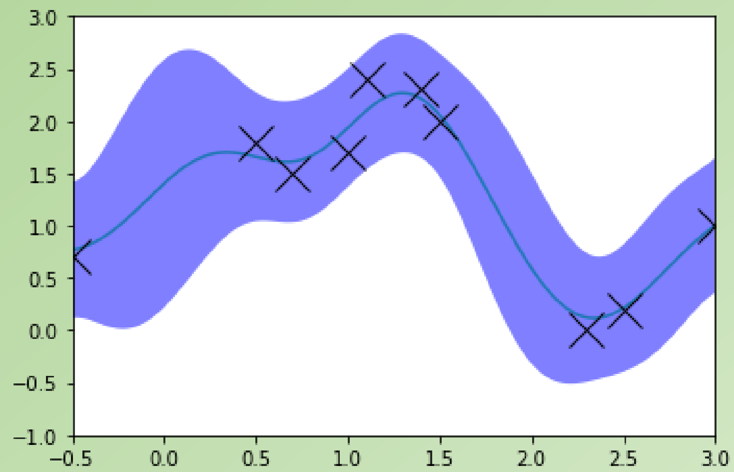
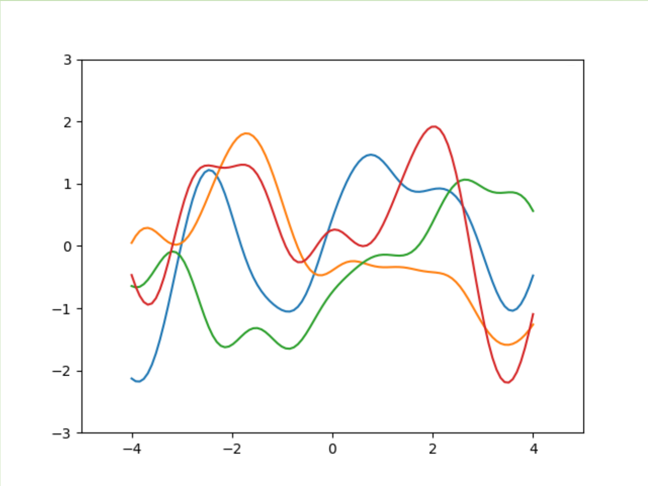
ガウス過程回帰による データ予測





2021年10月30日(土)-31日(日)

香取研究室

佐藤 歩未，西沢 瞭

中央大学理工学部物理学科統計物理学・数理物理学研究室

はじめに

このスライドは, 今年度の前期に

[1] 持橋大地,大羽成征. ガウス過程と機械学習, 機械学習プロフェショナルシリーズ, 講談社, 2019.

で学習したことをもとに, まとめたものである. また, これに加えて, [1]のサポートページ[2]を参考にプログラムを実際に組んでみた.

まずは, [1]で学んだことの中でも「ガウス過程回帰」を中心に述べる. そのあとで, 実際に組んでみたプログラムを用いて行ったデータ予測を示す.

多変量ガウス分布

・D次元ベクトル$(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \cdots, \boldsymbol{x}_D)$ が平均$\boldsymbol{\mu}$,共分散行列$\boldsymbol{\Sigma}$の多変量ガウス分布に従うとき、その確率密度関数は

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

で表される．　$\boldsymbol{\Sigma}$はベクトル同士の共分散

$$\sigma_{ij} = E[(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})]$$

を並べた行列になっている：

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1D} \\ \vdots & \ddots & \vdots \\ \sigma_{D1} & \cdots & \sigma_{DD} \end{pmatrix}$$

ガウス過程とは

・入力 $(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \cdots, \boldsymbol{x}_N)$ に対して、出力

$$\boldsymbol{y} = (f(\boldsymbol{x}_1), f(\boldsymbol{x}_2), f(\boldsymbol{x}_3), \cdots, f(\boldsymbol{x}_N))$$

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.

$$\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

・どんなNに対しても上は成り立つ.（∞でもOK）

→ガウス分布の周辺分布もガウス分布であるから、
無限次元ガウス分布を周辺化したものとも考えられる.

ガウス過程回帰では、この性質を利用して未知のデータの予測を行う.（ガウス過程回帰を用いた予測分布に続く.）

ガウス過程回帰

N個の観測値, すなわち入力$\boldsymbol{x} \in \chi$と出力$y \in \mathbb{R}$のN個の組

$$\mathcal{D} = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_N, y_N)\}$$

このとき, $y = f(\boldsymbol{x})$

の関係があり, $f \sim \mathrm{GP}\left(\boldsymbol{0}, k(\boldsymbol{x}, \boldsymbol{x}')\right)$ から生成されているとする.

