Gaussian process regression

Naoya Kawakami

July 8, 2023
version 1.0

1 Regression

Regression is an analysis to estimate the relationship between the data. The most simple example is the simple linear regression:

$$\hat{y} = a + bx. \tag{1}$$

Equation 1 regression equation of y for x. The value with hat ($\hat{y}$) indicates that the value is obtained from the regression equation.[1] The coefficient a and b are adjusted to fit the data.

If the input is multidimensional ($\boldsymbol{x}^T = (x_1, x2, \ldots, x_D)$), the simple linear regression is expressed as the linear combination of each input as follows:[2][3]

$$\begin{align} \hat{y} &= w_0 + w_1 x_1 + \cdots + w_D x_D \tag{2}\\ &= (x_0, \ldots, x_D) \begin{pmatrix} w_0 \\ \vdots \\ w_D \end{pmatrix} \tag{3}\\ &= \boldsymbol{x}^T \boldsymbol{w}. \tag{4} \end{align}$$

This eqution yields the output $\hat{y}_n$ for all the possible data points $\boldsymbol{x}_n^T = (x_{n1}, \ldots, x_{nD})$ as follows:

$$\begin{align} \hat{y}_1 &= \boldsymbol{x}_1^T \boldsymbol{w}, \tag{5}\\ \hat{y}_2 &= \boldsymbol{x}_2^T \boldsymbol{w}, \tag{6}\\ &\vdots \tag{7}\\ \hat{y}_N &= \boldsymbol{x}_N^T \boldsymbol{w}. \tag{8}\\ & \tag{9} \end{align}$$

[1] Simple y is used to express the experimental data, which may differ from the regression value because of the noise.

[2] The superscript ^{T} represents the transpose of the matrix. The vector without transpose is the vertical vector in this document.

[3] The input $\boldsymbol{x}$ is extended to include x_0. The x_0 is a dummy input to express the offset w_0. Note that $x_0 = 1$.

The regression for all these points $\boldsymbol{x}_n^T$ is summarized as follows:

$$\begin{aligned}
\hat{\boldsymbol{y}} = \begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_N \end{pmatrix} &= \begin{pmatrix} \boldsymbol{x}_{\boldsymbol{1}}^T \\ \boldsymbol{x}_2^T \\ \vdots \\ \boldsymbol{x}_N^T \end{pmatrix} \boldsymbol{w} & (10) \\
&= \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1D} \\ 1 & x_{21} & x_{22} & \dots & x_{2D} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{N1} & x_{22} & \dots & x_{ND} \end{pmatrix} \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_D \end{pmatrix} & (11) \\
&= \boldsymbol{X}\boldsymbol{w}. & (12)
\end{aligned}$$

Here, $\boldsymbol{X}$ represents the data points of regression, called the **design matrix**.

2 Linar regression

Linear regression is the extension of the simple linear regression. The regression is performed by the linear combination of arbitrary functions $\phi_n(\boldsymbol{x})$ as follows:

$$\hat{y} = w_0 + w_1\phi_1(\boldsymbol{x}) + \cdots + w_H\phi_H(\boldsymbol{x}) \qquad (13)$$

Comparing eq. 13 with eq. 3, it is the replacement of $x_n \to \phi_n$. Thus, the design matrix of this regression is expressed by[4]

$$\begin{aligned}
\hat{\boldsymbol{y}} &= \boldsymbol{X}\boldsymbol{w} & (14) \\
\boldsymbol{X} &= \begin{pmatrix} \phi_0(\boldsymbol{x}_1) & \phi_1(\boldsymbol{x}_1) & \dots & \phi_H(\boldsymbol{x}_1) \\ \phi_0(\boldsymbol{x}_2) & \phi_1(\boldsymbol{x}_2) & \dots & \phi_H(\boldsymbol{x}_2) \\ \vdots & \vdots & & \vdots \\ \phi_0(\boldsymbol{x}_D) & \phi_1(\boldsymbol{x}_D) & \dots & \phi_H(\boldsymbol{x}_D) \end{pmatrix} & (15)
\end{aligned}$$

The horizontal components $\boldsymbol{\phi}(\boldsymbol{x})^T = (\phi_0(\boldsymbol{x}), \ldots, \phi_H(\boldsymbol{x}))^T$ is called **feature vector**, representing the value of each function at $\boldsymbol{x}$.

3 Gaussian distribution

The Gaussian distribution (with a single variable) is defined as follows:

$$N(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \qquad (16)$$

Here, μ and σ represent the mean value and standard deviation, respectively.[5]

The Gaussian distribution (with a single variable) is defined as follows:

$$N(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \qquad (17)$$

Here, μ and σ represent the mean value and standard deviation, respectively.[6]

[4] $\phi_0(\boldsymbol{x}) = 1$

[5] The expression here is normalized to express the probability.

[6] The expression here is normalized to express the probability.

