ガウス過程回帰

Gaussian Process Regression

大阪府立大学 工学研究科
清水　悠生

本記事の内容と対象

- ✓ ガウス過程による回帰を
  なんとなく理解するための記事

- ✓ この順番で説明
  ① ガウス分布とは
  ② ガウス過程回帰をざっくり理解する
  ③ ガウス過程の導出と意味
  ④ ガウス過程を用いた回帰分析

①ガウス分布とは

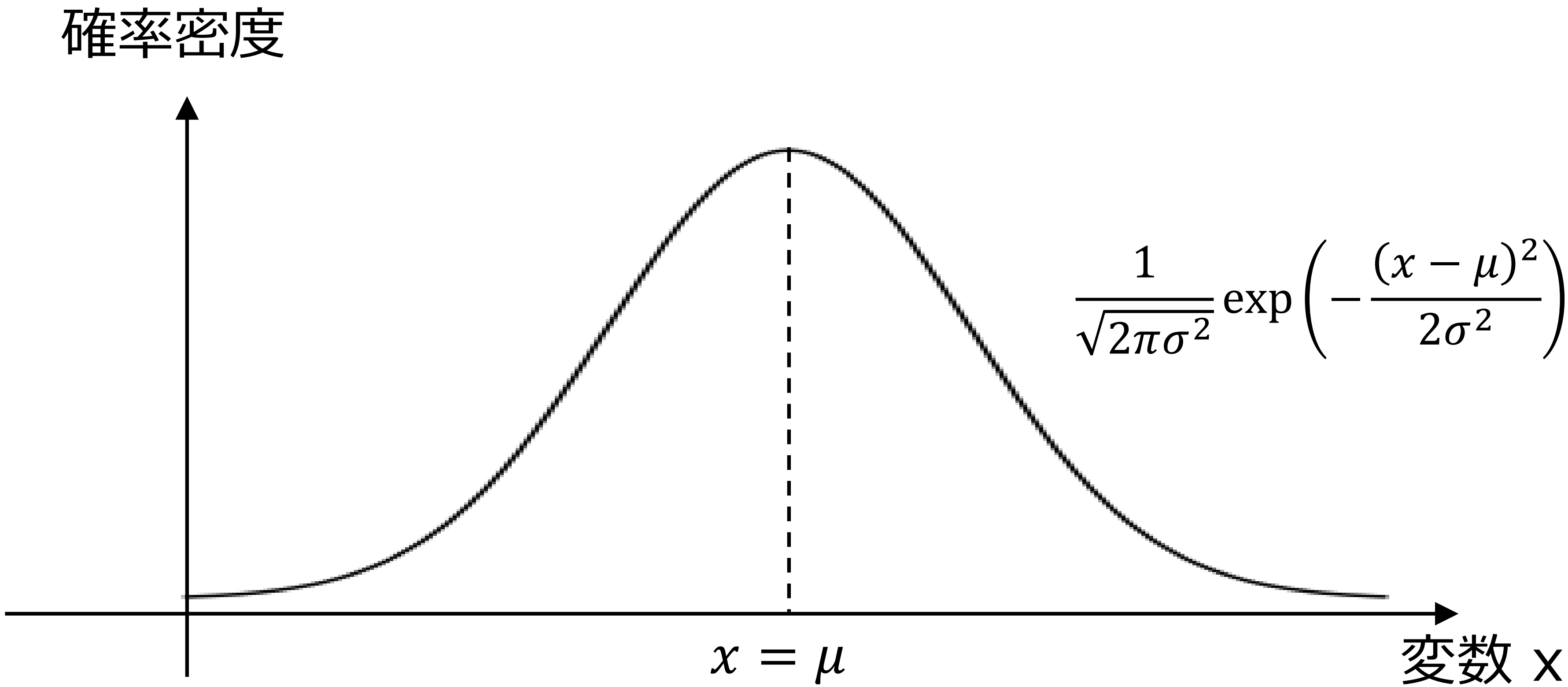
- ✓ まずガウス過程回帰の名前の由来でもある
  ガウス分布と呼ばれる確率分布の話からスタート

ガウス分布（正規分布）とは

- ✓ 確率密度が次式で与えられる分布を **ガウス分布（正規分布）** と呼び，$N(\mu, \sigma^2)$ で表す

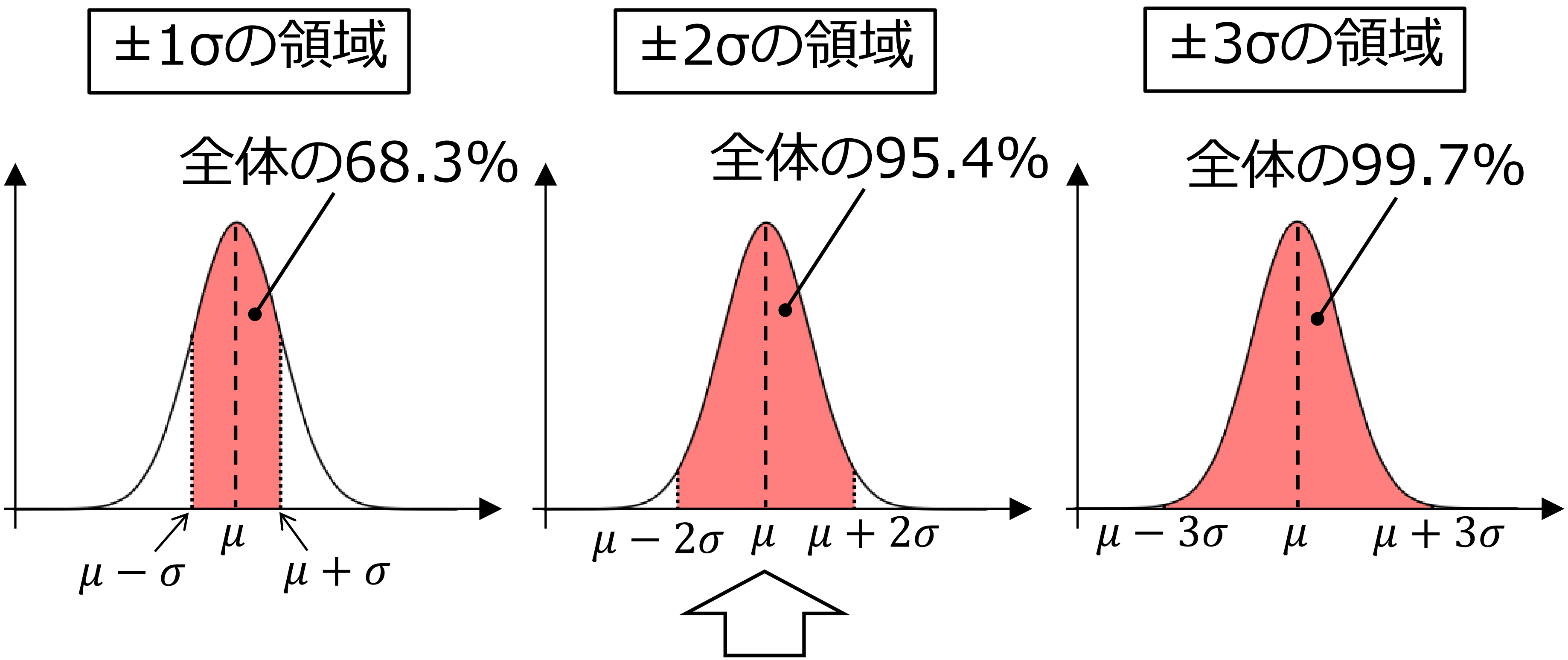
$$N(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

