A Brief Note on Gaussian Process Regression

Di Wang
280868861@qq.com
PhD Candidate of Xi'an Jiaotong University, China
04/30/2020

Assuming we have the training set $\mathfrak{D} = \{\mathbf{x}_i, y_i\}_{i=1}^{n}, \mathbf{x}_i \in \mathbb{R}^D, y_i \in \mathbb{R}$, we can also formulate the training set as compact as $\mathfrak{D} = (\mathbf{X}, \mathbf{y}), \mathbf{X} \in \mathbb{R}^{D\times n}, \mathbf{y} \in \mathbb{R}^n$. For brevity, we discard the input $\mathbf{X}$ in the following sections. The readers should keep in mind that the input $\mathbf{X}$ is deterministic parameter, while the target $\mathbf{y}$ and latent variable $\mathbf{f}$ are random variables.

Given the new input $\mathbf{X}_* \in \mathbb{R}^{D\times n_*}$, our purpose is to infer $\mathbf{y}_* \in \mathbb{R}^{n_*}$ as follows

$$p(\mathbf{y}_* \mid \mathbf{y}) = ? \tag{1.1}$$

An intuitive idea is to firstly learn a linear model on training set:

$$\begin{aligned} & y_i = \mathbf{w}^T \boldsymbol{\varphi}_i + \varepsilon_i, i = 1, 2, \cdots, n. \\ & \varepsilon_i \sim \mathcal{N}(0, \sigma_n), \mathbf{w} \in \mathbb{R}^M, \boldsymbol{\varphi} \triangleq \begin{bmatrix} 1 & x_i & \cdots & x_i^{M-1} \end{bmatrix}^T \in \mathbb{R}^M \\ & \mathbf{z} \triangleq \begin{bmatrix} \mathbf{w} \\ \sigma_n \end{bmatrix} \in \mathbb{R}^{M+1}, \end{aligned} \tag{1.2}$$

where $\mathbf{z}$ denotes the latent variable. The predictive distribution can be accordingly computed:

$$p(\mathbf{y}_* \mid \mathbf{y}) = \int p(\mathbf{y}^* \mid \mathbf{z})\, p(\mathbf{z} \mid \mathbf{y})\, d\mathbf{z}. \tag{1.3}$$

The main limitation of aforementioned linear model lies in that the feature vector $\boldsymbol{\varphi}_i$ is hard to be selected. The Gaussian process is equivalent to the linear model with an infinite feature vector, and the computation of feature vector is fully specified by positive definite covariance function $k(\bullet,\bullet): \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}^+$.

The regression model based on Gaussian process is stated as follows:

$$y_i = f_i + \varepsilon_i, i = 1, 2, \cdots, n, \cdots, n + n_*, \tag{1.4}$$

where

$$\mathbf{f} \sim \mathrm{GP}(\mathbf{0}, k(\bullet,\bullet)) \tag{1.5}$$

denotes the latent variable and its length can be infinity. The Gaussian process states that $\mathbf{f}$ is a joint Gaussian distribution. Here we don't assume the noise ε_i is Gaussian, that is to say, the likelihood function $p(y_i \mid f_i)$ is not necessarily to be a Gaussian. Robust distribution like Student’s t or Laplacian can be utilized when the target value y_i contains outliers.

Utilizing the joint Gaussian assumption, we can directly compute the posterior as follows according to the Appendix of GPML:

$$\begin{aligned} & (\mathbf{f}, \mathbf{f}_*) \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{K} & \mathbf{K}_* \\ \mathbf{K}_*^T & \mathbf{K}_{**} \end{bmatrix}\right) \\ & \Rightarrow \mathbf{f}_* \mid \mathbf{f} \sim \mathcal{N}\left(\mathbf{K}_*^T \mathbf{K}^{-1} \mathbf{f}, \mathbf{K}_{**} - \mathbf{K}_*^T \mathbf{K}^{-1} \mathbf{K}^T\right) \end{aligned} \tag{1.6}$$

