ページ末に App. を載せています。今回レジュメは内容の理解不足のため出来が悪いです。先に謝っておきます。

前回はカーネル関数の定義と構成方法について学んだ。カーネル関数は最も簡単な形では特徴量の基底関数ベクトルの内積として書くことができて、ここからさらに（6.13）〜（6.22）で整理されたフレームワークを用いることで拡張された表現を得ることができたのであった。我々は今しばらくカーネル関数を用いたモデル表現を取り扱うことにする。

今回やることは以下の 3 ポイント。

・基底関数を特定の形 RBF[*1]で与えることによって入力に対する予測関数（回帰モデル）を構成する手法（Nadaraya-Watson モデル）を学ぶ。

・Parzen 推定法を用いることで Nadaraya-Watson モデルを導出する。

・カーネル関数を用いた形でガウス過程を表現する[*2]。

## 1　RBF ネットワーク

これまでも[*3]回帰モデルを基底関数の線形結合の形で導出した。ここでは基底関数を RBF に定めたうえで、回帰モデルがどのような形で表せるのか考えたい。

ここで RBF について簡単に説明しておく。基底関数が、その中心 $\boldsymbol{\mu}_j$ からの動径のみに依存した形式 $\phi(\boldsymbol{r}) = h(|\boldsymbol{r} - \boldsymbol{\mu}_j|)$ で表されるとき、この基底関数を RBF と呼ぶ[*4]。この RBF にパラメタの重みをつけて、線形結合させてやることで、入力と出力を対応させる関数を次のように表す[*5]ことができる。

$$f(\boldsymbol{x}) = \sum_{n=1}^{N} w_n h(|\boldsymbol{x} - \boldsymbol{x}_n|) \tag{1}$$

$\boldsymbol{x}_n$ は入力ベクトル、$t_n$ は目的変数ベクトルの第 $n$ 成分である。

過学習についての議論は自信がないのだが、以下自分の理解した内容を記す。どうも、ここでの議論の前提として、入力出力においてはノイズが入り込んでいない場合について考えていた模様？一般的には出力、つまり目標変数にはノイズが入っているものである。式（6.38）のモデルで学習を行うことはノイズも拾い上げることになり、それは過学習につながってしまうよという話をしていると思われる。

RBF と正則化理論との関係はよくわからないため、教科書の説明をすることは諦めた。ここではオマケのトピックスとしてグリーン関数が挙げられているため、これについて紹介を行う[*6]。

本題に戻ろう。ここからは入力変数にノイズが含まれる場合の学習について考える。入力変数 $\boldsymbol{x}$ に含まれるノイズが確率分布 $\nu(\boldsymbol{\xi})$ に従う確率変数 $\boldsymbol{\xi}$ によって表されるとき、二乗和誤差関数は次のように期待値を

[*1] Radial Basis Function；動径基底関数のこと。

[*2] もう少しナイーブに言うとガウス分布の平均や共分散をカーネル関数で表すということです。

[*3] 例えば上巻第三章を見よ。とはいうものの既に線形結合で表されたモデルは見飽きただろうが。

[*4] この RBF は物理において非常によく出てくる。それは物理モデルの多くが動径成分のみに依存した形式で表現されるからだ。例えば電磁気学における静電ポテンシャル、力学における万有引力ポテンシャルなどのいわゆる中心力ポテンシャルと呼ばれるものたちが、この形式に当てはまっている。電磁気学の例はこの後、グリーン関数の説明とともに紹介したい。動径成分のみ取り出して考えられるのは便利だよね、という話になる。

[*5] この手法を「関数補間」と呼ぶらしい。本レジュメでは「学習」と一括りにして呼んでしまうことにする。両者とも概念的には異なるものではないと理解している。

[*6] https://www.phys.chuo-u.ac.jp/labs/nakano/suurikaiseki/sec4(suuri).pdf 参照

取った形で表される[*7]。

$$E = \frac{1}{2}\sum_{n=1}^{N}\int \{y(\boldsymbol{x}_n+\boldsymbol{\xi})-t_n\}^2 \nu(\boldsymbol{\xi})d\boldsymbol{\xi} \tag{2}$$

