol{X}, \boldsymbol{y}) \mathrm{d}\boldsymbol{\omega} \\ &= \int \delta(f(\boldsymbol{x}_{n+1}) - \boldsymbol{x}_{n+1}^T \boldsymbol{\omega}) \mathcal{N}(\boldsymbol{\omega}, \boldsymbol{\mu}_{\boldsymbol{\omega}}, \boldsymbol{\Sigma}_{\boldsymbol{\omega}}) \mathrm{d}\boldsymbol{\omega} \\ &= \mathcal{N}(f(\boldsymbol{x}_{n+1}); \boldsymbol{x}_{n+1}^T \boldsymbol{\mu}_{\boldsymbol{\omega}}, \boldsymbol{x}_{n+1}^T \boldsymbol{\Sigma}_{\boldsymbol{\omega}} \boldsymbol{x}_{n+1}) \qquad (18) \end{aligned}$$

the observed value y_{n+1} is therefore

$$\begin{aligned} p(y_{n+1}|\boldsymbol{x}, \boldsymbol{X}, \boldsymbol{y}) &= \int p(y_{n+1}|f(\boldsymbol{x})) p(f(\boldsymbol{x})|\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{x}) \mathrm{d}f(\boldsymbol{x}) \\ &= \mathcal{N}(y_{n+1}; \boldsymbol{x}_{n+1}^T \boldsymbol{\mu}_{\boldsymbol{\omega}}, \boldsymbol{x}_{n+1}^T \boldsymbol{\Sigma}_{\boldsymbol{\omega}} \boldsymbol{x}_{n+1} + \sigma_n^2 \boldsymbol{I}) \qquad (19) \end{aligned}$$

4 Non-linear interpolation

Consider a basis function of the form $\phi(|\boldsymbol{x} - \boldsymbol{x_i}|)$, where $\phi()$ is some non-linear function, and $|\boldsymbol{x} - \boldsymbol{x_i}|$ is the distance between vector $\boldsymbol{x}$ and the prototype vector $\boldsymbol{x_i}$, then we can define the mapping function $f(x)$ as

$$f(\boldsymbol{x}) = \sum_{i=1}^{n} \omega_i \phi(|\boldsymbol{x} - \boldsymbol{x}_i|) = \boldsymbol{\omega}^T \phi(\boldsymbol{x}) \qquad (20)$$

we still have $y = f(\boldsymbol{x}) + \epsilon$, or $\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{\omega} + \boldsymbol{\epsilon}$ in terms of training data,

$$\boldsymbol{\Phi} = \begin{bmatrix} \phi(|\boldsymbol{x}_1 - \boldsymbol{x}_1|) & \phi(|\boldsymbol{x}_1 - \boldsymbol{x}_2|) & \cdots & \phi(|\boldsymbol{x}_1 - \boldsymbol{x}_n|) \\ \phi(|\boldsymbol{x}_2 - \boldsymbol{x}_1|) & \phi(|\boldsymbol{x}_2 - \boldsymbol{x}_2|) & \cdots & \phi(|\boldsymbol{x}_2 - \boldsymbol{x}_n|) \\ \vdots & \vdots & \ddots & \vdots \\ \phi(|\boldsymbol{x}_n - \boldsymbol{x}_1|) & \phi(|\boldsymbol{x}_n - \boldsymbol{x}_2|) & \cdots & \phi(|\boldsymbol{x}_n - \boldsymbol{x}_n|) \end{bmatrix} = \begin{bmatrix} \phi(\boldsymbol{x}_1)' \\ \phi(\boldsymbol{x}_2)' \\ \vdots \\ \phi(\boldsymbol{x}_n)' \end{bmatrix} \qquad (21)$$

Below is a list of commonly used forms of $\phi(z)$, where $z = \frac{x - x_i}{\sigma}$

- Gaussian: $\exp(-z^2/2)$
- Exponential: $\exp(-z)$
- Quadratic: $z^2 + \alpha z + \beta$
- Inverse Quadratic: $1/(1 + z^2)$
- Thin plate spine: $z^\alpha \log(z)$
- Trigonometric: $\sin(z)$

where σ controls the smoothness of the non-linear interpolation function.

