Kernel VAE: カーネル法を用いた次元削減

内海 佑麻 *[1]

[1] 慶應義塾大学理工学部情報工学科

2019/10/11

Abstract

...

Contents

1 Introduction

...

*uchiumi@ailab.ics.keio.ac.jp

2 ガウス過程

2.1 ガウス過程の定義

入力空間 $\mathcal{X}$ 上の関数 $f:\mathcal{X}\to\mathbb{R}$ がガウス過程 (Gaussian Process, GP) に従うとは，$\mathcal{X}$ 上の任意の n 点 $\{x_i\}_{i=1}^n$ に対して，ベクトル $\mathbf{f}=(f(x_1),\ldots,f(x_n))^{\mathrm{T}}\in\mathbb{R}^n$ が n 次元ガウス分布に従うことをいう．ここで，確率変数 $\mathbf{f}\in\mathbb{R}^n$ が n 次元ガウス分布に従う時，その確率密度関数 $p(\mathbf{f})$ は，平均関数 $m(\cdot)$ と共分散関数 $v(\cdot,\cdot)$ を用いて

$$p(\mathbf{f})=\frac{1}{(2\pi)^{n/2}|V(\mathbf{x})|^{1/2}}\exp\left(-\frac{1}{2}(\mathbf{f}-m(\mathbf{f}))^{\mathrm{T}}V(\mathbf{f})^{-1}(\mathbf{f}-m(\mathbf{f}))\right) \tag{1}$$

と定められる．ただし，$V(\mathbf{f})$ は共分散 $v(\mathbf{f}_i,\mathbf{f}_j)$ を ij 要素にもつ共分散行列である．ゆえに，関数 $f:\mathcal{X}\to\mathbb{R}$ がガウス過程 (Gaussian Process, GP) に従うとき，その挙動は平均関数 $m(\cdot)$ と共分散関数 $v(\cdot,\cdot)$ によって定められ，これを以下のように記述する．

$$f(\cdot)\sim\mathcal{GP}(m(\cdot),v(\cdot,\cdot)) \tag{2}$$

2.2 条件つき分布の計算

一般に，2 つのベクトル $\mathbf{f}_n\in\mathbb{R}^n,\mathbf{f}_m\in\mathbb{R}^m$ に対して，

$$\begin{pmatrix}\mathbf{f}_n\\ \mathbf{f}_m\end{pmatrix}\sim\mathcal{N}\left(\begin{pmatrix}\boldsymbol{\mu}_n\\ \boldsymbol{\mu}_m\end{pmatrix},\begin{bmatrix}\Sigma_{nn} & \Sigma_{nm}\\ \Sigma_{nm}^{\mathrm{T}} & \Sigma_{mm}\end{bmatrix}\right) \tag{3}$$

が成り立つとき，

$$\mathbf{f}_m|\mathbf{f}_n\sim\mathcal{N}\left(\boldsymbol{\mu}_{m|n},\Sigma_{m|n}\right) \tag{4}$$

$$where\ \begin{cases}\boldsymbol{\mu}_{m|n}=\boldsymbol{\mu}_m+\Sigma_{nm}^{\mathrm{T}}\Sigma_{nn}^{-1}(\mathbf{f}_n-\boldsymbol{\mu}_n)\\ \Sigma_{m|n}=\Sigma_{mm}-\Sigma_{nm}^{\mathrm{T}}\Sigma_{nn}^{-1}\Sigma_{nm}\end{cases}$$

2.3 ガウス過程回帰

確率変数 $X\in\mathbb{R}^d, Y\in\mathbb{R}$ の実現値からなる n 個のデータサンプル $\mathcal{D}=\{\mathbf{x}_i,y_i\}_{i=1}^n$ を用いて，X の値から Y の値を推定するモデル $f:X\to Y$ を特定することを回帰問題という．すべての $(\mathbf{x},y)$ に対して，モデルの出力値 $f(\mathbf{x})$ と y との誤差を ε とおき，これが正規分布 $\mathcal{N}(0,\sigma^2)$ に従うと仮定すると回帰モデルは.

$$y=f(\mathbf{x})+\varepsilon,\quad \varepsilon\sim\mathcal{N}(0,\sigma^2) \tag{5}$$