μ：平均
σ^2：分散
σ：標準偏差



標準偏差と確率

- ✓ 平均を中心とした標準偏差の整数倍以内のデータが出力される確率は下図のとおり

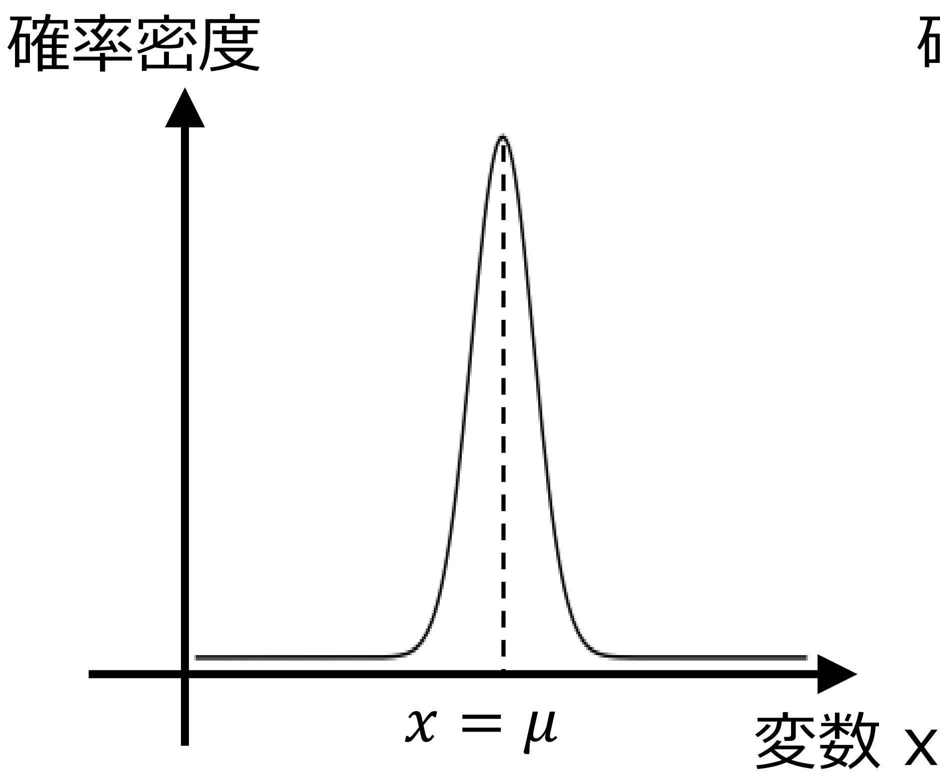


95.4%の確率で $\mu \pm 2\sigma$ の間の値が出力される！ という意味

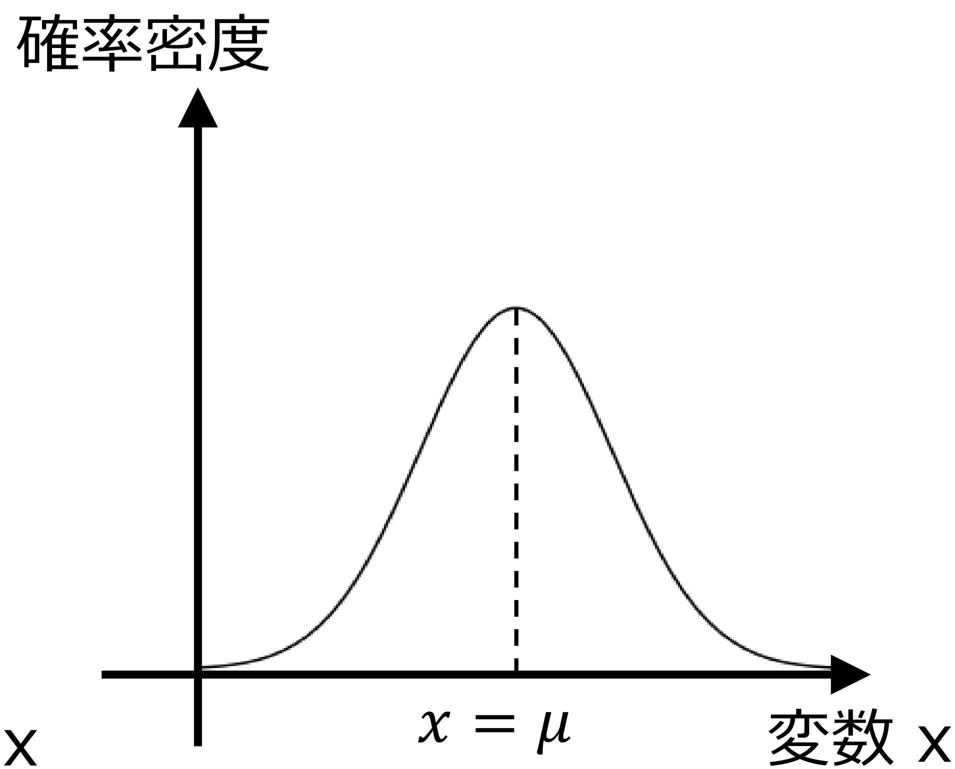
分散のイメージ

- ✓ 分散はばらつき具合を表す指標であり
  分散が小さいほど平均値に近い値が得られやすくなる

分散：小



分散：大



2次元のガウス分布

- ✓ ガウス分布を2次元に拡張すると次式のようになる

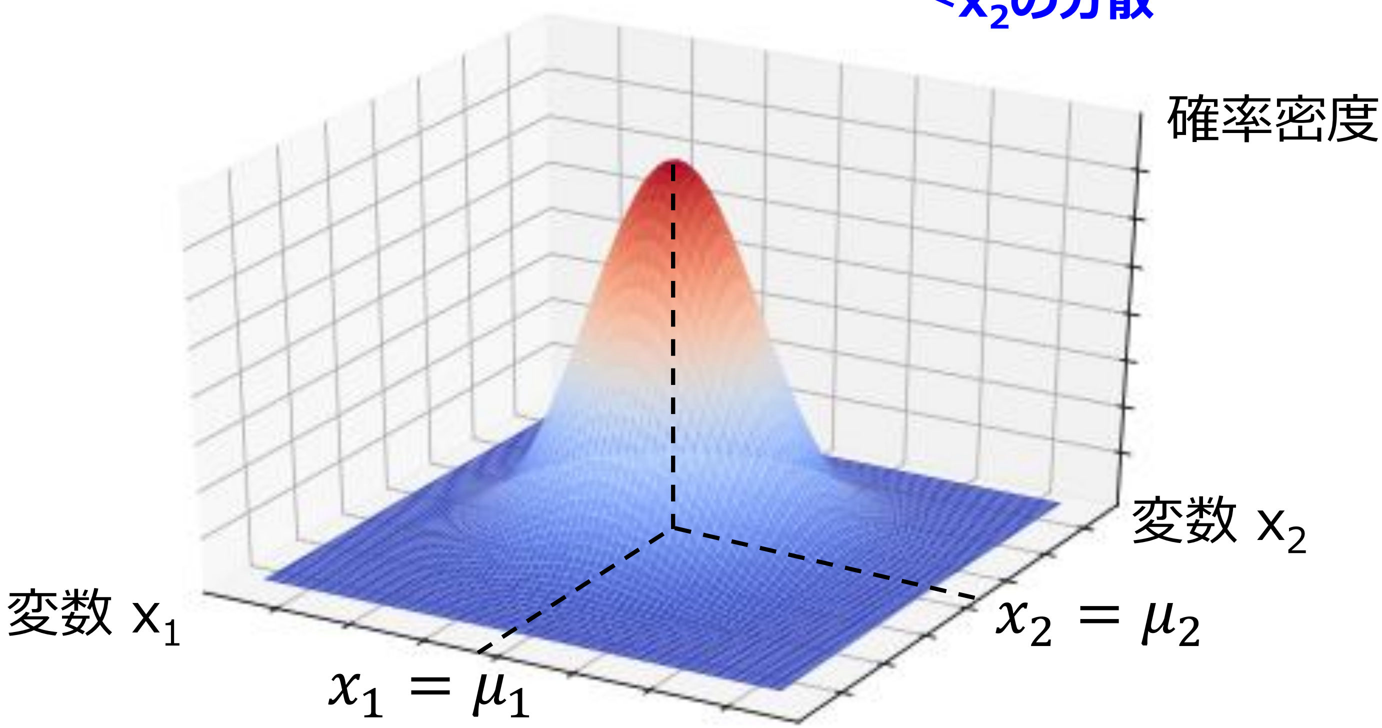
$$N(\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{2\pi\sqrt{|\boldsymbol{\Sigma}|}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$: 確率変数ベクトル,

x_1の分散　　**x_1とx_2の共分散 ($\sigma_{12} = \sigma_{21}$)**

$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$: 平均ベクトル, $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix}$: 分散共分散行列

