ベイズ最適化・ベイズ推定

強化学習の1つ。

1. 既知データを学習
2. 所望の特性を出す確率の高い記述子候補を推薦
3. 推薦に応じでデータを追加、1へ戻る

ベイズ最適化の手順

a. カーネル回帰: カーネルは自由に選ぶ
b. 回帰変数が正規分布 (ベイズ的考え方)
c. 測定値(誤差)、予測値なども正規分布で連結:
ガウス過程
d. 予測値の分布関数のパラメータの統計分布
(つまり、予測値の平均と分散が得られる)

統計学

頻度主義 (frequentism) 統計学

確率が結果 (データ) を決める

ベイズ主義 (Bayesianism) 統計学

データは定数、確率分布 (確率分布関数に含まれる定数) が分布をもつ

客観確率 (objective probability)

実験または理論的考察（思考実験）から求められ、客観的な観測結果と比較できるランダムな事象についての確率

主観確率 (subjective probability)

人間の主観的な信念あるいは信頼の度合

回帰のベイズ的考え方

観測データ (x_i, y_i) が以下のようにあらわされるとする。

$$y_i = w_0 + w_1 x_i$$

頻度主義:「w_iは決定していて、
(誤差があって) y_i に不確定性(確率密度分布)がある」

ベイズ主義:「データ (x_i, y_i) は決定している。w_iに確率密度分布がある」

・w_iの分布が決まれば、w_i をサンプリングすれば $y_i = w_0 + w_1 x_i$ が計算できる。

例: w_i が正規分布であれば、正規分布の乱数によりサンプリングできる

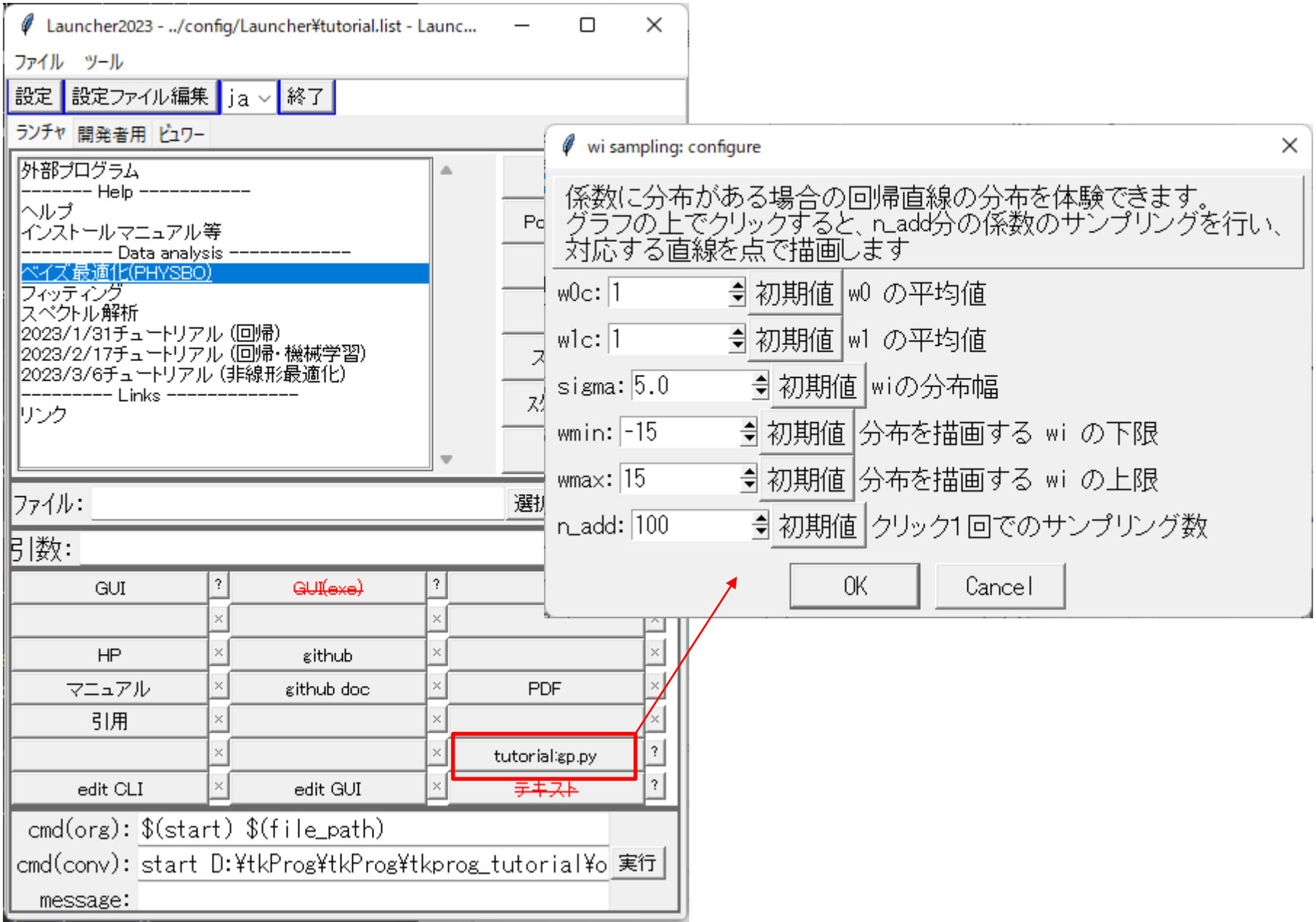
独立な正規分布 (平均 $\boldsymbol{\mu}$、分散 $\boldsymbol{\sigma}$): $p(\boldsymbol{w}) = \frac{1}{2\pi\sqrt{\sigma_0{}^2\sigma_1{}^2}} \exp\left(-\sum \frac{(w_i-\mu_i)^2}{2\sigma_i{}^2}\right)$