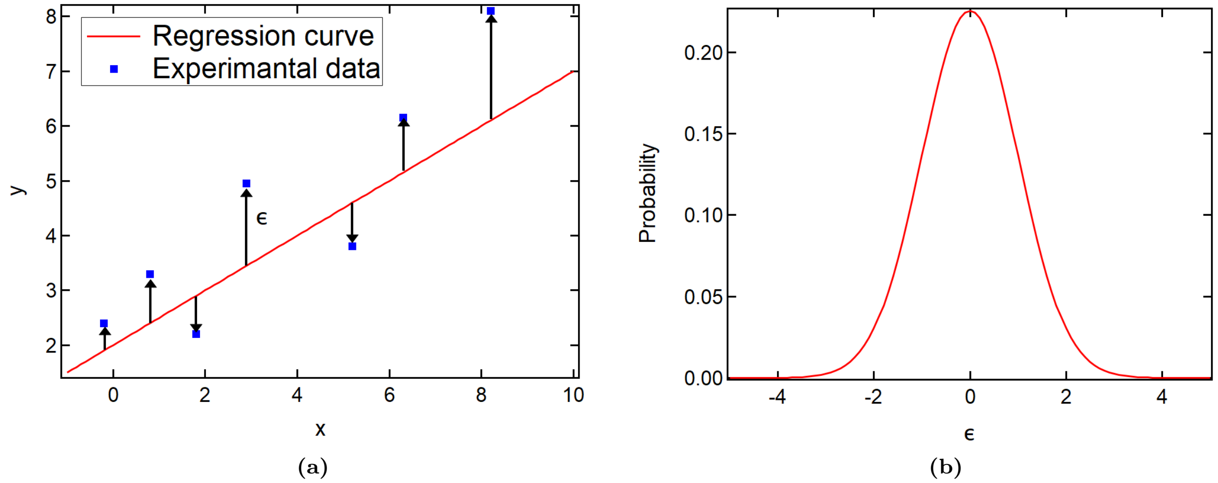


Figure 1: (a) The relationship between the regression curve (red) and experimental data (blue dots). Experimental data are displaced from the regression value by an error ϵ. (b) Gaussian distribution, showing the distribution of errors.

The Gaussian distribution is often used to express the probabilistic event. Here, we assume that the experimental data includes probabilistic noise as follows:

$$y = \boldsymbol{x}^T \boldsymbol{w} + \epsilon, \tag{18}$$

$$p(\epsilon) = N(0, \sigma^2). \tag{19}$$

$\boldsymbol{x}^T \boldsymbol{w}$ is the ideal value obtained from the regression curve. ϵ represents the probabilistic noise, which follows the Gaussian distribution. This relation is schematically represented in Fig. 1. The probability of getting y at $\boldsymbol{x}$ is

$$\begin{aligned} p(y|x) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\epsilon^2}{2\sigma^2}\right) & (20) \\ &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y - \boldsymbol{x}^T \boldsymbol{w})^2}{2\sigma^2}\right). & (21) \end{aligned}$$

If we have N data sets of $(y_n, \boldsymbol{x}_n)$, the probability of yielding this data set is expressed as follows:

$$\begin{aligned} p(\boldsymbol{y}|\boldsymbol{x}) &= \prod_{n=1}^{N} p(y_n|\boldsymbol{x}) & (22) \\ &= \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_n - \boldsymbol{x}_n^T \boldsymbol{w})^2}{2\sigma^2}\right) & (23) \end{aligned}$$

Taking the logarithm of both sides yields

$$\log(p(\boldsymbol{y}|\boldsymbol{x})) = -N \log(\sqrt{2\pi}\sigma) - \sum_{n=1}^{N} \frac{(y_n - \boldsymbol{x}_n^T \boldsymbol{w})^2}{2\sigma^2}. \tag{24}$$

The best regression curve should maximize the probability, which is realized when

$$\sum_{n=1}^{N} (y_n - \boldsymbol{x}_n^T \boldsymbol{w})^2 \tag{25}$$

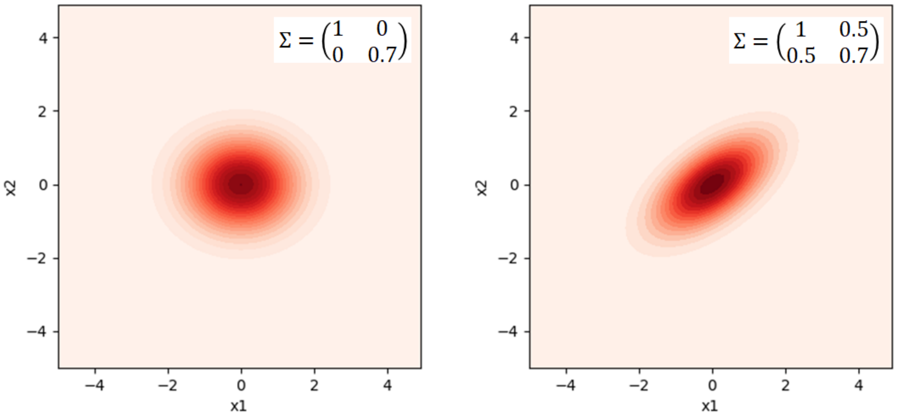


Figure 2: 2D Gaussian distribution with different variance-covariance matrix. The matrix used for drawing the figures is shown on the upper right side.

is minimized. Equation 25 is the same as the mean square error. We often use mean square error to evaluate the fitting accuracy. Now we know that using mean square error equals considering the error with Gaussian distribution.