・・・「関数fは平均$\boldsymbol{0}$, 共分散$k(\boldsymbol{x}, \boldsymbol{x}')$の
ガウス過程(Gaussian process)に従う」

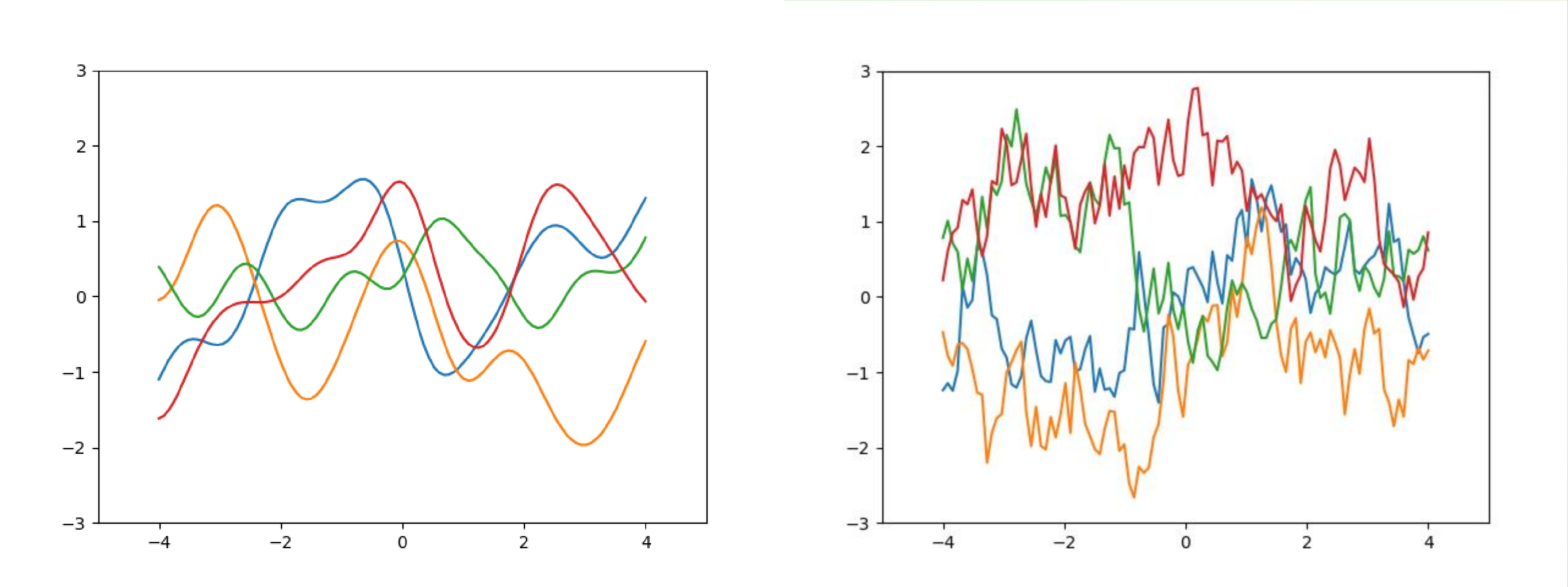
$\boldsymbol{y} = (y_1, y_2, \cdots, y_N)^T$ とおくと, この$\boldsymbol{y}$はガウス分布に従う.
入力のすべてのペア $(\boldsymbol{x}_n, \boldsymbol{x}_{n'})$の類似度を表すカーネル関数$k(\boldsymbol{x}, \boldsymbol{x}')$
用いて $K_{nn'} = k(\boldsymbol{x}_n, \boldsymbol{x}_{n'})$ で与えられるカーネル行列$\boldsymbol{K}$を使って,

$$\boxed{\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{K})}$$ と書ける.