注: w_i が独立であるとして、共分散を0にしている

1. 最初は w_i に関する情報は無い: 主観確率

適当な事前分布を仮定する: $p(\boldsymbol{w}) = \frac{1}{2\pi\sigma^2} \exp\left(-\sum \frac{(w_i-\mu_i)^2}{2\sigma^2}\right) = N(\boldsymbol{w}; \mu, \sigma)$

w_i の分布と回帰


Launcher2023 - ../config/Launcher¥tutorial.list - Launc...
ファイル ツール
設定
設定ファイル編集
ja
終了
ランチャ
開発者用
ビュワー
外部プログラム
-------- Help ----------
ヘルプ
インストールマニュアル等
---------- Data analysis ----------
ベイズ最適化(PHYSBO)
フィッティング
スペクトル解析
2023/1/31チュートリアル (回帰)
2023/2/17チュートリアル (回帰・機械学習)
2023/3/6チュートリアル (非線形最適化)
---------- Links ----------
リンク
ファイル:
引数:
GUI
GUI(exe)
HP
github
マニュアル
github doc
PDF
引用
tutorial:gp.py
edit CLI
edit GUI
テキスト
cmd(org): $(start) $(file_path)
cmd(conv): start D:¥tkProg¥tkProg¥tkprog_tutorial¥o
実行
message:
wi sampling: configure
係数に分布がある場合の回帰直線の分布を体験できます。
グラフの上でクリックすると、n_add分の係数のサンプリングを行い、
対応する直線を点で描画します
w0c: 1
初期値
w0 の平均値
w1c: 1
初期値
w1 の平均値
sigma: 5.0
初期値
wiの分布幅
wmin: -15
初期値
分布を描画する wi の下限
wmax: 15
初期値
分布を描画する wi の上限
n_add: 100
初期値
クリック1回でのサンプリング数
OK
Cancel