4 Multivariable Gaussian distribution

We will introduce the basics of Gaussian distribution with multiple variables. The D-dimensional multivariable Gaussian distribution is expressed as follows:

$$N(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(\sqrt{2})^D\sqrt{|\boldsymbol{\Sigma}|}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right). \tag{26}$$

Here, Σ is called the **variance-covariance matrix**. We will express each component of the matrix as follows:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1D} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2D} \\ \vdots & \vdots & & \vdots \\ \sigma_{D1} & \sigma_{D2} & \dots & \sigma_D^2 \end{pmatrix} \tag{27}$$

The diagonal component is the variance of each variable, the same as in the 1D case. The covariance is defined by the average of the multiple of two variances as follows:

$$\begin{aligned} \Sigma_{ij} &= \boldsymbol{E}[(x_i - \boldsymbol{E}[x_i])(x_j - \boldsymbol{E}[x_j]] && (28) \\ &= \boldsymbol{E}[x_i x_j - x_i\boldsymbol{E}[x_j] - x_j\boldsymbol{E}[x_i] + \boldsymbol{E}[x_i]\boldsymbol{E}[x_j]] && (29) \\ &= \boldsymbol{E}[x_i x_j] - \boldsymbol{E}[x_i]\boldsymbol{E}[x_j] && (30) \\ \therefore \boldsymbol{\Sigma} &= \boldsymbol{E}[\boldsymbol{x}\boldsymbol{x}^T] - \boldsymbol{E}[\boldsymbol{x}]\boldsymbol{E}[\boldsymbol{x}]^T && (31) \end{aligned}$$

$\boldsymbol{E}[x]$ represents the average of all x. When there is a tendency for the x_i and x_j to take a similar value, the covariance gets larger. In other words, the covariance represents the correlation between the two variables. The inverse of the covariance matrix $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$ is called the **precision matrix**.

Figure 2 shows two examples of Gaussian distribution with two variables (x_1 and x_2) using a different matrix. The covariance is zero in Fig. 2(a), while the variance of x_1 and x_2 is set to 1 and 0.7, resulting in the eclipse shape distribution. The positive covariance indicates that the two variables are positively correlated. Thus, the distribution in 2(b) is tilted.

The following subsection introduces several properties of multivariance Gaussian distribution.

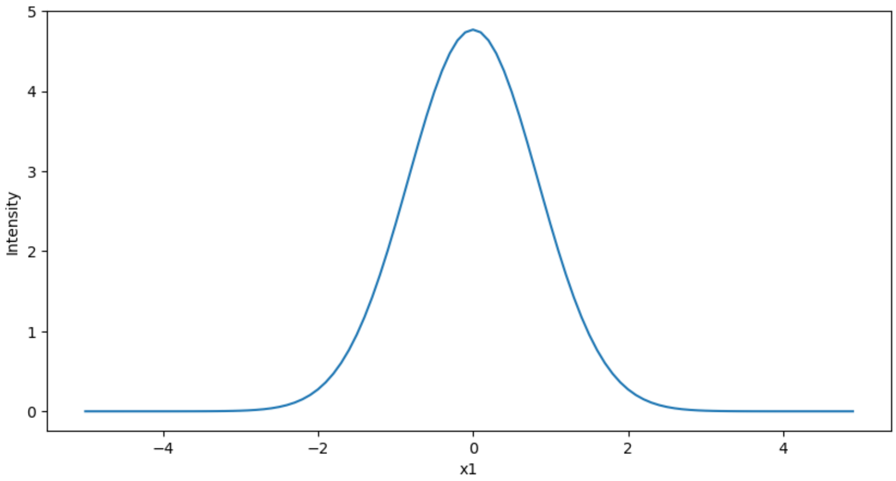


Figure 3: Marginalized distribution of Fig. 2(b). Integration is performed for x_2.

