カーネル法

中田和秀

東京科学大学 工学院 経営工学系

機械学習入門

https://www.nakatalab.iee.e.titech.ac.jp/text/nakata.html

概要

ここではカーネルを使うことによって、単純な計算によって複雑は分析を行う一般的な方法論について説明を行う。その後で、カーネル回帰、カーネル Ridge 回帰、ガウス過程回帰について説明する。なお、サポートベクターマシンやカーネル主成分分析は別スライドで説明を行っている。

目次：

記号の使い方：

- $A := B$ は、B で A を定義する、B を A に代入することを意味する
- $[n]$ は n までのインデックスの集合を表し $[n] := \{1, 2, \cdots, n\}$

高次元空間への写像

- 入力空間を高次元の特徴空間に写像して、その特徴空間上で単純な分析
- 写像 $\phi : \mathbb{R}^n \to V$ を使って、$\boldsymbol{x}_d$ の代わりに $\phi(\boldsymbol{x}_d)$ を使う
- 元の空間 $\mathbb{R}^n$ でみると複雑な分析となる

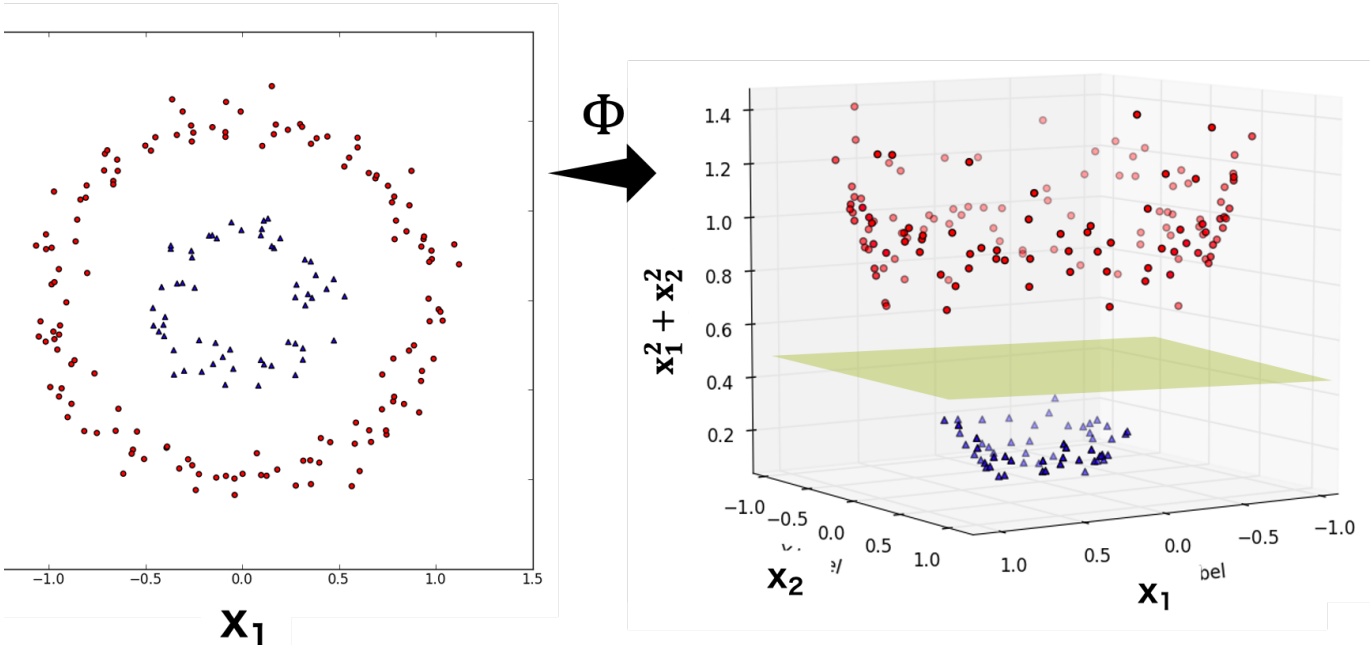


Figure: https://axa.biopapyrus.jp/machine-learning/svm/kernel-svm.html

カーネルトリック

写像 $\boldsymbol{\phi} : \mathbb{R}^n \to V$ をどう決めるかが難しい

もし

(1) $\boldsymbol{x}_i$ と $\boldsymbol{x}_j$ から直接 $\boldsymbol{\phi}(\boldsymbol{x}_i)^T \boldsymbol{\phi}(\boldsymbol{x}_j)$ の値を計算できる

(2) 高次元空間上での内積 $\boldsymbol{\phi}(\boldsymbol{x}_i)^T \boldsymbol{\phi}(\boldsymbol{x}_j)$ にのみ基づいて分析

$\Longrightarrow$ 写像 $\boldsymbol{\phi}(\boldsymbol{x})$ を陽に定めることなく分析が可能

まず、(1) について議論する。

$\boldsymbol{\phi}(\boldsymbol{x}_a)^T \boldsymbol{\phi}(\boldsymbol{x}_b)$ に相当するカーネル関数 $(\mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R})$ を導入

$$\mathrm{Ker}(\boldsymbol{x}_a, \boldsymbol{x}_b)$$

論点：
写像 ϕ の存在が担保されるように、カーネル関数を構築する必要がある

マーサーの条件

$K_{ij} := \mathrm{Ker}(\boldsymbol{x}_i, \boldsymbol{x}_j) \quad (i, j \in [D])$ とする。

$\boldsymbol{K} = \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^T$

$$\widetilde{\boldsymbol{X}} := \begin{pmatrix} \boldsymbol{\phi}(\boldsymbol{x}_1)^T \\ \boldsymbol{\phi}(\boldsymbol{x}_2)^T \\ \vdots \\ \boldsymbol{\phi}(\boldsymbol{x}_D)^T \end{pmatrix} \in \mathbb{R}^{D \times |V|}$$

必要条件： カーネル行列 $\boldsymbol{K} = \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^T$ は半正定値行列

実は必要十分条件である

マーサーの条件

任意の $i, j \in [D]$ に対し、$K_{ij} := \mathrm{Ker}(\boldsymbol{x}_i, \boldsymbol{x}_j)$ で作られる カーネル行列 $\boldsymbol{K} \in \mathbb{R}^{D \times D}$ が半正定値となるとき、写像 $\boldsymbol{\phi}$ の存在が保証される

カーネルの例

- $\mathrm{Ker}(\boldsymbol{x}_a, \boldsymbol{x}_b) := \boldsymbol{x}_a^T \boldsymbol{x}_b$ 線形カーネル
- $\mathrm{Ker}(\boldsymbol{x}_a, \boldsymbol{x}_b) := (\boldsymbol{x}_a^T \boldsymbol{x}_b + 1)^d \qquad (d \in \mathbb{N})$ 多項式カーネル
- $\mathrm{Ker}(\boldsymbol{x}_a, \boldsymbol{x}_b) := \exp\{-\gamma \|\boldsymbol{x}_a - \boldsymbol{x}_b\|^2\}$ RBF カーネル、ガウシアンカーネル

カーネルの証明

線形カーネル：

$K_{ij} := \boldsymbol{x}_i^T \boldsymbol{x}_j$ のとき、$\boldsymbol{K} = \boldsymbol{X}\boldsymbol{X}^T$ より半正定値行列

多項式カーネル：

$K_{ij} := (\boldsymbol{x}_i^T \boldsymbol{x}_j + 1)^d$ から作られるカーネル行列を $\boldsymbol{K}_d$ とする。

- $\boldsymbol{K}_1 = \boldsymbol{X}\boldsymbol{X}^T + \boldsymbol{e}\boldsymbol{e}^T$ より半正定値行列
- $\boldsymbol{K}_d, \boldsymbol{K}_1$ が半正定値のとき，$\boldsymbol{K}_{d+1} = \boldsymbol{K}_d \circ \boldsymbol{K}_1$ より，$\boldsymbol{K}_{d+1}$ も半正定値

  （証明は次スライド）

  $\boldsymbol{C} = \boldsymbol{A} \circ \boldsymbol{B}$ は Hadamard（アダマール）積を意味し $C_{ij} := A_{ij} B_{ij}$