x_2の分散

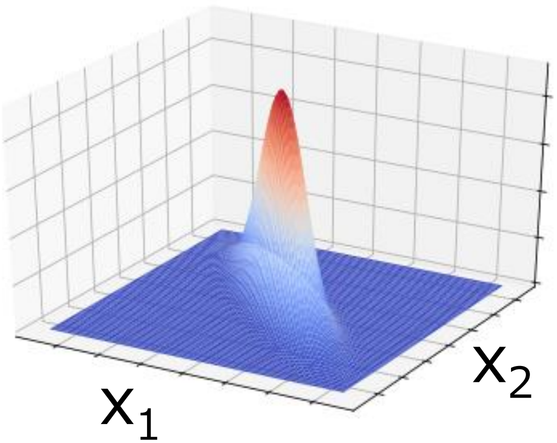


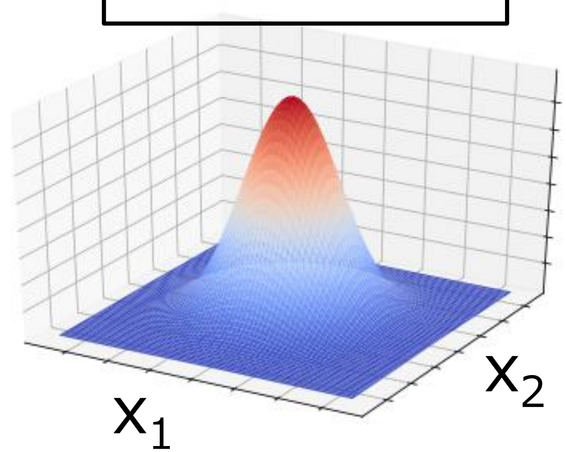
共分散のイメージ

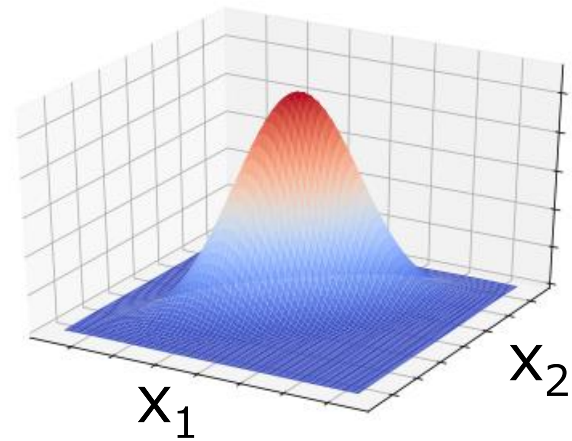
✓ 変数間に相関があるとき，共分散の絶対値は大きくなる

$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & -0.9 \\ -0.9 & 1 \end{bmatrix}$　$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$　$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.7 \\ 0.7 & 1 \end{bmatrix}$

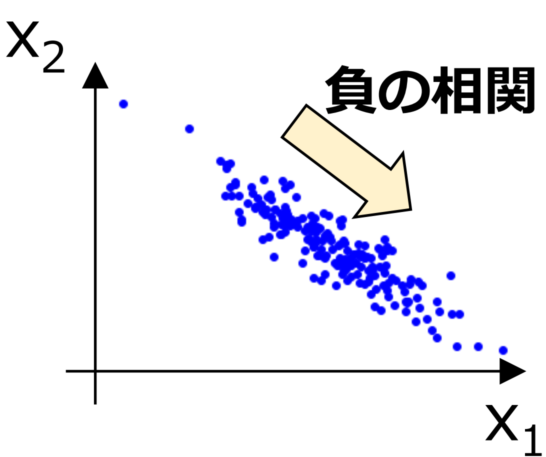
確率密度分布

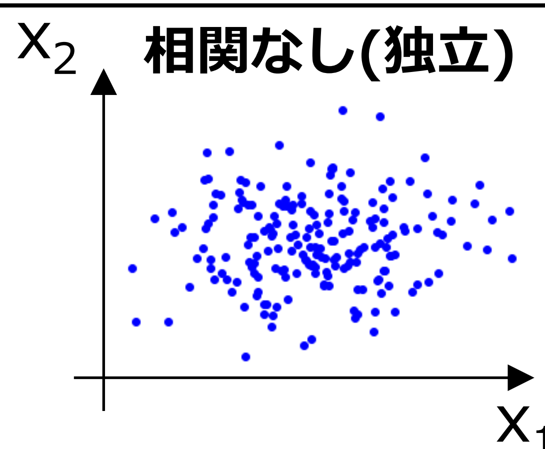


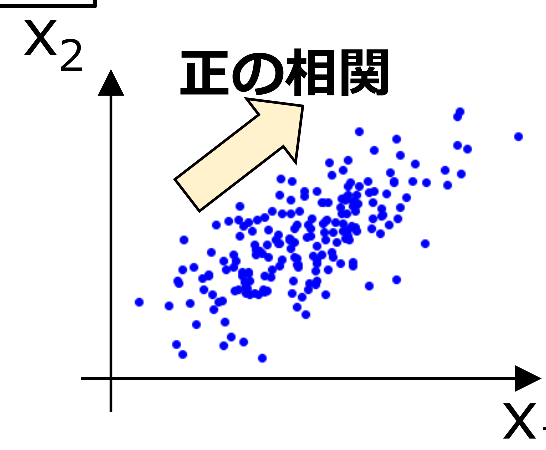




ガウス分布からのサンプル







N次元のガウス分布（多変量正規分布）

- ✓ ガウス分布をN次元に拡張すると次式のようになる

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

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix}$$: 確率変数ベクトル，$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_N \end{bmatrix}$$: 平均ベクトル，

共分散

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1N} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{N1} & \sigma_{N2} & \cdots & \sigma_N^2 \end{bmatrix}$$: 分散共分散行列

分散

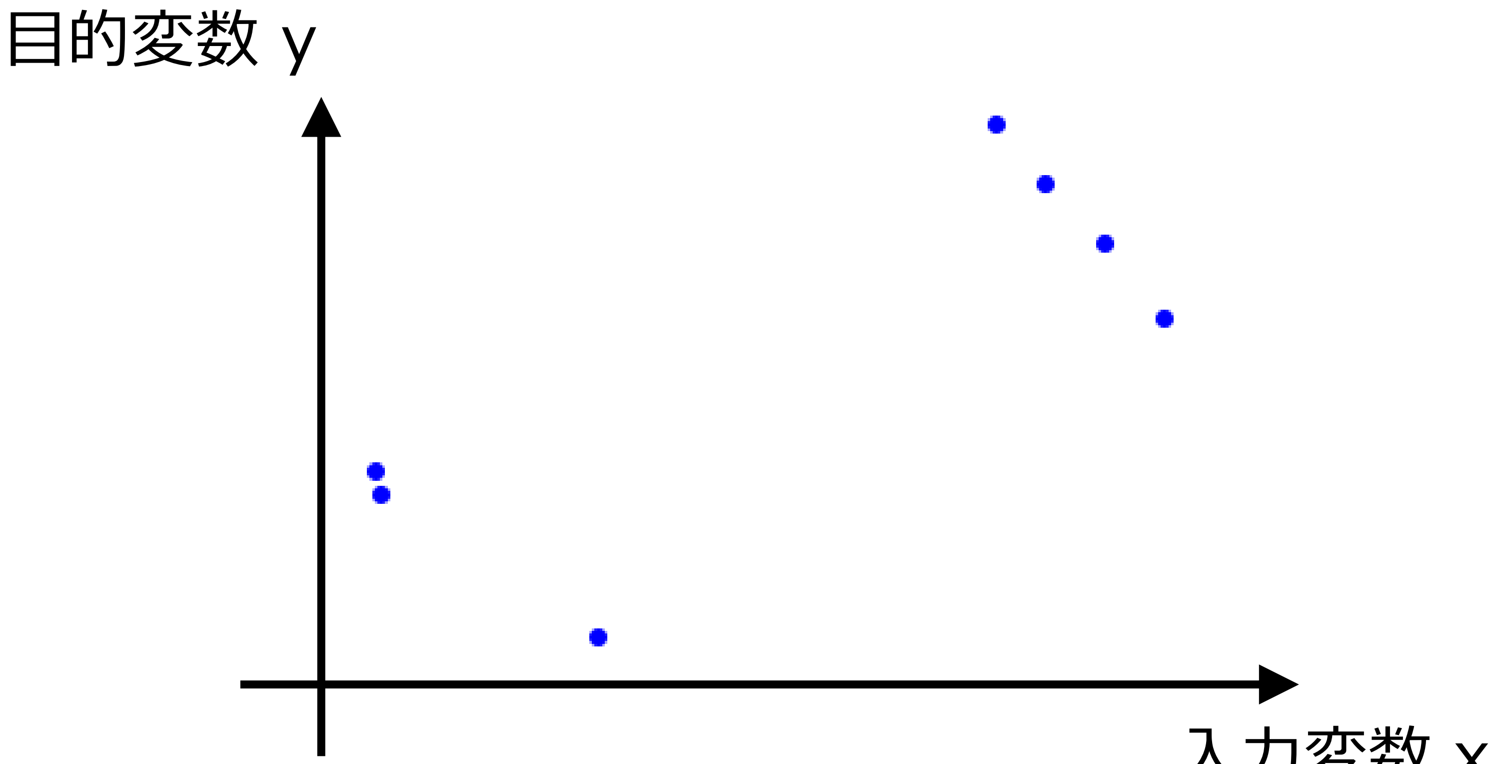
- ✓ 多変量正規分布とよぶ

②ガウス過程回帰をざっくり理解する

- ✓ ガウス過程回帰は回帰分析手法の一つ
- ✓ カーネル関数を適切に選択することで
  非線形な回帰分析が可能
- ✓ **どれだけその予測が信頼できるか**，という情報も算出可能

