Kernel VAE: Gaussian Process Latent Variable Model on Deep Neural Network

Yuma Uchiumi*[1]

[1]Department of Information and Computer Science, Faculty of Science and Technology, Keio University, Japan

2019/10/11

Abstract

...

Contents

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

1 Introduction

...

2 ガウス過程と確率モデル

2.1 ガウス分布

分布の再生性 任意の $n \in \mathbb{N}$ と $\{a_i, b_i \in \mathbb{R}\}_{i=1}^{n}$ に対して，次が成り立つ．

$$X_i \sim i.i.d.\ \mathcal{N}(\mu_i, \sigma_i^2), \quad (i = 1, \ldots, n) \tag{1}$$

$$\Rightarrow \sum_{i=1}^{n} a_i X_i + b_i \sim \mathcal{N}\left(\sum_{i=1}^{n} a_i \mu_i + b_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2\right) \tag{2}$$

条件つき分布の計算 任意の 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，

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

ならば，次が成り立つ．

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

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

2.2 ガウス過程 (GP)

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

$$p(\mathbf{f}) = \frac{1}{(2\pi)^{n/2}|V(\mathbf{f})|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{f}-m(\mathbf{f}))^{\mathrm{T}} V(\mathbf{f})^{-1}(\mathbf{f}-m(\mathbf{f}))\right) \tag{5}$$

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．

$$f(\cdot) \sim \mathcal{GP}(m(\cdot), v(\cdot,\cdot)) \tag{6}$$

2.3 ガウス過程回帰 (GPR)

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．

$$y = f(\mathbf{x}) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0,\sigma^2) \tag{7}$$

あるいは，正規分布の再生性より，

$$y|\mathbf{x} \sim \mathcal{N}(f(\mathbf{x}), \sigma^2) \tag{8}$$

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．

$$p(f|\mathcal{D}) = \frac{p(\mathcal{D}|f)p(f)}{p(\mathcal{D})}, \quad i.e. \quad p(f|Y,X) = \frac{p(Y|X,f)p(f)}{p(Y|X)} \tag{9}$$

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，

$$f(\cdot) \sim \mathcal{GP}(m(\cdot), k(\cdot,\cdot)) \tag{10}$$

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，

$$\mathbf{f}_n \sim \mathcal{N}(m(X_n), K(X_n,X_n)) \tag{11}$$

$$\mathbf{y}_n \sim \mathcal{N}(m(X_n), K(X_n,X_n) + \sigma^2 I_n) \tag{12}$$

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 (10) は，式 (9) 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 (8) 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 (10) より，

$$\begin{pmatrix} \mathbf{f}_n \\ \mathbf{f}_m \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} m(X_n) \\ m(X_m) \end{pmatrix}, \begin{bmatrix} K(X_n, X_n) & K(X_n, X_m) \\ K(X_n, X_m)^{\mathrm{T}} & K(X_m, X_m) \end{bmatrix}\right) \tag{13}$$

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，

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

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

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

$$\mathbf{f}_m|\mathbf{y}_n \sim \mathcal{N}\left(E[\mathbf{f}_m|\mathbf{y}_n], V[\mathbf{f}_m|\mathbf{y}_n]\right) \tag{15}$$

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

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

$$\mathbf{y}_m|\mathbf{y}_n \sim \mathcal{N}\left(E[\mathbf{y}_m|\mathbf{y}_n], V[\mathbf{y}_m|\mathbf{y}_n]\right) \tag{16}$$

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

となる．

2.4 ガウス過程潜在変数モデル (GP-LVM)

確率変数 $X \in \mathbb{R}^Q, Y \in \mathbb{R}^D$ に対して，Y の N 個のデータサンプル $\{\mathbf{y}_i\}_{i=1}^n$ から，生成モデル $p(Y|X)$ を特定することを考える．このとき，Y を観測変数，X を潜在変数と呼ぶ．観測変数の観測値 (計画行列) を $\mathbf{Y}^n = \mathbb{R}^{n\times D}$，それを表現する潜在変数の観測値 (計画行列) を $\mathbf{X}^n \in \mathbb{R}^{n\times Q}$ とおく．すべての $i \in \{1,\ldots,N\}$ と $d \in \{1,\ldots,D\}$ に対して，関数 $f : \mathbb{R}^Q \to \mathbb{R}$ を用いて，生成過程：

$$\mathbf{y}_{id} = f(\mathbf{x}_i) + \varepsilon_i, \quad \varepsilon_i \sim i.i.d.\ \mathcal{N}(0, \sigma_\varepsilon^2) \tag{17}$$