この $E$ について $\boldsymbol{\xi}$ に関する変分問題を考える（対応；演習 6.17）。形式的には式 2 を $y$ で偏微分してゼロをとるだけである。ただし準備として次の変数変換を施しておく。

$$\boldsymbol{x}_n + \boldsymbol{\xi} = \boldsymbol{x} \tag{3}$$

これによって $E$ は次の形になる。

$$E = \frac{1}{2}\sum_{n=1}^{N}\int (y(\boldsymbol{x})-t_n)^2 \nu(\boldsymbol{x}-\boldsymbol{x}_n)d\boldsymbol{x} \tag{4}$$

計算は以下の通り。

$$\begin{aligned}\frac{\delta E}{\delta y} &= \frac{1}{2}\sum_{n=1}^{N}\int 2(y(\boldsymbol{x})-t_n)\nu(\boldsymbol{x}-\boldsymbol{x}_n)d\boldsymbol{x} \\ &= \int \sum_{n=1}^{N}(y(\boldsymbol{x})-t_n)\nu(\boldsymbol{x}-\boldsymbol{x}_n)d\boldsymbol{x} \end{aligned} \tag{5}$$

$\frac{\delta E}{\delta y}=0$ が恒等的に成立するためには

$$\begin{aligned}\sum_{n=1}^{N}(y(\boldsymbol{x})-t_n)\nu(\boldsymbol{x}-\boldsymbol{x}_n) &= 0 \\ \sum_{n=1}^{N}y(\boldsymbol{x})\nu(\boldsymbol{x}-\boldsymbol{x}_n) - \sum_{n=1}^{N}t_n\nu(\boldsymbol{x}-\boldsymbol{x}_n) &= 0 \\ y(\boldsymbol{x})\sum_{n=1}^{N}\nu(\boldsymbol{x}-\boldsymbol{x}_n) &= \sum_{n=1}^{N}t_n\nu(\boldsymbol{x}-\boldsymbol{x}_n) \\ y(\boldsymbol{x}) &= \frac{\sum_{n=1}^{N}t_n\nu(\boldsymbol{x}-\boldsymbol{x}_n)}{\sum_{n=1}^{N}\nu(\boldsymbol{x}-\boldsymbol{x}_n)} \end{aligned} \tag{6}$$

したがって予測関数は次のように表すことができる。

$$y(\boldsymbol{x}) = \sum_{n=1}^{N} t_n h(\boldsymbol{x}-\boldsymbol{x}_n) \tag{7}$$

ここで RBF として $h(\boldsymbol{x}-\boldsymbol{x}_n)$ を以下で定義する。

$$h(\boldsymbol{x}-\boldsymbol{x}_n) = \frac{\nu(\boldsymbol{x}-\boldsymbol{x}_n)}{\sum_{n=1}^{N}\nu(\boldsymbol{x}-\boldsymbol{x}_n)} \tag{8}$$

これは Nadaraya-Watson モデルとして知られている。ノイズ分布 $\nu(\boldsymbol{\xi})$ が等方的（Radial）であるならば、$|\boldsymbol{\xi}|$ のみに依存した関数になるので、基底関数 $h(\boldsymbol{x}-\boldsymbol{x}_n)$ も Radial となる。

[*7] なぜ積分による期待値で表したのか理由が明確にわかっていない。「二乗和の期待損失」と呼ばれるのならば、多少はピンと来るのだが...