・・・$\boldsymbol{y}$は平均$\boldsymbol{0}$, 共分散行列$\boldsymbol{K}$のガウス分布に従う 5

カーネル関数

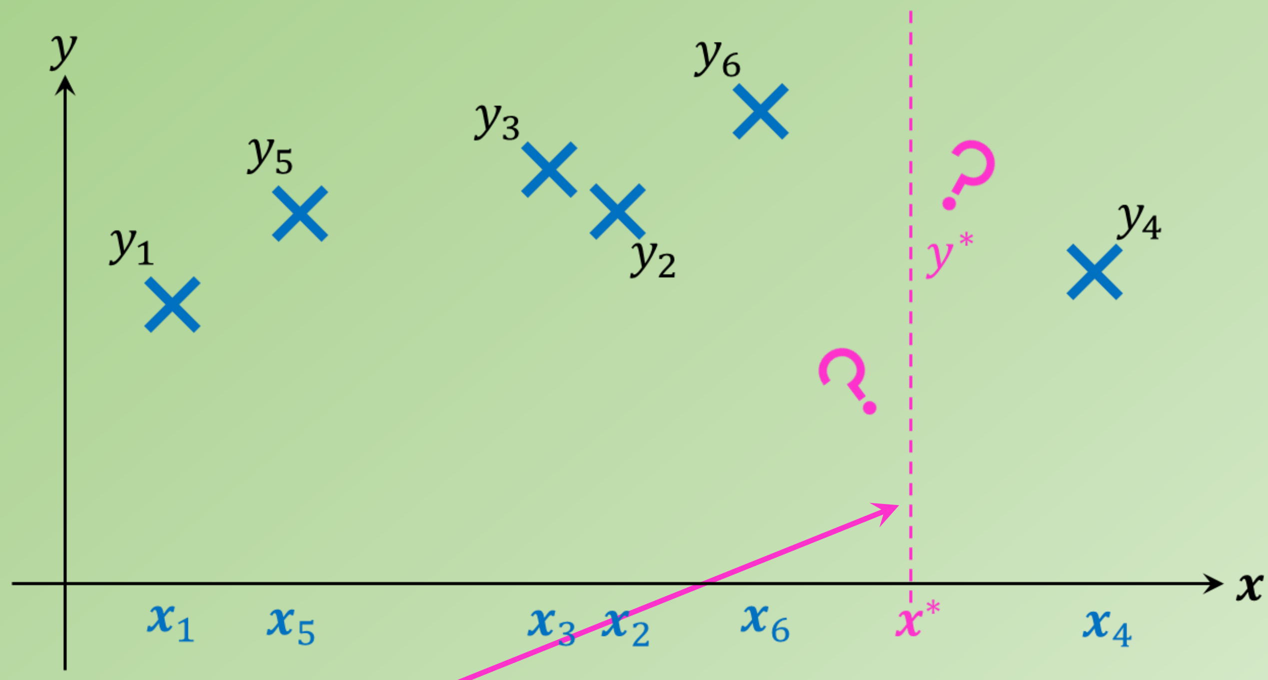
xとx'の類似度（距離）を表す**カーネル関数$k(x, x')$**には、以下のようなものがある.



Gaussカーネル: $\theta_1 \exp\left(-\frac{|x-x'|^2}{\theta_2}\right)$　　指数カーネル: $\exp\left(-\frac{|x-x'|}{\theta_2}\right)$

図1: カーネル関数ごとのガウス過程からのサンプル

ガウス過程回帰を用いた予測分布



このようにデータに含まれていないものが知りたい！

ではどうするのか？

⟶ $\boldsymbol{y}$にy^*を新しく含めて$\boldsymbol{y}$'をつくろう！

$$\boldsymbol{y}' = (y_1, y_2, \cdots, y_N, y^*)^T$$

$\boldsymbol{K}$'： $\boldsymbol{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N)$ および$\boldsymbol{x}^*$から計算される

$(N+1)\times(N+1)$次元カーネル行列

これら全体がガウス分布に従う ⟶ $\boldsymbol{y}' \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{K}')$

これら全体がガウス分布に従う

$\longrightarrow \quad \boldsymbol{y}' \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{K}')$

$\Downarrow$

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

$\boldsymbol{k}_* = \left(k(\boldsymbol{x}^*, \boldsymbol{x}_1), \cdots\cdots, k(\boldsymbol{x}^*, \boldsymbol{x}_N)\right)^T$

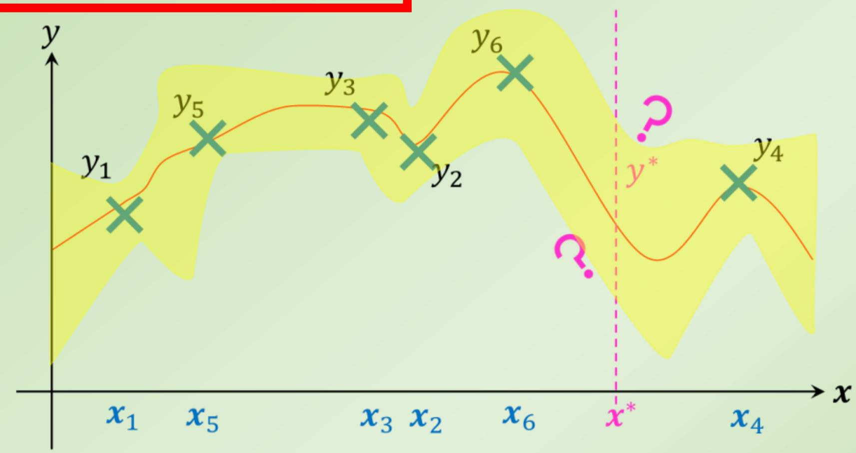
・・・新しい入力$\boldsymbol{x}^*$と学習データの入力の類似度を並べたベクトル