RBF カーネル：

後述

Hadamard 積の半正定値性

$\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{n \times n}$ が半正定値のとき、$\boldsymbol{Z} := \boldsymbol{X} \circ \boldsymbol{Y} \in \mathbb{R}^{n \times n}$ も半正定値

Hadamard 積　$Z_{ij} := X_{ij} Y_{ij} \quad (i, j \in [n])$

証明：
$\boldsymbol{X} = \boldsymbol{X}^{1/2} \boldsymbol{X}^{1/2}$, $\boldsymbol{Y} = \boldsymbol{Y}^{1/2} \boldsymbol{Y}^{1/2}$ なので、$\boldsymbol{X}^{1/2}, \boldsymbol{Y}^{1/2}$ の i 列目を $\boldsymbol{x}_i, \boldsymbol{y}_i$ とする。

$$
\begin{aligned}
\boldsymbol{v}^T (\boldsymbol{X} \circ \boldsymbol{Y}) \boldsymbol{v} &= \sum_{i \in [n]} \sum_{j \in [n]} v_i v_j (\boldsymbol{x}_i^T \boldsymbol{x}_j)(\boldsymbol{y}_i^T \boldsymbol{y}_j) = \sum_{i \in [n]} \sum_{j \in [n]} v_i v_j (\boldsymbol{x}_i^T \boldsymbol{x}_j)(\boldsymbol{y}_j^T \boldsymbol{y}_i) \\
&= \sum_{i \in [n]} \sum_{j \in [n]} v_i v_j \mathsf{Tr}(\boldsymbol{x}_i^T \boldsymbol{x}_j \boldsymbol{y}_j^T \boldsymbol{y}_i) = \sum_{i \in [n]} \sum_{j \in [n]} v_i v_j \mathsf{Tr}(\boldsymbol{y}_i \boldsymbol{x}_i^T \boldsymbol{x}_j \boldsymbol{y}_j^T) \\
&= \mathsf{Tr} \left(\sum_{i \in [n]} \sum_{j \in [n]} v_i v_j \boldsymbol{y}_i \boldsymbol{x}_i^T \boldsymbol{x}_j \boldsymbol{y}_j^T \right) = \mathsf{Tr} \left(\left(\sum_{i \in [n]} v_i \boldsymbol{y}_i \boldsymbol{x}_i^T \right) \left(\sum_{j \in [n]} v_j \boldsymbol{x}_j \boldsymbol{y}_j^T \right) \right) \\
&= \mathsf{Tr} \left(\boldsymbol{M} \boldsymbol{M}^T \right) = \sum_{i \in [n]} \sum_{j \in [n]} (M_{ij})^2 \geq 0.
\end{aligned}
$$

写像 ϕ の例： 多項式カーネル

$\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^2$ に対する 3 次の多項式カーネル $\mathrm{Ker}(\boldsymbol{x}, \boldsymbol{y}) := (\boldsymbol{x}^T \boldsymbol{y} + 1)^3$ を考える。

$$
\begin{aligned}
\mathrm{Ker}(\boldsymbol{x}, \boldsymbol{y}) &:= (\boldsymbol{x}^T \boldsymbol{y} + 1)^3 \\
&= (x_1 y_1 + x_2 y_2 + 1)^3 \\
&= x_1^3 y_1^3 + 3 x_1^2 y_1^2 x_2 y_2 + 3 x_1 y_1 x_2^2 y_2^2 + x_2^3 y_2^3 \\
&\quad + 3 x_1^2 y_1^2 + 6 x_1 y_1 x_2 y_2 + 3 x_2^2 y_2^2 + 3 x_1 y_1 + 3 x_2 y_2 + 1
\end{aligned}
$$

よって、$\boldsymbol{\phi}(\boldsymbol{x}) : \mathbb{R}^2 \to \mathbb{R}^{10}$ を

$$
\boldsymbol{\phi}(\boldsymbol{x}) := \left(x_1^3,\ \sqrt{3} x_1^2 x_2,\ \sqrt{3} x_1 x_2^2,\ x_2^3,\ \sqrt{3} x_1^2,\ \sqrt{6} x_1 x_2,\ \sqrt{3} x_2^2,\ \sqrt{3} x_1,\ \sqrt{3} x_2,\ 1 \right)^T
$$

と定めると、次の関係が成り立つ。

$$
\boldsymbol{\phi}(\boldsymbol{x})^T \boldsymbol{\phi}(\boldsymbol{y}) = (\boldsymbol{x}^T \boldsymbol{y} + 1)^3 = \mathrm{Ker}(\boldsymbol{x}, \boldsymbol{y})
$$

$\boldsymbol{\phi}(\boldsymbol{x}) \in \mathbb{R}^{10}$ に対する線形な判別は、$\boldsymbol{x} \in \mathbb{R}^2$ に対する非線形な判別（3 次多項式）

無限次元ベクトルと関数

(1) 有限次元のベクトル： $\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix} \in \mathbb{R}^m$

これは次のように定義した f で表すことができる。

- $f(1) = x_1, \quad f(2) = x_2, \cdots, \quad f(m) = x_m$

  関数の定義域は $1, 2, \cdots, m$

(2) 無限次元のベクトル： $\boldsymbol{x} \in V$

これは次のように定義した f で表すことができる（２例挙げる）。

- $f(s) = x_s$ 関数の定義域は $s \in \mathbb{R}$
- $f(\boldsymbol{s}) = x_{\boldsymbol{s}}$ 関数の定義域は $\boldsymbol{s} \in \mathbb{R}^n$

無限次元のベクトルは関数で表すことができる。

無限次元ベクトルと関数

- 有限次元のベクトルの和とスカラー倍を自然に拡張した形で、関数の和とスカラー倍が定義できる。 $\cdots\cdots$ 関数空間は線形空間
- 内積も自然な形で拡張すると、$\boldsymbol{x}_a{}^T\boldsymbol{x}_b \ \rightarrow \ \int f_a(s)f_b(s)ds, \ \int f_a(\boldsymbol{s})f_b(\boldsymbol{s})d\boldsymbol{s}$

内積の説明：

M 次元の場合、$\boldsymbol{x}_a^T\boldsymbol{x}_b = \sum_{k\in[M]} f_a(k)f_b(k)$ であるが、無限次元に拡張するとき、無限個の和とすると発散する可能性があるため、要素数で割ることにする。

$$\text{内積} \ = \ \frac{1}{M}\sum_{k\in[M]} f_a(k)f_b(k)$$

関数の定義域 $\mathcal{S}_k$ を $-k$ から k までを $\frac{1}{k}$ の間隔で取ったものとする。

$$\mathcal{S}_k := \left\{-\frac{k^2}{k}, -\frac{k^2-1}{k}, \cdots, \frac{0}{k}, \cdots, \frac{k^2-1}{k}, \frac{k^2}{k}\right\}$$

このとき、 $\lim_{k\to\infty}\frac{1}{|\mathcal{S}_k|}\sum_{k\in\mathcal{S}_k} f_a(k)f_b(k) = \int_{\mathbb{R}} f_a(s)f_b(s)ds$

写像 ϕ の例： RBF カーネル

$\boldsymbol{\phi} : \mathbb{R}^n \to V$　　ベクトル $\boldsymbol{x}$ が入力されると、関数 $f(\boldsymbol{s})$ を出力

$$f(\boldsymbol{s}) := \beta \exp\{-2\gamma \|\boldsymbol{x} - \boldsymbol{s}\|^2\} \qquad (\alpha, \beta \text{は定数})$$

$\boldsymbol{\phi}(\boldsymbol{x}_a)$ と $\boldsymbol{\phi}(\boldsymbol{x}_b)$ の内積 :

$$\int \beta \exp\{-2\gamma \|\boldsymbol{x}_a - \boldsymbol{s}\|^2\} \, \beta \exp\{-2\gamma \|\boldsymbol{x}_b - \boldsymbol{s}\|^2\} \, d\boldsymbol{s}$$

$$= \beta^2 \left(\frac{\pi}{4\gamma}\right)^{n/2} \exp\left\{-\gamma \|\boldsymbol{x}_a - \boldsymbol{x}_b\|^2\right\}$$