4.1 Frequentist solution

$$\hat{\boldsymbol{\omega}} = \boldsymbol{\Phi}^{-1}\boldsymbol{\omega} \tag{22}$$

4.2 Bayesian solution

Comparing to its linear equivalent eq(13), Bayesian version for posterior distribution of ω is

$$p(\boldsymbol{\omega}|\boldsymbol{X}, \boldsymbol{y}) = \mathcal{N}(\boldsymbol{\omega}; \boldsymbol{\mu}_{\boldsymbol{\omega}}, \boldsymbol{\Sigma}_{\omega}) \tag{23}$$

where

$$\begin{aligned} \boldsymbol{\mu}_{\boldsymbol{\omega}} &= \frac{1}{\sigma_n^2}\left(\frac{1}{\sigma_n^2}\boldsymbol{\Phi}^{\boldsymbol{T}}\boldsymbol{\Phi} + \frac{1}{\sigma_{\boldsymbol{\omega}}^2}\boldsymbol{I}\right)^{-1}\boldsymbol{\Phi}^{\boldsymbol{T}}\boldsymbol{y} \\ \boldsymbol{\Sigma}_{\boldsymbol{\omega}} &= \left(\frac{1}{\sigma_n^2}\boldsymbol{\Phi}^{\boldsymbol{T}}\boldsymbol{\Phi} + \frac{1}{\sigma_{\boldsymbol{\omega}}^2}\boldsymbol{I}\right)^{-1} \end{aligned}$$

4.3 Bayesian prediction

For simplicity, assuming $E(f(\boldsymbol{x})) = 0$, the prior distribution for $\boldsymbol{\omega}$ is $\mathcal{N}(0, \sigma_{\boldsymbol{\omega}}^2\boldsymbol{I})$, the likelihood funtion of $p(\boldsymbol{y}|\boldsymbol{X})$ is $\mathcal{N}(0, \Sigma_{\boldsymbol{y}})$ where

$$\boldsymbol{\Sigma}_{\boldsymbol{y}} = \sigma_{\boldsymbol{\omega}}^2\boldsymbol{\Phi}^{\boldsymbol{T}}\boldsymbol{\Phi} + \sigma_n^2\boldsymbol{I} = \boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X}) + \sigma_n^2\boldsymbol{I} \tag{24}$$

with

1. $$\boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X}) = \boldsymbol{\Phi}^{\boldsymbol{T}}(\boldsymbol{X})\sigma_{\boldsymbol{\omega}}^2\boldsymbol{\Phi}(\boldsymbol{X}) \tag{25}$$

2. $$\boldsymbol{k}(\boldsymbol{x}_{n+1}, \boldsymbol{X}) = \boldsymbol{\Phi}(\boldsymbol{X})\sigma_{\boldsymbol{\omega}}^2\boldsymbol{\Phi}(\boldsymbol{x}_{n+1}) \tag{26}$$

3. $$k(\boldsymbol{x}_{n+1}, \boldsymbol{x}_{n+1}) = \boldsymbol{\Phi}(\boldsymbol{x}_{n+1})\sigma_{\boldsymbol{\omega}}^2\boldsymbol{\Phi}(\boldsymbol{x}_{n+1}) \tag{27}$$

we can write the joint distribution of $f(\boldsymbol{x})$ and y as

$$\begin{bmatrix} f(\boldsymbol{x}_{n+1}) \\ \boldsymbol{D}_{(1:n)} \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} k(\boldsymbol{x}_{n+1}, \boldsymbol{x}_{n+1}) & \boldsymbol{k}(\boldsymbol{x}_{n+1}, \boldsymbol{X})' \\ \boldsymbol{k}(\boldsymbol{x}_{n+1}, \boldsymbol{X}) & \boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X}) + \sigma_n^2\boldsymbol{I} \end{bmatrix}\right) \tag{28}$$

where

$$\boldsymbol{k}(x_{n+1}, \boldsymbol{X}) = \begin{bmatrix} k(x, x_1) \\ k(x, x_2) \\ \vdots \\ k(x, x_n) \end{bmatrix}, \qquad \boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X}) = \begin{bmatrix} k(x_1, x_1) & k(x_1, x_2) & \cdots & k(x_1, x_n) \\ k(x_2, x_1) & k(x_2, x_2) & \cdots & k(x_2, x_n) \\ \vdots & \vdots & \ddots & \vdots \\ k(x_n, x_1) & k(x_n, x_2) & \cdots & k(x_n, x_n) \end{bmatrix}$$