$k_{**} = k(\boldsymbol{x}^*, \boldsymbol{x}^*)$ ・・・$\boldsymbol{x}^*$自身との類似度

これはy^*と$\boldsymbol{y}$の同時分布の式

ガウス過程の予測分布

$$\underbrace{p(y^* | \boldsymbol{x}^*, \mathcal{D})}_{\boldsymbol{x}^*\text{とDが与えられたときの}y^*\text{の分布}} = \mathcal{N}(\underbrace{\boldsymbol{k}_*^T \boldsymbol{K}^{-1} \boldsymbol{y}}_{\text{平均}}, k_{**} - \boldsymbol{k}_*^T \boldsymbol{K}^{-1} \boldsymbol{k}_*)$$



：fの事後分布の平均

：ガウス事後分布の±2σの誤差範囲

ガウス過程回帰の計算

ガウス過程回帰は, 次の 1) と 2) で定義される:

1) 入力$\boldsymbol{x}$と出力yのペアからなる学習データ$\{(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_N, y_N)\}$

2) 入力$\boldsymbol{x}$と入力x'の間の類似度, すなわち,
ガウス分布の共分散を与えるカーネル関数$k(\boldsymbol{x}, \boldsymbol{x}')$

学習データ$\boldsymbol{X} = (\boldsymbol{x}_1, \cdots\cdots, \boldsymbol{x}_N)^T$および, $\boldsymbol{y} = (y_1, \cdots\cdots, y_N)^T$が与えられたとき, 新しい入力$\boldsymbol{x}$*に対する出力$y$*を

$$p(y^*|\boldsymbol{x}^*, \boldsymbol{X}, \boldsymbol{y}) = \mathcal{N}(\boldsymbol{k}_*^T \boldsymbol{K}^{-1} \boldsymbol{y}, \boldsymbol{k}_{**} - \boldsymbol{k}_*^T \boldsymbol{K}^{-1} \boldsymbol{k}_*)$$

として求める.

↓

この分布は, カーネルが与えられれば一意に定まる

⟶「学習」は, ガウス過程回帰においては
カーネルのハイパーパラメータ以外には依存しない
:推論や予測の枠組みの中で決定されないパラメータ

あらかじめ, 値の候補を用意しておいて, それらをいったん採用し,
予測や推論の中で, 最もよいモデルのハイパーパラメータを採用する

ハイパーパラメータ推定する方法を考えてみる.

$\longrightarrow$ $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$ と置いてみる.

カーネルはこれに依存するので,

$$k(\boldsymbol{x}, \boldsymbol{x}'|\boldsymbol{\theta}) = \theta_1 \exp\left(-\frac{|\boldsymbol{x}-\boldsymbol{x}'|^2}{\theta_2}\right) + \theta_3 \delta(\boldsymbol{x}, \boldsymbol{x}')$$

$\downarrow$

もちろん, カーネル行列$\boldsymbol{K}$は, このカーネルkから計算されるので,
$\boldsymbol{\theta}$に依存する $\rightarrow$ $\boldsymbol{K_\theta}$と書こう.

学習データの確率

$$p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{0}, \boldsymbol{K_\theta}) = \frac{1}{(2\pi)^{N/2}} \frac{1}{|\boldsymbol{K_\theta}|^{1/2}} \exp\left(-\frac{1}{2}\boldsymbol{y}^T \boldsymbol{K_\theta}^{-1} \boldsymbol{y}\right)$$

の両辺, 対数をとって

$$\log p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta}) = \underline{-\frac{N}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{K_\theta}| - \frac{1}{2}\boldsymbol{y}^T \boldsymbol{K_\theta}^{-1} \boldsymbol{y}} = L \text{とおく}$$

$\longrightarrow$ Lを最大にする$\boldsymbol{\theta}$を求める

勾配法(gradient method)を用いて考えてみる.