α, β を適切に設定すると、RBF カーネル

$$\mathrm{Ker}(\boldsymbol{x}_a, \boldsymbol{x}_b) := \exp\{-\gamma \|\boldsymbol{x}_a - \boldsymbol{x}_b\|^2\}$$

に対応する高次元写像となる

※ なお、無限次元は扱いにくいので、カーネル回帰以降は有限次元のイメージで話を進める

式変形

$$
\begin{aligned}
\boldsymbol{\phi}(\boldsymbol{x}_1)^T\boldsymbol{\phi}(\boldsymbol{x}_2) &:= \int \beta\exp\{-2\gamma\|\boldsymbol{s}-\boldsymbol{x}_1\|^2\}\ \beta\exp\{-2\gamma\|\boldsymbol{s}-\boldsymbol{x}_2\|^2\}\ d\boldsymbol{s} \\
&= \beta^2\int \exp\{-2\gamma(\|\boldsymbol{s}-\boldsymbol{x}_1\|^2+\|\boldsymbol{s}-\boldsymbol{x}_2\|^2)\}\ d\boldsymbol{s} \\
&= \beta^2\int \exp\left\{-\gamma\left(4\left\|\boldsymbol{s}-\frac{\boldsymbol{x}_1+\boldsymbol{x}_2}{2}\right\|^2+\|\boldsymbol{x}_1-\boldsymbol{x}_2\|^2\right)\right\}\ d\boldsymbol{s} \\
&= \beta^2\exp\left\{-\gamma\|\boldsymbol{x}_1-\boldsymbol{x}_2\|^2\right\}\int \exp\left\{-4\gamma\left\|\boldsymbol{s}-\frac{\boldsymbol{x}_1+\boldsymbol{x}_2}{2}\right\|^2\right\}\ d\boldsymbol{s} \\
&= \beta^2\left(\frac{\pi}{4\gamma}\right)^{n/2}\exp\left\{-\gamma\|\boldsymbol{x}_1-\boldsymbol{x}_2\|^2\right\}
\end{aligned}
$$

$$
\begin{aligned}
2\|\boldsymbol{s}-\boldsymbol{x}_1\|^2+2\|\boldsymbol{s}-\boldsymbol{x}_2\|^2 &= 4\|\boldsymbol{s}\|^2-4\boldsymbol{s}^T(\boldsymbol{x}_1+\boldsymbol{x}_2)+2\|\boldsymbol{x}_1\|^2+2\|\boldsymbol{x}_2\|^2 \\
&= 4\left\|\boldsymbol{s}-\frac{\boldsymbol{x}_1+\boldsymbol{x}_2}{2}\right\|^2-\|\boldsymbol{x}_1+\boldsymbol{x}_2\|^2+2\|\boldsymbol{x}_1\|^2+2\|\boldsymbol{x}_2\|^2 \\
&= 4\left\|\boldsymbol{s}-\frac{\boldsymbol{x}_1+\boldsymbol{x}_2}{2}\right\|^2+\|\boldsymbol{x}_1-\boldsymbol{x}_2\|^2
\end{aligned}
$$

$4\gamma = \dfrac{1}{2\sigma^2}$ より、$\dfrac{1}{2\pi\sigma^2} = \dfrac{4\gamma}{\pi}$

RBF カーネルのカーネル行列

RBF カーネル： $K_{ij} = \mathrm{Ker}(\boldsymbol{x}_i, \boldsymbol{x}_j) := \exp\{-\gamma\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2\}$

すでに写像 $\boldsymbol{\phi}(\boldsymbol{x})$ は構築できているため、$\boldsymbol{K}$ が半正定値行列であることを示す必要はないが、以下のように写像 $\boldsymbol{\phi}(\boldsymbol{x})$ を使うと示すことができる。

任意の $\boldsymbol{v} \in \mathbb{R}^D$ に対し、

$$
\begin{aligned}
\boldsymbol{v}^T \boldsymbol{K} \boldsymbol{v} &= \sum_{i,j \in [D]} v_i v_j \exp\{-\gamma\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2\} \\
&= \sum_{i,j \in [D]} v_i v_j \left(\frac{4\gamma}{\pi}\right)^{n/2} \int \exp\{-2\gamma\|\boldsymbol{x}_i - \boldsymbol{s}\|^2\} \ \exp\{-2\gamma\|\boldsymbol{x}_j - \boldsymbol{s}\|^2\} \, d\boldsymbol{s} \\
&= \left(\frac{4\gamma}{\pi}\right)^{n/2} \int \sum_{i,j \in [D]} v_i v_j \exp\{-2\gamma\|\boldsymbol{x}_i - \boldsymbol{s}\|^2\} \ \exp\{-2\gamma\|\boldsymbol{x}_j - \boldsymbol{s}\|^2\} \, d\boldsymbol{s} \\
&= \left(\frac{4\gamma}{\pi}\right)^{n/2} \int \left(\sum_{i \in [D]} v_i \exp\{-2\gamma\|\boldsymbol{x}_i - \boldsymbol{s}\|^2\} \right)^2 d\boldsymbol{s} \\
&\geq 0.
\end{aligned}
$$

カーネル法

再掲　もし

(1) $\boldsymbol{x}_i$ と $\boldsymbol{x}_j$ から直接 $\boldsymbol{\phi}(\boldsymbol{x}_i)^T\boldsymbol{\phi}(\boldsymbol{x}_j)$ の値を計算できる

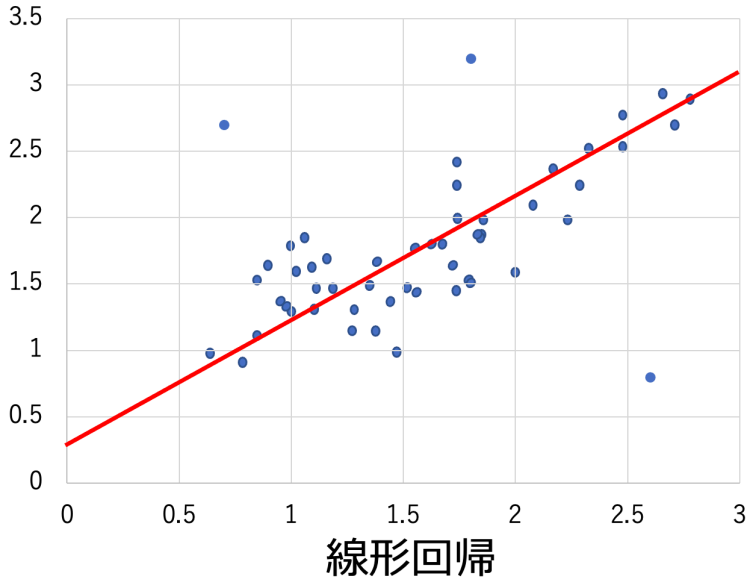
(2) 高次元空間上での内積 $\boldsymbol{\phi}(\boldsymbol{x}_i)^T\boldsymbol{\phi}(\boldsymbol{x}_j)$ にのみ基づいて分析

$\Longrightarrow$ 写像 $\boldsymbol{\phi}(\boldsymbol{x})$ を陽に定めることなく分析が可能

(2) の主な分析法

- カーネル回帰、カーネル Ridge 回帰
- ガウス過程回帰
- サポートベクターマシン　（サポートベクターマシンのスライド参照）
- カーネル主成分分析　（特徴抽出のスライド参照）

線形回帰



データ： $\{(\boldsymbol{x}_d, y_d)\}_{d\in[D]}$, $\quad \boldsymbol{x}_d \in \mathbb{R}^n,\ y_d \in \mathbb{R}$

線形関数（アフィン関数）：$y = \boldsymbol{w}^T\boldsymbol{x} + b$

$$\tilde{\boldsymbol{x}} := \begin{pmatrix} 1 \\ \boldsymbol{x} \end{pmatrix} \in \mathbb{R}^{1+n},\ \tilde{\boldsymbol{w}} := \begin{pmatrix} b \\ \boldsymbol{w} \end{pmatrix} \in \mathbb{R}^{1+n} \text{ とすると、} \boldsymbol{w}^T\boldsymbol{x} + b = \tilde{\boldsymbol{w}}^T\tilde{\boldsymbol{x}}$$