基底関数が正規化されていることは式 8 の両辺において $n$ に関する総和を取ることで自明であると言える。一方でもし基底関数が正規化されていない場合、2 つ問題点が生じることになる。これは全ての基底関数の値が小さな値を持ってしまうような入力空間の領域が存在するとき

・その領域内における予測値の値も小さくなる

・予測値がほとんどバイアスパラメタによって定まってしまう。

というものである。正規化はこの問題を解決してくれる（See. 頁 11 図 6.2)。

## 1.1 Nadaraya-Watson モデル

先のセクションでは RBF を基底関数として設定することで Nadaraya-Watson モデルと呼ばれる回帰モデルを導出した。本サブセクションも Nadaraya-Watson モデルを導出するが、ここでは Parzen 推定法を用いてカーネル関数を導入した形式でアプローチを行う。

ここで Parzen 推定法[*8]について復習する。Parzen 推定において、あるデータ $\boldsymbol{x}$ での推定密度の表式は

$$p(\boldsymbol{x}) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{h^D}k(\frac{\boldsymbol{x}-\boldsymbol{x}_n}{h}) \tag{9}$$

であった。$k(\frac{\boldsymbol{x}-\boldsymbol{x}_n}{h})$ は $\boldsymbol{x}$ を中心とする一辺が $h$ の立方体の中に、データ点 $\boldsymbol{x}_n$ があれば 1、なければ 0 をとるカーネル関数となっている。このカーネル関数のことを Parzen 窓と呼ぶのであった。そして、Parzen 窓を用いた確率密度の推定方法は Parzen 推定として知られている。この手法の利点は確率密度を推定する際に、関数領域の分割が入力データに対して適応的に行えることにある[*9]。

今回は同時分布 $p(\boldsymbol{x},t)$ を推定するため、次のような Parzen 推定の式を用いる。

$$p(\boldsymbol{x},t) = \frac{1}{N}\sum_{n=1}^{N}f(\boldsymbol{x}-\boldsymbol{x}_n, t-t_n) \tag{10}$$

今 Parzen 窓が $f(\boldsymbol{x}-\boldsymbol{x}_n, t-t_n)$ の部分に相当している。

ここから予測値、すなわち回帰関数 $y(\boldsymbol{x})$ を求めるためには、入力データ $\boldsymbol{x}$ で条件つけられた目標値 $t$ の条件付き期待値 $E(t|\boldsymbol{x})$ を計算してやれば良い。この計算は初等的であるため、詳細については割愛。

$$\begin{aligned} y(\boldsymbol{x}) &= E(t|\boldsymbol{x}) = \int tp(t|\boldsymbol{x})dt \\ &= \ldots \\ &= \frac{\sum_n \int tf(\boldsymbol{x}-\boldsymbol{x}_n, t-t_n)dt}{\sum_m \int f(\boldsymbol{x}-\boldsymbol{x}_m, t-t_m)dt} \end{aligned} \tag{11}$$

以下では、簡単のため $f(\boldsymbol{x},t)$ の $t$ についての平均はゼロとする。すなわち

$$\int f(\boldsymbol{x},t)tdt = 0 \tag{12}$$

とする。ここから次のような変数変換を施す。

$$t-t_n = s,\ t-t_m = s' \tag{13}$$

[*8] 上巻第二章；頁 119〜122

[*9] ヒストグラム法との対比を見よ。事前に送付した「Parzen 推定補足資料」を参照されたい。

このもとでさらに式変形を進めると

$$
\begin{aligned}
y(\boldsymbol{x}) &= \frac{\sum_n \int (s+t_n) f(\boldsymbol{x}-\boldsymbol{x}_n, s) ds}{\sum_m \int f(\boldsymbol{x}-\boldsymbol{x}_m, s') ds'} \\
&= \frac{\sum_n (\int f(\boldsymbol{x}-\boldsymbol{x}_n, s) s ds + \int f(\boldsymbol{x}-\boldsymbol{x}_n, s) ds \; t_n)}{\sum_m \int f(\boldsymbol{x}-\boldsymbol{x}_m, s') ds'} \\
&= \frac{\sum_n \int f(\boldsymbol{x}-\boldsymbol{x}_n, s) ds \; t_n}{\sum_m \int f(\boldsymbol{x}-\boldsymbol{x}_m, s') ds'} \quad (\because \text{式 } 12) \\
&= \frac{\sum_n g(\boldsymbol{x}-\boldsymbol{x}_n) t_n}{g(\boldsymbol{x}-\boldsymbol{x}_m)} \\
&= \sum_{n=1} k(\boldsymbol{x}, \boldsymbol{x}_n) t_n
\end{aligned}
\tag{14}
$$