Lを$\boldsymbol{\theta}$のある要素θで微分は,

$$\frac{\partial L}{\partial \theta} = -\frac{\partial}{\partial \theta}\log|\boldsymbol{K}_{\boldsymbol{\theta}}| - \boldsymbol{y}^T\left(\frac{\partial}{\partial \theta}\boldsymbol{K}_{\boldsymbol{\theta}}^{-1}\right)\boldsymbol{y}$$

$$= -\mathrm{tr}\left(\boldsymbol{K}_{\boldsymbol{\theta}}^{-1}\frac{\partial \boldsymbol{K}_{\boldsymbol{\theta}}}{\partial \theta}\right) + \left(\boldsymbol{K}_{\boldsymbol{\theta}}^{-1}\boldsymbol{y}\right)^T\frac{\partial \boldsymbol{K}_{\boldsymbol{\theta}}}{\partial \theta}\left(\boldsymbol{K}_{\boldsymbol{\theta}}^{-1}\boldsymbol{y}\right)$$

$\frac{\partial \boldsymbol{K}_{\boldsymbol{\theta}}}{\partial \theta}$は, カーネル行列$\boldsymbol{K}_{\boldsymbol{\theta}}$の各要素に注目しているパラメータ$\theta \in \boldsymbol{\theta}$で微分した行列である.

$$k(\boldsymbol{x}, \boldsymbol{x}'|\boldsymbol{\theta}) = \theta_1 \exp\left(-\frac{|\boldsymbol{x}-\boldsymbol{x}'|^2}{\theta_2}\right) + \theta_3\delta(\boldsymbol{x}, \boldsymbol{x}')$$

のガウスカーネルおよびガウス観測誤差の場合は, $\theta_1=1$, $\theta_2=0.4$, $\theta_3=0.1$より, $\theta_1>0$, $\theta_2>0$, $\theta_3>0$なので, **最適化**(optimization)のために$\theta_1=\mathrm{e}^{\tau}$, $\theta_2=\mathrm{e}^{\sigma}$, $\theta_3=\mathrm{e}^{\eta}$ すなわち, $\tau=\log\theta_1$, $\sigma=\log\theta_2$, $\eta=\log\theta_3$とおき, カーネルを書き直すと,

$$k(\boldsymbol{x}_n, \boldsymbol{x}_{n'}|\boldsymbol{\theta}) = \mathrm{e}^{\tau}\exp\left(-\frac{|\boldsymbol{x}_n-\boldsymbol{x}_{n'}|^2}{\mathrm{e}^{\sigma}}\right) + \mathrm{e}^{\eta}\delta(n, n')$$

これを, τ, σ, ηでそれぞれ微分すると,

$$\frac{\partial}{\partial \tau} k(\boldsymbol{x}_n, \boldsymbol{x}_{n'}) = k(\boldsymbol{x}_n, \boldsymbol{x}_{n'}) - \mathrm{e}^{\eta} \delta(n, n')$$

$$\frac{\partial}{\partial \sigma} k(\boldsymbol{x}_n, \boldsymbol{x}_{n'}) = \left(k(\boldsymbol{x}_n, \boldsymbol{x}_{n'}) - \mathrm{e}^{\eta} \delta(n, n')\right) \cdot \mathrm{e}^{-\sigma} |\boldsymbol{x}_n - \boldsymbol{x}_{n'}|^2$$

$$\frac{\partial}{\partial \eta} k(\boldsymbol{x}_n, \boldsymbol{x}_{n'}) = \mathrm{e}^{\eta} \delta(n, n')$$

となる. これを $\frac{\partial L}{\partial \theta} = -\mathrm{tr}\left(\boldsymbol{K}_{\boldsymbol{\theta}}^{-1} \frac{\partial \boldsymbol{K}_{\boldsymbol{\theta}}}{\partial \theta}\right) + \left(\boldsymbol{K}_{\boldsymbol{\theta}}^{-1} \boldsymbol{y}\right)^{\boldsymbol{T}} \frac{\partial \boldsymbol{K}_{\boldsymbol{\theta}}}{\partial \theta} \left(\boldsymbol{K}_{\boldsymbol{\theta}}^{-1} \boldsymbol{y}\right)$ の $\frac{\partial \boldsymbol{K}_{\boldsymbol{\theta}}}{\partial \theta}$ の要素として代入すれば, ハイパーパラメータに対する勾配 $\frac{\partial L}{\partial \theta}$ が求まる.

→これらをゼロとおいた連立方程式を解けば, Lを最大にするτ, σ, ηが求まる.

$$\frac{\partial L}{\partial \tau} = 0, \frac{\partial L}{\partial \sigma} = 0, \frac{\partial L}{\partial \eta} = 0$$

Pythonのガウス過程パッケージ(GPy)が有名だが,
今回は用いずに,[1]で学んだこと,それからサポートページ[2]のコードを参考にし,実際に自分たちでプログラムを組んでみた.
ハイパーパラメータの最適化には、Pythonのライブラリであるscipy.optimize.minimizeでBFGS法を利用した.
次頁以降にこれを用いていくつかデータをプロットしてみたものを示す.