を仮定し，さらに関数 f に対してガウス過程

$$f(\cdot) \sim \mathcal{GP}(0, k(\cdot,\cdot)) \tag{18}$$

を仮定すると，

$$p(\mathbf{y}_i|\mathbf{X}^n) = \prod_{d=1}^{D} p(\mathbf{y}_{id}) = \prod_{d=1}^{D} \mathcal{N}(\mathbf{y}_{id}|f(\mathbf{x}_i), \sigma_\varepsilon^2) \tag{19}$$

$$p(\mathbf{y}_d|\mathbf{X}^n) = \prod_{i=1}^{n} p(\mathbf{y}_{id}) = \mathcal{N}(\mathbf{y}_d|\mathbf{0}, \mathbf{\Sigma}_{nn}) \tag{20}$$

となるから，生成モデルの尤度は，次式で与えられる．

$$p(\mathbf{Y}^n|\mathbf{X}^n) = \prod_{d=1}^{D} p(\mathbf{y}_d|\mathbf{X}^n) = \prod_{d=1}^{D} \mathcal{N}(\mathbf{y}_d|\mathbf{0}, \mathbf{\Sigma}_{nn}) \tag{21}$$

$$= \prod_{d=1}^{D} \frac{1}{(2\pi)^{\frac{n}{2}} |\mathbf{\Sigma}_{nn}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\mathbf{y}_d^{\mathrm{T}} \mathbf{\Sigma}_{nn}^{-1} \mathbf{y}_d\right) \tag{22}$$

$$= \frac{1}{(2\pi)^{\frac{Dn}{2}} |\mathbf{\Sigma}_{nn}|^{\frac{D}{2}}} \exp\left(-\frac{1}{2}\mathrm{tr}(\mathbf{\Sigma}_{nn}^{-1} \mathbf{Y}^n \mathbf{Y}^{n\mathrm{T}})\right) \tag{23}$$

ただし，

$$\mathbf{\Sigma}_{nn} = \mathbf{K}_{nn} + \sigma_\varepsilon^2 \mathbf{I}_n \tag{24}$$

$$\mathbf{K}_{nn}(i,j) = k(\mathbf{x}_i, \mathbf{x}_j), \quad \forall (i,j) \tag{25}$$

とする．

最尤推定　$\mathbf{Y}^n$ が与えられたときの $\mathbf{X}^n$ の対数尤度は，

$$\log p(\mathbf{Y}^n|\mathbf{X}^n) = -\frac{Dn}{2}\log(2\pi) - \frac{D}{2}\log|\mathbf{\Sigma}_{nn}| - \frac{1}{2}\mathrm{tr}(\mathbf{\Sigma}_{nn}^{-1}\mathbf{Y}^n\mathbf{Y}^{n\mathrm{T}}) \tag{26}$$

となるから，観測値によって固定された $\mathbf{Y}^n$ の下で，これを目的関数として最大化すれば，$\mathbf{X}^n$ の最尤推定量 $\hat{\mathbf{X}}^n_{ML}$ が求められる．

$$\hat{\mathbf{X}}^n_{ML} = \underset{\mathbf{X}^n}{\mathrm{argmax}}\ \log p(\mathbf{Y}^n|\mathbf{X}^n) \tag{27}$$

ここで，対数尤度において $\mathbf{X}^n$ に依存する変数は $\mathbf{\Sigma}_{nn}$ のみであることに注意する．対数尤度の $\mathbf{\Sigma}_{nn}$ に対する勾配は，次のように計算される．

$$\frac{\partial \log p(\mathbf{Y}^n|\mathbf{X}^n)}{\partial \mathbf{\Sigma}_{nn}} = \mathbf{\Sigma}_{nn}^{-1}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\mathbf{\Sigma}_{nn}^{-1} - D\mathbf{\Sigma}_{nn}^{-1} \tag{28}$$