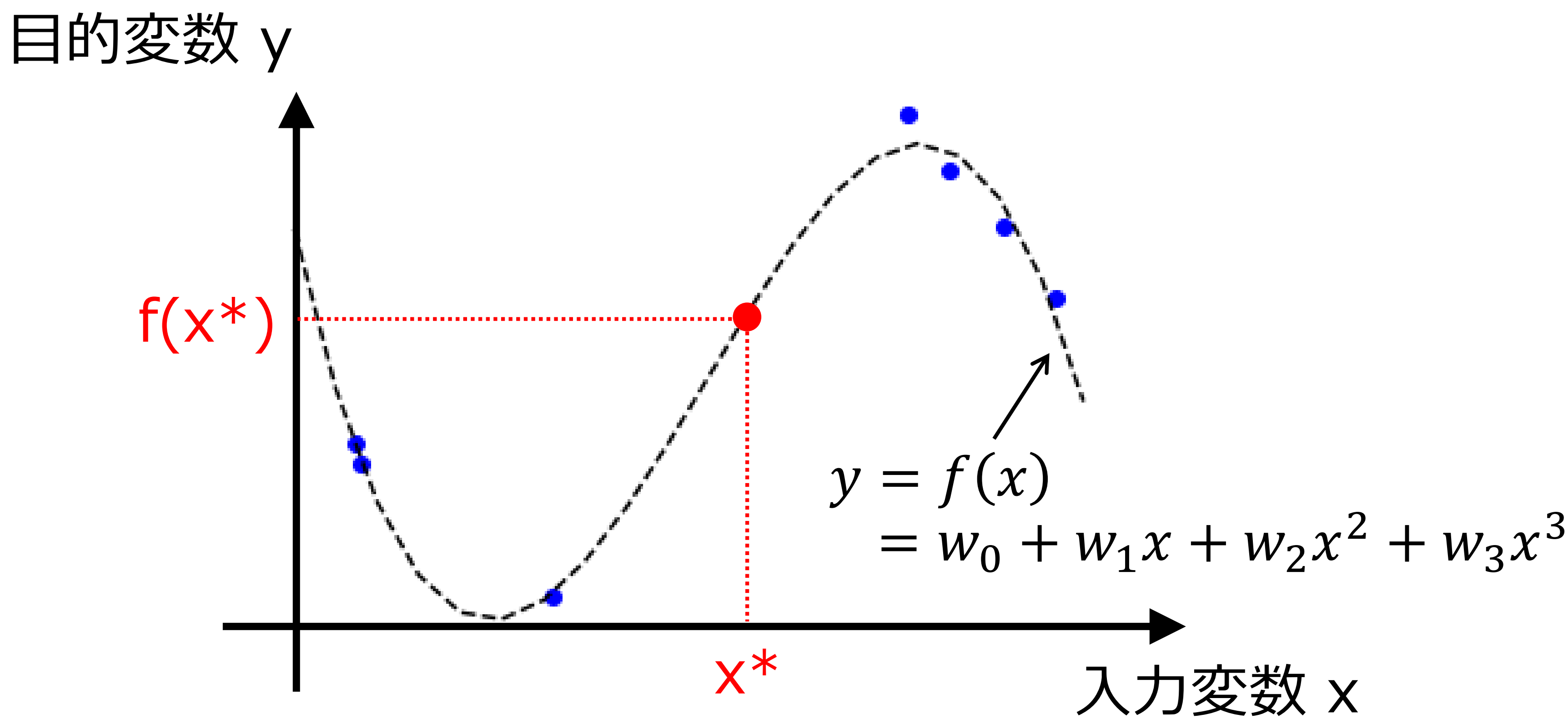
回帰分析の前提

✓ 下図のデータが与えられている状況を考える



最小二乗法の場合

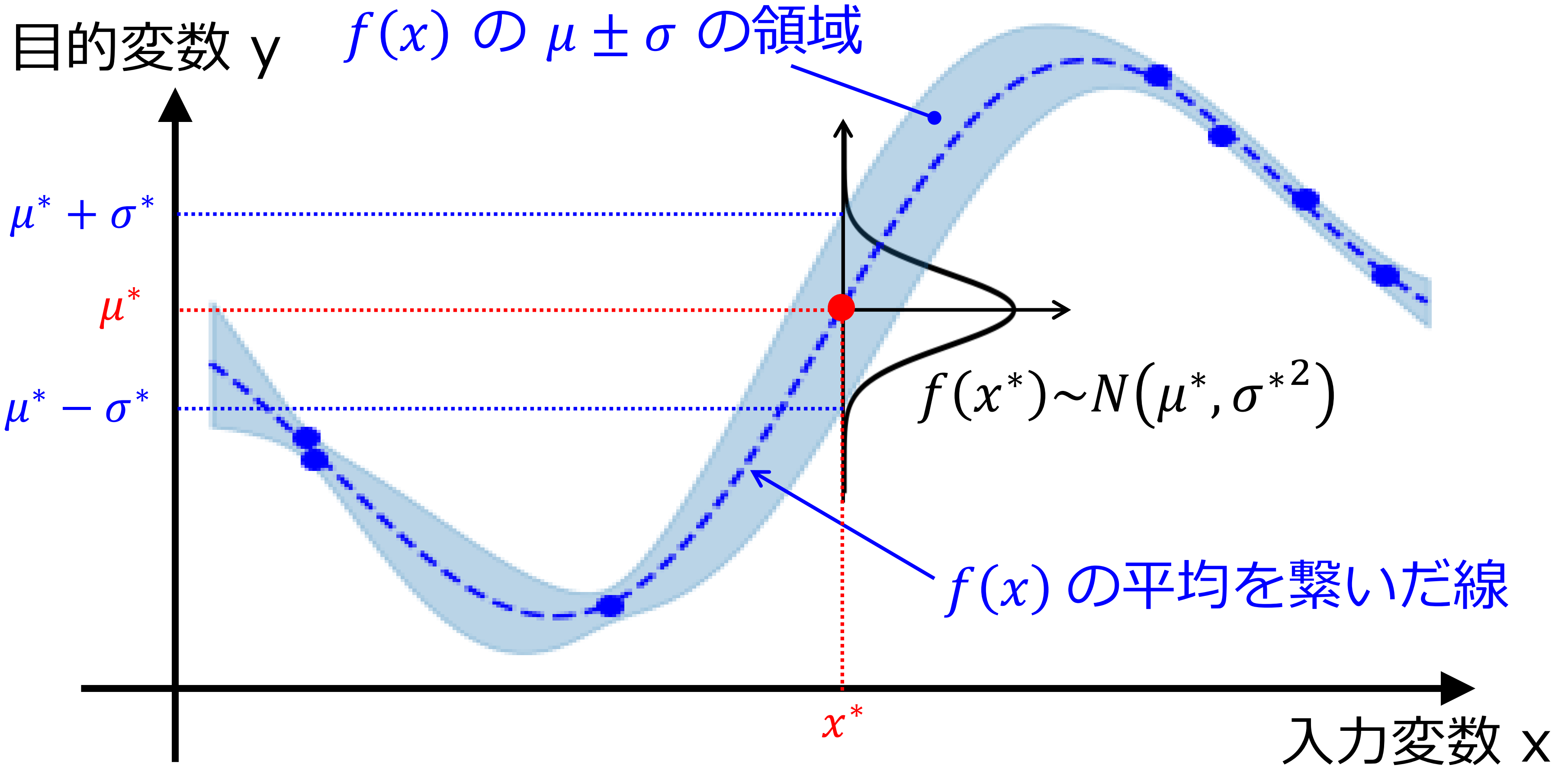
- ✓ 最小二乗法により求めた3次の多項式近似曲線は下図の通り
- ✓ 未知データ x* の予測 f(x*) は定数として与えられる