以下では、１次元を増やした $\tilde{\boldsymbol{x}}_d, \tilde{\boldsymbol{w}}$ を $\boldsymbol{x}_d, \boldsymbol{w} \in \mathbb{R}^n$ として考える。

線形回帰

予測器

$$y \simeq \hat{y} = \boldsymbol{w}^T \boldsymbol{x} \qquad (\text{パラメタは } \boldsymbol{w})$$

誤差関数

予測誤差の２乗

$$L(y, \hat{y}) := \frac{1}{2}(y - \hat{y})^2 \qquad y\text{: 目標} \quad \hat{y}\text{: 予測値}$$

学習

$$\min_{\boldsymbol{w}} \sum_{d \in [D]} (y_d - \boldsymbol{w}^T \boldsymbol{x}_d)^2$$

目的関数の $1/2D$ は省略

線形回帰は最小２乗法や重回帰とも呼ばれる

学習アルゴリズム

$$\boldsymbol{X} := \begin{pmatrix} \boldsymbol{x}_1^T \\ \boldsymbol{x}_2^T \\ \vdots \\ \boldsymbol{x}_D^T \end{pmatrix} \in \mathbb{R}^{D \times n}, \quad \boldsymbol{y} := \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_D \end{pmatrix} \in \mathbb{R}^D \quad \text{とする。}$$

学習の行列表現

$$\min_{\boldsymbol{w}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|^2$$

$\boldsymbol{X}$ が列フルランク $\mathrm{rank}(\boldsymbol{X}) = n$ と仮定

学習アルゴリズム

最適解は解析的に与えられる

$$\boldsymbol{w}^* := (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

証明

$$F(\boldsymbol{w}) := \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|^2 = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})$$
$$= \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{y}^T\boldsymbol{X}\boldsymbol{w} + \boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w}$$

最適解の必要条件 $\nabla F(\boldsymbol{w}^*) = \boldsymbol{0}$ を考える。

$$\nabla F(\boldsymbol{w}^*) = -2\boldsymbol{X}^T\boldsymbol{y} + 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w}^* = \boldsymbol{0}$$
$$\iff \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w}^* = \boldsymbol{X}^T\boldsymbol{y}$$
$$\iff \boldsymbol{w}^* = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \qquad (\because \mathrm{rank}(\boldsymbol{X}) = n \text{ より逆行列は存在})$$

また、$\nabla^2 F(\boldsymbol{w})$ は半正定値である。なぜなら、任意の $\boldsymbol{v} \in \mathbb{R}^n$ に対して

$$\boldsymbol{v}^T\nabla^2 F(\boldsymbol{w})\boldsymbol{v} = 2\boldsymbol{v}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{v} = 2(\boldsymbol{X}\boldsymbol{v})^T(\boldsymbol{X}\boldsymbol{v}) \geq 0$$

よって、$F(\boldsymbol{w})$ は凸関数となり、$\nabla F(\boldsymbol{w}^*) = \boldsymbol{0}$ は最適解の必要十分条件

以上より、$\boldsymbol{w}^* = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ が最適解。

カーネル回帰 1

$\boldsymbol{X} := \begin{pmatrix} \boldsymbol{x}_1^T \\ \boldsymbol{x}_2^T \\ \vdots \\ \boldsymbol{x}_D^T \end{pmatrix}$ の代わりに $\widetilde{\boldsymbol{X}} := \begin{pmatrix} \boldsymbol{\phi}(\boldsymbol{x}_1)^T \\ \boldsymbol{\phi}(\boldsymbol{x}_2)^T \\ \vdots \\ \boldsymbol{\phi}(\boldsymbol{x}_D)^T \end{pmatrix}$ を使う

カーネル回帰

$$\min_{\widetilde{\boldsymbol{w}}} \|\boldsymbol{y} - \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{w}}\|^2$$

$\widetilde{\boldsymbol{w}}^* := (\widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{X}})^{-1}\widetilde{\boldsymbol{X}}^T\boldsymbol{y}$ としたいが，問題がある。

- 高次元空間の次元 $|V|$ がデータ数 D より大きいと、解が一意に定まらない。
  （$\widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{X}}$ が正則にならない）
- 我々が使えるのはカーネル行列 $\boldsymbol{K} := \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^T$
  $K_{ij} := \mathrm{Ker}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{\phi}(\boldsymbol{x}_i)^T\boldsymbol{\phi}(\boldsymbol{x}_j)$

カーネル回帰 2

以下では、$D < |V|$ で、$\boldsymbol{K}$ は正則を仮定する（RBF カーネルなど）

カーネル回帰

多数ある最適解の中で、$\widetilde{\boldsymbol{w}}$ のノルムが小さなものを採用する。

$$\begin{aligned} \min_{\widetilde{\boldsymbol{w}}} \quad & \frac{1}{2}\|\widetilde{\boldsymbol{w}}\|^2 \\ \text{s.t.} \quad & \boldsymbol{y} = \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{w}} \end{aligned}$$

ラグランジュの未定乗数法

$$L(\widetilde{\boldsymbol{w}}, \boldsymbol{\lambda}) := \frac{1}{2}\widetilde{\boldsymbol{w}}^T\widetilde{\boldsymbol{w}} + \boldsymbol{\lambda}^T(\boldsymbol{y} - \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{w}})$$

最適解の必要十分条件

$$\nabla_{\widetilde{\boldsymbol{w}}} L(\widetilde{\boldsymbol{w}}, \boldsymbol{\lambda}) := \widetilde{\boldsymbol{w}} - \widetilde{\boldsymbol{X}}^T\boldsymbol{\lambda} = \boldsymbol{0}$$

$$\nabla_{\boldsymbol{\lambda}} L(\widetilde{\boldsymbol{w}}, \boldsymbol{\lambda}) := \boldsymbol{y} - \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{w}} = \boldsymbol{0}$$

学習

これを解くと、

$$\widetilde{\boldsymbol{w}}^* = \widetilde{\boldsymbol{X}}^T \boldsymbol{K}^{-1} \boldsymbol{y}$$

新しい $\boldsymbol{x}$ に対する $f(\boldsymbol{x}) = (\widetilde{\boldsymbol{w}}^*)^T \widetilde{\boldsymbol{x}}$ の値は次のように計算できる。

$$\begin{aligned} f(\boldsymbol{x}) &= (\widetilde{\boldsymbol{w}}^*)^T \widetilde{\boldsymbol{x}} \\ &= \widetilde{\boldsymbol{x}}^T \widetilde{\boldsymbol{X}}^T \boldsymbol{K}^{-1} \boldsymbol{y} \end{aligned}$$

$\boldsymbol{\alpha} := \boldsymbol{K}^{-1} \boldsymbol{y}$ とすると

$$\begin{aligned} &= \widetilde{\boldsymbol{x}}^T \widetilde{\boldsymbol{X}}^T \boldsymbol{\alpha} \\ &= \widetilde{\boldsymbol{x}}^T \sum_{d \in [D]} \alpha_d \widetilde{\boldsymbol{x}_d} \\ &= \sum_{d \in [D]} \alpha_d \mathrm{Ker}(\boldsymbol{x}, \boldsymbol{x}_d) \end{aligned}$$

基底関数による意味付け

基底関数： $f_1(\boldsymbol{x}),\ f_2(\boldsymbol{x}), \cdots,\ f_M(\boldsymbol{x})$

予測器（$\boldsymbol{w} \in \mathbb{R}^M$ はパラメタ）： $f(\boldsymbol{x}) := \sum_{k \in [M]} w_k f_k(\boldsymbol{x})$

$\boldsymbol{\phi}(\boldsymbol{x}) := \begin{pmatrix} f_1(\boldsymbol{x}) \\ f_2(\boldsymbol{x}) \\ \vdots \\ f_M(\boldsymbol{x}) \end{pmatrix}$ とすると、$f(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x})$