よって，対数尤度の $\mathbf{x}_{ij}$ に対する勾配 $\partial \log p(\mathbf{Y}^n|\mathbf{X}^n)/\partial \mathbf{x}_{ij}$ は，$\partial\mathbf{\Sigma}_{nn}/\partial\mathbf{x}_{ij} = \partial\mathbf{K}_{nn}/\partial\mathbf{x}_{ij}$ と連鎖律を用いて計算される．

$$\frac{\partial \log p(\mathbf{Y}^n|\mathbf{X}^n)}{\partial \mathbf{x}_{ij}} = \frac{\partial \log p(\mathbf{Y}^n|\mathbf{X}^n)}{\partial \mathbf{\Sigma}_{nn}}\frac{\partial \mathbf{\Sigma}_{nn}}{\partial \mathbf{x}_{ij}} \tag{29}$$

$$= \left(\mathbf{\Sigma}_{nn}^{-1}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\mathbf{\Sigma}_{nn}^{-1} - D\mathbf{\Sigma}_{nn}^{-1}\right)\frac{\partial \mathbf{K}_{nn}}{\partial \mathbf{x}_{ij}} \tag{30}$$

これによって，$\mathbf{x}_{ij}$ を逐次更新できる．

$$\mathbf{x}_{ij} \leftarrow \mathbf{x}_{ij} - \eta \cdot \frac{\partial \log p(\mathbf{Y}^n|\mathbf{X}^n)}{\partial \mathbf{x}_{ij}} \tag{31}$$

MAP 推定　$\mathbf{X}^n$ の事前分布として，

$$p(\mathbf{X}^n) = \prod_{q=1}^{Q} p(\mathbf{x}_q) = \prod_{q=1}^{Q}\mathcal{N}(\mathbf{x}_q|\mathbf{0}, \sigma_x^2 I_N) \tag{32}$$

$$= \prod_{q=1}^{Q}\frac{1}{(2\pi\sigma_x^2)^{\frac{n}{2}}}\exp\left(-\frac{\mathbf{x}_q^{\mathrm{T}}\mathbf{x}_q}{2\sigma_x^2}\right) \tag{33}$$

$$= \frac{1}{(2\pi\sigma_x^2)^{\frac{QN}{2}}}\exp\left(-\frac{1}{2\sigma_x^2}\mathrm{tr}(\mathbf{X}^n\mathbf{X}^{n\mathrm{T}})\right) \tag{34}$$

$$\tag{35}$$

を仮定すると，ベイズの定理より，X の事後分布は，

$$p(X|Y) = \frac{1}{Z} p(Y|X) p(X) \tag{36}$$

となる．ただし，Z は規格化定数とする，よって，$\mathbf{X}^n$ の事後確率は，

$$\begin{aligned} p(\mathbf{X}^n|\mathbf{Y}^n) &\propto \log p(\mathbf{Y}^n|\mathbf{X}^n) p(\mathbf{X}^n) & (37) \\ &= -\frac{Dn}{2}\log(2\pi) - \frac{D}{2}\log|\mathbf{\Sigma}_{nn}| - \frac{1}{2}\mathrm{tr}(\mathbf{\Sigma}_{nn}^{-1}\mathbf{Y}^n\mathbf{Y}^{n\mathrm{T}}) & (38) \\ &\quad - \frac{Qn}{2}\log(2\pi\sigma_x^2) - \frac{1}{2}\mathrm{tr}(\mathbf{X}^n\mathbf{X}^{n\mathrm{T}}) & (39) \end{aligned}$$

となるから，これを目的関数として最大化すれば，$\mathbf{X}^n$ の MAP 推定量は，次のように求められる．

$$\hat{\mathbf{X}}^n_{MAP} = \underset{X^n}{\mathrm{argmax}}\ \log p(\mathbf{Y}^n|\mathbf{X}^n) p(\mathbf{X}^n) \tag{40}$$

3 Deep Neural Network のカーネル法による近似

3.1 線形写像に対応するガウス過程

3.1.1 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 $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^N$，出力ベクトルを $\mathbf{y} \in \mathcal{Y} \subseteq \mathbb{R}^M$，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 $\phi : \mathbb{R} \to \mathbb{R}$ とすると，各ノード y_i への推論処理は以下のように構成される．

$$y_i = \phi(z_i(\mathbf{x})), \quad z_i(\mathbf{x}) = {\mathbf{w}_i}^{\mathrm{T}}\mathbf{x} = \sum_{j=1}^{N} w_{ij}x_j, \quad (i = 1, \ldots, M) \tag{41}$$