4.1 Linear transformation

Suppose a series of variables $\boldsymbol{x}$ follows the multivariable Gaussian distribution, and $\boldsymbol{x}$ is transformed by a matrix $\boldsymbol{A}$ as follows:

$$\begin{aligned} \boldsymbol{x} &= N(0, \boldsymbol{\Sigma}), \quad (32) \\ \boldsymbol{y} &= \boldsymbol{A}\boldsymbol{x}. \quad (33) \end{aligned}$$

The resulting distribution $\boldsymbol{y}$ also obeys the Gaussian distribution with the precision matrix of

$$\boldsymbol{\Lambda} = (\boldsymbol{A}^{-1})^T \boldsymbol{\Sigma}^{-1} \boldsymbol{A}^{-1}. \quad (34)$$

4.2 Marginalizing

We can get a new Gaussian distribution with fewer variables by integrating certain variables. This process is called **marginalizing**. For example, let us consider two variables case. The integration of a variable yield

$$p(\boldsymbol{x_1}) = \int p(\boldsymbol{x}_1, \boldsymbol{x}_2) d\boldsymbol{x}_2 = N(\mu_1, \Sigma_{11}). \quad (35)$$

Figure 3 shows the example of marginalization, in which the x_2 is marginalized, resulting in the distribution of x_1. Marginalization corresponds to ignoring the variables. A set (x_1, x_2) would be obtained in data sampling. By plotting the distribution of x_1 with ignoring x_2, the distribution shown in Fig. 3 will be obtained. Marginalization can be performed for multiple variables.

4.3 Conditional probability distribution

If we apply a condition like $x_i = a$, a new distribution of remaining variables will be obtained. The resulting distribution is called **conditional probability distribution**, which is a new Gaussian distribution. The distribution of $\boldsymbol{x}_2$ when $\boldsymbol{x}_1$ is fixed is expressed as follows:

$$p(\boldsymbol{x}_2|\boldsymbol{x}_1) = N(\mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(\boldsymbol{x}_1 - \boldsymbol{\mu}), \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}) \quad (36)$$

Figure 4 shows an example. The x_2 of the distribution Fig. 2 is fixed at $x_2 = 1$, and x_1 shows the new Gaussian distribution.

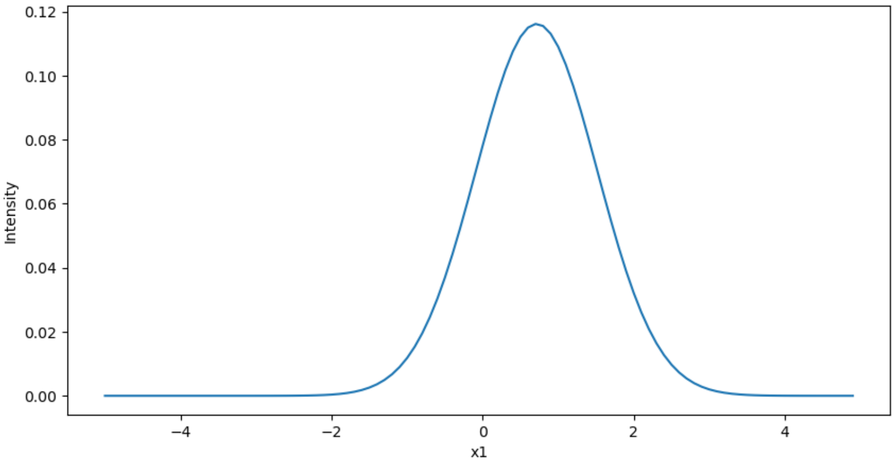


Figure 4: Distribution of x_1 when $x_2 = 1$ for the distribution in Fig. 2(b).

5 Gaussian process

In the traditional regression process, the design matrix and weights determine the regression curve.

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{w} \tag{37}$$

$$\boldsymbol{X} = \begin{pmatrix} \phi_0(\boldsymbol{x}_1) & \phi_1(\boldsymbol{x}_1) & \dots & \phi_H(\boldsymbol{x}_1) \\ \phi_0(\boldsymbol{x}_2) & \phi_1(\boldsymbol{x}_2) & \dots & \phi_H(\boldsymbol{x}_2) \\ \vdots & \vdots & \dots & \vdots \\ \phi_0(\boldsymbol{x}_D) & \phi_1(\boldsymbol{x}_D) & \dots & \phi_H(\boldsymbol{x}_D) \end{pmatrix}, \boldsymbol{w} = \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_H \end{pmatrix}. \tag{38}$$

Note that the design matrix $\boldsymbol{X}$ is a constant matrix if the data points $\boldsymbol{x}_1 \sim \boldsymbol{x}_D$ are given. Suppose each weight is generated from a multivariable Gaussian distribution as follows:[7]

$$\boldsymbol{w} = N(0, \lambda^2 \boldsymbol{I}). \tag{39}$$

Equation 38 indicates that $\hat{\boldsymbol{y}}$ is the linear transform of $\boldsymbol{w}$. Thus, as discussed in subsection 4.1, $\hat{\boldsymbol{y}}$ also follows the multivariable Gaussian distribution as follows:

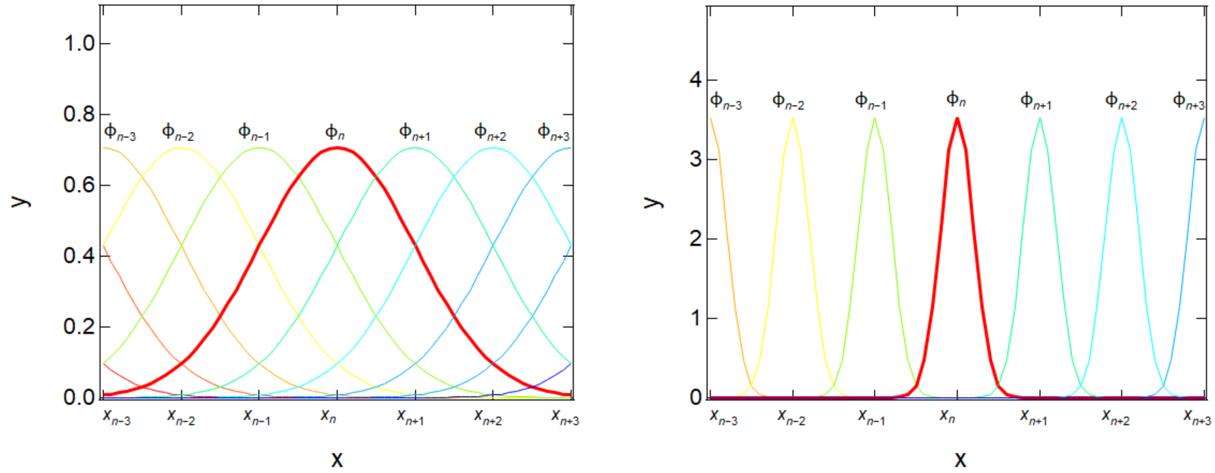
$$\hat{\boldsymbol{y}} = N(0, \boldsymbol{K}) \tag{40}$$

$$\boldsymbol{K} = \lambda^2 \boldsymbol{X}\boldsymbol{X}^T \tag{41}$$

Note that the distribution of $\hat{\boldsymbol{y}}$ is determined without the information of $\boldsymbol{w}$.

When the output $\boldsymbol{y} = (y_1, y_2, \dots, y_n)$ always follows the multivariable Gaussian distribution for all the $\boldsymbol{x} = (x_1, x_2, \dots, x_n)$, we say that $\boldsymbol{x}$ and $\boldsymbol{y}$ follows the **Gaussain process**.

[7] $\boldsymbol{I}$ represents the primitive matrix.



(a)

(b)

Figure 5: The basis functions when a series of Gaussian functions are placed at each grid point. The σ is set at (a) large and (b) small values.

5.1 The meaning of Gaussian process

The nn' component of the cvariance-covariance matrix $\boldsymbol{K}$ is expressed as follows:

$$\begin{aligned}
\boldsymbol{K} &= \lambda^2 \boldsymbol{X}\boldsymbol{X}^T &(42)\\
&= \lambda^2 \begin{pmatrix} \phi_0(\boldsymbol{x}_1) & \phi_1(\boldsymbol{x}_1) & \dots & \phi_H(\boldsymbol{x}_1) \\ \phi_0(\boldsymbol{x}_2) & \phi_1(\boldsymbol{x}_2) & \dots & \phi_H(\boldsymbol{x}_2) \\ \vdots & \vdots & \dots & \vdots \\ \phi_0(\boldsymbol{x}_D) & \phi_1(\boldsymbol{x}_D) & \dots & \phi_H(\boldsymbol{x}_D) \end{pmatrix} \begin{pmatrix} \phi_0(\boldsymbol{x}_1) & \phi_0(\boldsymbol{x}_2) & \dots & \phi_0(\boldsymbol{x}_D) \\ \phi_1(\boldsymbol{x}_1) & \phi_1(\boldsymbol{x}_2) & \dots & \phi_1(\boldsymbol{x}_D) \\ \vdots & \vdots & \dots & \vdots \\ \phi_H(\boldsymbol{x}_1) & \phi_H(\boldsymbol{x}_2) & \dots & \phi_H(\boldsymbol{x}_D) \end{pmatrix} &(43)\\
&= \lambda^2 \begin{pmatrix} \boldsymbol{\phi}^T(\boldsymbol{x}_1) \\ \boldsymbol{\phi}^T(\boldsymbol{x}_2) \\ \vdots \\ \boldsymbol{\phi}^T(\boldsymbol{x}_H) \end{pmatrix} (\boldsymbol{\phi}(\boldsymbol{x}_1), \boldsymbol{\phi}(\boldsymbol{x}_2), \dots, \boldsymbol{\phi}(\boldsymbol{x}_H)) &(44)\\
\therefore K_{nn'} &= \lambda^2 \boldsymbol{\phi}(\boldsymbol{x}_n)^T \boldsymbol{\phi}(\boldsymbol{x}_{n'}) &(45)
\end{aligned}$$

Equation 45 indicates that the covariance of $y(\boldsymbol{x}_n)$ and $y(\boldsymbol{x}_{n'})$ is determined by the inner product of the feature vectors. If x_n and $x_{n'}$ are close to each other, it is natural to think that $y(\boldsymbol{x}_n$ and $y(\boldsymbol{x}_{n'})$ are also the similar value, that should result in the large covariance between y_n and $y_{n'}$. The similarity is determined by what basis function to use.