- ✓ 最小二乗法の詳細はこちら
- ✓ https://yuyumoyuyu.com/2020/12/13/simpleregressionwithols/

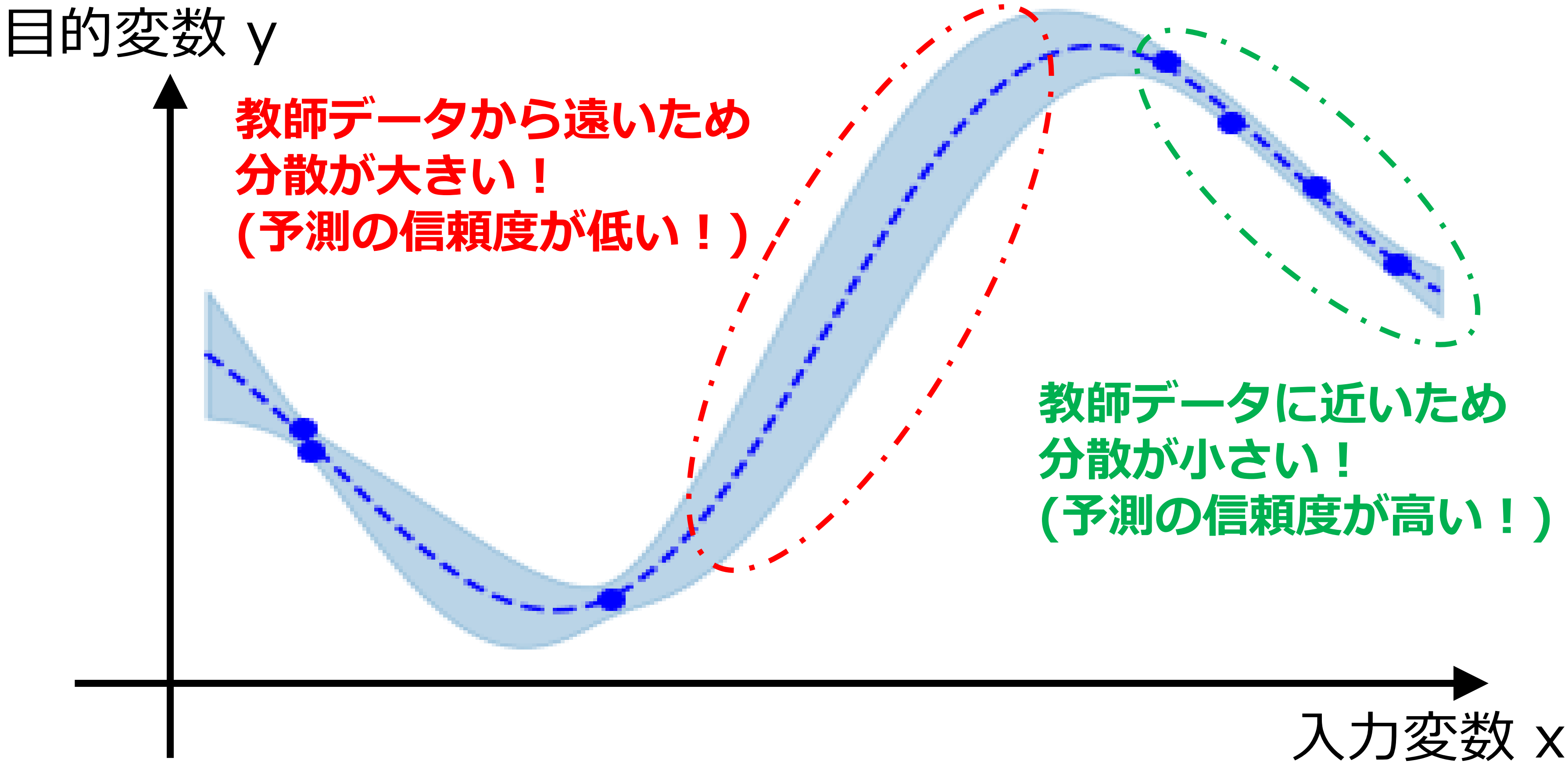
ガウス過程回帰の場合

- ✓ ガウス過程回帰の場合は**予測が確率として与えられる**！
- ✓ 未知データ x* の予測 f(x*) は**ガウス分布に従う**



ガウス過程回帰の特徴

✓ ガウス過程回帰の予測結果は，入力(教師)データから遠ければ遠いほど分散が大きくなる (予測の信頼度が低下)



ガウス過程回帰のざっくりまとめ

✓ ガウス過程回帰の特徴は下記の通り

- モデルの予測値が確定した値ではなく**確率で与えられる**
- その予測値は**ガウス分布**に従い，
  分散の値により予測の信頼度が求められる
- **教師データに近い入力変数**が与えられた場合は，
  予測値の分散が小さくなる（予測の信頼度が高くなる）

✓ なんとなく理解出来たら，
先にプログラムを動かしてみるのも手↓

✓ https://github.com/yshimizu12/GaussianProcessRegression

③ガウス過程とは

- ✓ まず，ガウス過程とは何か説明

ガウス過程の定義

- ✓ ガウス過程の定義は以下の通り
- ✓ 入力 $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N$ が与えられたとき
  対応する**出力ベクトル f が多変量ガウス分布に従うとき**
  f はガウス過程に従う，という

$$\boldsymbol{f} = \begin{pmatrix} f\left(x^{(1)}\right) \\ f\left(x^{(2)}\right) \\ \vdots \\ f\left(x^{(N)}\right) \end{pmatrix} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

- ✓ 次ページから，この導出と意味について説明

線形回帰モデルの引用

- ✓ n組のデータが与えられた場合の線形回帰モデルは以下の通り

$$
\begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(n)} \end{bmatrix} = w_0 + w_1 \begin{bmatrix} x_1^{(1)} \\ x_1^{(2)} \\ \vdots \\ x_1^{(n)} \end{bmatrix} + \cdots + w_m \begin{bmatrix} x_m^{(1)} \\ x_m^{(2)} \\ \vdots \\ x_m^{(n)} \end{bmatrix} = \begin{bmatrix} 1 & x_1^{(1)} & \cdots & x_m^{(1)} \\ 1 & x_1^{(2)} & \cdots & x_m^{(2)} \\ \vdots & \vdots & & \vdots \\ 1 & x_1^{(n)} & \cdots & x_m^{(n)} \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_m \end{bmatrix}
$$

$$\Leftrightarrow \boldsymbol{y} = \boldsymbol{X}\boldsymbol{w}$$

$y^{(i)}$ ：i番目の出力データ(i = 1,⋯,n)

$x_1^{(i)}, \dots, x_m^{(i)}$ ：i番目のm種類の入力データ(i = 1,⋯,n)

$w_0, \dots, w_m$ ：重み係数

- ✓ こちら↓で定義したモデル
- ✓ https://yuyumoyuyu.com/2020/12/20/multipleregression/

非線形写像

✓ 関数の表現力をあげるため,
入力ベクトルを関数化して表現する

$$\begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(n)} \end{bmatrix} = \begin{bmatrix} \phi_0(\boldsymbol{x}^{(1)}) & \phi_1(\boldsymbol{x}^{(1)}) & \cdots & \phi_m(\boldsymbol{x}^{(1)}) \\ \phi_0(\boldsymbol{x}^{(2)}) & \phi_1(\boldsymbol{x}^{(2)}) & \cdots & \phi_m(\boldsymbol{x}^{(2)}) \\ \vdots & \vdots & & \vdots \\ \phi_0(\boldsymbol{x}^{(n)}) & \phi_1(\boldsymbol{x}^{(n)}) & \cdots & \phi_m(\boldsymbol{x}^{(n)}) \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_m \end{bmatrix}$$

$$\Leftrightarrow \boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{w}$$

$y^{(i)}$: i番目の出力データ (i = 1,⋯,n)
$\boldsymbol{x}^{(i)}$: i番目の入力ベクトル (i = 1,⋯,n)
ϕ_j : j番目の関数 (j = 0,⋯,m)
$w_0, \ldots, w_m$: 重み係数