ただし，w_{ij} は行列 $\mathbf{W}$ の ij 要素とし，$\mathbf{w}_i = (w_{i1}, \ldots, w_{iN})^{\mathrm{T}} \in \mathbb{R}^N$ とする．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，重み行列 $\mathbf{W}$ の各要素 w_{ij} に対して，

$$w_{ij} \ \ i.i.d. \sim p_w(\cdot), \quad E[w_{ij}] = 0,\ V[w_{ij}] = \sigma_w^2 \tag{42}$$

を仮定すると，任意の入力 $\mathbf{x}$ に対して，$x_j \perp\!\!\!\perp x_{j'}$ $(j \neq j')$ だから，中心極限定理 (Central Limit Therem) より，

$$z_i(\mathbf{x}) \ \sim \ \mathcal{N}(0, \sigma_w^2) \quad (N \to \infty) \tag{43}$$

が成り立つ．すなわち n 個の入力 $\{\mathbf{x}^{(i)}\}_{i=1}^n$ が与えられたとき，$\mathbf{z}_i = (z_i(\mathbf{x}^{(1)}), \ldots, z_i(\mathbf{x}^{(n)}))^{\mathrm{T}} \in \mathbb{R}^n$ は n 次元ガウス分布に従うから，

$$z_i(\cdot) \sim \mathcal{GP}(0, k(\cdot, \cdot)) \tag{44}$$

が与えられる．このガウス過程を用いて上述の推論処理を解く際には，共分散関数 (カーネル関数)$k(\cdot, \cdot) : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$ を特定する必要がある．実際，任意の入力 $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$ に対して，

$$\begin{aligned}
k(\mathbf{x}, \mathbf{x}') &= Cov[z_i(\mathbf{x}), z_i(\mathbf{x}')] = E_w[z_i(\mathbf{x}) z_i(\mathbf{x}')] && (45)\\
&= E_w\left[\left(\sum_{j=1}^N w_{ij} x_j\right)\left(\sum_{j=1}^N w_{ij} {x'}_j\right)\right] && (46)\\
&= E_w\left[\sum_{j=1}^N \sum_{k=1}^N w_{ij} w_{ik} x_j {x'}_k\right] && (47)\\
&= \sum_{j=1}^N \sum_{k=1}^N E_w[w^2] x_j {x'}_k && (48)\\
&= \sigma_w^2 \sum_{j=1}^N \sum_{k=1}^N x_j {x'}_k && (49)\\
& && (50)
\end{aligned}$$

となり．さらに，任意の $\mathbf{x} = (x_1, \ldots, x_N)^{\mathrm{T}} \in \mathcal{X}$ に対して，

$$x_j \ \ i.i.d. \sim p_x(\cdot), \quad E[x_j] = 0, \quad (i = 1, \ldots, N) \tag{51}$$

を仮定すると，

$$E_{\mathbf{x}, \mathbf{x}'}[k(\mathbf{x}, \mathbf{x}')] = \sigma_w^2 \sum_{j=1}^N E[x_j {x'}_j] = \sigma_w^2 E[\mathbf{x}^{\mathrm{T}} \mathbf{x}'] \tag{52}$$

となることに注意する．

3.1.2 多層の Feed Forward Neural Network

活性化関数 ϕ をもち L 層からなる Feed Forward Neural Network を考え，$l = 1, \ldots, L$ に対して，第 l 層の活性化前の状態を $\mathbf{z}^{(l)}$，ノード数を $N^{(l)}$ とそれぞれおき，上の議論を多層に拡張する．すなわち，第 l 層の任意のノード値 $z_i^{(l)}(\cdot)$ に対して，ガウス過程

$$z_i^{(l)}(\cdot) \sim \mathcal{GP}(0, k^{(l)}(\cdot, \cdot)) \tag{53}$$

を仮定し，任意の入力 $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$ に対して，この分散関数 $k^{(l)}$ を特定すればよい（図 1 参照）．上の議論から，任意の第 l 層に対して，カーネル関数 $k^{(l)}$ は