w_i の分布と回帰結果: 広い分布

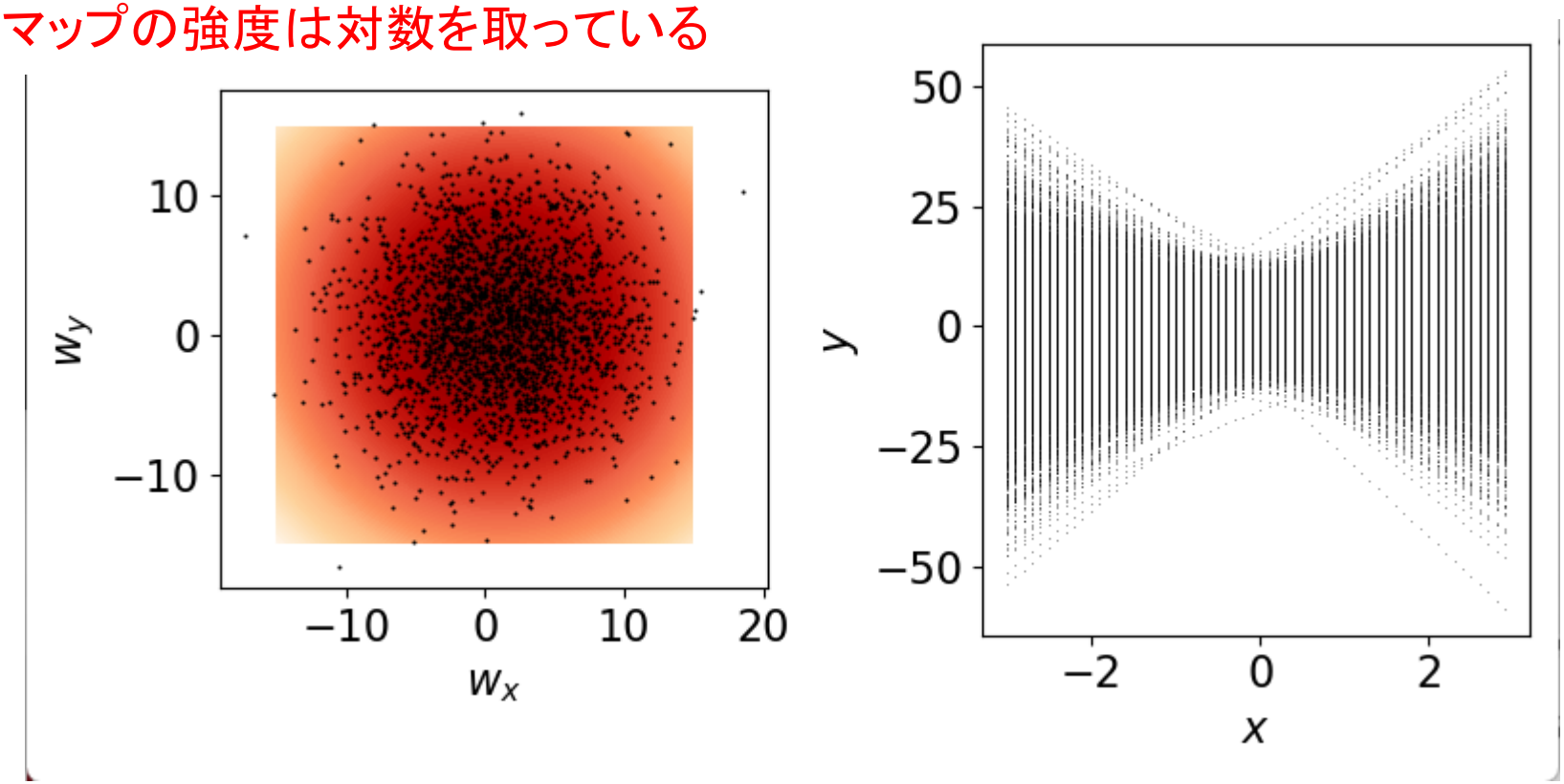
[tkProg]¥tkprog_tutorial¥gp¥gp.py

python gp.py

$y = w_0 + w_1 * x$

w_i: 中心 (1, 1)、σ = 5.0

マップの強度は対数を取っている



ベイズ的: データを使って w_i の分布を狭める

1. 最初は w_i に関する情報は無い: 事前分布 (主観確率)

適当な事前分布を仮定する: $p(\boldsymbol{w}) = \frac{1}{2\pi\sqrt{{\sigma_0}^2{\sigma_1}^2}} \exp\left(-\sum \frac{(w_i-\mu_i)^2}{2{\sigma_i}^2}\right) = N(\boldsymbol{w}; \boldsymbol{\mu}, \boldsymbol{\sigma})$

2. データ $\boldsymbol{D} = (x_i, y_i)$ を追加: w_i の範囲を狭めていく

事後分布 (データ $\boldsymbol{D}$ が与えられた場合に $\boldsymbol{w}$ の確率密度分布がどうなるか):