関数の例

$$\phi(x) = x^2, x^3, \sin x, \log x, \ldots$$

様々な非線形表現を用いることでモデルの表現力を向上！

重みをガウス分布に従って生成

✓ 重み **w** が以下のガウス分布から生成されるものとする

平均0　分散は全て λ^2　共分散は全て0⇒独立！

$$\boldsymbol{w} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_m \end{bmatrix} \sim N(\mathbf{0}, \lambda^2 \mathbf{I}) = N\left(\begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} \lambda^2 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \lambda^2 \end{bmatrix} \right)$$

✓ 行列 **Φ** は定数行列であるため,
出力ベクトル **y** = **Φw** も同様にガウス分布に従う

出力ベクトルの平均と共分散行列

✓ 出力ベクトルの期待値は以下のとおり

$$\mathbb{E}[\boldsymbol{y}] = \mathbb{E}[\boldsymbol{\Phi w}] = \boldsymbol{\Phi}\mathbb{E}[\boldsymbol{w}] = \boldsymbol{0}$$

$\mathbb{E}[X]$: Xの期待値

Φは定数　**w**の期待値は**0**

✓ 出力ベクトルの共分散行列は以下のとおり

$$\begin{aligned}\Sigma &= \mathbb{E}[(\boldsymbol{y} - \mathbb{E}[\boldsymbol{y}])(\boldsymbol{y} - \mathbb{E}[\boldsymbol{y}])^T] \\ &= \mathbb{E}[\boldsymbol{y}\boldsymbol{y}^T] \\ &= \mathbb{E}[(\boldsymbol{\Phi w})(\boldsymbol{\Phi w})^T] \\ &= \mathbb{E}[\boldsymbol{\Phi w w}^T\boldsymbol{\Phi}^T] \\ &= \boldsymbol{\Phi}\mathbb{E}[\boldsymbol{w w}^T]\boldsymbol{\Phi}^T \\ &= \boldsymbol{\Phi}\,\lambda^2\mathbf{I}\,\boldsymbol{\Phi}^T = \lambda^2\boldsymbol{\Phi\Phi}^T\end{aligned}$$

E[**y**]=**0**

y=**Φw**を代入

転置行列の積の公式

Φは定数

wの分散をV[**w**]とすると
V[**w**]=E[**ww**^{T}]-E[**w**]E[**w**]^T
=E[**ww**^{T}]

まとめると…

- ✓ これまでの計算をまとめると次のようになり
これは最初のガウス過程の定義に一致

$$\boldsymbol{y} \sim N(\mathbf{0}, \lambda^2 \mathbf{\Phi}\mathbf{\Phi}^T)$$

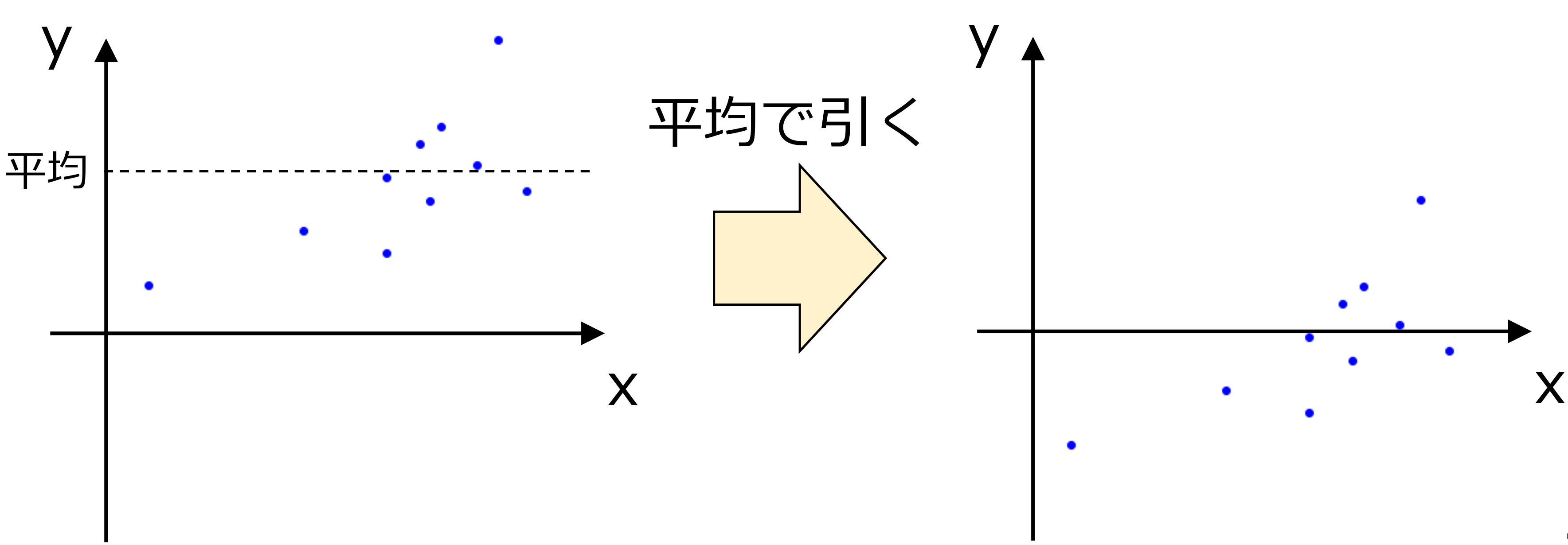
- ✓ この計算には**重みベクトル w が出てこない！**

- ✓ つまり，他の機械学習手法のように重みを学習する
必要がなく，共分散行列を計算するだけでよい！

出力データの平均は0でよいのか？

✓ 平均ベクトルが **0** になっているが
観測データ y はあらかじめ平均を引いておけば
平均が **0** になるため，以下では平均 **0** のガウス過程を扱う

$$\boldsymbol{y} \sim N(\boldsymbol{0}, \lambda^2 \boldsymbol{\Phi}\boldsymbol{\Phi}^T)$$



ガウス過程の共分散行列の解釈

- ✓ 共分散は特徴ベクトル $\boldsymbol{\Phi}$ の内積で計算できるため
  類似した入力ベクトル x に対応する出力 y も似た値をとる

$$\boldsymbol{\Sigma} = \lambda^2 \boldsymbol{\Phi}\boldsymbol{\Phi}^T = \lambda^2 \begin{bmatrix} & \vdots & \\ \phi_0\left(\boldsymbol{x}^{(i)}\right) & \cdots & \phi_m\left(\boldsymbol{x}^{(i)}\right) \\ & \vdots & \end{bmatrix} \begin{bmatrix} \cdots & \begin{matrix} \phi_0\left(\boldsymbol{x}^{(j)}\right) \\ \vdots \\ \phi_m\left(\boldsymbol{x}^{(j)}\right) \end{matrix} & \cdots \end{bmatrix}$$

$$= \lambda^2 \begin{bmatrix} \vdots \\ \boldsymbol{\phi}\left(\boldsymbol{x}^{(i)}\right)^T \\ \vdots \end{bmatrix} \begin{bmatrix} \cdots & \boldsymbol{\phi}\left(\boldsymbol{x}^{(j)}\right) & \cdots \end{bmatrix}$$

$$\Rightarrow \underline{\boldsymbol{\Sigma}_{ij}} = \lambda^2 \boldsymbol{\phi}\left(\boldsymbol{x}^{(i)}\right)^T \boldsymbol{\phi}\left(\boldsymbol{x}^{(j)}\right)$$

共分散行列の(i,j)要素

$\boldsymbol{x}^{(i)}$ と $\boldsymbol{x}^{(j)}$ が似たベクトル ⇒ **$\boldsymbol{\phi}\left(\boldsymbol{x}^{(i)}\right)$と $\boldsymbol{\phi}\left(\boldsymbol{x}^{(j)}\right)$の内積が大きい** ⇒ **$\Sigma_{ij}$が大きい** ⇒ **$y^{(i)}$と$y^{(j)}$が似た値となりやすい (p.8参照)**

カーネル関数の導入

- ✓ 共分散行列の各要素は特徴ベクトル $\boldsymbol{\Phi}$ の内積によって決まるため，特徴ベクトルを明示的に求める必要はない
- ✓ 適当な**カーネル関数**を用いることで，共分散行列を表現
- ✓ **カーネル行列**や**グラム行列**と呼ばれる