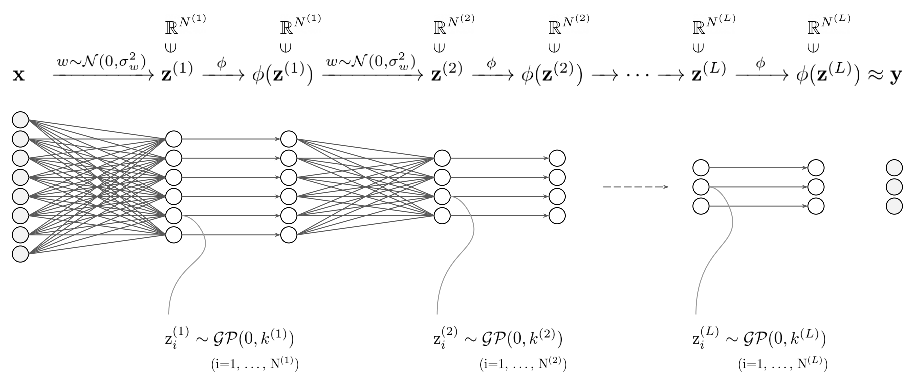

Figure 1: ガウス過程と DNN

次のように計算される．

$$
\begin{aligned}
k^{(l)}(\mathbf{x}, \mathbf{x}') &= Cov[z_i^{(l)}(\mathbf{x}), z_i^{(l)}(\mathbf{x}')] && (54)\\
&= E_w[z_i^{(l)}(\mathbf{x}) z_i^{(l)}(\mathbf{x}')] && (55)\\
&= E_w\left[\left(\sum_{j=1}^{N^{(l-1)}} w_{ij}\phi(z_j^{(l-1)}(\mathbf{x}))\right)\left(\sum_{j=1}^{N^{(l-1)}} w_{ij}\phi(z_j^{(l-1)}(\mathbf{x}'))\right)\right] && (56)\\
&= E_w\left[\sum_{j=1}^{N^{(l-1)}}\sum_{k=1}^{N^{(l-1)}} w_{ij}w_{ik}\phi(z_j^{(l-1)}(\mathbf{x}))\phi(z_k^{(l-1)}(\mathbf{x}'))\right] && (57)\\
&= \sum_{j=1}^{N^{(l-1)}}\sum_{k=1}^{N^{(l-1)}} E_w[w^2]\phi(z_j^{(l-1)}(\mathbf{x}))\phi(z_k^{(l-1)}(\mathbf{x}')) && (58)\\
&= \sigma_w^2 \sum_{j=1}^{N^{(l-1)}}\sum_{k=1}^{N^{(l-1)}} \phi(z_j^{(l-1)}(\mathbf{x}))\phi(z_k^{(l-1)}(\mathbf{x}')) \quad (l \geq 2) && (59)\\
& && (60)\\
k^{(1)}(\mathbf{x}, \mathbf{x}') &= Cov[z_i^{(1)}(\mathbf{x}), z_i^{(1)}(\mathbf{x}')] && (61)\\
&= E_w[z_i^{(1)}(\mathbf{x}) z_i^{(1)}(\mathbf{x}')] && (62)\\
&= \sigma_w^2 \sum_{j=1}^{N^{(1)}}\sum_{k=1}^{N^{(1)}} x_j {x'}_k && (63)
\end{aligned}
$$

さらに，任意の $\mathbf{z}^{(l)} = (z^{(l)}(\mathbf{x}_1), \ldots, z^{(l)}(\mathbf{x}_N))^{\mathrm{T}} \in \mathbb{R}^N$ に対して，

$$
z^{(l)}(\mathbf{x}_i) \ \ i.i.d. \sim p_z(\cdot), \quad E[z^{(l)}(\mathbf{x}_i)] = 0, \quad (i = 1, \ldots, N) \tag{64}
$$

を仮定すると，

$$
\phi(z^{(l)}(\mathbf{x}_i)) \ \ i.i.d. \sim p_{\phi(z)}(\cdot), \quad E[\phi(z^{(l)}(\mathbf{x}_i))] = 0, \quad (i = 1, \ldots, N) \tag{65}
$$