あるいは，正規分布の再生性より，

$$y|\mathbf{x} \sim \mathcal{N}(f(\mathbf{x}), \sigma^2) \tag{6}$$

となる．また一般の回帰問題において，データサンプル $\mathcal{D} = \{X_i, Y_i\}_{i=1}^n$ とモデル $f : X \to Y$ に対して下式が成り立つことから，これはモデル f の分布に関するベイズ推論へ拡張できる．

$$p(f|\mathcal{D}) = \frac{p(\mathcal{D}|f)p(f)}{p(\mathcal{D})}, \quad i.e. \quad p(f|Y, X) = \frac{p(Y|X, f)p(f)}{p(Y|X)} \tag{7}$$

上のモデルにおいて，関数 f がガウス過程に従う場合，これをガウス過程回帰という．たとえば，関数 f に対して，

$$f(\cdot) \sim \mathcal{GP}(m(\cdot), k(\cdot, \cdot)) \tag{8}$$

を仮定すると，$\mathbf{f}_n = (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n))^{\mathrm{T}}$ と $\mathbf{y}_n = (y_1, \ldots, y_n)^{\mathrm{T}}$ に対して，

$$\mathbf{f}_n \sim \mathcal{N}(m(X_n), K(X_n, X_n)) \tag{9}$$

$$\mathbf{y}_n \sim \mathcal{N}(m(X_n), K(X_n, X_n) + \sigma^2 I_n) \tag{10}$$

が成り立つ．ただし，$m(X_n) = (m(\mathbf{x}_1), \ldots, m(\mathbf{x}_n))^{\mathrm{T}}$，$K(X_n, K(X_n))_{ij} = k(f(\mathbf{x}_i), f(\mathbf{x}_j))$，$I_n$ は $n \times n$ の単位行列とする．式 (8) は，式 (7) でモデル f の事前分布 $p(f)$ を定めることに対応する．さらに，式 (6) と正規分布の共役性より，ガウス過程回帰では，事前分布 $p(f)$ と事後分布 $p(f|Y, X)$ が共に正規分布に従うため，データサンプル $\mathcal{D} = \{X_i, Y_i\}_{i=1}^n$ に基づく平均関数 $m(\cdot)$ と共分散関数 $k(\cdot, \cdot)$ の行列計算 ($O(n^2)$) のみで事後分布の形状が求められる．

さらに，未知の m 個のデータ $\{\mathbf{x}_i\}_{i=n+1}^{n+m}$ に対して，対応する $\{y_i\}_{i=n+1}^{n+m}$ の同時分布 (予測分布) を求めることもできる．式 (8) より，

$$\begin{pmatrix} \mathbf{f}_n \\ \mathbf{f}_m \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} m(X_n) \\ m(X_m) \end{pmatrix}, \begin{bmatrix} K(X_n, X_n) & K(X_n, X_m) \\ K(X_n, X_m)^{\mathrm{T}} & K(X_m, X_m) \end{bmatrix}\right) \tag{11}$$

が成り立つから，式 (4) より，$\mathbf{f}_m$ の $\mathbf{f}_n$ に対する予測分布は，

$$\mathbf{f}_m|\mathbf{f}_n \sim \mathcal{N}\left(E[\mathbf{f}_m|\mathbf{f}_n], V[\mathbf{f}_m|\mathbf{f}_n]\right) \tag{12}$$

$$where \ \begin{cases} E[\mathbf{f}_m|\mathbf{f}_n] = m(X_m) + K(X_n, X_m)^{\mathrm{T}} K(X_n, X_n)^{-1}(\mathbf{f}_n - m(X_n)) \\ V[\mathbf{f}_m|\mathbf{f}_n] = K(X_m, X_m) - K(X_n, X_m)^{\mathrm{T}} K(X_n, X_n)^{-1} K(X_n, X_m) \end{cases}$$

となり，$\mathbf{f}_m$ の $\mathbf{y}_n$ に対する予測分布は，

$$\mathbf{f}_m|\mathbf{y}_n \sim \mathcal{N}\left(E[\mathbf{f}_m|\mathbf{y}_n], V[\mathbf{f}_m|\mathbf{y}_n]\right) \tag{13}$$

$$
where \begin{cases} E[\mathbf{f}_m|\mathbf{y}_n] = m(X_m) + K(X_n, X_m)^{\mathrm{T}}(K(X_n, X_n) + \sigma^2 I_n)^{-1}(\mathbf{y}_n - m(X_n)) \\ V[\mathbf{f}_m|\mathbf{y}_n] = K(X_m, X_m) - K(X_n, X_m)^{\mathrm{T}}(K(X_n, X_n) + \sigma^2 I_n)^{-1} K(X_n, X_m) \end{cases}
$$