が導かれる。ここで、カーネル関数 $k(\boldsymbol{x}, \boldsymbol{x}_n)$ は以下の形で与えられる。

$$
k(\boldsymbol{x}, \boldsymbol{x}_n) = \frac{g(\boldsymbol{x}-\boldsymbol{x}_n)}{\sum_m g(\boldsymbol{x}-\boldsymbol{x}_m)} \tag{15}
$$

式 14 は Nadaraya-Watson モデル、あるいはカーネル回帰と呼ばれる。

このモデルは次のように条件付き確率分布も定義することができて

$$
p(t|\boldsymbol{x}) = \frac{p(t, \boldsymbol{x})}{\int p(t, \boldsymbol{x}) dt} = \frac{\sum_n f(\boldsymbol{x}-\boldsymbol{x}_n, t-t_n)}{\sum_m \int f(\boldsymbol{x}-\boldsymbol{x}_m, t-t_n) dt} \tag{16}
$$

ここから条件付き期待値や分散を求めることができる。そこで、簡単な例についてカーネル関数を用いた形で表式を得ることを考える（対応；演習 6.18）。

例の設定として、共分散行列が $\sigma^2 \boldsymbol{I}$ で与えられるようなガウス基底を持つ Nadaraya-Watson モデルを考える。入力変数 $x$ と目標変数 $t$ は、どちらも一次元であるとする。このとき次の三つ：条件付き確率密度 $p(t|x)$、条件付き期待値 $E(t|x)$、条件付き分散 $var(t|x)$ について、それぞれカーネル関数 $k(x, x_n)$ を用いた表現を得たい（手書き App. 参照）。

# 2 ガウス過程

このセクション[*10]では回帰・分類モデルを予測値がガウス過程に従う場合について求める。これらのモデルの予測値について、期待値や分散がカーネル関数を用いた形で表せることを見ていく[*11]。

## 2.1 線形回帰再訪

ここでは線形回帰の例で予測関数 $y(\boldsymbol{x}, \boldsymbol{w})$ を再導出し、さらにそれらを要素とする $\boldsymbol{y}$ がガウス過程に従うことを確かめる。そこで、この $\boldsymbol{y}$ に関して期待値や共分散をカーネル関数を用いた形で表現してやることを考える。これによって入力 $\boldsymbol{x}$ に対する予測関数 $\boldsymbol{y}$ を評価することができる。この評価は $\boldsymbol{y}$ の確率分布を求めることに他ならない。

[*10] セクション 6.3 との関連は見出せなかった。強いていうならカーネル関数を用いて期待値や分散を表現しようとするところは共通点と言えるだろうか。

[*11] このセクションについて、もっと本当は大局的な視点で整理したい気持ちはあります。とはいうものの、後半のサブセクションを見ることができず、クリアなストーリーはまだ描けていない状態です。

まずガウス過程が何かということを確認しよう。ガウス過程の定義は「関数 $y(\boldsymbol{x})$ 上の確率分布として定義され、任意の点集合 $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ に対する $y(\boldsymbol{x})$ の値の同時分布がガウス分布に従うこと[*12]」とされている[*13]。

$\boldsymbol{y}$ の 1 成分である $y(\boldsymbol{x}_1)$ は $\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_1)$ であり全成分で表示すると次のようになる（手書き App. 参照）。