だから，

$$E_{\mathbf{x},\mathbf{x}'}\left[k^{(l)}(\mathbf{x},\mathbf{x}')\right] \tag{66}$$

$$=E_{\mathbf{x},\mathbf{x}'}\left[\sigma_w^2\sum_{j=1}^{N^{(l-1)}}\sum_{k=1}^{N^{(l-1)}}\phi(z_j^{(l-1)}(\mathbf{x}))\phi(z_k^{(l-1)}(\mathbf{x}'))\right] \tag{67}$$

$$=\sigma_w^2 E_{\mathbf{z}^{(l-1)},\mathbf{z}'^{(l-1)}}\left[\sum_{j=1}^{N^{(l-1)}}\phi(z_i^{(l-1)})\phi({z'}_i^{(l-1)})\right] \tag{68}$$

$$=\sigma_w^2 E_{\mathbf{z}^{(l-1)},\mathbf{z}'^{(l-1)}}\left[\boldsymbol{\phi}(\mathbf{z}^{(l-1)})^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{z}'^{(l-1)})\right] \tag{69}$$

$$=\sigma_w^2 N^{(l-1)} E_{z^{(l-1)},z'^{(l-1)}\sim\mathcal{GP}(0,k^{(l-1)})}\left[\phi(z^{(l-1)})\phi(z'^{(l-1)})\right] \tag{70}$$

となることに注意する．

3.2 カーネル法

3.2.1 正定値カーネルによる近似

式 (41) の推論処理をまとめて関数 $f_i:\mathbb{R}^N\to\mathbb{R}$ とおく．

$$y_i = f_i(\mathbf{x}) = \phi({\mathbf{w}_i}^{\mathrm{T}}\mathbf{x}) \tag{71}$$

ここで，関数 f_i に対する内積 $f_i(\cdot)^{\mathrm{T}}f_i(\cdot):\mathbf{x}\times\mathbf{x}'\mapsto\mathbb{R}$ 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

$$\mathrm{I}(x)=\begin{cases}1\ (x\geq 0)\\ 0\ (x<0)\end{cases} \tag{72}$$

を定義すると，

$$f_i(\mathbf{x})f_i(\mathbf{x}')=\mathrm{I}(\mathbf{w}_i^{\mathrm{T}}\mathbf{x})\mathrm{I}(\mathbf{w}_i^{\mathrm{T}}\mathbf{x}')(\mathbf{w}_i^{\mathrm{T}}\mathbf{x})(\mathbf{w}_i^{\mathrm{T}}\mathbf{x}') \tag{73}$$

$$=\mathrm{I}(\sum_{j=1}^{N}w_{ij}x_j)\mathrm{I}(\sum_{j=1}^{N}w_{ij}{x'}_j)(\sum_{j=1}^{N}w_{ij}x_j)(\sum_{j=1}^{N}w_{ij}{x'}_j) \tag{74}$$

となる．中心極限定理 (Central Limit Theorem) より，任意の $\mathbf{w}\in\mathbb{R}^N$ の関数 $C(\mathbf{w})$ に対して，

$$E\left[\frac{1}{M}\sum_{i=1}^{M}C(\mathbf{w}_i)\right]\to\int d\mathbf{w}\frac{\exp(\frac{-||\mathbf{w}||^2}{2})}{(2\pi\sigma_w^2)^{N/2}}C(\mathbf{w})\quad(M\to\infty) \tag{75}$$

が満たされるから，$\mathbf{x}, \mathbf{x}'$ を固定したとき，

$$\begin{aligned} &E\left[\frac{1}{M}\sum_{i=1}^{M} f_i(\mathbf{x})f_i(\mathbf{x}')\right] = E\left[\frac{1}{M}\sum_{i=1}^{M} \mathrm{I}(\mathbf{w}_i^{\mathrm{T}}\mathbf{x})\mathrm{I}(\mathbf{w}_i^{\mathrm{T}}\mathbf{x}')(\mathbf{w}_i^{\mathrm{T}}\mathbf{x})(\mathbf{w}_i^{\mathrm{T}}\mathbf{x}')\right] \\ &\to \int d\mathbf{w}\frac{\exp(\frac{-\|\mathbf{w}\|^2}{2})}{(2\pi\sigma_w^2)^{N/2}}\mathrm{I}(\mathbf{w}^{\mathrm{T}}\mathbf{x})\mathrm{I}(\mathbf{w}^{\mathrm{T}}\mathbf{x}')(\mathbf{w}^{\mathrm{T}}\mathbf{x})(\mathbf{w}^{\mathrm{T}}\mathbf{x}') \quad (M \to \infty) \end{aligned} \tag{76}$$