カーネル関数：x$^{(i)}$とx$^{(j)}$の類似度を表現する関数

$$k\left(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}\right) = \lambda^2 \boldsymbol{\phi}\left(\boldsymbol{x}^{(i)}\right)^T \boldsymbol{\phi}\left(\boldsymbol{x}^{(j)}\right)$$

共分散行列をカーネル行列（グラム行列）K を用いて表現

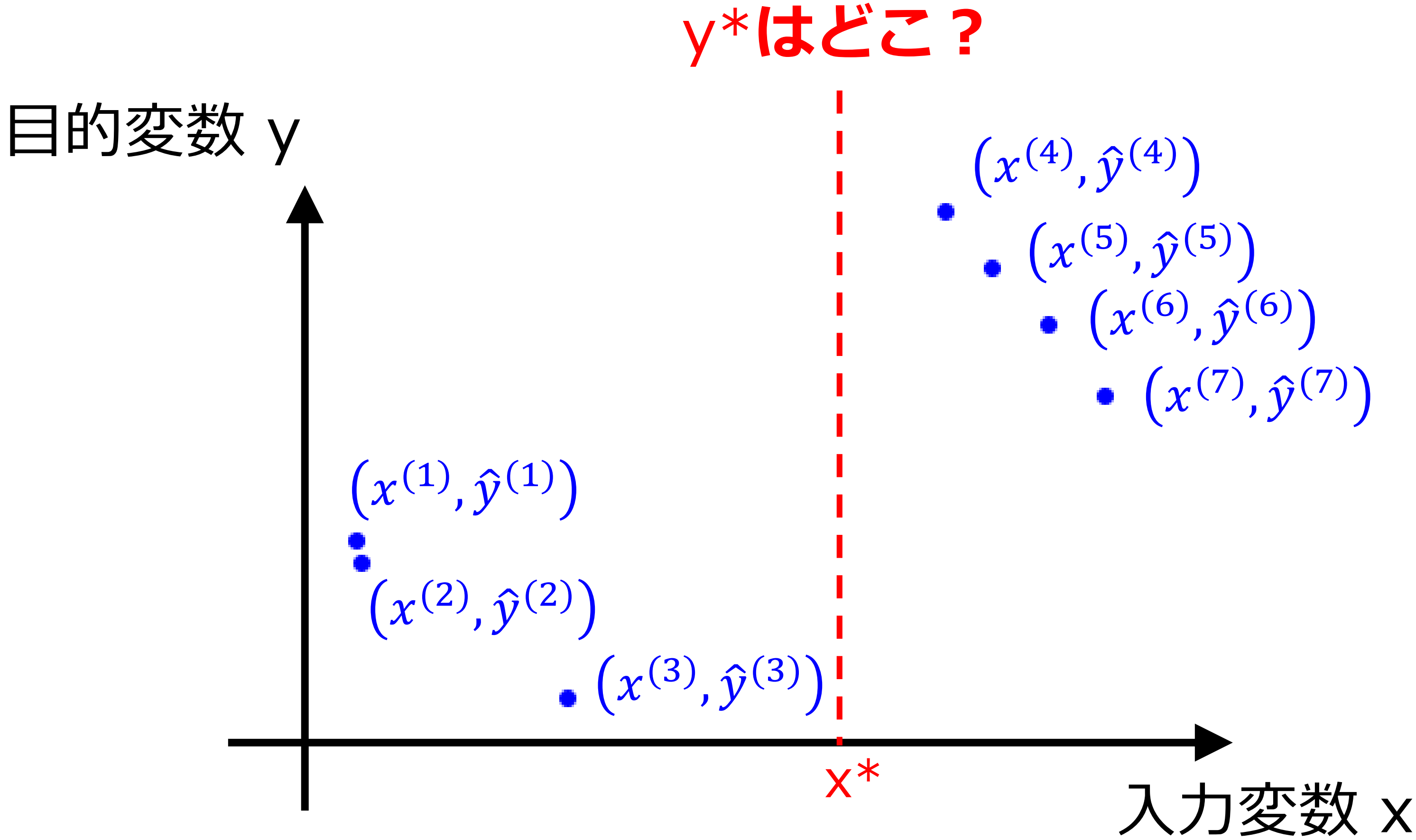
$$\boldsymbol{\Sigma} = \mathbf{K} = \begin{bmatrix} k\left(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(1)}\right) & k\left(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}\right) & \cdots & k\left(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(N)}\right) \\ k\left(\boldsymbol{x}^{(2)}, \boldsymbol{x}^{(1)}\right) & k\left(\boldsymbol{x}^{(2)}, \boldsymbol{x}^{(2)}\right) & \cdots & k\left(\boldsymbol{x}^{(2)}, \boldsymbol{x}^{(N)}\right) \\ \vdots & \vdots & \ddots & \vdots \\ k\left(\boldsymbol{x}^{(N)}, \boldsymbol{x}^{(1)}\right) & k\left(\boldsymbol{x}^{(N)}, \boldsymbol{x}^{(2)}\right) & \cdots & k\left(\boldsymbol{x}^{(N)}, \boldsymbol{x}^{(N)}\right) \end{bmatrix}$$