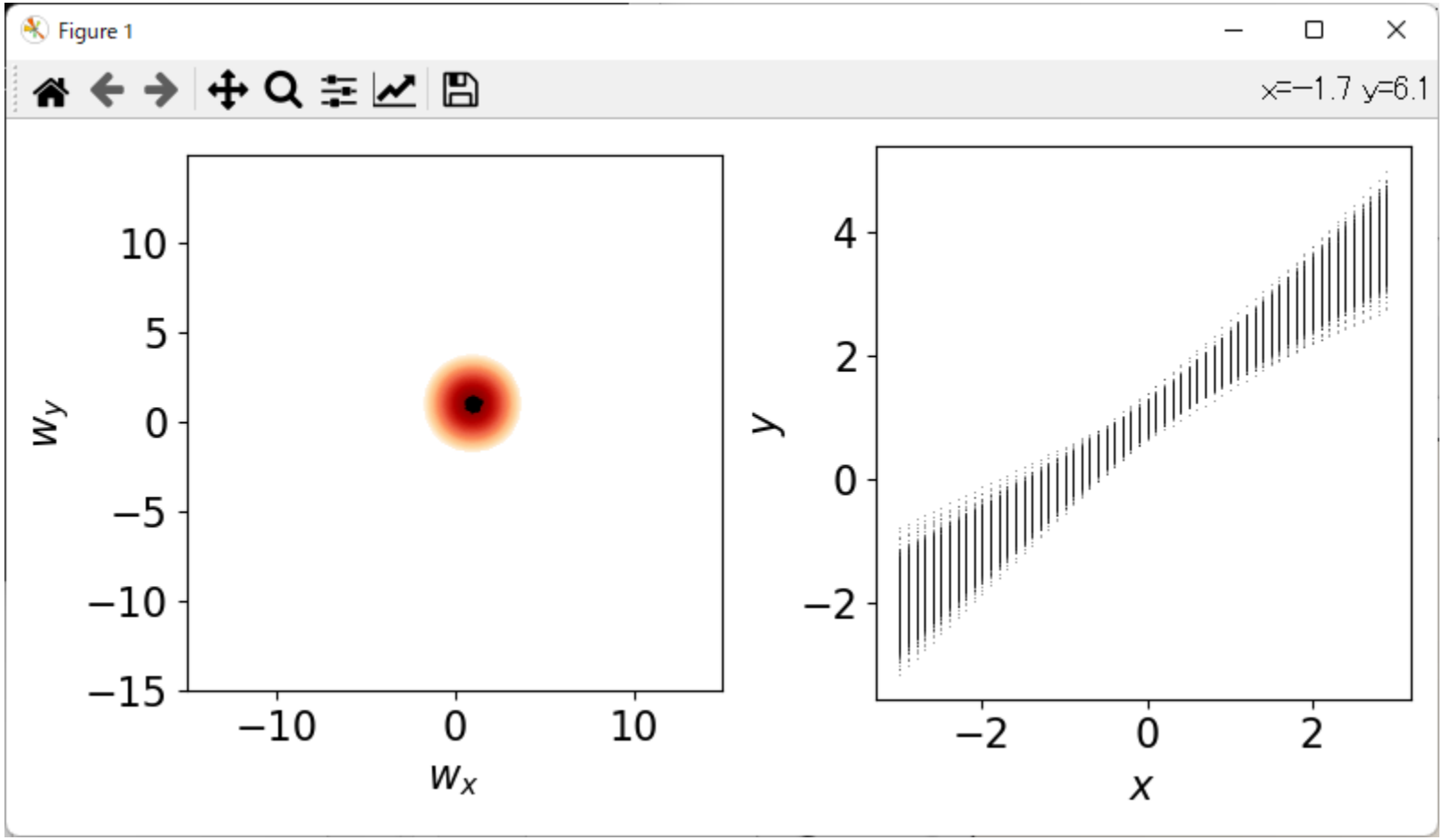
$$p(\boldsymbol{w}|\boldsymbol{D}) = \frac{1}{2\pi\sqrt{{\sigma_0}'^2\sigma_1\&^2}} \exp\left(-\frac{(\boldsymbol{w}-\boldsymbol{\mu}')^2}{2\boldsymbol{\sigma}'^2}\right) = N(\boldsymbol{w}; \boldsymbol{\mu}', \boldsymbol{\sigma}')$$

w_i の分布と回帰結果: 狭い分布

[tkProg]¥tkprog_tutorial¥gp¥gp.py

python gp.py 1 1 0.1

w_i: 中心 (1, 1)、σ = 0.1



多変量ガウス分布と線形変換

D次元の確率変数ベクトル $\boldsymbol{x}$ の正規分布

$$p(\boldsymbol{x}) = \frac{1}{\sqrt{2\pi}^{D}\sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \langle \boldsymbol{x} \rangle)^{T}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \langle \boldsymbol{x} \rangle)\right)$$

$\langle \boldsymbol{x} \rangle$: $\boldsymbol{x}$ の平均

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1D} \\ \sigma_{21} & \sigma_{22} & & \\ \vdots & & \ddots & \\ \sigma_{D1} & & & \sigma_{DD} \end{pmatrix}$$: 共分散行列

$$\sigma_{ij} = \frac{\sum(\boldsymbol{x_i} - \langle \boldsymbol{x} \rangle)(\boldsymbol{x_j} - \langle \boldsymbol{x} \rangle)}{n-1}$$