が成り立つ．式 (76) 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，

$$\theta = \cos^{-1}\left(\frac{\mathbf{x}^{\mathrm{T}}\mathbf{x}'}{\|\mathbf{x}\|\|\mathbf{x}'\|}\right) \tag{77}$$

$$J(\theta) = \sin\theta + (\pi - \theta)\cos\theta \tag{78}$$

とおき，*Arc-Cosine* カーネル

$$k^{(1)}(\mathbf{x}, \mathbf{x}') = \frac{1}{\pi}\|\mathbf{x}\|\|\mathbf{x}'\|J(\theta) \tag{79}$$

を定義する．

定理 1 各要素 w_{ij} が *i.i.d.* で，平均が 0，分散が σ_w^2 となるランダム行列 $\mathbf{W} \in \mathbb{R}^{M\times N}$ 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 $f_i : \mathbb{R}^N \to \mathbb{R}$*:*

$$f_i(\mathbf{x}) = \phi(\mathbf{w}_i^{\mathrm{T}}\mathbf{x}), \quad (i = 1, \ldots, M) \tag{80}$$

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

$$E\left[\frac{1}{M}\sum_{i=1}^{M} f_i(\mathbf{x})f_i(\mathbf{x}')\right] \to \frac{1}{\sigma_w^N}k^{(1)}(\mathbf{x}, \mathbf{x}') \quad (M \to \infty) \tag{81}$$

が成り立つ．

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

$$k^{(1)}(\mathbf{x}, \mathbf{x}') = \frac{1}{\pi}\|\mathbf{x}\|\|\mathbf{x}'\|J(\theta) \tag{82}$$

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，

$$K_{ij} = \langle f(\mathbf{x}_i), f(\mathbf{x}_j)\rangle_{\mathcal{F}} = \sum_{k=1}^{n} f(\mathbf{x}_i)_k f(\mathbf{x}_j)_k \tag{83}$$

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 (82) 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．

3.2.2 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，

$$\forall l \in [1, L] \subset \mathbb{N}, \quad \mathbf{W}_{ij}^{(l)} \sim i.i.d.\ \mathcal{N}(0, \sigma_w^2) \tag{84}$$

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．

$$\begin{aligned}
k^{(1)}(\mathbf{x}, \mathbf{x}') &= Cov\left[f^{(1)}(\mathbf{x}), f^{(1)}(\mathbf{x}')\right] \\
&= Cov\left[\phi(\mathbf{W}^{(1)}\mathbf{x}), \phi(\mathbf{W}^{(1)}\mathbf{x}')\right] \\
&= \frac{1}{\pi\sigma_w^{N^{(1)}}} \|\mathbf{x}\|\|\mathbf{x}'\| J(\theta^{(0)}) \qquad (85) \\
k^{(l+1)}(\mathbf{x}, \mathbf{x}') &= Cov\left[f^{(l+1)}(\mathbf{x}), f^{(l+1)}(\mathbf{x}')\right] \\
&= Cov\left[\phi(\mathbf{W}^{(l+1)} \cdots \phi(\mathbf{W}^{(1)}\mathbf{x})), \phi(\mathbf{W}^{(l+1)} \cdots \phi(\mathbf{W}^{(1)}\mathbf{x}'))\right] \\
&= \frac{1}{\pi\sigma_w^{N^{(l+1)}}} \sqrt{k^{(l)}(\mathbf{x}, \mathbf{x})}\sqrt{k^{(l)}(\mathbf{x}', \mathbf{x}')} J(\theta^{(l)}) \qquad (86)
\end{aligned}$$

ただし，

$$J(\theta^{(l)}) = \sin\theta^{(l)} + (\pi - \theta^{(l)})\cos\theta^{(l)} \tag{87}$$

$$\theta^{(l)} = \begin{cases} \cos^{-1}\left(\frac{\mathbf{x}^{\mathrm{T}}\mathbf{x}'}{\|\mathbf{x}\|\|\mathbf{x}'\|}\right) & (l = 0) \\ \cos^{-1}\left(\frac{k^{(l)}(\mathbf{x},\mathbf{x}')}{\sqrt{k^{(l)}(\mathbf{x},\mathbf{x})}\sqrt{k^{(l)}(\mathbf{x}',\mathbf{x}')}}\right) & (l \neq 0) \end{cases} \tag{88}$$