Following eq(5), the prediction for $f(\boldsymbol{x}_{n+1})$ is

$$p(f(\boldsymbol{x}_{n+1})|\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{x}) \propto \mathcal{N}(\mu_f, \sigma_f^2) \tag{29}$$

where

$$\begin{aligned}
\mu_f &= \boldsymbol{k}(\boldsymbol{x}_{n+1}, \boldsymbol{X})^T \left[\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X}) + \sigma_n^2 \boldsymbol{I}\right]^{-1} \boldsymbol{y} \\
&= \boldsymbol{\Phi}^T(\boldsymbol{x}_{n+1})\boldsymbol{\mu} \\
\sigma_f^2 &= k(\boldsymbol{x}_{n+1}, \boldsymbol{x}_{n+1}) - \boldsymbol{k}(\boldsymbol{x}_{n+1}, \boldsymbol{X})^T \left[\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X}) + \sigma_n^2 \boldsymbol{I}\right]^{-1} \boldsymbol{k}(\boldsymbol{x}_{n+1}, \boldsymbol{X}) \\
&= \boldsymbol{\Phi}^T(\boldsymbol{x}_{n+1})^T \boldsymbol{\Sigma} \boldsymbol{\Phi}^T(\boldsymbol{x}_{n+1})
\end{aligned}$$

where

$$\begin{aligned}
\boldsymbol{\mu} &= \left(\boldsymbol{\Phi}(\boldsymbol{X})^T \boldsymbol{\Phi}(\boldsymbol{X}) + \frac{\sigma_n^2}{\sigma_{\boldsymbol{\omega}}^2} \boldsymbol{I}\right)^{-1} \boldsymbol{\Phi}(\boldsymbol{X})^T \boldsymbol{y} \\
\boldsymbol{\Sigma} &= \left(\frac{1}{\sigma_n^2} \boldsymbol{\Phi}(\boldsymbol{X})^T \boldsymbol{\Phi}(\boldsymbol{X}) + \frac{1}{\sigma_{\boldsymbol{\omega}}^2} \boldsymbol{I}\right)^{-1}
\end{aligned}$$

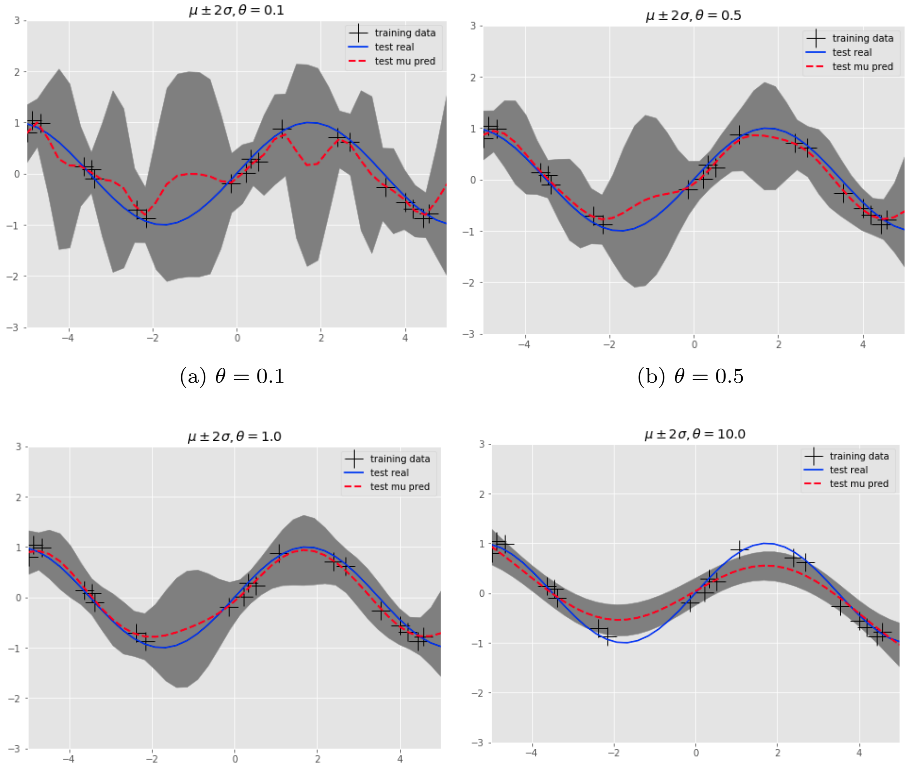
Further more, the predictive distribution for y is