Let us think about a simple example. We assume that Gaussian functions, placed at each mesh in the x axis, are used as the basis functions. Figures 5 (a) and (b) are examples of the basis functions with large and small σ. When σ is large, each Gaussian basis function extends well over the neighbor points. Therefore, when the value at x_n is large, it is expected that the neighbor points x_{n-1} and x_{n+1} also have relatively large values, indicating the large correlation between the neighbor points. When σ is small, each basis function is localized around each point. Therefore, even when the value at x_n is large, it does not mean the neighbor points also have a large value, indicating a slight correlation. In this case, the correlation between each point is determined by the σ of the basis function. The correlation between each point is determined by what basis functions to use, as mathematically expressed as eq. 45.

5.2 Kernel trick

We have shown that the covariance of $K_{nn'} = \lambda^2 \boldsymbol{\phi}(\boldsymbol{x}_n)^T \boldsymbol{\phi}(\boldsymbol{x}_{n'})$ determines the dispersion of $\boldsymbol{y}$. $k(\boldsymbol{x}_n, \boldsymbol{x}_{n'}) = \boldsymbol{\phi}(\boldsymbol{x}_n)^T \boldsymbol{\phi}(\boldsymbol{x}_{n'})$ is called **kernel function**. $k_{nn'}$ can be calculated from the feature vectors. However, knowing the feature vector itself is unnecessary to calculate $\boldsymbol{y}$. It is called the **kernel trick**.

For example, let us consider the following kernel function:

$$k(\boldsymbol{x}, \boldsymbol{x'}) = (\boldsymbol{x}^T \boldsymbol{x'} + 1)^2, \tag{46}$$

$$\boldsymbol{x} = (x_1, x_2), \boldsymbol{x'} = (x'_1, x'_2). \tag{47}$$

This kernel function comes from the feature function

$$\boldsymbol{\phi} = (x_1^2, , x_2^2, \sqrt{2}x_1x_2, \sqrt{2}x_1, \sqrt{2}x_2, 1)^T. \tag{48}$$

The covariance of $\boldsymbol{y}$ can be determined from the eq. 47. Thus, it is unnecessary to calculate the feature function itself. What kernel function to use is one of the hyperparameters in the Gaussian process. However, the basis function is often not apparently shown; instead, the kernel function is. Therefore, the relationship between the kernel function and the basis function behind is better understood in advance. We will introduce several famous kernel functions in the next subsection.

5.3 Kernels

What kernel to use is one of the hyper-parameters in the Gaussian process. We will introduce several famous kernels.

Firstly, let us consider that a series of Gaussian functions are used as the basis functions, as shown in Fig. 5. The basis functions are expressed as

$$\phi_h = \tau \exp\left(-\frac{(x - h/H)^2}{r^2}\right) \quad (h = -H^2, -H^2 + 1, \ldots, H^2). \tag{49}$$

Equation 49 corresponds that Gaussian functions are placed at each $1/H$ point along the x axis from $-H$ to $+H$. The feature vector is expressed as

$$\boldsymbol{\phi} = (\phi_{-H^2}(x), \ldots, \phi_{H^2}(x)). \tag{50}$$

The kernel function is calculated to be

$$k(x, x') = \theta_1 \exp(-\frac{1}{\theta^2}(x - x')^2), \tag{51}$$

$$\theta_1 = \tau^2 \sqrt{\pi r^2}/2, \theta_2 = 2r^2. \tag{52}$$

Equation 52 is called Gauss kernel or radius basis function (RBF) kernel. Again, for calculating the covariance, the kernel function is enough, and the original basis function is not necessary to be shown. When the RBF kernel is used, note that it assumes the eq. 49 as the basis functions.

The followings are the other representative kernels:

$$k(x, x') = x^T x' \tag{53}$$

$$k(x, x') = \exp(-\frac{|x - x'|}{\theta}) \tag{54}$$

$$k(x, x') = \exp(\theta_1 \cos(\frac{|x - x'|}{\theta_2})) \tag{55}$$

$$k(x, x') = \frac{2^{1-\gamma}}{\Gamma(\gamma)}(\frac{\sqrt{2}\gamma r}{\theta})K_\gamma(\frac{\sqrt{2}\gamma r}{\theta}) \tag{56}$$

The kernels in eq. 54 to eq. 56 is called the linear kernel, exponential kernel, periodic kernel, and Matern kernel. The Matern kernel equals the exponential kernel when $\gamma = 1/2$, and the RBF kernel when $\gamma \to \infty$. Each kernel corresponds to the regressions using different basis functions. In machine learning, the RBF kernel is often used.