$|\Sigma|$: Σの固有値

公式1：変数変換 $\boldsymbol{y} - \langle \boldsymbol{y} \rangle = \boldsymbol{A}(\boldsymbol{x} - \langle \boldsymbol{x} \rangle)$

$$p(\boldsymbol{y}) \propto \exp\left(-\frac{1}{2}(\boldsymbol{y} - \langle \boldsymbol{y} \rangle)^{T} A^{-1T}\Sigma^{-1}A^{-1}(\boldsymbol{y} - \langle \boldsymbol{y} \rangle)\right)$$

$$p(\boldsymbol{y}) = \frac{1}{\sqrt{2\pi}^{D}\sqrt{|\Lambda|}} \exp\left(-\frac{1}{2}(\boldsymbol{y} - \langle \boldsymbol{y} \rangle)^{T}\Lambda^{-1}(\boldsymbol{y} - \langle \boldsymbol{y} \rangle)\right)$$

$\Lambda^{-1} = A^{-1T}\Sigma^{-1}A^{-1}$

xが正規分布をすれば、線形変換しても正規分布になる

w_i の分布をどうやって決めるか

$y = w_0 + w_1 x = \boldsymbol{w}^T \boldsymbol{x}$　データ (x_i, y_i)

$$p(\boldsymbol{w}|\Sigma^{-1}) = \frac{1}{\sqrt{2\pi}^D \sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}\boldsymbol{w}^T \Sigma^{-1} \boldsymbol{w}\right)$$

($\boldsymbol{w}$は平均 0、共分散 Σ の正規分布と仮定)

誤差ベクトル $\boldsymbol{\varepsilon} = \boldsymbol{w}^T \boldsymbol{x_i} - \boldsymbol{y_i}$ も正規分布とする (平均 0 、共分散 Λ)

$$p(\boldsymbol{\varepsilon}|\Lambda^{-1}) = \frac{1}{\sqrt{2\pi}^D \sqrt{|\Lambda|}} \exp\left(-\frac{1}{2}\sum \boldsymbol{\varepsilon}^T \Lambda^{-1} \boldsymbol{\varepsilon}\right)$$

最尤推定法: 尤度関数 $p(\boldsymbol{\varepsilon}|\Lambda^{-1})$ を最大化する $\boldsymbol{w}$、Λ を決めればよいので、

対数尤度 $-\frac{1}{2}\sum(\boldsymbol{w}^T \boldsymbol{x_i} - \boldsymbol{y_i})^T \Lambda^{-1} (\boldsymbol{w}^T \boldsymbol{x_i} - \boldsymbol{y_i})$ を 最大化

$\Rightarrow \sum(\boldsymbol{w}^T \boldsymbol{x_i} - \boldsymbol{y_i})^T \Lambda^{-1} (\boldsymbol{w}^T \boldsymbol{x_i} - \boldsymbol{y_i})$ を最小化する

Λ^{-1}を定数と考えて w を最適化すると

線形最小二乗法と同じになってしまい、意味がない

- $\boldsymbol{w}$ に正規分布を仮定しガウス過程を利用すると、
  正規分布の分散に出てくる $\boldsymbol{w}$ を平均化 (周辺化) して
  Λ^{-1} を求める問題にできる

尤度関数

事象 (x_k) が起こる確率を、既知のパラメータ (a_k)の確率密度関数 (PDF)

$$P(X = x_i|a_k) = \prod_i \left\{ \frac{1}{\sqrt{2\pi\sigma_i}} \exp\left[-\frac{\varepsilon_i(x_i|a_k)^2}{2{\sigma_i}^2} \right] \right\} = \prod_i \left(\frac{1}{\sqrt{2\pi\sigma_i}} \right) \cdot \exp\left[-\sum_i \frac{\varepsilon_i(x_i|a_k)^2}{2{\sigma_i}^2} \right]$$

などとする。　　($\varepsilon_i(x_i|a_k)$ は 誤差。$(x_i|a_k)$ は、x_iが確率変数でa_kがパラメータであることを示す)

頻度主義の考え方:

データ x_i が得られる確率は$P(X = x_i|a_k)$。パラメータ a_k は先に定まっている

ベイズ的な考え方:

データ x_i は確定した事象として考え、

確率分布 (確率変数 a_k) が確率密度関数を持つ と考える

$X = (x_i)$ がわかっているとし、上記の確率密度関数を

パラメータ (a_k) がどれだけ尤もらしいか (尤度) を表す確率密度関数とみなし、

変数 (a_k) の関数として

$P(a_i) = P(x_i|a_i)$ を尤度関数と呼ぶ

(x_iがパラメータ(データ)でa_kが確率変数)

