A Technical Introduction to Gaussian Process Regression

Tzu-Kuo Huang[1]

1 Definition

In regression problems, we are given a sample $S = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_l, y_l)\}$ in which $\mathbf{x}_i$ denotes the ith observation and y_i is the corresponding target value. The relationship between $\mathbf{x}_i$ and y_i is formulated as

$$y_i = f(\mathbf{x}_i) + \epsilon(\mathbf{x}_i),$$

that is, a function f maps the input vector $\mathbf{x}_i$ to the true target, which, being corrupted by noise $\epsilon(\mathbf{x}_i)$, is measured as y_i. Gaussian Process Regression (GPR) is a non-parametric model that assumes

$$\mathbf{f} \equiv [f(\mathbf{x}_1), f(\mathbf{x}_2), \ldots, f(\mathbf{x}_l)]^T \sim N(\mathbf{0}, K), \tag{1}$$

where K is the covariance matrix whose (i, j)th element is given by a kernel function $\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j)$, and

$$(\mathbf{y}|\mathbf{f}) = \big([y_1, y_2, \ldots, y_l]^T|\mathbf{f}\big) \sim N(\mathbf{f}, \sigma^2 I), \tag{2}$$

which means the noise follows a zero-mean and independent joint Gaussian distribution. Note that (2) also implies $\mathbf{y}$'s conditional independence of $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_l\}$ given $\mathbf{f}$. For a new instance $\mathbf{x}^*$, the goal is to estimate $P(f(\mathbf{x}^*)|\mathbf{x}^*, S)$. In the sequel we assume K to be invertible.

2 Derivation of The Predictive Distribution

For convenience, we denote $f(\mathbf{x}^*)$ as f^*. Following a standard Bayesian approach, we write

$$\begin{aligned} P(f^*|\mathbf{x}^*, S) &= \int P(f^*, \mathbf{f}|\mathbf{x}^*, S) d\mathbf{f} \\ &= \int P(f^*|\mathbf{f}, \mathbf{x}^*, S) P(\mathbf{f}|\mathbf{x}^*, S) d\mathbf{f}. \end{aligned} \tag{3}$$

We then derive $P(f^*|\mathbf{f}, \mathbf{x}^*, S)$ in Section 2.1, $P(\mathbf{f}|\mathbf{x}^*, S)$ in Section 2.2, and finally $P(f^*|\mathbf{x}^*, S)$ in Section 2.3.

[1]Department of Computer Science, National Taiwan University, Taipei 106, Taiwan.

2.1 Derivation of $P(f^*|\mathbf{f}, \mathbf{x}^*, S)$

Define $\mathbf{k} \equiv [\mathcal{K}(\mathbf{x}^*, \mathbf{x}_1), \mathcal{K}(\mathbf{x}^*, \mathbf{x}_2), \ldots, \mathcal{K}(\mathbf{x}^*, \mathbf{x}_l)]^T$. Then the joint distribution of $[\mathbf{f}\ f^*]^T$ is

$$\begin{bmatrix} \mathbf{f} \\ f^* \end{bmatrix} \sim N\left(\mathbf{0}, \begin{bmatrix} K & \mathbf{k} \\ \mathbf{k}^T & \mathcal{K}(\mathbf{x}^*, \mathbf{x}^*) \end{bmatrix}\right). \tag{4}$$

Since conditions on $\mathbf{x}^*$ and S are embedded in the covariance matrix in (4), $P(f^*|\mathbf{f}, \mathbf{x}^*, S)$ is equivalent to

$$P(f^*|\mathbf{f}) = \frac{P(f^*, \mathbf{f})}{P(\mathbf{f})}. \tag{5}$$

Let

$$\begin{bmatrix} A & \mathbf{b} \\ \mathbf{b}^T & c \end{bmatrix} \equiv \begin{bmatrix} K & \mathbf{k} \\ \mathbf{k}^T & \mathcal{K}(\mathbf{x}^*, \mathbf{x}^*) \end{bmatrix}^{-1}, \tag{6}$$

then according to (5)