5.4 Observation noise

In a real experiment, the observed value always accompanies noise. We assume that the observed signal is composed of regression value and Gaussian noise as follows:

$$y_n = f(x_n) + \epsilon_n, \tag{57}$$

$$\epsilon_n \sim N(0, \sigma^2). \tag{58}$$

In this case, the distribution of $\boldsymbol{y}$ is expressed by the Gaussian distribution as follows:

$$p(\boldsymbol{y}|\boldsymbol{x}) \quad = \quad N(\mu, k + \sigma \underline{\mathrm{I}}), \tag{59}$$

$$k'(x_n, x_{n'}) \quad = \quad k(x_n, x_{n'}) + \sigma^2 \delta(n, n'). \tag{60}$$

The variance of $\boldsymbol{y}$ is the sum of the original kernel and noise distribution. As such, the effect of noise in the Gaussian process is included in the kernel.

5.5 Regression in Gaussian process

We assume that we have the observed data set D:

$$D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}. \tag{61}$$

We also assume that the mean value of all y_n is subtracted. We will see how we get the expected distribution of unknown data point (x^*, y^*). We define the following vectors:

$$\boldsymbol{y} \quad = \quad (y_1, y_2, \ldots, y_N), \tag{62}$$

$$\boldsymbol{y'} \quad = \quad (y_1, y_2, \ldots, y_N, y^*) \tag{63}$$

The variance-covariance matrix $\boldsymbol{K'}$ for the data points can be calculated if $x_1, x_2, \ldots, x_N, x^*$ is given. The $\boldsymbol{y'}$ floows the Gaussian distribution, $\boldsymbol{y'} \sim N(0, K')$. The matrix K' is composed of the matrix for observed data points (K) and additional components as follows:

$$\boldsymbol{K'} = \begin{pmatrix} \boldsymbol{K} & \boldsymbol{k}_* \\ \boldsymbol{k}_*^T & \boldsymbol{k}_{**} \end{pmatrix}. \tag{64}$$

Here, $\boldsymbol{k}_*$ is the covariance related to x^*, and k_{**} is the variance at x^*. If the kernel function is given, the matrix is calculated from $\boldsymbol{x}$. To summarise, $\boldsymbol{y'}$ follows the following distribution:

$$\begin{pmatrix} \boldsymbol{y} \\ y* \end{pmatrix} \sim N\left(0, \begin{pmatrix} \boldsymbol{K} & \boldsymbol{k}_* \\ \boldsymbol{k}_*^T & \boldsymbol{k}_{**} \end{pmatrix}\right) \tag{65}$$

The condition for the data $\boldsymbol{y}$ is given by the experiment. Thus, we can consider the conditional probability as introduced in subsection 4.3. Using eq. 36, the distribution of y^* after the data points $\boldsymbol{y}$ were given are obtained as

$$p(y^*|\boldsymbol{y}) = N(\boldsymbol{k}_*^T \boldsymbol{K}^{-1} \boldsymbol{y}, \boldsymbol{k}_{**} - \boldsymbol{k}_*^T \boldsymbol{K}^{-1} k_*) \tag{66}$$

Equation 66 represents the expected distribution of unknown datapoint (y^*, x^*). The Gaussian process gives the expected value and the distribution of unknown data points. It means the regression result yields how reliable the regression is at unknown data points. If the variance of y^* is large, it indicates that the expected value still has a huge error and is not so reliable.