①ここまで述べたことを用いてプロットしてみる.

1) まずは, [1]のサポートページ[2]にあるデータを用いる.

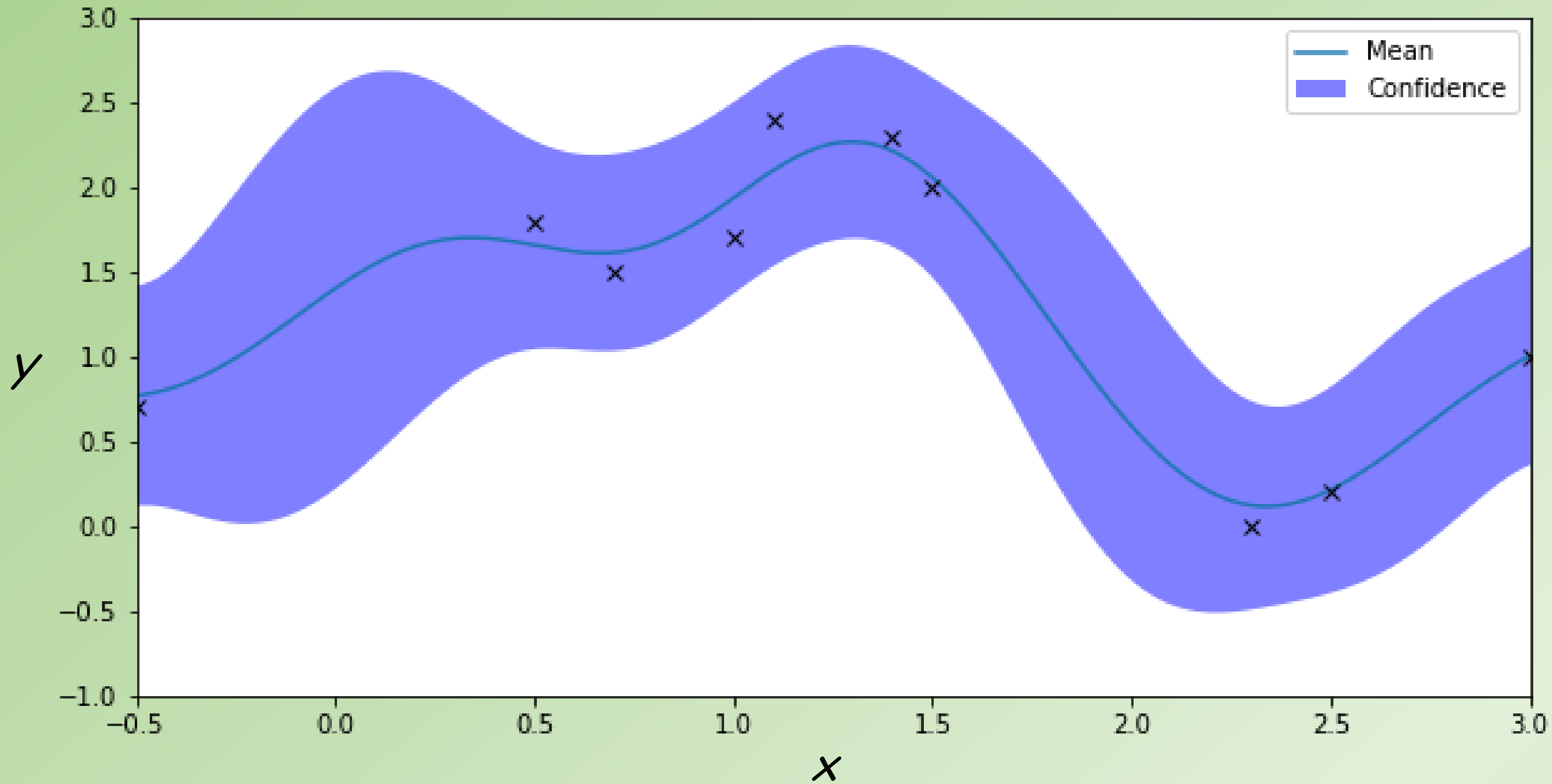


図2: ガウス過程回帰による予測分布

なお, データは[1]のサポートページ[2]にある「gpr.dat」を用いたが, プログラムのコードは[2]に掲載されているものを参考にプログラム組んでプロットしてみた.

2) 次に, 実データ（東京の2020年の気温）を用いてプロットする.