ここから $\boldsymbol{y}$ の期待値はゼロ、共分散はグラム行列 $\boldsymbol{K}$ で表されて、その行列の要素は $K_{nm}$ は

$$K_{nm} = k(\boldsymbol{x}_n, \boldsymbol{x}_m) = \frac{1}{\alpha}\boldsymbol{\phi}(\boldsymbol{x}_m)^T\boldsymbol{\phi}(\boldsymbol{x}_m) \tag{17}$$

となることは p.16 上部の式変形より明らかである。

こうして $\boldsymbol{y}$ の分布については完全な情報を得ることができた。この情報はのちに真の目標値ベクトル $\boldsymbol{t} = (t_1, \ldots, t_N)^T$ の分布を得るために用いられることになる。もう少し具体的に説明しておくと、目標値ベクトルの分布を周辺化積分によって求めるときに、$\boldsymbol{y}$ で周辺化を行うことになるのである。この $\boldsymbol{y}$ の分布に関する情報がわかっている前提で公式 (2.115)[*14]を用いるという手続きを辿ることになるが、それは次のサブセクションに譲ることにしよう[*15]。

[*12] もっと端的に言うと $y(\boldsymbol{x}_1), \cdots, y(\boldsymbol{x}_N)$ がいずれもガウス分布に従うということだと理解している。
[*13] p.16 中段参照。
[*14] 上巻第二章：頁 90 下部
[*15] 式（6.61）の導出のことを言っている。

App.

[条件付 p.d.f.、期待値、分散をカーネル関数で表示]

入力 $x$ が1次元のとき (6.48) より

$$p(t|x) = \frac{\sum_n f(x-x_n, t-t_n)}{\sum_m \int f(x-x_m, t-t_m)\,dt}$$

見やすさのため $z = (x, t)$ と約束しておく。

$$p(t|x) = \frac{\sum_n f(z-z_n)}{\sum_m \int f(z-z_m)\,dt}$$

$$= \frac{\sum_n N(z|z_n, \sigma^2 I)}{\sum_m \int N(z|z_m, \sigma^2 I)\,dt}$$

($\because$ $f(z)$ は平均0, 分散 $\sigma^2 I$ のガウス分布

$f(z-z_m)$ は平均 $z_m$, 分散 $\sigma^2 I$ のガウス分布)

$$= \frac{\sum_n \frac{1}{2\pi\sigma^2} \exp\left(-\frac{1}{2}(z-z_n)^T(\sigma^2 I)^{-1}(z-z_n)\right)}{\sum_m N(x|x_m, \sigma^2)}$$

($\because$ $t$ で周辺化)

$$= \frac{\sum_n \frac{1}{2\pi\sigma^2} \exp\left(-\frac{1}{2}\boxed{(z-z_n)^T(\sigma^2 I)^{-1}(z-z_n)}\right)}{\sum_m \left(\frac{1}{2\pi\sigma^2}\right)^{1/2} \exp\left(-\frac{(x-x_m)^2}{2\sigma^2}\right)}$$

$$\rightarrow (x-x_n, t-t_n)\frac{1}{\sigma^2}I\begin{pmatrix} x-x_n \\ t-t_n \end{pmatrix}$$

$$= \left(\frac{1}{\sigma^2}\right)\{(x-x_n)^2 + (t-t_n)^2\}$$

$$= \frac{(x-x_n)^2}{\sigma^2} + \frac{(t-t_n)^2}{\sigma^2}$$

$$= \sum_n \left(\frac{1}{2\pi\sigma^2}\right)^{1/2} \frac{\exp\left[-\frac{1}{2}\left\{\frac{(x-x_n)^2}{\sigma^2} + \frac{(t-t_n)^2}{\sigma^2}\right\}\right]}{\sum_m \exp\left(-\frac{(x-x_m)^2}{2\sigma^2}\right)}$$

$$= \sum_n \frac{\exp\left(-\frac{(x-x_n)^2}{2\sigma^2}\right)}{\sum_m \exp\left(-\frac{(x-x_m)^2}{2\sigma^2}\right)} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(t-t_n)^2}{2\sigma^2}\right)$$