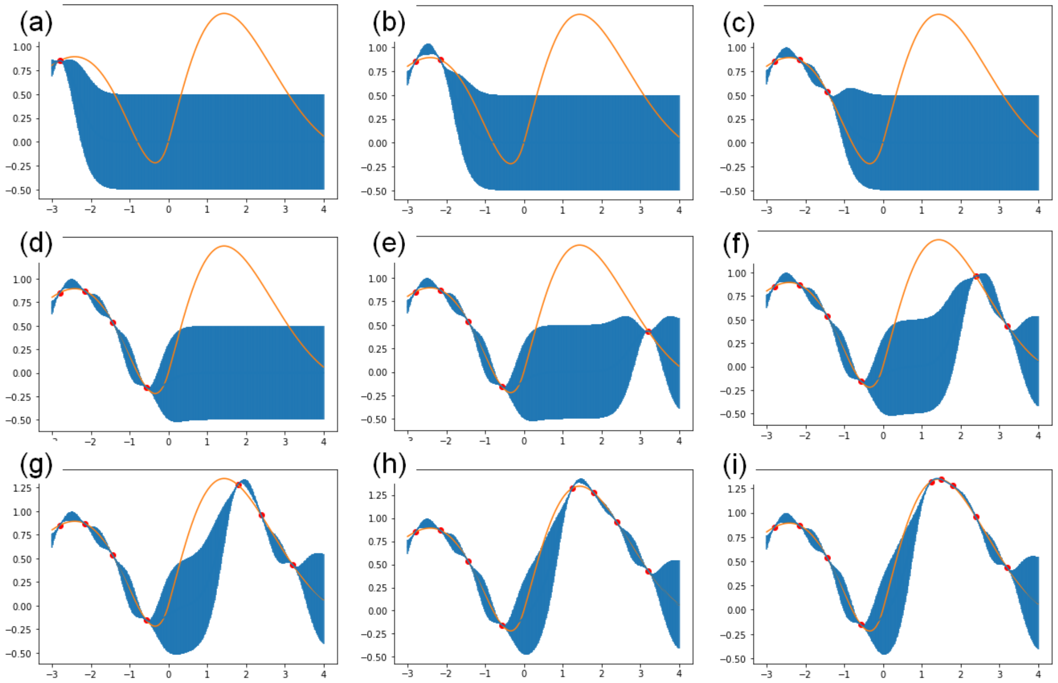


Figure 6: The Bayesian optimization process. The orange curve is the true function. The red point is the observed data point. The blue band is the variance at each point suggested by Gaussian process regression.

5.6 Leaning in Gaussian process

The variance-covariance matrix is determined only from the kernel function. Thus, the learning target is the parameters in the kernel functions. For example, the RBF kernel with noise is expressed as

$$k(x,x') = \theta_1 \exp\left(-\frac{(x-x')^2}{\theta_2}\right) + \theta_3\delta(x,x'). \tag{67}$$

In this kernel, there are three parameters ($\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$). These three parameters will be adjusted to yield the data with the highest probability.

Generally, what kernel to choose is one of the hyperparameters. However, if we combine several kernels like

$$k(x,x') = \theta_1 + \theta_2 \boldsymbol{x}^T \boldsymbol{x}' + \theta_3 \exp\left(-\frac{(x-x')^2}{\theta_2}\right) + \theta_5\delta(x,x') \tag{68}$$

, then the choice of the kernel is unnecessary. However, increasing the number of parameters leads to a higher computational cost.

6 Bayesian optimization

Sometimes we should find maxima of unknown functions in the experiment. [8] A simple method to find out the best parameter is to investigate all the parameters in a grid. However, investigating all the grid points is time- and cost-consuming. What is the practical method to search for the maxima efficiently?

[8] For example, we want to find a temperature to synthesize a material with the highest crystallinity, but we do not know the dependence of crystallinity on temperature.

An important characteristic in searching the maxima in unknown functions is that the searching is not trapped in local maxima. For example, see Fig. 6(d). The orange curve is the actual function, and the red point indicates the observed point. The four observed points look to form a hill showing maxima at around $x = -2.2$. We may misunderstand it as the global maxima while it is just a local maxima. We should have a strategy to search the global maxima without being trapped by the local maxima.

Here we will see that Bayesian optimization based on Gaussian process regression (GPR) is effective. As we have discussed, GPR gives the prediction with the variance (or we can call it confidence). The global maxima can be found at the point with the largest prediction. However, a new maximum can be hidden at the point with low confidence. Therefore, the prediction and confidence points should be considered to determine the search point. It can be done by using a function like

$$a(x) = \mu_x + \epsilon\sigma_x \tag{69}$$

, where mu_x and $sigma_x$ is the mean and variance values at x. By using eq. 69, both the high prediction and low confidence (high σ) points will be the candidate for the next point. ϵ is the coefficient to determine the balance. If ϵ is small, the program preferentially searches the points with the largest prediction. Thus, it soon reaches the maxima nearby the observed point, but local maxima can trap it. If ϵ is large, a point with low confidence is preferred. However, it takes a long time to be converged. The function to determine the next search parameter based on prediction and variance is called the **acquisition function**.[9]

Figure 6 shows the example of the Bayesian optimization, using $\epsilon = 2.3$. In the beginning, all the y distributes at around zero. After getting the data shown in (a), the program shows high confidence around the point. Because the value of the first point is large, the points nearby have relatively large expected values. Therefore, the program proposes a point nearby, as shown in (b). After getting new data, the GPR result is updated, and the next point is proposed based on the acquisition function.

After getting four data points in (d), the shape of the curve at the negative x side is almost revealed with high confidence. However, the values at the positive x side are almost unknown, thus showing a large variance. Therefore, the program proposes to get the value at positive x, as shown in (e). By repeating this process, the process successfully finds the global maxima as in (i) with a relatively small number of data points.

[9] Various acquisition functions with different characteristics have been proposed. Check it if necessary.