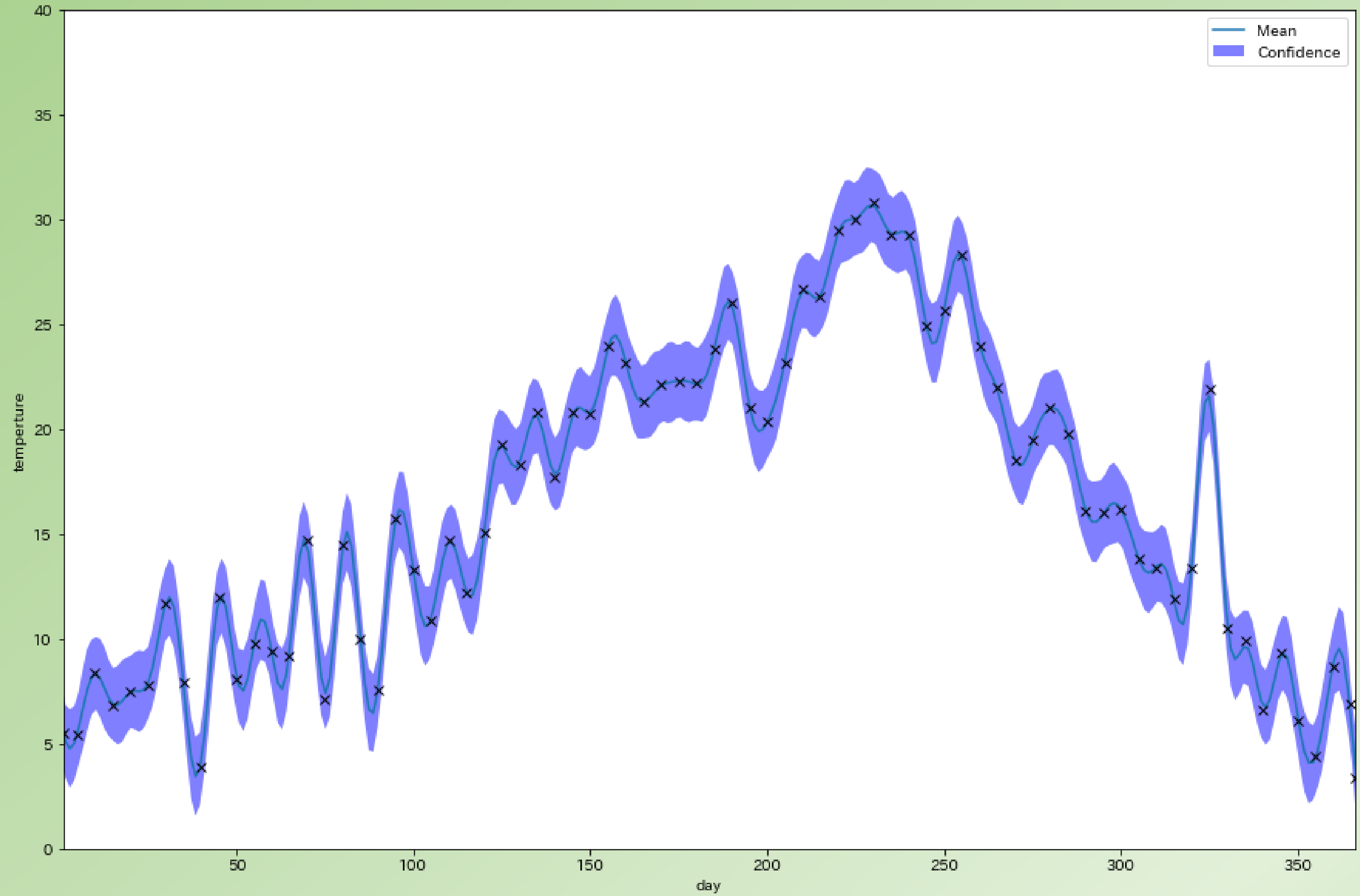


図3: 東京の2020年の5日ごとの気温の分布（横軸は1月1日から日数(日), 縦軸は気温(度)）
なお, データは気象庁のHP[3]で公開されているものを用いた.

② ここまでの内容の応用として, 周期カーネルを用いて予測してみた.