Short Answer

Since the likelihood model is latent variable plus a noise term as follows:

$$\begin{aligned} & \mathbf{f} \sim \mathrm{GP}(\mathbf{0}, \mathbf{K}), \boldsymbol{\varepsilon} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{W}^{-1}\right) \\ & \mathbf{y} = \mathbf{f} + \boldsymbol{\varepsilon}, \end{aligned} \tag{1.7}$$

where $\mathbf{W} = \frac{1}{\sigma_n^2} \mathbf{I}$ denotes the precision matrix for Gaussian noise term in GPML. Therefore, the target value is also a Gaussian distribution as follow:

$$\begin{aligned} & \mathbf{y} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{K} + \mathbf{W}^{-1}\right), \mathbf{y}^* \sim \mathcal{N}\left(\mathbf{0}, \mathbf{K}_* + \mathbf{W}_*^{-1}\right) \\ & \Rightarrow \begin{bmatrix} \mathbf{y}_* \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} \mathbf{K} + \mathbf{W}^{-1} & \mathbf{K}_* \\ \mathbf{K}_*^T & \mathbf{K}_{**} + \mathbf{W}_{**}^{-1} \end{bmatrix} \end{aligned} \tag{1.8}$$

According to the conditional Gaussian assumption, we have:

$$\mathbf{y}_* \mid \mathbf{y} \sim \mathcal{N}\left(\mathbf{K}_*^T \left(\mathbf{K} + \mathbf{W}^{-1}\right)^{-1} \mathbf{y}, \mathbf{K}_{**} + \mathbf{W}_*^{-1} - \mathbf{K}_*^T \left(\mathbf{K} + \mathbf{W}^{-1}\right)^{-1} \mathbf{K}^T\right) \tag{1.9}$$

The marginal likelihood of $\mathbf{y} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{K} + \mathbf{W}^{-1}\right)$ is as follows:

$$\log p(\mathbf{y}) = -\frac{1}{2} \mathbf{y}^T \left(\mathbf{K} + \mathbf{W}^{-1}\right)^{-1} \mathbf{y} - \frac{1}{2} \log \left|\mathbf{K} + \mathbf{W}^{-1}\right| - \frac{n}{2} \log 2\pi \tag{1.10}$$

The marginal likelihood is utilized to optimize the hyper-parameters in Gaussian process.

Long Answer

The long answer does not leverages the structure of likelihood function (latent variable plus a noise term), thus it is more general but more complex than short answer. The long answer is extensively used in the book of GPML.

The predictive distribution is as follows:

$$\begin{aligned} & p(\mathbf{y}_* \mid \mathbf{y}) = \int p(\mathbf{y}_* \mid \mathbf{f}_*)\, p(\mathbf{f}_* \mid \mathbf{y}) d\mathbf{f}_* \\ & p(\mathbf{f}_* \mid \mathbf{y}) = \int p(\mathbf{f}_* \mid \mathbf{f})\, p(\mathbf{f} \mid \mathbf{y}) d\mathbf{f} \end{aligned} \tag{1.11}$$

As we can see, the conditional Gaussian $p\left(\mathbf{f}_* \mid \mathbf{f}\right)$ and the posterior over latent variable $p\left(\mathbf{f} \mid \mathbf{y}\right)$ are required in order to compute the predictive distribution $p\left(\mathbf{f}_* \mid \mathbf{y}\right)$ or $p\left(\mathbf{y}_* \mid \mathbf{y}\right)$.

Posterior over Latent Variable

$$\begin{aligned} p(\mathbf{f} \mid \mathbf{y}) &= \frac{p(\mathbf{f})\, p(\mathbf{y} \mid \mathbf{f})}{p(\mathbf{y})} \\ p(\mathbf{f} \mid \mathbf{y}) &\propto p(\mathbf{f})\, p(\mathbf{y} \mid \mathbf{f}) \end{aligned} \tag{1.12}$$

Gaussian likelihood

When the likelihood $p\left(y_i \mid f_i\right)$ is a Gaussian $\mathbf{y} \mid \mathbf{f} \sim \mathcal{N}\left(\mathbf{f}, \mathbf{W}^{-1}\right), \mathbf{W} = \frac{1}{\sigma_n^2}\mathbf{I}$, $\mathbf{W}$ denotes the precision matrix, the posterior $p\left(\mathbf{f} \mid \mathbf{y}\right)$ is also a Gaussian as follows:

$$\begin{aligned} &\log p(\mathbf{f} \mid \mathbf{y}) = \log p(\mathbf{f}) + \log p(\mathbf{y} \mid \mathbf{f}) = -\frac{1}{2}\mathbf{f}^T\mathbf{K}^{-1}\mathbf{f} - \frac{1}{2}(\mathbf{y}-\mathbf{f})^T \mathbf{W}(\mathbf{y}-\mathbf{f}) \\ &= -\frac{1}{2}\left\{\mathbf{f}^T\left(\mathbf{K}^{-1}+\mathbf{W}\right)\mathbf{f} - 2(\mathbf{W}\mathbf{y})^T\mathbf{f}\right\} \\ &\Rightarrow \mathbf{f} \mid \mathbf{y} \sim \mathcal{N}\left(\hat{\mathbf{f}}, \left(\mathbf{K}^{-1}+\mathbf{W}\right)^{-1}\right), \left(\mathbf{K}^{-1}+\mathbf{W}\right)\hat{\mathbf{f}} = \mathbf{W}\mathbf{y} \end{aligned} \tag{1.13}$$

Accordingly, the predictive distribution $p\left(\mathbf{f}_* \mid \mathbf{y}\right) = \int p\left(\mathbf{f}_* \mid \mathbf{f}\right) p\left(\mathbf{f} \mid \mathbf{y}\right) d\mathbf{f}$ is a Gaussian. Its mean is:

$$\begin{aligned} &\mathbb{E}_{\mathbf{f}_*}\left[\mathbf{f}_*\right] = \int p\left(\mathbf{f}_* \mid \mathbf{y}\right)\mathbf{f}_* d\mathbf{f}_* \\ &= \int\left\{\int \mathcal{N}\left(\mathbf{f}_* \mid \mathbf{K}_*^T\mathbf{K}^{-1}\mathbf{f}, \mathbf{K}_{**} - \mathbf{K}_*^T\mathbf{K}^{-1}\mathbf{K}^T\right)\mathbf{f}_* d\mathbf{f}_*\right\}\mathcal{N}\left(\mathbf{f} \mid \hat{\mathbf{f}}, \left(\mathbf{K}^{-1}+\mathbf{W}\right)^{-1}\right)d\mathbf{f} \\ &= \mathbf{K}_*^T\mathbf{K}^{-1}\int \mathbf{f}\mathcal{N}\left(\mathbf{f} \mid \hat{\mathbf{f}}, \left(\mathbf{K}^{-1}+\mathbf{W}\right)^{-1}\right)d\mathbf{f} \\ &= \mathbf{K}_*^T\mathbf{K}^{-1}\hat{\mathbf{f}} = \mathbf{K}_*^T\mathbf{K}^{-1}\left(\mathbf{K}^{-1}+\mathbf{W}\right)^{-1}\mathbf{W}\mathbf{y} = \mathbf{K}_*^T\left\{\mathbf{W}^{-1}\left(\mathbf{K}^{-1}+\mathbf{W}\right)\mathbf{K}\right\}^{-1}\mathbf{y} \\ &= \mathbf{K}_*^T\left(\mathbf{K}+\mathbf{W}^{-1}\right)^{-1}\mathbf{y} \end{aligned} \tag{1.14}$$

Its variance is as follows:

$$
\begin{aligned}
&\mathbb{E}_{\mathbf{f}_*}\left[\mathbf{f}_*^2\right]=\int p\left(\mathbf{f}_*\,|\,\mathbf{y}\right)\mathbf{f}_*^2 d\mathbf{f}_* \\
&=\int\left\{\int\mathcal{N}\left(\mathbf{f}_*\,|\,\mathbf{K}_*^T\mathbf{K}^{-1}\mathbf{f},\mathbf{K}_{**}-\mathbf{K}_*^T\mathbf{K}^{-1}\mathbf{K}^T\right)\mathbf{f}_*\mathbf{f}_*^T d\mathbf{f}_*\right\}\mathcal{N}\left(\mathbf{f}\,|\,\hat{\mathbf{f}},\left(\mathbf{K}^{-1}+\mathbf{W}\right)^{-1}\right)d\mathbf{f} \\
&=\int\left\{\mathbf{K}_{**}-\mathbf{K}_*^T\mathbf{K}^{-1}\mathbf{K}^T+\mathbf{K}_*^T\mathbf{K}^{-1}\mathbf{f}\mathbf{f}^T\mathbf{K}^{-1}\mathbf{K}_*\right\}\mathcal{N}\left(\mathbf{f}\,|\,\hat{\mathbf{f}},\left(\mathbf{K}^{-1}+\mathbf{W}\right)^{-1}\right)d\mathbf{f} \\
&=\mathbf{K}_{**}-\mathbf{K}_*^T\mathbf{K}^{-1}\mathbf{K}^T+\mathbf{K}_*^T\mathbf{K}^{-1}\left\{\int\mathbf{f}\mathbf{f}^T\mathcal{N}\left(\mathbf{f}\,|\,\hat{\mathbf{f}},\left(\mathbf{K}^{-1}+\mathbf{W}\right)^{-1}\right)d\mathbf{f}\right\}\mathbf{K}^{-1}\mathbf{K}_* \\
&=\mathbf{K}_{**}-\mathbf{K}_*^T\mathbf{K}^{-1}\mathbf{K}^T+\mathbf{K}_*^T\mathbf{K}^{-1}\left\{\left(\mathbf{K}^{-1}+\mathbf{W}\right)^{-1}+\hat{\mathbf{f}}\hat{\mathbf{f}}^T\right\}\mathbf{K}^{-1}\mathbf{K}_* \\
&\Rightarrow \operatorname{cov}\left[\mathbf{f}_*\right]=\mathbb{E}_{\mathbf{f}_*}\left[\mathbf{f}_*^2\right]-\mathbf{K}_*^T\mathbf{K}^{-1}\hat{\mathbf{f}}\hat{\mathbf{f}}^T\mathbf{K}^{-1}\mathbf{K}_* \\
&=\mathbf{K}_{**}-\mathbf{K}_*^T\mathbf{K}^{-1}\mathbf{K}^T+\mathbf{K}_*^T\mathbf{K}^{-1}\left(\mathbf{K}^{-1}+\mathbf{W}\right)^{-1}\mathbf{K}^{-1}\mathbf{K}_* \\
&\mathbf{K}_*^T\mathbf{K}^{-1}\left(\mathbf{K}^{-1}+\mathbf{W}\right)^{-1}\mathbf{K}^{-1}\mathbf{K}_*=\mathbf{K}_*^T\mathbf{K}^{-1}\left\{\mathbf{K}-\mathbf{K}\left(\mathbf{K}+\mathbf{W}^{-1}\right)^{-1}\mathbf{K}\right\}\mathbf{K}^{-1}\mathbf{K}_* \\
&=\mathbf{K}_*^T\left\{\mathbf{K}^{-1}-\left(\mathbf{K}+\mathbf{W}^{-1}\right)^{-1}\right\}\mathbf{K}_*=\mathbf{K}_*^T\mathbf{K}^{-1}\mathbf{K}_*-\mathbf{K}_*^T\left(\mathbf{K}+\mathbf{W}^{-1}\right)^{-1}\mathbf{K}_* \\
&\Rightarrow \operatorname{cov}\left[\mathbf{f}_*\right]=\mathbf{K}_{**}-\mathbf{K}_*^T\left(\mathbf{K}+\mathbf{W}^{-1}\right)^{-1}\mathbf{K}_*
\end{aligned}
\tag{1.15}
$$

Where we extensively utilize the property of $\operatorname{cov}\left[\mathbf{x}\right]=\mathbb{E}\left[\mathbf{x}^2\right]-\mathbb{E}\left[\mathbf{x}\right]\left\{\mathbb{E}\left[\mathbf{x}\right]\right\}^T$.

Marginal Likelihood

It is a product of prior and likelihood:

$$
\begin{aligned}
&\log p\left(\mathbf{f}\right)=-\frac{1}{2}\mathbf{f}^T\mathbf{K}^{-1}\mathbf{f}-\frac{1}{2}\log\left|\mathbf{K}\right|-\frac{n}{2}\log 2\pi \\
&\log p\left(\mathbf{y}\,|\,\mathbf{f}\right)=-\frac{1}{2}\left(\mathbf{f}-\mathbf{y}\right)^T\mathbf{W}\left(\mathbf{f}-\mathbf{y}\right)+\frac{1}{2}\log\left|\mathbf{W}\right|-\frac{n}{2}\log 2\pi \\
&\log p\left(\mathbf{y}\,|\,\mathbf{f}\right)+\log p\left(\mathbf{f}\right) \\
&=-\frac{1}{2}\mathbf{f}^T\left(\mathbf{K}^{-1}+\mathbf{W}\right)\mathbf{f}^T-2\left(\mathbf{W}\mathbf{y}\right)^T\mathbf{f}-\frac{1}{2}\mathbf{y}^T\mathbf{W}\mathbf{y}+\frac{1}{2}\log\left|\mathbf{W}\right|-\frac{1}{2}\log\left|\mathbf{K}\right|-n\log 2\pi \\
&=-\frac{1}{2}\left(\mathbf{f}-\hat{\mathbf{f}}\right)^T\left(\mathbf{K}^{-1}+\mathbf{W}\right)\left(\mathbf{f}-\hat{\mathbf{f}}\right)^T+\frac{1}{2}\hat{\mathbf{f}}^T\left(\mathbf{K}^{-1}+\mathbf{W}\right)\hat{\mathbf{f}}-\frac{1}{2}\mathbf{y}^T\mathbf{W}\mathbf{y}+\frac{1}{2}\log\left|\mathbf{W}\right|-\frac{1}{2}\log\left|\mathbf{K}\right|-n\log 2\pi \\
&=\left\{-\frac{1}{2}\left(\mathbf{f}-\hat{\mathbf{f}}\right)^T\left(\mathbf{K}^{-1}+\mathbf{W}\right)\left(\mathbf{f}-\hat{\mathbf{f}}\right)^T-\frac{n}{2}\log 2\pi+\log\left|\mathbf{K}^{-1}+\mathbf{W}\right|\right\} \\
&+\frac{1}{2}\hat{\mathbf{f}}^T\left(\mathbf{K}^{-1}+\mathbf{W}\right)\hat{\mathbf{f}}+\mathbf{y}^T\mathbf{W}\mathbf{y}+\frac{1}{2}\log\left|\mathbf{W}\right|-\frac{1}{2}\log\left|\mathbf{K}\right|-\log\left|\mathbf{K}^{-1}+\mathbf{W}\right|-\frac{n}{2}\log 2\pi
\end{aligned}
\tag{1.16}
$$

Thus

$$
\begin{aligned}
&\log p\left(\mathbf{y}\,|\,\mathbf{f}\right)+\log p\left(\mathbf{f}\right) \\
&=\log\mathcal{N}\left(\mathbf{f}\,|\,\hat{\mathbf{f}},\left(\mathbf{K}^{-1}+\mathbf{W}\right)^{-1}\right)+\frac{1}{2}\hat{\mathbf{f}}^T\left(\mathbf{K}^{-1}+\mathbf{W}\right)\hat{\mathbf{f}}+\mathbf{y}^T\mathbf{W}\mathbf{y}-\frac{1}{2}\log\left|\mathbf{K}+\mathbf{W}^{-1}\right|-\frac{n}{2}\log 2\pi \\
&\Rightarrow \log p\left(\mathbf{y}\right)=\frac{1}{2}\hat{\mathbf{f}}^T\left(\mathbf{K}^{-1}+\mathbf{W}\right)\hat{\mathbf{f}}-\frac{1}{2}\mathbf{y}^T\mathbf{W}\mathbf{y}+\frac{1}{2}\log\left|\mathbf{W}\right|-\frac{1}{2}\log\left|\mathbf{K}\right|-\log\left|\mathbf{K}^{-1}+\mathbf{W}\right|-\frac{n}{2}\log 2\pi
\end{aligned}
\tag{1.17}
$$

We only need to compute the following term:

$$\left(\mathbf{K}^{-1}+\mathbf{W}\right)\hat{\mathbf{f}}=\mathbf{W}\mathbf{y}$$

$$\Rightarrow \frac{1}{2}\hat{\mathbf{f}}^{T}\left(\mathbf{K}^{-1}+\mathbf{W}\right)\hat{\mathbf{f}}-\frac{1}{2}\mathbf{y}^{T}\mathbf{W}\mathbf{y}=\frac{1}{2}\mathbf{y}^{T}\left\{\mathbf{W}\left(\mathbf{K}^{-1}+\mathbf{W}\right)^{-1}\mathbf{W}-\mathbf{W}\right\}\mathbf{y}=-\frac{1}{2}\mathbf{y}^{T}\left(\mathbf{K}+\mathbf{W}^{-1}\right)^{-1}\mathbf{y}$$

(1.18)

Thus the marginal likelihood is as follows:

$$\log p\left(\mathbf{y}\right)=-\frac{1}{2}\mathbf{y}^{T}\left(\mathbf{K}+\mathbf{W}^{-1}\right)^{-1}\mathbf{y}-\frac{1}{2}\log\left|\mathbf{K}+\mathbf{W}^{-1}\right|-\frac{n}{2}\log 2\pi \qquad (1.19)$$