とする．以上の結果から，入力から第 l 層の各ユニットの出力値を得る関数 $f_i^{(l)} : \mathcal{X} \to \mathbb{R}$ は以下のガウス過程で近似できる．

$$f_i^{(l)}(\cdot) \sim \mathcal{GP}(0, k^{(l)}(\cdot,\cdot)), \quad (i = 1, \ldots, N^{(l)}) \tag{89}$$

4 Kernel Variational AutoEncoder

前章での議論から，GP-LVM における潜在変数 X と観測関数 Y に対する確率モデル $p(Y|X)$ を，多層の Feed Forward Neural Network に相当するモデル

$$y = f^{(l)}(\mathbf{x}) = \phi({\mathbf{w}^{(l)}}^{\mathrm{T}}\phi(\mathbf{W}^{(l-1)}\ldots\phi(\mathbf{W}^{(1)}\mathbf{x}))) \tag{90}$$

によって表現することを考える．すなわち，

$$k^{(l)}(\mathbf{x},\mathbf{x}') = \langle f^{(l)}(\mathbf{x}), f^{(l)}(\mathbf{x}')\rangle_{\mathcal{F}^{(l)}} \tag{91}$$

に対応するガウス過程

$$f^{(l)}(\cdot) \sim \mathcal{GP}(0, k^{(l)}(\cdot,\cdot)) \tag{92}$$

を考える．

4.1 変分推論

4.2 びぶん

$$\frac{\partial k^{(l)}(\mathbf{x},\mathbf{x}')}{\partial \mathbf{x}} = \frac{1}{\pi\sigma}\left\{\frac{\partial k^{(l-1)}(\mathbf{x},\mathbf{x}')}{\partial \mathbf{x}} J(\theta^{(l-1)}) + k^{(l-1)}(\mathbf{x},\mathbf{x}')\frac{\partial J(\theta^{(l-1)})}{\partial \mathbf{x}}\right\} \tag{93}$$

$$\frac{\partial k^{(1)}(\mathbf{x},\mathbf{x}')}{\partial \mathbf{x}} = \frac{\|\mathbf{x}'\|}{\pi\sigma}\left\{\frac{\mathbf{x}}{\|\mathbf{x}\|} J(\theta^{(0)}) + \|\mathbf{x}\|\frac{\partial J(\theta^{(0)})}{\partial \mathbf{x}}\right\} \tag{94}$$

$$\frac{\partial J(\theta^{(l)})}{\partial \mathbf{x}} = \frac{\pi-\theta^{(l)}}{\sin\theta^{(l)}}\left\{\frac{1}{\sqrt{k^{(l)}(\mathbf{x},\mathbf{x})k^{(l)}(\mathbf{x}',\mathbf{x}')}}\frac{\partial k^{(l)}(\mathbf{x},\mathbf{x}')}{\partial \mathbf{x}} - \frac{\cos\theta^{(l)}}{2k^{(l)}(\mathbf{x},\mathbf{x}')}\frac{\partial k^{(l)}(\mathbf{x},\mathbf{x})}{\partial \mathbf{x}}\right\} \tag{95}$$

$$\frac{\partial J(\theta^{(0)})}{\partial \mathbf{x}} = \frac{\pi-\theta^{(0)}}{\sin\theta^{(0)}}\left\{\frac{1}{\|\mathbf{x}\|\|\mathbf{x}'\|}\mathbf{x}' - \frac{\|\mathbf{x}'\|\cos\theta^{(0)}}{\|\mathbf{x}\|^2}\mathbf{x}\right\} \tag{96}$$

$$\frac{\partial k^{(l)}(\mathbf{x},\mathbf{x})}{\partial \mathbf{x}} = \frac{1}{\pi\sigma}\left\{\frac{\partial k^{(l-1)}(\mathbf{x},\mathbf{x})}{\partial \mathbf{x}} + k^{(l-1)}(\mathbf{x},\mathbf{x})\frac{\partial J(\theta^{(l-1)})}{\partial \mathbf{x}}\right\} \tag{97}$$

$$\frac{\partial k^{(1)}(\mathbf{x},\mathbf{x})}{\partial \mathbf{x}} = \frac{\sin\theta^{(0)} + (\pi-\theta^{(0)})\cos\theta^{(0)}}{\pi\sigma} \tag{98}$$

5 Experiments

6 Conclusion

文献 [?] [?] [?] [?] [?] [?] [?] [?] [?]