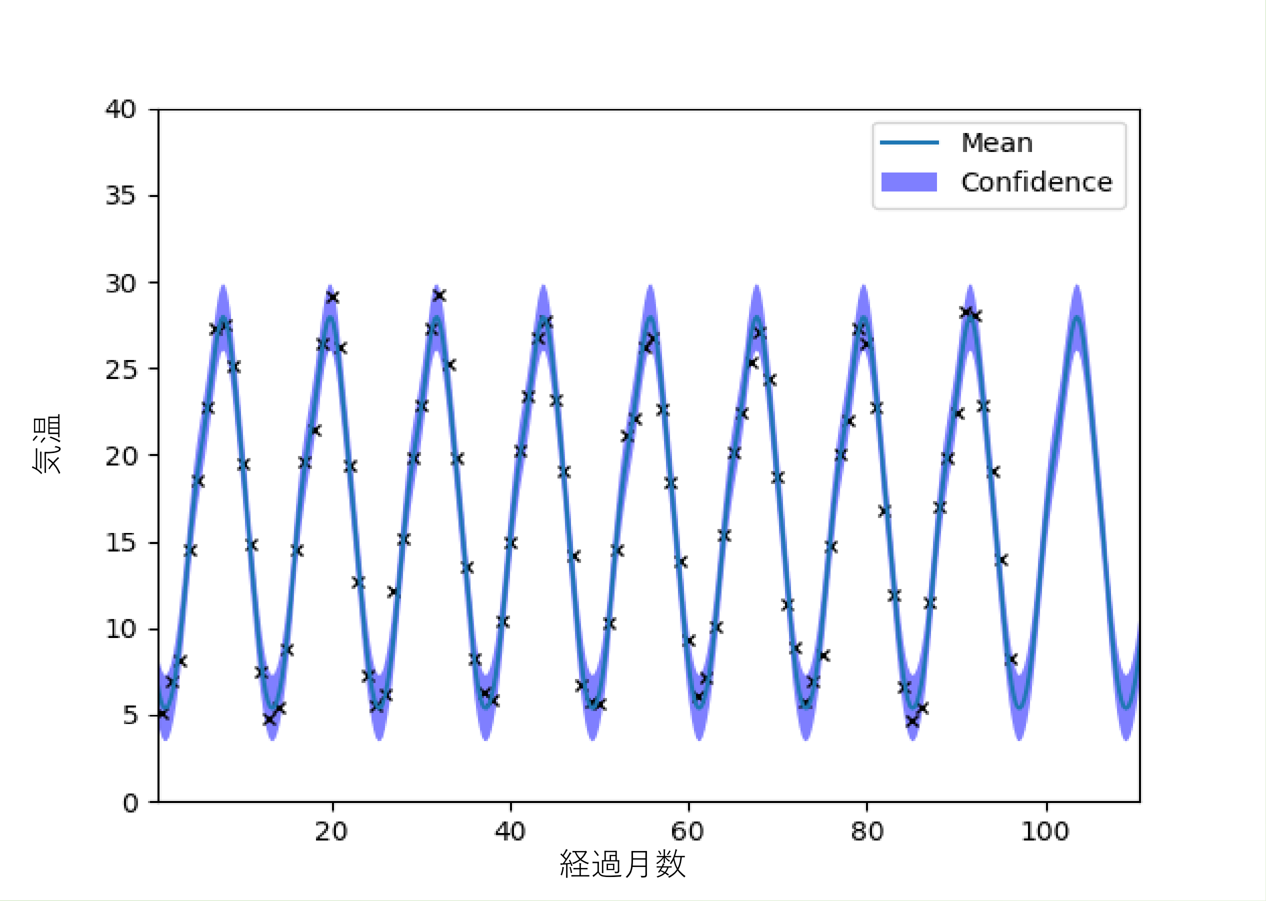


図4: 2011年以降の月平均気温の変動

2011年～2019年（96か月）のデータを与えてそれ以降は予測させた.
周期カーネルで四季による気温変動を捉えることができた。

なお, データは気象庁のHP[3]で公開されているものを用いた.

今後の目標

・より複雑なデータについての回帰を行っていきたい.

→カーネルを組み合わせて回帰する

例) 周期カーネル+ガウスカーネル
周期カーネル+線形カーネル

・効率的なパラメータの探索方法を検討したい.

今回の周期カーネルでの回帰では、初期値によって局所解に
落ちることが多かった.
→MCMCなど他の探索方法を使ったらどうなるか.

・回帰プログラムの改良

最適化においてオーバーフローが起こり、正しく計算できないことがあった.
計算機を扱う上での難しさを実感した.

参考文献

[1] 持橋大地,大羽成征. ガウス過程と機械学習, 機械学習プロフェショナルシリーズ, 講談社, 2019.

[2] 持橋大地,大羽成征.『ガウス過程と機械学習』サポートページ. 2020-10-31. http://chasen.org/~daiti-m/gpbook/ ,（参照 2021-10-14）

[3] 気象庁HP ホーム > 各種データ･資料 > 過去の気象データ･ダウンロード http://www.data.jma.go.jp/risk/obsdl/index.php,（参照 2021-10-15）

･このスライドの内容は主に [1]の文献を参考にした.
･スライド中の図はサポートページ[2]を参考にし､Pythonで出力した.