④ガウス過程を用いた回帰分析

- ✓ つづいて，ガウス過程を用いた回帰分析の方法について説明

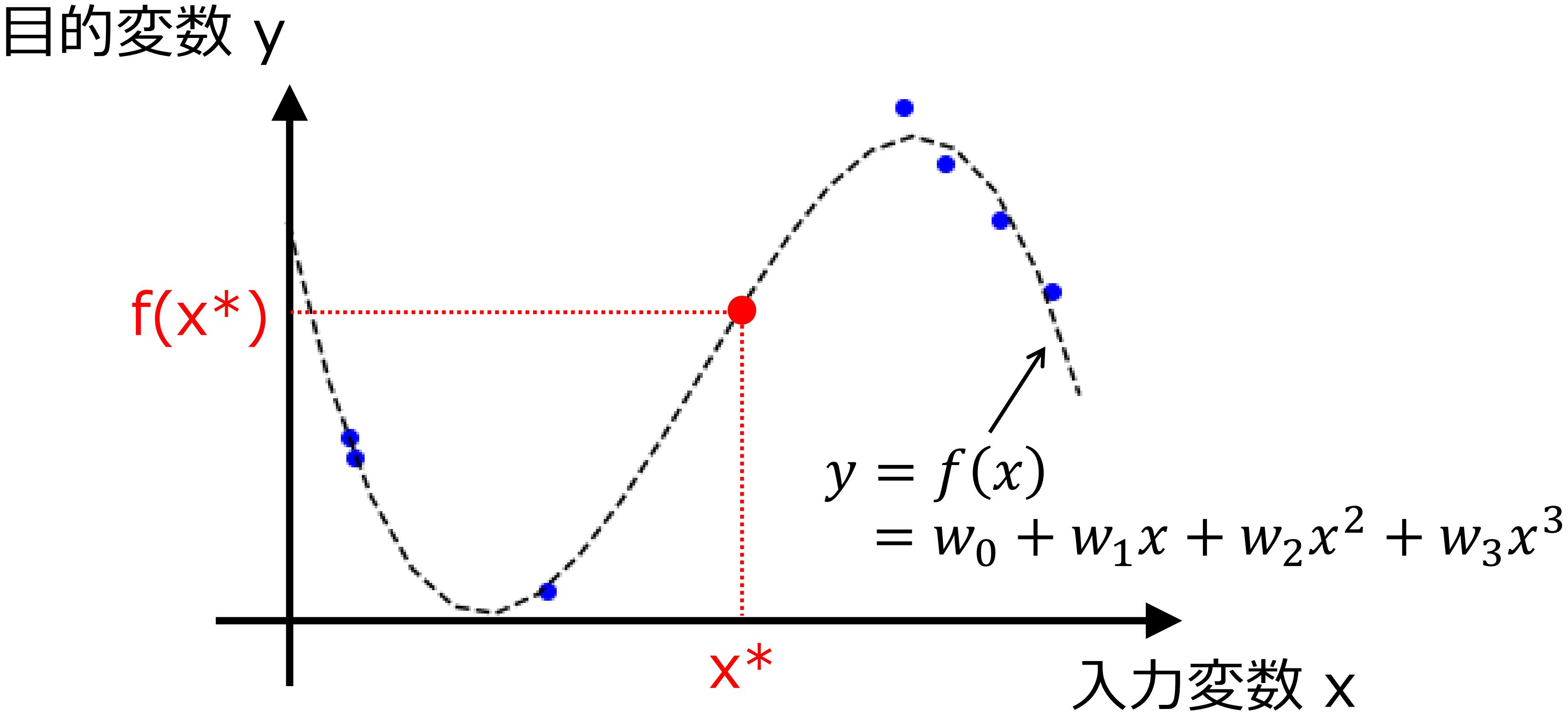
未知のデータをどう予測するか？

✓ 図のように既知のデータセットが与えられた前提で
未知の入力変数 x* に対応する y* を予測する



最小二乗法なら簡単

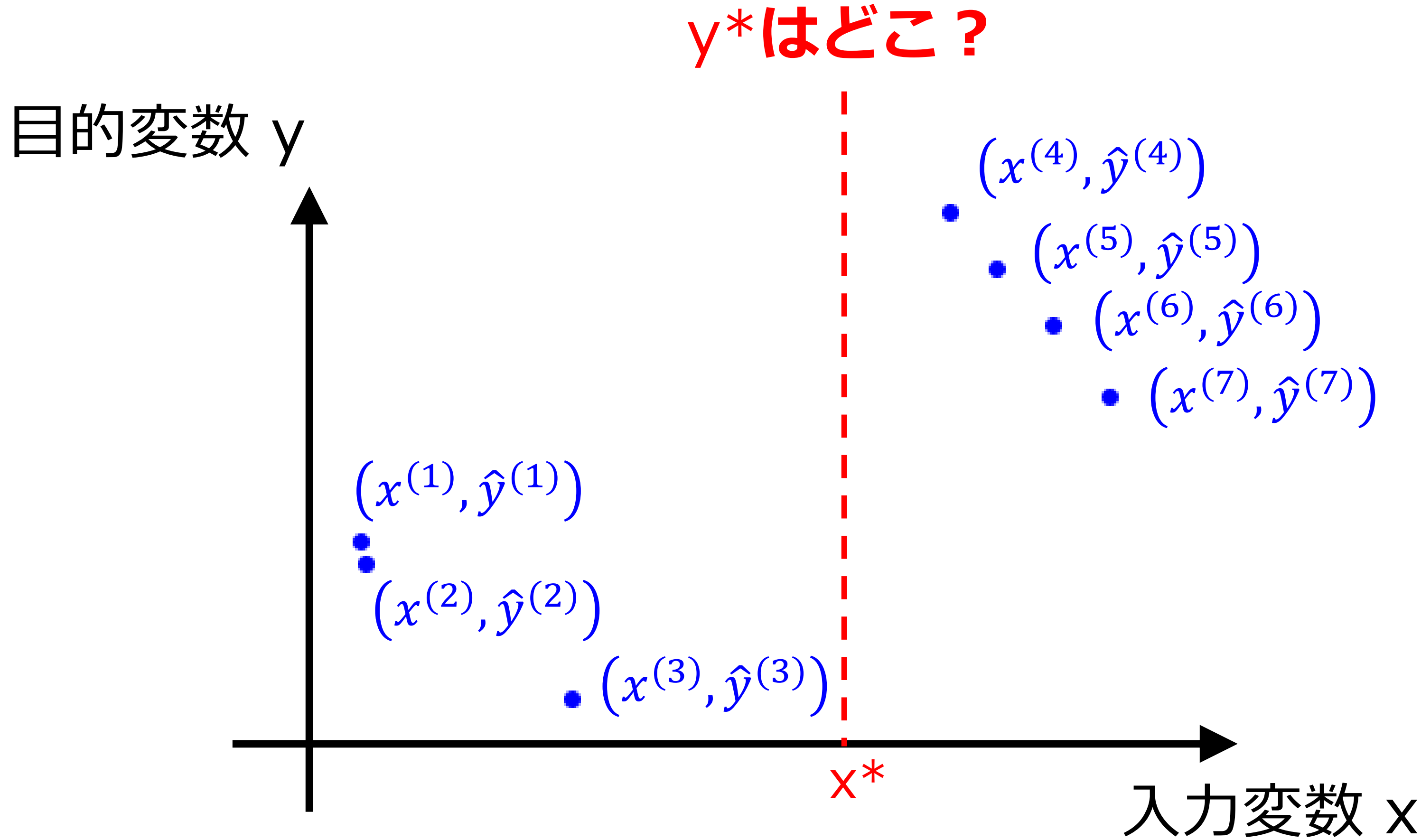
✓ 以下のように，最小二乗法により3次多項式を求めた場合
重み係数 w_0, w_1, w_2, w_3 が全て計算済みなので
未知の入力データ x* を代入するだけでよい



ガウス過程では一筋縄ではいかない

- ✓ ガウス過程の場合，既知データの確率分布だけが与えられ重み係数は明示的に計算されておらず，予測ができない！

$$\boldsymbol{y} = [y^{(1)} \quad y^{(2)} \quad y^{(3)} \quad y^{(4)} \quad y^{(5)} \quad y^{(6)} \quad y^{(7)}]^T \sim N(\boldsymbol{0}, \boldsymbol{K})$$



未知の入力データを含んだ分布を考える

✓ ガウス過程回帰では，未知の入力データを含んだ分布を再定義して考える！

$\boldsymbol{y}$　　$\mathbf{K}$　　$\boldsymbol{k}_*$

$$\boldsymbol{y}' = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(7)} \\ y^* \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} k\left(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(1)}\right) & \cdots & k\left(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(7)}\right) & k\left(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{*}\right) \\ \vdots & \ddots & \vdots & \vdots \\ k\left(\boldsymbol{x}^{(7)}, \boldsymbol{x}^{(1)}\right) & \cdots & k\left(\boldsymbol{x}^{(7)}, \boldsymbol{x}^{(7)}\right) & k\left(\boldsymbol{x}^{(7)}, \boldsymbol{x}^{*}\right) \\ k\left(\boldsymbol{x}^{*}, \boldsymbol{x}^{(1)}\right) & \cdots & k\left(\boldsymbol{x}^{*}, \boldsymbol{x}^{(7)}\right) & k(\boldsymbol{x}^{*}, \boldsymbol{x}^{*}) \end{bmatrix} \right)$$

$\boldsymbol{k}_*^T$

未知データと既知データのカーネル関数

✓ この多変量ガウス分布から，予測分布は次式で与えられる

（証明は参考文献参照）

$$\underline{p(y^*|x^*, \boldsymbol{x}, \boldsymbol{y})} = N(\underbrace{\boldsymbol{k}_*^T \mathbf{K}^{-1} \boldsymbol{y}}_{\text{期待値}}, \underbrace{k(\boldsymbol{x}^*, \boldsymbol{x}^*) - \boldsymbol{k}_*^T \mathbf{K}^{-1} \boldsymbol{k}_*}_{\text{分散}})$$

$x^*, \boldsymbol{x}, \boldsymbol{y}$ が与えられた時の y^* の確率

予測分布の期待値と分散の解釈

- ✓ 予測分布の解釈は次の通り

期待値 $\boldsymbol{k}_*^T \mathbf{K}^{-1} \boldsymbol{y}$

$\boldsymbol{K}^{-1}$によって線形変換された $\boldsymbol{y}$ と $\boldsymbol{k}_*^T$ の内積と解釈可能

$\boldsymbol{k}_*^T$ は未知データと既知データの類似度を表しているため

x^* に類似した $x^{(i)}$ の出力 $y^{(i)}$ に期待値が類似すると解釈できる

分散 $k(\boldsymbol{x}^*, \boldsymbol{x}^*) - \boldsymbol{k}_*^T \mathbf{K}^{-1} \boldsymbol{k}_*$

$\boldsymbol{K}^{-1}$によって線形変換された $\boldsymbol{k}_*$ と $\boldsymbol{k}_*^T$ の内積が $k(\boldsymbol{x}^*, \boldsymbol{x}^*)$ から

差し引かれる

$\boldsymbol{k}_*^T$ は未知データと既知データの類似度を表しているため

x^* に類似した $x^{(i)}$ が多ければ多いほど $\boldsymbol{k}_*^T \mathbf{K}^{-1} \boldsymbol{k}_*$ が大きくなり

分散が小さくなり，予測の信頼度が向上すると解釈できる