$$\begin{aligned} P(f^*|\mathbf{f}) &\propto \exp\left(-\frac{1}{2} \begin{bmatrix} \mathbf{f} \\ f^* \end{bmatrix}^T \begin{bmatrix} A & \mathbf{b} \\ \mathbf{b}^T & c \end{bmatrix} \begin{bmatrix} \mathbf{f} \\ f^* \end{bmatrix} + \frac{1}{2}\mathbf{f}^T K^{-1}\mathbf{f} \right) \\ &= \exp\left(-\frac{1}{2}\Big(c(f^*)^2 + 2(\mathbf{b}^T\mathbf{f})f^* + \mathbf{f}^T A \mathbf{f} \Big) + \frac{1}{2}\mathbf{f}^T K^{-1}\mathbf{f} \right) \\ &\propto \exp\left(-\frac{1}{2c^{-1}}\Big(f^* + \frac{\mathbf{b}^T\mathbf{f}}{c} \Big)^2 \right). \end{aligned} \tag{7}$$

From (6), we have

$$K\mathbf{b} + c\mathbf{k} = \mathbf{0} \text{ and } \mathbf{k}^T\mathbf{b} + c\mathcal{K}(\mathbf{x}^*, \mathbf{x}^*) = 1,$$

which in turn implies

$$\mathbf{b} = -\frac{K^{-1}\mathbf{k}}{\mathcal{K}(\mathbf{x}^*, \mathbf{x}^*) - \mathbf{k}^T K^{-1}\mathbf{k}}, \tag{8}$$

$$c = \frac{1}{\mathcal{K}(\mathbf{x}^*, \mathbf{x}^*) - \mathbf{k}^T K^{-1}\mathbf{k}}. \tag{9}$$

Plugging (8) and (9) into (7) yields

$$P(f^*|\mathbf{f}) \propto \exp\left(-\frac{\big(f^* - \mathbf{k}^T K^{-1}\mathbf{f}\big)^2}{2\big(\mathcal{K}(\mathbf{x}^*, \mathbf{x}^*) - \mathbf{k}^T K^{-1}\mathbf{k}\big)} \right),$$

that is,

$$(f^*|\mathbf{f}) \sim N\Big(\mathbf{k}^T K^{-1}\mathbf{f}, \mathcal{K}(\mathbf{x}^*, \mathbf{x}^*) - \mathbf{k}^T K^{-1}\mathbf{k}\Big). \tag{10}$$

2.2 Derivation of $P(\mathbf{f}|\mathbf{x}^*, S)$

Since $\mathbf{f}$ does not depend on $\mathbf{x}^*$, it suffices to derive $P(\mathbf{f}|S)$. Again, we use standard Bayesian techniques to have

$$\begin{aligned} P(\mathbf{f}|S) &\propto P(S|\mathbf{f})P(\mathbf{f}) \\ &= P(\mathbf{y}, \mathbf{x}|\mathbf{f})P(\mathbf{f}) \\ &= P(\mathbf{y}|\mathbf{x}, \mathbf{f})P(\mathbf{x}|\mathbf{f})P(\mathbf{f}) \qquad (11) \\ &\propto P(\mathbf{y}|\mathbf{f})P(\mathbf{f}). \qquad (12) \end{aligned}$$

From (11) to (12), we use the fact that $P(\mathbf{y}|\mathbf{x}, \mathbf{f}) = P(\mathbf{y}|\mathbf{f})$ and $P(\mathbf{x}|\mathbf{f}) = P(\mathbf{x})$, which is assumed to have a uniform distribution. According to (1), (2) and (12),

$$\begin{aligned} P(\mathbf{f}|S) &\propto \exp\left(-\frac{(\mathbf{y}-\mathbf{f})^T(\mathbf{y}-\mathbf{f})}{2\sigma^2} - \frac{\mathbf{f}^T K^{-1}\mathbf{f}}{2} \right) \\ &\propto \exp\left(-\frac{\mathbf{f}^T\big(K^{-1}+\sigma^{-2}I\big)\mathbf{f} - 2\sigma^{-2}\mathbf{y}^T\mathbf{f}}{2} \right) \\ &\propto \exp\left(-\frac{\big(\mathbf{f}-\mathbf{u}\big)\Sigma^{-1}\big(\mathbf{f}-\mathbf{u}\big)}{2} \right), \end{aligned}$$

where

$$\begin{aligned} \Sigma &= \big(K^{-1}+\sigma^{-2}I\big)^{-1} \\ &= \big(K^{-1}+\sigma^{-2}KK^{-1}\big)^{-1} \\ &= \Big(\big(I+\sigma^{-2}K\big)K^{-1}\Big)^{-1} \\ &= \sigma^2 K\big(K+\sigma^2 I\big)^{-1} \end{aligned}$$

and

$$\mathbf{u} = \sigma^{-2}\Sigma\mathbf{y} = K\big(K+\sigma^2 I\big)^{-1}\mathbf{y}.$$