最小二乗法の統計学的基盤: 最尤推定法

最尤推定法

誤差 $\varepsilon_i = f(x_i, a_i) - y_i$ が分散 σ_i の正規分布に従うとする。

データ (x_i, y_i) に対するパラメータ (a_i) の尤度関数は

$$P(a_i) = \prod_i \left(\frac{1}{\sqrt{2\pi\sigma_i}}\right) \cdot \exp\left[-\sum_i \frac{\varepsilon_i(x_i|a_k)^2}{2{\sigma_i}^2}\right]$$

尤度を最大化するパラメータ a_i を求めるのが「最尤推定法」。

$$\max P(a_i) = \max \ln P(a_i)$$

$$= \min \sum_i \frac{{\varepsilon_i}^2}{{\sigma_i}^2}$$: 最小二乗法に一致する

ガウス過程 (Gauss process)

例: 観測データ $\{x_i, y_i\}$ を基底関数 $\phi_k(x_i)$ で展開

観測データに誤差 ε_i があることを考慮すると、

$$t_i = \sum_k w_k \phi_k(x_i)$$

$$y_i = t_i + \varepsilon_i$$

とあらわされるような場合。

このような一連の関数 w_k、t_i、y_i、ε_iが
正規分布に従うとき、これをガウス過程と呼ぶ

Wikipedia:

確率過程 $\{X_t\}_{t\in T}$ は、任意に(有限個の)$X_{t1}, ..., X_{tk}$ を選んで作った線型結合(あるいはより一般に、$\{X_t\}_{t\in T}$ を標本関数 X_t 全体からなる連続濃度の函数空間と見たときの、任意の線型汎関数が正規分布に従うとき、ガウス過程という。

ガウス過程を利用したベイズ推定の流れ

1. 観測データ $\{x_i, y_i\}$

   $y_i(x_i) = t_i(x_i) + \varepsilon_i$　　ε_i:誤差

2. 予測値を基底関数$\varphi(x_i)$ で展開

   $t_i = \sum_{k'} w_{k'} \varphi(x_{i,k'})$

3. カーネルトリックを用いて計算を簡単化

   カーネルは、対称性・正定値性を満たしていればなんでもいい

4. w_k、ε_i が正規分布に従うとする: y_i は ε_i と同じ正規分布に従う

   分散にでてくる w_k で平均を取り、w_kを消去。

   カーネル回帰ではデータ数と同数の w_k の計算、逆行列の計算が必要だったが、不要になる。

   平均、分散を求めることを目的にするため、計算が軽くなる

5. 最尤推定法により、**w_kの平均、分散**が求まる

6. ベイズ推定により、**y_i を推定**

   w_kの具体的な値は使わない

正規分布の表記と公式

平均ベクトル $\boldsymbol{\mu}$、共分散行列 Σ の正規分布を

$$N(\boldsymbol{\mu}, \Sigma) = N(\text{平均}, \text{共分散行列}) = \frac{1}{\sqrt{2\pi}^{D}\sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}\boldsymbol{x}^{T}\Sigma^{-1}\boldsymbol{x}\right)$$

$$N(\boldsymbol{x}|\boldsymbol{\mu}, \Sigma) = N(\text{確率変数}|\text{平均}, \text{共分散行列})$$

と書く（ところどころ $N(\boldsymbol{\mu}, \Sigma^{-1})$ で書いたりもするので、柔軟に対応してください）。

公式1：線形変換： $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ の変化により、

$$p(\boldsymbol{y}) = N\left(\langle \boldsymbol{y} \rangle, \left((A^{-1})^{T}\Sigma^{-1}A^{-1}\right)^{-1}\right)$$

同時分布： ベクトル $\boldsymbol{x}$ が正規分布に従うとき、すべての x_i の同時分布である

$$p(\boldsymbol{x}) = N(\boldsymbol{\mu}, \Sigma)$$

公式2：条件付分布： $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2)$ のうち $\boldsymbol{x}_1$ を固定したときに 残りの $\boldsymbol{x}_2$ が従う分布

$$p(\boldsymbol{x_2}|\boldsymbol{x_1}) = N\left(\boldsymbol{\mu}_2 + \Sigma_{21}{\Sigma_{11}}^{-1}(\boldsymbol{x_1} - \boldsymbol{\mu}_1), \Sigma_{22} - \Sigma_{21}{\Sigma_{11}}^{-1}\Sigma_{12}\right)$$