２乗損失を使ったパラメタ $\boldsymbol{w}$ を学習は線形回帰となる。

$$\min_{\boldsymbol{w}} \sum_{d \in [D]} (y_d - \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}))^2 \qquad \Longleftrightarrow \qquad \min_{\boldsymbol{w}} \|\boldsymbol{y} - \widetilde{\boldsymbol{X}} \boldsymbol{w}\|^2$$

$\mathcal{F} = \{f(\boldsymbol{x}) : \exists \boldsymbol{w} \in \mathbb{R}^M,\ f(\boldsymbol{x}) = \sum_{k \in [M]} w_k f_k(\boldsymbol{x})\}$ とすると、

$$\min_{f \in \mathcal{F}} \sum_{d \in [D]} (y_d - f(\boldsymbol{x}_d))^2$$

基底関数の例

例１： 線形回帰　　特徴量 $\boldsymbol{x} \in \mathbb{R}^n$

$$f_k(x) = \boldsymbol{e}_k \quad (k = 1, 2, \cdots, n)$$

$\boldsymbol{e}_k \in \mathbb{R}^n$ は、k 番目の要素が 1 で、残りの要素が 0 のベクトル

例２： 多項式回帰　　特徴量 $x \in \mathbb{R}$

$$f_0(x) = 1,\ f_1(x) = x,\ f_2(x) = x^2, \cdots, f_M(x) = x^M$$

例３：RBF カーネルを用いたカーネル回帰　　特徴量 $\boldsymbol{x} \in \mathbb{R}^n$

$$f_{\boldsymbol{s}}(\boldsymbol{x}) := \exp\{-\gamma \|\boldsymbol{x} - \boldsymbol{s}\|^2\} \qquad (\boldsymbol{s} \in \mathbb{R}^n)$$

無限個の基底関数

カーネル Ridge 回帰

正則化項を加えた線形回帰

Ridge 回帰

$$\min_{\boldsymbol{w}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|^2 + \lambda\|\boldsymbol{w}\|^2$$

$$\boldsymbol{w}^* = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

カーネル Ridge 回帰

$$\min_{\widetilde{\boldsymbol{w}}} \|\boldsymbol{y} - \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{w}}\|^2 + \lambda\|\widetilde{\boldsymbol{w}}\|^2$$

$$\widetilde{\boldsymbol{w}}^* = (\widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{X}} + \lambda\boldsymbol{I})^{-1}\widetilde{\boldsymbol{X}}^T\boldsymbol{y}$$

- $\lambda > 0$ で、$\widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{X}} + \lambda\boldsymbol{I}$ は正則　$\rightarrow$　最適解は一つ
- $\boldsymbol{K} = \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^T$ を使って計算できない

カーネル Ridge 回帰

$\boldsymbol{K} = \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^T$ に対して、次の性質が成り立つ（$\lambda > 0$）。

$$\begin{aligned}\widetilde{\boldsymbol{w}}^* &= (\widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{X}} + \lambda \boldsymbol{I})^{-1}\widetilde{\boldsymbol{X}}^T \boldsymbol{y} \\ &= \widetilde{\boldsymbol{X}}^T(\boldsymbol{K} + \lambda \boldsymbol{I})^{-1}\boldsymbol{y}\end{aligned}$$

※ 上式では $\boldsymbol{I} \in \mathbb{R}^{|V|\times|V|}$, 下式では $\boldsymbol{I} \in \mathbb{R}^{D\times D}$ であることに注意

新しい $\boldsymbol{x}$ に対する $f(\boldsymbol{x}) = (\widetilde{\boldsymbol{w}}^*)^T\widetilde{\boldsymbol{x}}$ の値は次のように計算できる。

$$\begin{aligned}f(\boldsymbol{x}) &= (\widetilde{\boldsymbol{w}}^*)^T\widetilde{\boldsymbol{x}} \\ &= \widetilde{\boldsymbol{x}}^T\widetilde{\boldsymbol{X}}^T(\boldsymbol{K} + \lambda \boldsymbol{I})^{-1}\boldsymbol{y}\end{aligned}$$

$\boldsymbol{\alpha} := (\boldsymbol{K} + \lambda \boldsymbol{I})^{-1}\boldsymbol{y}$ とすると

$$\begin{aligned}&= \widetilde{\boldsymbol{x}}^T \sum_{d\in[D]} \alpha_d \widetilde{\boldsymbol{x}_d} \\ &= \sum_{d\in[D]} \alpha_d \mathrm{Ker}(\boldsymbol{x}, \boldsymbol{x}_d)\end{aligned}$$

式変形

Sherman–Morrison–Woodbury の公式より、

$$
\begin{aligned}
(\widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{X}} + \lambda\boldsymbol{I})^{-1} &= \frac{1}{\lambda}\boldsymbol{I} - \frac{1}{\lambda}\boldsymbol{I}\widetilde{\boldsymbol{X}}^T\left(\boldsymbol{I} + \widetilde{\boldsymbol{X}}\frac{1}{\lambda}\boldsymbol{I}\widetilde{\boldsymbol{X}}^T\right)^{-1}\widetilde{\boldsymbol{X}}\frac{1}{\lambda}\boldsymbol{I} \\
&= \frac{1}{\lambda}\boldsymbol{I} - \frac{1}{\lambda}\widetilde{\boldsymbol{X}}^T\left(\lambda\boldsymbol{I} + \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^T\right)^{-1}\widetilde{\boldsymbol{X}}
\end{aligned}
$$

よって、

$$
\begin{aligned}
(\widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{X}} + \lambda\boldsymbol{I})^{-1}\widetilde{\boldsymbol{X}}^T\boldsymbol{y} &= \left(\frac{1}{\lambda}\boldsymbol{I} - \frac{1}{\lambda}\widetilde{\boldsymbol{X}}^T\left(\lambda\boldsymbol{I} + \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^T\right)^{-1}\widetilde{\boldsymbol{X}}\right)\widetilde{\boldsymbol{X}}^T\boldsymbol{y} \\
&= \frac{1}{\lambda}\widetilde{\boldsymbol{X}}^T\boldsymbol{y} - \frac{1}{\lambda}\widetilde{\boldsymbol{X}}^T\left(\lambda\boldsymbol{I} + \boldsymbol{K}\right)^{-1}\boldsymbol{K}\boldsymbol{y} \\
&= \frac{1}{\lambda}\widetilde{\boldsymbol{X}}^T\left(\lambda\boldsymbol{I} + \boldsymbol{K}\right)^{-1}\left(\lambda\boldsymbol{I} + \boldsymbol{K}\right)\boldsymbol{y} - \frac{1}{\lambda}\widetilde{\boldsymbol{X}}^T\left(\lambda\boldsymbol{I} + \boldsymbol{K}\right)^{-1}\boldsymbol{K}\boldsymbol{y} \\
&= \frac{1}{\lambda}\widetilde{\boldsymbol{X}}^T\left(\lambda\boldsymbol{I} + \boldsymbol{K}\right)^{-1}\lambda\boldsymbol{I}\boldsymbol{y} \\
&= \widetilde{\boldsymbol{X}}^T\left(\lambda\boldsymbol{I} + \boldsymbol{K}\right)^{-1}\boldsymbol{y}
\end{aligned}
$$

カーネル回帰とカーネル Ridge 回帰

カーネル回帰で得られる重み：

$$\begin{aligned}\widetilde{\boldsymbol{w}}^* := \underset{\widetilde{\boldsymbol{w}}}{\mathrm{argmin}} \quad & \frac{1}{2}\|\widetilde{\boldsymbol{w}}\|^2 \\ \text{s.t.} \quad & \boldsymbol{y} = \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{w}}\end{aligned}$$

カーネル Ridge 回帰で得られる重み：

$$\widetilde{\boldsymbol{w}}^*(\lambda) := \underset{\widetilde{\boldsymbol{w}}}{\mathrm{argmin}} \, \|\boldsymbol{y} - \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{w}}\|^2 + \lambda\|\widetilde{\boldsymbol{w}}\|^2$$