Therefore,

$$(\mathbf{f}|S) \sim N\Big(K\big(K+\sigma^2 I\big)^{-1}\mathbf{y}, \sigma^2 K\big(K+\sigma^2 I\big)^{-1}\Big). \qquad (13)$$

2.3 Derivation of $P(f^*|\mathbf{x}^*, S)$

For the ease of presentation, we define

$$\begin{aligned} \mathbf{a} &\equiv K^{-1}\mathbf{k}, \\ \triangle &\equiv \mathcal{K}(\mathbf{x}^*, \mathbf{x}^*) - \mathbf{k}^T K^{-1}\mathbf{k}, \\ \mathbf{b} &\equiv K\big(K+\sigma^2 I\big)^{-1}\mathbf{y}, \\ \Sigma &\equiv \sigma^2 K(K+\sigma^2 I)^{-1}. \end{aligned}$$

Then according to (3), (10) and (13),

$$
\begin{aligned}
& P(f^*|\mathbf{x}^*, S) \\
\propto & \int \exp\left(-\frac{\left(f^* - \mathbf{a}^T\mathbf{f}\right)^2}{2\triangle} - \frac{\left(\mathbf{f}-\mathbf{b}\right)^T\Sigma^{-1}\left(\mathbf{f}-\mathbf{b}\right)}{2}\right) d\mathbf{f} \\
= & \int \exp\left(-\frac{(f^*)^2}{2\triangle} - \frac{1}{2}\Big(\left(\mathbf{f}-\mathbf{b}\right)^T\Sigma^{-1}\left(\mathbf{f}-\mathbf{b}\right) - \frac{2f^*\mathbf{a}^T\mathbf{f}}{\triangle} + \mathbf{f}^T\frac{\mathbf{a}\mathbf{a}^T}{\triangle}\mathbf{f}\Big)\right) d\mathbf{f} \\
\propto & \int \exp\left(-\frac{(f^*)^2}{2\triangle} - \frac{1}{2}\Big(\mathbf{f}^T\big(\Sigma^{-1} + \frac{\mathbf{a}\mathbf{a}^T}{\triangle}\big)\mathbf{f} - 2\big(\Sigma^{-1}\mathbf{b} + \frac{f^*\mathbf{a}}{\triangle}\big)^T\mathbf{f}\Big)\right) d\mathbf{f} \\
\propto & \exp\left(-\frac{(f^*)^2}{2\triangle} + \frac{1}{2}\Big(\frac{f^*\mathbf{a}}{\triangle} + \Sigma^{-1}\mathbf{b}\Big)^T\Big(\Sigma^{-1} + \frac{\mathbf{a}\mathbf{a}^T}{\triangle}\Big)^{-1}\Big(\frac{f^*\mathbf{a}}{\triangle} + \Sigma^{-1}\mathbf{b}\Big)\right) \\
\propto & \exp\left(-\frac{1}{2}\Big(\frac{1 - \mathbf{a}^T\left(\triangle\Sigma^{-1} + \mathbf{a}\mathbf{a}^T\right)^{-1}\mathbf{a}}{\triangle}(f^*)^2 - 2\mathbf{a}^T\left(\triangle\Sigma^{-1} + \mathbf{a}\mathbf{a}^T\right)^{-1}\Sigma^{-1}\mathbf{b}f^*\Big)\right).
\end{aligned}
$$

Therefore,

$$
(f^*|\mathbf{x}^*, S) \sim N\big(\mu^*, (\sigma^*)^2\big)
$$

where

$$
\begin{aligned}
\mu^* &= \frac{\triangle\mathbf{a}^T\left(\triangle\Sigma^{-1} + \mathbf{a}\mathbf{a}^T\right)^{-1}\Sigma^{-1}\mathbf{b}f^*}{1 - \mathbf{a}^T\left(\triangle\Sigma^{-1} + \mathbf{a}\mathbf{a}^T\right)^{-1}\mathbf{a}}, \\
(\sigma^*)^2 &= \frac{\triangle}{1 - \mathbf{a}^T\left(\triangle\Sigma^{-1} + \mathbf{a}\mathbf{a}^T\right)^{-1}\mathbf{a}}.
\end{aligned}
$$