となる．よって，$\mathbf{y}_m$ の $\mathbf{y}_n$ に対する予測分布は,

$$
\mathbf{y}_m|\mathbf{y}_n \sim \mathcal{N}\left(E[\mathbf{y}_m|\mathbf{y}_n], V[\mathbf{y}_m|\mathbf{y}_n]\right) \tag{14}
$$

$$
where \begin{cases} E[\mathbf{y}_m|\mathbf{y}_n] = m(X_m) + K(X_n, X_m)^{\mathrm{T}}(K(X_n, X_n) + \sigma^2 I_n)^{-1}(\mathbf{y}_n - m(X_n)) \\ V[\mathbf{y}_m|\mathbf{y}_n] = K(X_m, X_m) - K(X_n, X_m)^{\mathrm{T}}(K(X_n, X_n) + \sigma^2 I_n)^{-1} K(X_n, X_m) + \sigma^2 I_m \end{cases}
$$

となる．

3　Deep Neural Network のカーネル法による近似

3.1　1 層の Feed Forward Neural Network

ニューラルネットワークの全結合層や畳み込み層における推論計算は，線形写像と非線形の活性化関数の組み合わせによって構成される．ここで，ある層の入力ベクトルを $\mathbf{x} \in \mathbb{R}^N$，出力ベクトルを $\mathbf{y} \in \mathbb{R}^M$，線形写像に対応する変換行列を $\mathbf{W} \in \mathbb{R}^{M \times N}$，活性化関数を $\phi : \mathbb{R}^M \to \mathbb{R}^M$ とすると，非線形関数 $f : \mathbf{x} \mapsto \mathbf{y}$ は以下のように構成される．

$$
\mathbf{y} = f(\mathbf{x}) = \phi(\mathbf{W}\mathbf{x}) \tag{15}
$$

なお，全結合層に置けるバイアスベクトルの追加に関しては，$\mathbf{x}$ と $\mathbf{W}$ に次元を 1 つ追加することで上式と等価となり，畳み込み層に置けるフィルタ演算も im2col によって上式と等価となることに注意する．ここで，関数 f に対する内積 $f(\cdot)^{\mathrm{T}} f(\cdot) : \mathbf{x} \times \mathbf{x}' \mapsto \mathbb{R}$ を考える．活性化関数 ϕ として ReLU を仮定して，対応する指示関数

$$
\mathrm{I}(x) = \begin{cases} 1 \ (x \geq 0) \\ 0 \ (x < 0) \end{cases} \tag{16}
$$

を定義すると，$\forall \mathbf{x}, \mathbf{x}'$ に対する内積は,

$$
f(\mathbf{x})^{\mathrm{T}} f(\mathbf{x}') = \sum_{i=1}^{M} \mathrm{I}(\mathbf{w}_i^{\mathrm{T}}\mathbf{x})\mathrm{I}(\mathbf{w}_i^{\mathrm{T}}\mathbf{x}')(\mathbf{w}_i^{\mathrm{T}}\mathbf{x})(\mathbf{w}_i^{\mathrm{T}}\mathbf{x}') \tag{17}
$$

となる．ただし，$\mathbf{w}_i = (\mathbf{W}_{i1}, \ldots, \mathbf{W}_{iN})^{\mathrm{T}} \in \mathbb{R}^N$ とする．いま，行列 $\mathbf{W}$ の各要素 $\mathbf{W}_{ij}$ を i.i.d. となる確率変数と仮定すると，すべての i に対して $\mathbf{w}_i$ も

i.i.d. となる，また，確率変数 $\mathbf{W}_{ij}$ は平均と分散として，それぞれ 0，σ_w^2 を与える．

$$\forall i, j,\ \mathbf{W}_{ij} \sim i.i.d. \tag{18}$$

$$E[\mathbf{W}_{ij}] = 0,\ V[\mathbf{W}_{ij}] = \sigma_w^2 \tag{19}$$

すると，中心極限定理 (Central Limit Theorem) より，$\mathbf{x}$ を固定したとき，任意の $\mathbf{w}$ の関数 $C(\mathbf{w})$ に対して，

$$E\left[\frac{1}{M}\sum_{i=1}^{M} C(\mathbf{w}_i)\right] \to \int d\mathbf{w} \frac{\exp(\frac{-||\mathbf{w}||^2}{2})}{(2\pi\sigma_w^2)^{N/2}} C(\mathbf{w}_i) \quad (M \to \infty) \tag{20}$$