このとき、次の関係が成り立つ。

$$\widetilde{\boldsymbol{w}}^* = \lim_{\lambda \to +0} \widetilde{\boldsymbol{w}}^*(\lambda)$$

説明：

$$\widetilde{\boldsymbol{w}}^* = \widetilde{\boldsymbol{X}}^T \boldsymbol{K}^{-1} \boldsymbol{y}$$

$$\widetilde{\boldsymbol{w}}^*(\lambda) = \widetilde{\boldsymbol{X}}^T (\boldsymbol{K} + \lambda \boldsymbol{I})^{-1} \boldsymbol{y} \qquad (\lambda > 0)$$

表現定理

- カーネル回帰もカーネル Ridge 回帰も、$\boldsymbol{w}^* = \widetilde{\boldsymbol{X}}^T \boldsymbol{\alpha}$ という形で表現できる。
- カーネル法では多くの場合同様の性質が成り立ち、表現定理という。

$\boldsymbol{w}^* = \widetilde{\boldsymbol{X}}^T \boldsymbol{\alpha}$ を前提とすれば、

カーネル回帰：

$$\min_{\boldsymbol{\alpha}} \|\boldsymbol{y} - \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^T \boldsymbol{\alpha}\|^2$$

カーネル Ridge 回帰：

$$\min_{\boldsymbol{\alpha}} \|\boldsymbol{y} - \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^T \boldsymbol{\alpha}\|^2 + \lambda \|\widetilde{\boldsymbol{X}}^T \boldsymbol{\alpha}\|^2$$

カーネル行列 $\boldsymbol{K}$ が正則とすれば、最適解は簡単に導出できる。

$$\boldsymbol{\alpha}^* = \boldsymbol{K}^{-1}\boldsymbol{y}$$
$$\boldsymbol{\alpha}^*(\lambda) = (\boldsymbol{K} + \lambda \boldsymbol{I})^{-1}\boldsymbol{y}$$

確率モデルに基づいた解析

データ $\{(\boldsymbol{x}_d, y_d)\}_{d\in[D]}$ が次の確率モデルから生成されたサンプル（標本）とする。

$$y_d = \boldsymbol{w}^T \boldsymbol{x}_d + \epsilon_d \quad (d \in [D])$$
$$\iff \quad \boldsymbol{y} = \boldsymbol{X}\boldsymbol{w} + \boldsymbol{\epsilon}$$

$\boldsymbol{\epsilon}$ はノイズ（測定誤差など）を表す確率変数

ここで確率変数 $\boldsymbol{\epsilon} \in \mathbb{R}^D$ が従う仮定として、次のものを考える。

誤差項に対する仮定

- $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$

線形回帰・カーネル回帰と最尤推定

このページの話は、線形回帰とカーネル回帰で共通
（チルダがついているかどうかの違い）

仮定：

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$$

このとき、

$$\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{w}, \sigma^2 \boldsymbol{I})$$

尤度関数：　$l(\boldsymbol{y}; \boldsymbol{w}) = \dfrac{1}{(2\pi\sigma^2)^{D/2}} \exp\left\{ -\dfrac{1}{2\sigma^2} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) \right\}$

対数尤度関数：　$\log l(\boldsymbol{y}; \boldsymbol{w}) = -\dfrac{D}{2} \log 2\pi\sigma^2 - \dfrac{1}{2\sigma^2} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|^2$

対数尤度関数の最大化は線形回帰と等価

$$\max_{\boldsymbol{w}} \ \log l(\boldsymbol{y}; \boldsymbol{w})$$

つまり、線形回帰やカーネル回帰は最尤推定法とみなせる

MAP 推定

ベイズの定理を使うと、データ $\mathcal{D} = \{(\boldsymbol{x}_d, y_d)\}_{d\in[D]}$ のもとでのパラメタ $\boldsymbol{w}$ の確率は、次のようになる。

$$P(\boldsymbol{w}|\mathcal{D}) = \frac{P(\mathcal{D}|\boldsymbol{w})P(\boldsymbol{w})}{P(\mathcal{D})}$$

- $P(\boldsymbol{w})$：パラメタ $\boldsymbol{w}$ の事前確率
- $P(\boldsymbol{w}|\mathcal{D})$：データ $\mathcal{D}$ が与えられた状況でのパラメタ $\boldsymbol{w}$ の事後確率

MAP(maximum a posteriori) 推定

事後確率 $P(\boldsymbol{w}|\mathcal{D})$ が最大となるパラメタ $\boldsymbol{w}$ を採用する

$$\max_{\boldsymbol{w}} \ \log P(\boldsymbol{w}|\mathcal{D}) \quad \Longleftrightarrow \quad \max_{\boldsymbol{w}} \ \log P(\mathcal{D}|\boldsymbol{w}) + \log P(\boldsymbol{w})$$

カーネル回帰と MAP 推定

仮定

- パラメタの事前分布：$\widetilde{\boldsymbol{w}} \sim N(\boldsymbol{0}, \tau^2 \boldsymbol{I})$
- $\boldsymbol{y} = \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{w}}$ $\cdots\cdots$ 誤差は無し

$$\log P(\mathcal{D}|\widetilde{\boldsymbol{w}}) + \log P(\widetilde{\boldsymbol{w}}) = \begin{cases} -\dfrac{n}{2}\log 2\pi\tau^2 - \dfrac{1}{2\tau^2}\widetilde{\boldsymbol{w}}^T\widetilde{\boldsymbol{w}} & (\boldsymbol{y} = \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{w}}) \\ -\infty & (\boldsymbol{y} \neq \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{w}}) \end{cases}$$

この関数の最大化は、$\boldsymbol{y} = \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{w}}$ のもとでの $\|\widetilde{\boldsymbol{w}}\|^2$ の最小化と等価

つまり、カーネル回帰は上記の仮定の元での MAP 推定とみなせる

Ridge 回帰・カーネル Ridge 回帰と MAP 推定

このページの話は、Ridge 回帰とカーネル Ridge 回帰で共通
（チルダがついているかどうかの違い）

仮定

- パラメタの事前分布：$\boldsymbol{w} \sim N(\boldsymbol{0}, \tau^2 \boldsymbol{I})$
- $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w} + \boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$

$$\begin{aligned}\log P(\mathcal{D}|\boldsymbol{w}) + \log P(\boldsymbol{w}) &= -\frac{D}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w})^T(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w}) \\ &\quad -\frac{n}{2}\log 2\pi\tau^2 - \frac{1}{2\tau^2}\boldsymbol{w}^T\boldsymbol{w} \\ &= -\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w})^T(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w}) - \frac{1}{2\tau^2}\boldsymbol{w}^T\boldsymbol{w} + \text{定数}\end{aligned}$$

この関数の最大化は、λ を適当に定めた $\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w}\|^2 + \lambda\|\boldsymbol{w}\|^2$ の最小化と等価

つまり、（カーネル）Ridge 回帰は上記の仮定の元での MAP 推定とみなせる

ガウス過程回帰

訓練データ：$\{(\boldsymbol{x}_d, y_d)\}_{d\in[D]}$

- y_d の平均は 0 となるように定数を引いておく。

このデータで学習を行い、新しい $\boldsymbol{x}$ に対する y の「分布」を構築したい

仮定：

- $\boldsymbol{x}$ によって決まる y はガウス過程である

ガウス過程

$y \in \mathbb{R}$ は $\boldsymbol{x} \in \mathbb{R}^n$ によって定まる確率変数とする。
任意に有限個の $x_1, x_2, \cdots, x_k$ を選んだとき、$(y_1, y_2, \cdots, y_k)$ が多変量正規分布に従うとき、ガウス過程という。

ガウス過程回帰

ガウス過程回帰

データ $\{\boldsymbol{x}_d, y_d\}_{d\in[D]}$ と任意の $\boldsymbol{x}$ に対する y において、次の関係が成り立つとする。

$$\begin{pmatrix}\boldsymbol{y}\\ y\end{pmatrix} \sim N\left(\begin{pmatrix}\boldsymbol{0}\\ 0\end{pmatrix}, \begin{pmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \Sigma_{22}\end{pmatrix}\right)$$