Using the *Sherman-Morrison-Woodbury formula*:

$$
\left(A + UV^T\right)^{-1} = A^{-1} - A^{-1}U\left(I + V^TA^{-1}U\right)^{-1}V^TA^{-1}, \tag{14}
$$

we have

$$
\begin{aligned}
\left(\triangle\Sigma^{-1} + \mathbf{a}\mathbf{a}^T\right)^{-1} &= \frac{\Sigma}{\triangle} - \frac{\Sigma}{\triangle}\mathbf{a}\Big(1 + \frac{\mathbf{a}^T\Sigma\mathbf{a}}{\triangle}\Big)^{-1}\mathbf{a}^T\frac{\Sigma}{\triangle} \\
&= \frac{1}{\triangle}\Big(\Sigma - \frac{\Sigma\mathbf{a}\mathbf{a}^T\Sigma}{\triangle + \mathbf{a}^T\Sigma\mathbf{a}}\Big).
\end{aligned}
$$

Consequently,

$$
\begin{aligned}
\mu^* &= \frac{\mathbf{a}^T\Big(\Sigma - \frac{\Sigma\mathbf{a}\mathbf{a}^T\Sigma}{\triangle+\mathbf{a}^T\Sigma\mathbf{a}}\Big)\Sigma^{-1}\mathbf{b}f^*}{1 - \frac{1}{\triangle}\mathbf{a}^T\Big(\Sigma - \frac{\Sigma\mathbf{a}\mathbf{a}^T\Sigma}{\triangle+\mathbf{a}^T\Sigma\mathbf{a}}\Big)\mathbf{a}} = \frac{1 - \frac{\mathbf{a}^T\Sigma\mathbf{a}}{\triangle+\mathbf{a}^T\Sigma\mathbf{a}}}{1 - \frac{\mathbf{a}^T\Sigma\mathbf{a}}{\triangle} + \frac{(\mathbf{a}^T\Sigma\mathbf{a})^2}{\triangle(\triangle+\mathbf{a}^T\Sigma\mathbf{a})}}\mathbf{a}^T\mathbf{b}f^* \\
&= \mathbf{a}^T\mathbf{b}f^* = \mathbf{k}^T\Big(K + \sigma^2 I\Big)^{-1}\mathbf{y},
\end{aligned}
$$

and

$$
\begin{aligned}
(\sigma^*)^2 &= \frac{\triangle}{1 - \frac{\mathbf{a}^T \Sigma \mathbf{a}}{\triangle} + \frac{(\mathbf{a}^T \Sigma \mathbf{a})^2}{\triangle(\triangle + \mathbf{a}^T \Sigma \mathbf{a})}} \\
&= \triangle + \mathbf{a}^T \Sigma \mathbf{a} \\
&= \triangle + \sigma^2 \mathbf{k}^T \left(\sigma^2 K + KK\right)^{-1} \mathbf{k} \qquad (15) \\
&= \triangle + \sigma^2 \mathbf{k}^T \left(\sigma^{-2} K^{-1} - \sigma^{-4} \left(K \sigma^{-2} + I\right)^{-1}\right) \mathbf{k} \qquad (16) \\
&= \triangle + \mathbf{k}^T K^{-1} \mathbf{k} - \mathbf{k}^T \left(K + \sigma^2 I\right)^{-1} \mathbf{k} \\
&= \mathcal{K}(\mathbf{x}^*, \mathbf{x}^*) - \mathbf{k}^T \left(K + \sigma^2 I\right)^{-1} \mathbf{k}.
\end{aligned}
$$

From (15) to (16), we use (14) with $A = \sigma^2 K$ and $U = V = K$. Finally, we are able to give the predictive distribution:

$$
(f^* | \mathbf{x}^*, S) \sim N\left(\mathbf{k}^T \left(K + \sigma^2 I\right)^{-1} \mathbf{y}, \mathcal{K}(\mathbf{x}^*, \mathbf{x}^*) - \mathbf{k}^T \left(K + \sigma^2 I\right)^{-1} \mathbf{k}\right).
$$