が満たされるから，

$$E\left[\frac{1}{M} f(\mathbf{x})^{\mathrm{T}} f(\mathbf{x}')\right] = E\left[\frac{1}{M}\sum_{i=1}^{M} \mathrm{I}(\mathbf{w}_i^{\mathrm{T}}\mathbf{x})\mathrm{I}(\mathbf{w}_i^{\mathrm{T}}\mathbf{x}')(\mathbf{w}_i^{\mathrm{T}}\mathbf{x})(\mathbf{w}_i^{\mathrm{T}}\mathbf{x}')\right] \to$$

$$\int d\mathbf{w} \frac{\exp(\frac{-||\mathbf{w}||^2}{2})}{(2\pi\sigma_w^2)^{N/2}} \mathrm{I}(\mathbf{w}^{\mathrm{T}}\mathbf{x})\mathrm{I}(\mathbf{w}^{\mathrm{T}}\mathbf{x}')(\mathbf{w}^{\mathrm{T}}\mathbf{x})(\mathbf{w}^{\mathrm{T}}\mathbf{x}') \quad (M \to \infty) \tag{21}$$

が成り立つ．式 (21) 右辺を整理すると，以下の定理が導かれる．

定義 1 *(arc-cosine kernel) 任意の 2 つのベクトル* $\mathbf{x}, \mathbf{x}' \in \mathcal{X} \subset \mathbb{R}^N$ *に対して，*

$$\theta = \cos^{-1}\left(\frac{\mathbf{x}^{\mathrm{T}}\mathbf{x}'}{\|\mathbf{x}\|\|\mathbf{x}'\|}\right) \tag{22}$$

$$J(\theta) = \sin\theta + (\pi - \theta)\cos\theta \tag{23}$$

とおき，arc-cosine カーネル

$$k^{(1)}(\mathbf{x}, \mathbf{x}') = \frac{1}{\pi}\|\mathbf{x}\|\|\mathbf{x}'\| J(\theta) \quad (M \to \infty) \tag{24}$$

を定義する．

定理 1 *各要素* W_{ij} *が互いに独立に正規分布* $\mathcal{N}(0,1)$ *に従うランダム行列* $\mathbf{W} \in \mathbb{R}^{M\times N}$ *を考える．活性化関数として ReLU 関数* ϕ *を用いて，1 層のニューラルネットワークに相当する関数* $f : \mathbb{R}^N \to \mathbb{R}^M$*:*

$$f(\mathbf{x}) = \phi(\mathbf{W}\mathbf{x}) \quad (\forall \mathbf{x} \in \mathbb{R}^N) \tag{25}$$

を考える．このとき任意の $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^N$ *に対して，*

$$E\left[\frac{1}{M} f(\mathbf{x})^{\mathrm{T}} f(\mathbf{x}')\right] \to \frac{1}{\sigma_w^N} k^{(1)}(\mathbf{x}, \mathbf{x}') \quad (M \to \infty) \tag{26}$$

が成り立つ．

定理 2 $\mathcal{X} \times \mathcal{X}$ 上の関数

$$k^{(1)}(\mathbf{x}, \mathbf{x}') = \frac{1}{\pi}\|\mathbf{x}\|\|\mathbf{x}'\|J(\theta) \tag{27}$$

は半正定値関数である．

定理 3 *(カーネル関数存在定理)* 任意の対称行列 $K \in \mathbb{R}^{n \times n}$ が半正定値ならば，データ空間 $\mathcal{X}$ 上の n 点 $\{\mathbf{x}_i\}_{i=1}^n$ と，特徴空間 $\mathcal{F} \subset \mathbb{R}^n$ 上の n 次元特徴ベクトル $\{f(\mathbf{x}_i)\}_{i=1}^n$ がそれぞれ存在して，

$$K_{ij} = \langle f(\mathbf{x}_i), f(\mathbf{x}_j) \rangle_{\mathcal{F}} = \sum_{k=1}^{n} f(\mathbf{x}_i)_k f(\mathbf{x}_j)_k \tag{28}$$