$\boldsymbol{y}, \boldsymbol{0} \in \mathbb{R}^D,\ y, 0 \in \mathbb{R},\ \boldsymbol{\Sigma}_{11} \in \mathbb{R}^{D\times D},\ \boldsymbol{\Sigma}_{12} \in \mathbb{R}^{D\times 1},\ \boldsymbol{\Sigma}_{21} \in \mathbb{R}^{1\times D},\ \Sigma_{22} \in \mathbb{R}$

$\boldsymbol{y}$ が観測された状態での y の条件付き分布は次のようになる。

$$y|\boldsymbol{y} \sim N(\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{y}, \Sigma_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})$$

これを、$\boldsymbol{x}$ における y の予測分布とする。

分散・共分散行列

論点：

- $\{\boldsymbol{x}_d\}_{d\in[D]}$ から分散・共分散行列 $\boldsymbol{\Sigma}$ をどう作るのか
- $\boldsymbol{x}$ が「近い」とき共分散が大きい（より関係している）

カーネル関数を使う（半正定値性が保証されることに注意）。

$$\Sigma_{ij} := \mathrm{Ker}(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

カーネルの例

- $\mathrm{Ker}(\boldsymbol{x}_a, \boldsymbol{x}_b) := \boldsymbol{x}_a^T \boldsymbol{x}_b$ 線形カーネル
- $\mathrm{Ker}(\boldsymbol{x}_a, \boldsymbol{x}_b) := (\boldsymbol{x}_a^T \boldsymbol{x}_b + 1)^d \quad (d \in \mathbb{N})$ 多項式カーネル
- $\mathrm{Ker}(\boldsymbol{x}_a, \boldsymbol{x}_b) := \exp\{-\gamma \|\boldsymbol{x}_a - \boldsymbol{x}_b\|^2\}$ RBF カーネル、ガウシアンカーネル

- 線形カーネルだと自由度が足りず $\boldsymbol{y}$ を再現できない（$\boldsymbol{\Sigma}_{11}$ が非正則）
- 多くの場合 RBF カーネルを使う

計算手順

データ： $\{(\boldsymbol{x}_d, y_d)\}_{d\in[D]}$
予測したい点： $\boldsymbol{x}$

$\boldsymbol{K} \in \mathbb{R}^{D\times D}$, $\boldsymbol{k} \in \mathbb{R}^D$, $k \in \mathbb{R}$ を次のように定義する。

$$
\begin{aligned}
K_{ij} &:= \mathrm{Ker}(\boldsymbol{x}_i, \boldsymbol{x}_j) && (i, j \in [D]) \\
k_i &:= \mathrm{Ker}(\boldsymbol{x}_i, \boldsymbol{x}) && (i \in [D]) \\
k &:= \mathrm{Ker}(\boldsymbol{x}, \boldsymbol{x})
\end{aligned}
$$

多変量正規分布の分散・共分散行列は次のようになる。

$$
\begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \Sigma_{22} \end{pmatrix} = \begin{pmatrix} \boldsymbol{K} & \boldsymbol{k} \\ \boldsymbol{k}^T & k \end{pmatrix}
$$

多変量正規分布の条件付き確率の式より、データ $\boldsymbol{y}$ のもとでの y の分布の予測は

$$
y|\boldsymbol{y} \sim N(\boldsymbol{k}\boldsymbol{K}^{-1}\boldsymbol{y}, k - \boldsymbol{k}^T\boldsymbol{K}^{-1}\boldsymbol{k})
$$

となる。

ガウス過程回帰の意味づけ

$\boldsymbol{\phi}(\boldsymbol{x}) \in V$ とする。

モデル

- $\widetilde{w}_i \sim N(0,1) \quad (i \in |V|)$
- $y = \widetilde{\boldsymbol{w}}^T \boldsymbol{\phi}(\boldsymbol{x}), \qquad \boldsymbol{y} = \widetilde{\boldsymbol{X}} \widetilde{\boldsymbol{w}}$

$$\widetilde{\boldsymbol{X}} := \begin{pmatrix} \boldsymbol{\phi}(\boldsymbol{x}_1)^T \\ \boldsymbol{\phi}(\boldsymbol{x}_2)^T \\ \vdots \\ \boldsymbol{\phi}(\boldsymbol{x}_D)^T \end{pmatrix} \in \mathbb{R}^{D \times |V|},$$

$$\Longrightarrow \quad \boldsymbol{y} \sim N(\boldsymbol{0}, \widetilde{\boldsymbol{X}} \widetilde{\boldsymbol{X}}^T)$$

よって、

$$\boldsymbol{\Sigma} = \widetilde{\boldsymbol{X}} \widetilde{\boldsymbol{X}}^T \in \mathbb{R}^{D \times D}$$

$$\Sigma_{ij} := \boldsymbol{\phi}(\boldsymbol{x}_i)^T \boldsymbol{\phi}(\boldsymbol{x}_j) = \mathrm{Ker}(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

誤差を含むモデル

モデル

- $\widetilde{w}_i \sim N(0,1) \quad (i \in |V|), \quad \epsilon \sim N(0, \sigma^2),$
- $y = \widetilde{\boldsymbol{w}}^T \boldsymbol{\phi}(\boldsymbol{x}) + \epsilon \ , \qquad \boldsymbol{y} = \widetilde{\boldsymbol{X}} \widetilde{\boldsymbol{w}} + \boldsymbol{\epsilon}$

$$\widetilde{\boldsymbol{X}} := \begin{pmatrix} \boldsymbol{\phi}(\boldsymbol{x}_1)^T \\ \boldsymbol{\phi}(\boldsymbol{x}_2)^T \\ \vdots \\ \boldsymbol{\phi}(\boldsymbol{x}_D)^T \end{pmatrix} \in \mathbb{R}^{D \times |V|}, \quad \boldsymbol{\epsilon} := \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_D \end{pmatrix} \in \mathbb{R}^D,$$

$$\Longrightarrow \quad \boldsymbol{y} \sim N(\boldsymbol{0}, \widetilde{\boldsymbol{X}} \widetilde{\boldsymbol{X}}^T + \sigma^2 \boldsymbol{I})$$

よって、

$$\begin{aligned} \boldsymbol{\Sigma} &= \widetilde{\boldsymbol{X}} \widetilde{\boldsymbol{X}}^T + \sigma^2 \boldsymbol{I} \\ \Sigma_{ij} &= \boldsymbol{\phi}(\boldsymbol{x}_i)^T \boldsymbol{\phi}(\boldsymbol{x}_j) + \sigma^2 \delta_{ij} \\ &= \mathrm{Ker}(\boldsymbol{x}_i, \boldsymbol{x}_j) + \sigma^2 \delta_{ij} \end{aligned}$$

ベイズ推定１

ガウス過程回帰は、パラメータ w をベイズ推定したうえで、新しい $\boldsymbol{x}$ に対する y の分布を求めていると捉えることができる。

誤差付きの場合で説明：

まず、前述した条件付き確率の計算手順を $\tilde{\boldsymbol{X}}$ を使って書き直す。
（$\phi(\boldsymbol{x})$ を $\widetilde{\boldsymbol{x}}$ と表記する）

$$\begin{pmatrix} \boldsymbol{y} \\ y \end{pmatrix} \sim N\left(\begin{pmatrix} \boldsymbol{0} \\ 0 \end{pmatrix}, \begin{pmatrix} \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^T + \sigma^2 \boldsymbol{I} & \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{x}} \\ \widetilde{\boldsymbol{x}}^T\widetilde{\boldsymbol{X}}^T & \widetilde{\boldsymbol{x}}^T\widetilde{\boldsymbol{x}} + \sigma^2 \end{pmatrix} \right)$$

このとき、多変量正規分布の条件付き分布は、

$$\begin{aligned} y|\boldsymbol{y} &\sim N(\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{y}, \Sigma_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}) \\ &= N(\widetilde{\boldsymbol{x}}^T\widetilde{\boldsymbol{X}}^T(\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^T + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{y},\ \widetilde{\boldsymbol{x}}^T\widetilde{\boldsymbol{x}} + \sigma^2 - \widetilde{\boldsymbol{x}}^T\widetilde{\boldsymbol{X}}^T(\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^T + \sigma^2\boldsymbol{I})^{-1}\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{x}}) \end{aligned}$$