$$= \sum_n k(x, x_n) N(t|t_n, \sigma^2)$$

$$E(t|x) = \int t\, p(t|x)\,dt = \int t \sum_n k(x, x_n) N(t|t_n, \sigma^2)\,dt = \sum_n k(x, x_n) t_n$$

$$\mathrm{Var}(t|x) = E(t^2|x) - E^2(t|x)$$

これは自明

$$E(t^2|x) = \int t^2 p(t|x)\, dt$$

$$= \sum_n k(x, x_n) \int t^2 \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(t-t_n)^2}{2\sigma^2}\right) dt$$

$$= \int (\sigma s + t_n)^2 \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}s^2\right) \cdot \sigma\, ds \quad \left(\because s = \frac{t - t_n}{\sigma}\right)$$

$$= \frac{1}{\sqrt{2\pi}} \int (\sigma^2 s^2 + t_n^2) \exp\left(-\frac{1}{2}s^2\right) ds \quad (\because \text{odd function の性質})$$

$$= \frac{1}{\sqrt{2\pi}} \int \sigma^2 s^2 \exp\left(-\frac{1}{2}s^2\right) ds + \frac{1}{\sqrt{2\pi}} \int t_n^2 \exp\left(-\frac{1}{2}s^2\right) ds$$

$$= \frac{\sigma^2}{\sqrt{2\pi}} \cdot \frac{1}{2}\sqrt{\frac{\pi}{\left(\frac{1}{2}\right)^3}} + \frac{1}{\sqrt{2\pi}} \cdot t_n^2 \cdot \sqrt{\frac{\pi}{\frac{1}{2}}} \quad \left(\because \int s^2 e^{-as^2} ds = \frac{1}{2}\sqrt{\frac{\pi}{a^3}}\right)$$

$$= \sigma^2 + t_n^2$$

$$\therefore \mathrm{var}(t|x) = \sum_n k(x, x_n)(\sigma^2 + t_n^2) - \left(\sum_n k(x, x_n)\, t_n\right)^2$$

App.

[ $y(x) \longrightarrow \mathbf{y}$ への拡張 ]

カンタンな式変形であるが念のため確認。

$y(x_1) = \mathbf{w}^T \boldsymbol{\phi}(x_1), \cdots, y(x_N) = \mathbf{w}^T \boldsymbol{\phi}(x_N)$ を成分ごとに縦ベクトルで並べると

$$
\begin{pmatrix} y(x_1) \\ \vdots \\ y(x_N) \end{pmatrix}
= \begin{pmatrix} \mathbf{w}^T \boldsymbol{\phi}(x_1) \\ \vdots \\ \mathbf{w}^T \boldsymbol{\phi}(x_N) \end{pmatrix}
$$

$$
= \begin{pmatrix} (\mathbf{w}^T \boldsymbol{\phi}(x_1))^T \\ \vdots \\ (\mathbf{w}^T \boldsymbol{\phi}(x_N))^T \end{pmatrix} \quad (\because \text{各成分はスカラ})
$$

$$
= \begin{pmatrix} \boldsymbol{\phi}(x_1)^T \mathbf{w} \\ \vdots \\ \boldsymbol{\phi}(x_N)^T \mathbf{w} \end{pmatrix}
$$

$$
= \begin{pmatrix} \boldsymbol{\phi}(x_1)^T \\ \vdots \\ \boldsymbol{\phi}(x_N)^T \end{pmatrix} \mathbf{w}
$$

$$
= \boldsymbol{\Phi} \mathbf{w} \quad \left( \because \boldsymbol{\Phi} = \begin{pmatrix} \boldsymbol{\phi}(x_1)^T \\ \vdots \\ \boldsymbol{\phi}(x_N)^T \end{pmatrix} \right)
$$

p.3 の (6.3) で定義