$$\begin{aligned}
p(y|\boldsymbol{X}, \mathcal{M}) &= N(y, 0, \sigma_f^2 + \sigma_n^2 \boldsymbol{I}) && (30) \\
&= N(y, 0, \boldsymbol{\Phi}^T(\boldsymbol{x}_{n+1})^T \boldsymbol{\Sigma} \boldsymbol{\Phi}^T(\boldsymbol{x}_{n+1}) + \sigma_n^2 \boldsymbol{I}) && (31)
\end{aligned}$$

5 Hyper-parameter Optimization

The hidden parameter in $\phi()$ which controls the smoothness of fitting is defined as $\boldsymbol{\theta}$, the likelihood function is $p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta})$, we can obtain $\boldsymbol{\theta}$ through Maximum Likelihood

$$\begin{aligned}
\frac{\partial}{\partial \theta_j} \log(p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta})) &= \frac{1}{2} \boldsymbol{y'} \frac{\partial(\boldsymbol{K^{-1}})}{\partial \theta_j} \boldsymbol{y} - \frac{1}{2} \frac{\partial \log(|\boldsymbol{K}|)}{\partial \theta_j} && (32) \\
&= \frac{1}{2} \boldsymbol{y'} \boldsymbol{K^{-1}} \frac{\partial \boldsymbol{K}}{\partial \theta_j} \boldsymbol{K^{-1}} \boldsymbol{y} - \frac{1}{2} \mathrm{tr}\left(\boldsymbol{K^{-1}} \frac{\partial \boldsymbol{K}}{\partial \theta_j}\right) && (33) \\
&= \mathrm{tr}\left((\boldsymbol{K^{-1}} \boldsymbol{y} \boldsymbol{y'} \boldsymbol{K'^{-1}} - \boldsymbol{K^{-1}}) \frac{\partial \boldsymbol{K}}{\partial \theta_j}\right) && (34)
\end{aligned}$$



(a) $\theta = 0.1$

(b) $\theta = 0.5$

(c) $\theta = 1.0$

(d) $\theta = 10.0$

Figure 1: The hyper-parameter θ in kernel function influences fitting and prediction.

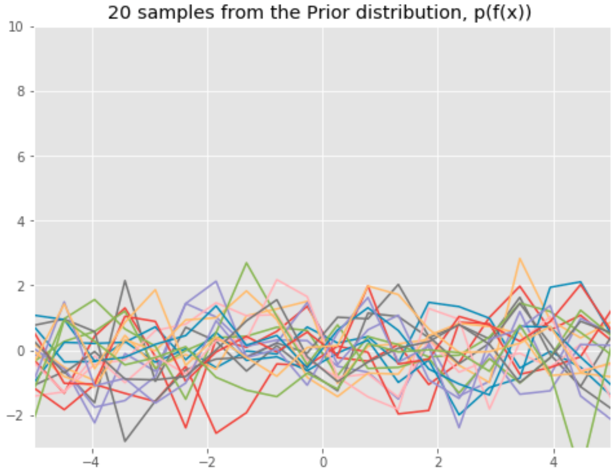


Figure 2: Sampling data from prior distribution of model $f(x)$

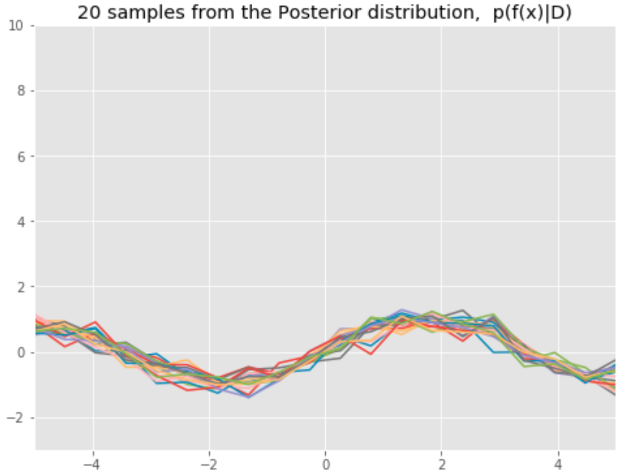


Figure 3: Sampling data from posterior distribution of model $f(x)$