である。

ベイズ推定２

一方、ベイズ推定より、パラメタ $\boldsymbol{w}$ の事後分布は次のようになる。

$$\widetilde{\boldsymbol{w}}|\boldsymbol{y} \sim N((\widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{X}} + \sigma^2\boldsymbol{I})^{-1}\widetilde{\boldsymbol{X}}^T\boldsymbol{y},\ \sigma^2(\widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{X}} + \sigma^2\boldsymbol{I})^{-1})$$

このとき、y の確率は

$$y|\boldsymbol{y} \sim N(\widetilde{\boldsymbol{x}}^T(\widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{X}} + \sigma^2\boldsymbol{I})^{-1}\widetilde{\boldsymbol{X}}^T\boldsymbol{y},\ \sigma^2\widetilde{\boldsymbol{x}}^T(\widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{X}} + \sigma^2\boldsymbol{I})^{-1}\widetilde{\boldsymbol{x}} + \sigma^2)$$

この２つの正規分布は平均も分散も等しいので、ガウス過程回帰のパラメタをベイズ推定しているといえる。

平均：

$$\widetilde{\boldsymbol{x}}^T\widetilde{\boldsymbol{X}}^T(\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^T + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{y} = \widetilde{\boldsymbol{x}}^T(\widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{X}} + \sigma^2\boldsymbol{I})^{-1}\widetilde{\boldsymbol{X}}^T\boldsymbol{y}$$

分散：

$$\widetilde{\boldsymbol{x}}^T\widetilde{\boldsymbol{x}} + \sigma^2 - \widetilde{\boldsymbol{x}}^T\widetilde{\boldsymbol{X}}^T(\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^T + \sigma^2\boldsymbol{I})^{-1}\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{x}} = \sigma^2\widetilde{\boldsymbol{x}}^T(\widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{X}} + \sigma^2\boldsymbol{I})^{-1}\widetilde{\boldsymbol{x}} + \sigma^2$$

式変形

平均：
すでに、 $\widetilde{\boldsymbol{X}}^T(\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^T + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{y} = (\widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{X}} + \sigma^2\boldsymbol{I})^{-1}\widetilde{\boldsymbol{X}}^T\boldsymbol{y}$ は説明しているので等式は成り立つ。

分散：
Sherman–Morrison–Woodbury の公式より、

$$(\widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{X}} + \sigma^2\boldsymbol{I})^{-1} = \frac{1}{\sigma^2}\boldsymbol{I} - \frac{1}{\sigma^2}\widetilde{\boldsymbol{X}}^T\left(\sigma^2\boldsymbol{I} + \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^T\right)^{-1}\widetilde{\boldsymbol{X}}$$

よって、

$$\begin{aligned}
&\sigma^2\widetilde{\boldsymbol{x}}^T(\widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{X}} + \sigma^2\boldsymbol{I})^{-1}\widetilde{\boldsymbol{x}} + \sigma^2 \\
&= \sigma^2\widetilde{\boldsymbol{x}}^T\left(\frac{1}{\sigma^2}\boldsymbol{I} - \frac{1}{\sigma^2}\widetilde{\boldsymbol{X}}^T\left(\sigma^2\boldsymbol{I} + \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^T\right)^{-1}\widetilde{\boldsymbol{X}}\right)\widetilde{\boldsymbol{x}} + \sigma^2 \\
&= \sigma^2\widetilde{\boldsymbol{x}}^T\widetilde{\boldsymbol{x}} - \widetilde{\boldsymbol{x}}^T\widetilde{\boldsymbol{X}}^T\left(\sigma^2\boldsymbol{I} + \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^T\right)^{-1}\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{x}} + \sigma^2
\end{aligned}$$

MAP 推定とベイズ推定

カーネル回帰・カーネル Ridge 回帰とガウス過程回帰の違いは、パラメタを MAP 推定しているかベイズ推定しているかの違いと捉えることができる

	カーネル	$\boldsymbol{w}$	誤差	推定
線形回帰・カーネル回帰	有・無	確定的	あり	最尤推定 [1]
カーネル回帰	有	確率的	なし	MAP 推定
（カーネル）Ridge 回帰	有・無	確率的	あり	MAP 推定
ガウス過程回帰	有	確率的	なし	ベイズ推定
ガウス過程回帰＋誤差	有・無	確率的	あり	ベイズ推定

※ $n < D < |V|$ を想定

[1] カーネル有りのときはノルム最小化も行う

参考：多変量正規分布

$N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$：平均ベクトル $\boldsymbol{\mu}$、分散共分散行列 $\boldsymbol{\Sigma}$ の正規分布

多変量正規分布

$\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ のとき、確率密度関数は

$$f(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) := \frac{1}{(2\pi)^{n/2}} \frac{1}{\det(\boldsymbol{\Sigma})^{1/2}} \exp\left\{ -\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \right\}$$

$\boldsymbol{x} \sim N(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I})$ のとき、確率密度関数は

$$f(\boldsymbol{x}; \boldsymbol{\mu}, \sigma^2 \boldsymbol{I}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{ -\frac{1}{2\sigma^2} (\boldsymbol{x} - \boldsymbol{\mu})^T (\boldsymbol{x} - \boldsymbol{\mu}) \right\}$$

参考：多変量正規分布

条件付き確率

$\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ のとき、

$$\boldsymbol{x} := \begin{pmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{pmatrix}, \quad \boldsymbol{\mu} := \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \quad \boldsymbol{\Sigma} := \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}$$

$\boldsymbol{x}_1$ のもとでの $\boldsymbol{x}_2$ の確率分布：

$$\boldsymbol{x}_2 | \boldsymbol{x}_1 \sim N(\boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{x}_1 - \boldsymbol{\mu}_1),\ \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})$$

ベイズ推定

$$\boldsymbol{w} \sim N(\boldsymbol{0}, \boldsymbol{I}), \quad \boldsymbol{y}|\boldsymbol{w} \sim N(\boldsymbol{X}\boldsymbol{w}, \sigma^2\boldsymbol{I}) \quad \text{のとき、}$$

$$\boldsymbol{y} \sim N(\boldsymbol{0}, \boldsymbol{X}\boldsymbol{X}^T + \sigma^2\boldsymbol{I})$$

$$\boldsymbol{w}|\boldsymbol{y} \sim N((\boldsymbol{X}^T\boldsymbol{X} + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X} + \sigma^2\boldsymbol{I})^{-1})$$

参考：多変量正規分布

ベイズ推定

N は平均ベクトルと共分散行列と引数にとる正規分布

$\boldsymbol{\Sigma}$：分散・共分散行列、$\boldsymbol{\Lambda}$：精度行列

$$\boldsymbol{w} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}_1)$$

$$\boldsymbol{y}|\boldsymbol{w} \sim N(\boldsymbol{X}\boldsymbol{w} + \boldsymbol{b}, \boldsymbol{\Sigma}_2) \quad \text{のとき、}$$

$$\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\mu} + \boldsymbol{b}, \boldsymbol{X}\boldsymbol{\Sigma}_1\boldsymbol{X}^T + \boldsymbol{\Sigma}_2)$$

$$\boldsymbol{w}|\boldsymbol{y} \sim N((\boldsymbol{\Lambda}_1 + \boldsymbol{X}^T\boldsymbol{\Lambda}_2\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{\Lambda}_2(\boldsymbol{y} - \boldsymbol{b}) + \boldsymbol{\Lambda}_1\boldsymbol{\mu}), (\boldsymbol{\Lambda}_1 + \boldsymbol{X}^T\boldsymbol{\Lambda}_2\boldsymbol{X})^{-1})$$

ただし、 $\boldsymbol{\Lambda}_1 := \boldsymbol{\Sigma}_1^{-1},\ \boldsymbol{\Lambda}_2 := \boldsymbol{\Sigma}_2^{-1}$