が成り立つ．さらに，$k(\mathbf{x}_i, \mathbf{x}_j) = K_{ij}$ とおけば，半正定値関数 $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ が定義される．

定理 3 から，カーネル関数 $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ を定義せずとも，与えられたデータサンプルから構成された半正定値対称行列を作ることで，それが何らかのカーネル関数によるグラム行列に相当することが保証される．よって，定理 1 と定理 2 より，式 (27) を共分散関数にもつガウス過程が 1 層の NeuralNetwork の近似となることがわかる．

3.2 多層の Feed Forward Neural Network

1 層の Feed Forward Neural Network に関する議論は，正定値カーネルの線形性より，容易に多層へと拡張することができる．一般性を失うことなく，$L > 0$ 層からなる Feed Forward Neural Network を考え，各層の線形写像に対応する重み行列 $\mathbf{W}^{(l)}$ に対して，固定された σ_w^2 を与えて，

$$\forall l \in [1, L] \subset \mathbb{N}, \quad \mathbf{W}_{ij}^{(l)} \sim i.i.d.\ \mathcal{N}(0, \sigma_w^2) \tag{29}$$

を仮定する．第 l 層のユニット数を N_l，第 l 層の空間を $\mathcal{T}^{(l)}$，第 l 層の出力を得る関数を $f^{(l)} : \mathcal{X} \to \mathcal{T}^{(l)}$ とおく．任意の入力データ $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$ に対して，Neural Network の第 l 層に対応する共分散関数 $k^{(l)} : \mathcal{T}^{(l)} \times \mathcal{T}^{(l)} \to \mathbb{R}$ は，次

の漸化式で求められる．

$$
\begin{aligned}
k^{(1)}(\mathbf{x},\mathbf{x}') &= Cov\left[f^{(1)}(\mathbf{x}), f^{(1)}(\mathbf{x}')\right] \\
&= Cov\left[\phi(\mathbf{W}^{(1)}\mathbf{x}), \phi(\mathbf{W}^{(1)}\mathbf{x}')\right] \\
&= \frac{1}{\pi\sigma_w^{N_1}}\|\mathbf{x}\|\|\mathbf{x}'\|J(\theta^{(0)}) \qquad (30) \\
k^{(l+1)}(\mathbf{x},\mathbf{x}') &= Cov\left[f^{(l+1)}(\mathbf{x}), f^{(l+1)}(\mathbf{x}')\right] \\
&= Cov\left[\phi(\mathbf{W}^{(l+1)}\cdots\phi(\mathbf{W}^{(1)}\mathbf{x})), \phi(\mathbf{W}^{(l+1)}\cdots\phi(\mathbf{W}^{(1)}\mathbf{x}'))\right] \\
&= \frac{1}{\pi\sigma_w^{N_{l+1}}}\sqrt{k^{(l)}(\mathbf{x},\mathbf{x})}\sqrt{k^{(l)}(\mathbf{x}',\mathbf{x}')}J(\theta^{(l)}) \qquad (31)
\end{aligned}
$$

ただし，

$$
J(\theta^{(l)}) = \sin\theta^{(l)} + (\pi-\theta^{(l)})\cos\theta^{(l)} \qquad (32)
$$

$$
\theta^{(l)} = \begin{cases} \cos^{-1}\left(\frac{\mathbf{x}^{\mathrm{T}}\mathbf{x}'}{\|\mathbf{x}\|\|\mathbf{x}'\|}\right) & (l=0) \\ \cos^{-1}\left(\frac{k^{(l)}(\mathbf{x},\mathbf{x}')}{\sqrt{k^{(l)}(\mathbf{x},\mathbf{x})}\sqrt{k^{(l)}(\mathbf{x}',\mathbf{x}')}}\right) & (l\neq 0) \end{cases} \qquad (33)
$$

とする．以上の結果から，第 l 層の出力 $f^{(l)}$ は以下のガウス過程で近似できる．

$$
f^{(l)}(\cdot) \sim \mathcal{GP}(\mathbf{0}, k^{(l)}(\cdot,\cdot)) \qquad (34)
$$

4 Kernel Variational AutoEncoder

5 Experiments

6 Conclusion