となる ($\boldsymbol{x}_1$ について積分を取る (周辺化))

基底関数 (特徴ベクトル) で展開、wの消去

- x の特徴ベクトルを $\boldsymbol{\varphi}(\boldsymbol{x})$ としてモデル t を展開

$$(t) = (\sum_k w_k \varphi_k(\boldsymbol{x}^{(i)})) = \Phi \boldsymbol{w}$$

$$\Phi = (\Phi_{ij}) = (\varphi_j(\boldsymbol{x}^{(i)})):\text{計画行列}$$

- $\boldsymbol{w}$ が正規分布 $N(0, \lambda^2 I)$ に従うとする。
  - 公式1より $\boldsymbol{t} = \Phi \boldsymbol{w}$ は正規分布 $\boldsymbol{N}(\Phi \boldsymbol{w}, \boldsymbol{w}^T \Phi^T \Phi \boldsymbol{w})$ に従う。
  - $\boldsymbol{w}^T \Phi^T \Phi \boldsymbol{w}$ の平均を取ると、

(Φ はデータ $\boldsymbol{x}^{(i)}$ が与えられると定数行列なので)

$$E(\Phi \boldsymbol{w}) = \Phi E(\boldsymbol{w}) = 0$$

$$E(\boldsymbol{w}^T \Phi^T \Phi \boldsymbol{w}) = \Phi^T \Phi E(\boldsymbol{w}^T \boldsymbol{w}) = \lambda^2 \Phi^T \Phi。$$

つまり、$\boldsymbol{t} = \Phi \boldsymbol{w}$ は正規分布 $\boldsymbol{N}(0, \lambda^2 \Phi^T \Phi)$ に従う。

※ $\boldsymbol{w}$について平均を取った(周辺化)ため、

分布関数から w が消えている

カーネルトリック

- $\boldsymbol{t} = \Phi \boldsymbol{w}$ は正規分布 $\boldsymbol{N}(0, \lambda^2 \Phi^T \Phi)$ に従う。

- **カーネルトリック: 分散 $\lambda^2 \Phi^T \Phi$ をカーネル K で置き換える**
  ＝＞ $\boldsymbol{t} = \Phi \boldsymbol{w}$ は正規分布 $\boldsymbol{N}(\mathbf{0}, \mathrm{K})$ に従う

$$\mathrm{K} = \left(K_{ij}\right) = \lambda^2 \Phi^T \Phi$$

$K_{ij} = \lambda^2 \varphi\left(\boldsymbol{x}^{(i)}\right)^T \varphi\left(\boldsymbol{x}^{(j)}\right)$: **カーネル関数 $k(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)})$ で置き換え**

カーネル関数の例:

$$k\left(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}\right) = \theta_1 \exp\left(-\frac{\left(\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)}\right)^2}{\theta_2}\right)$$

動径基底関数 (radial basis function, RBF) / ガウスカーネル

θ_1, θ_2 は**ハイパーパラメータ**

カーネルは任意のものを使える (ガウス関数でなくてもよい)

誤差（ノイズ）とRidge回帰の関係

- **観測値 y_i には誤差 (ノイズ) ε_i が含まれる。モデルを t_i とする。**

$$y_i = t_i + \varepsilon_i$$

$$(t_i) = (\sum_k w_k \varphi_k(\boldsymbol{x}^{(i)})) = \Phi \boldsymbol{w}$$

ε_i は正規分布 $N(0, \sigma^2 I)$ に従い、互いに独立 (共分散は 0) とする

- $\boldsymbol{y}$ の確率分布は $y_i = t_i + \varepsilon_i$ より (t_i は x が与えらると定数なので、分散は 0)

$$p(\boldsymbol{y}|\boldsymbol{t}) = N(\boldsymbol{t}, \sigma^2 I)$$

- **モデル $\boldsymbol{t} = \Phi(\boldsymbol{x}^{(i)})\boldsymbol{w}$ は正規分布 $N(\boldsymbol{0}, \mathrm{K})$ に従う**

- $\boldsymbol{x}$ が与えられた時の $\boldsymbol{y}$ の分布関数は $\boldsymbol{t}$ について平均を取って

$$p(\boldsymbol{y}|\boldsymbol{x}) = \int p(\boldsymbol{y}, \boldsymbol{t}|\boldsymbol{x}) d\boldsymbol{t} = \int p(\boldsymbol{y}|\boldsymbol{t}) p(\boldsymbol{t}|\boldsymbol{x}) d\boldsymbol{t}$$

$$= \int N(\boldsymbol{y}|\langle \boldsymbol{t} \rangle, \sigma^2 I) N(\boldsymbol{t}|\langle \boldsymbol{t} \rangle, K) d\boldsymbol{t}$$

ガウス関数のコンボリューションなので

$$\mathrm{p}(\boldsymbol{y}|\boldsymbol{x}) = N(\boldsymbol{y}|\langle \boldsymbol{t} \rangle, K + \sigma^2 I)$$

Ridge回帰で、正則化係数 $p\alpha$ が σ^2 に対応している
ノイズを考慮することで正則化が自動的に入る
σ^2 もハイパーパラメータ

- 以下、$K + \sigma^2 I$ をカーネル $k(\boldsymbol{x}, \boldsymbol{x}')$ と書き換える

学習データ (x_i) $(i = 1, 2, \cdots, N)$ と予測

- N個の観測値 $D = (x_i)$ $(i = 1, 2, \cdots, N)$
  $\boldsymbol{y}, \boldsymbol{t}$ が正規化されているとすると、
  $\boldsymbol{y}, \boldsymbol{t}$ は分散 $k(\boldsymbol{x}, \boldsymbol{x}')$ のガウス過程 $\mathrm{GP}(0, k(\boldsymbol{x}, \boldsymbol{x}'))$ から生成されている
- $\boldsymbol{x}^*$に対する未知の $\boldsymbol{y}^*$ を予測する。観測データ $\boldsymbol{y}$ と合わせた $(\boldsymbol{y}, \boldsymbol{y}^*)$も
  $\boldsymbol{y}$ と $\boldsymbol{y}^*$ の同時分布として、同じガウス過程に従う。

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

$$k^* = \left(k(\boldsymbol{x}^*, \boldsymbol{x_1}), k(\boldsymbol{x}^*, \boldsymbol{x_2}), \cdots, k(\boldsymbol{x}^*, \boldsymbol{x_N})\right)$$

$$k^{**} = (k(\boldsymbol{x}^*, \boldsymbol{x}^*))$$

- 条件付分布の公式2より

$$p(\boldsymbol{x_2}|\boldsymbol{x_1}) = N\left(\boldsymbol{\mu}_2 + \Sigma_{21}{\Sigma_{11}}^{-1}(\boldsymbol{x_1} - \boldsymbol{\mu}_1), \Sigma_{22} - \Sigma_{21}{\Sigma_{11}}^{-1}\Sigma_{12}\right)$$

D と $\boldsymbol{x}^*$ が与えられたときに $\mathbf{y}^*$ がとる条件付確率分布は

$$p(\boldsymbol{y}^*|\boldsymbol{x}^*, D) = N\left(k^{*T}K^{-1}\boldsymbol{y}, k^{**} - k^{*T}K^{-1}k^*\right)$$

予測値の平均値: $k^{*T}K^{-1}\boldsymbol{y}$

予測値の分散　: $k^{**} - k^{*T}K^{-1}k^*$

RBFの場合: $k\left(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}\right) = \theta_1 \exp\left(-\frac{\left(\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)}\right)^2}{\theta_2}\right) + \delta_{ij}\theta_3$

ハイパーパラメータの推定

学習データの確率:カーネルにハイパーパラメータ $\boldsymbol{\theta}$ が含まれる

$$p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}) = N(\boldsymbol{y}|0, K(\boldsymbol{\theta}))$$

最尤推定法:

$$p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}) = N(\boldsymbol{y}|0, K(\boldsymbol{\theta})) = \frac{1}{\sqrt{2\pi}^{D}\sqrt{|K(\boldsymbol{\theta})|}} \exp\left(-\frac{1}{2}\boldsymbol{y}^{T}K(\boldsymbol{\theta})^{-1}\boldsymbol{y}\right)$$

の対数

$$\log p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}) = -\frac{D}{2}\log 2\pi - \frac{1}{2}\log|K(\boldsymbol{\theta})| - \frac{1}{2}\boldsymbol{y}^{T}K(\boldsymbol{\theta})^{-1}\boldsymbol{y}$$

が最大になるように、つまり、

$$\log|K(\boldsymbol{\theta})| + \boldsymbol{y}^{T}K(\boldsymbol{\theta})^{-1}\boldsymbol{y}$$

が最小になるようにハイパーパラメータを決定する。

=> Cross Validationや非線形最小化 (勾配法など) を使う