第 45 卷 第 2 期 计 算 机 工 程 2019 年 2 月
Vol. 45 No. 2 Computer Engineering February 2019

·人工智能及识别技术· 文章编号：1000-3428(2019)02-0160-07 文献标志码：A 中图分类号：TP181

基于深度高斯过程的多元类别数据分布估计

刘姝君，李艳婷

（上海交通大学 机械与动力工程学院，上海 200240）

摘 要： 多元类别数据的可能取值会随向量长度的增长呈指数级增长，从而造成数据稀疏性问题。通过将观察数据嵌入到连续空间中训练识别数据之间的相似性，构建多元类别数据的线性高斯隐变量模型和类别隐高斯过程（CLGP）。在 CLGP 模型基础上，建立小样本多元类别数据分布估计的多元类别深度隐高斯过程模型，并结合蒙特卡洛采样的变分推断方法对模型进行参数优化。实验结果表明，与 CLGP 模型相比，该模型分布估计精确度有所提升。

关键词： 多元类别数据；生成式模型；深度高斯过程；无监督学习；变分推断

中文引用格式： 刘姝君，李艳婷. 基于深度高斯过程的多元类别数据分布估计[J]. 计算机工程，2019，45(2)：160-166.

英文引用格式： LIU Shujun，LI Yanting. Multivariate categorical data distribution estimation based on deep Gaussian process[J]. Computer Engineering，2019，45(2)：160-166.

Multivariate Categorical Data Distribution Estimation Based on Deep Gaussian Process

LIU Shujun，LI Yanting

（School of Mechanical Engineering，Shanghai Jiaotong University，Shanghai 200240，China）

【Abstract】 The possible value of multivariate categorical data increases exponentially with the length of the vector，resulting in data sparsity. The similarity between the identified data is trained by embedding the observation data into the continuous space，and the linear Gaussian hidden variable model and the Categorical Latent Gaussian Process（CLGP）of the multi-category data are constructed. Based on the CLGP model，a multi-class deep hidden Gaussian process model for small sample multi-class data distribution estimation is proposed，and the parameters are optimized by Monte Carlo sampling. Experimental results show that compared with the CLGP model，this model distribution estimation accuracy has improved.

【Key words】 multivariate categorical data；generative model；Deep Gaussian Process（DGP）；unsupervised learning；variational inference

DOI： 10.19678/j.issn.1000-3428.0049671

0 概述

高维多元类别数据的学习与推断是机器学习领域中重要的问题之一，其广泛应用于金融行业的离散选择模型、社会关系调查响应分析、推荐系统、自然语言处理[1]以及医疗诊断等子领域。

近年来，基于类别变量向量相似度原理在大规模有标签数据的研究中取得一定的进展[2-3]。基于大规模有标签数据集利用贝叶斯非参数模型作为非线性转换函数，进行小规模无标签数据集的分布估计。现有的离散有标签数据的监督式模型将观察数据嵌入到一个连续的空间中，从而获得类别变量之间的相似性。为了对稀疏多模态类别数据的分布建模，首先利用一个简单的隐空间建立连续性分布，然后对空间中的点做非线性转换得到概率分布。文献[4]对隐空间的先验分布设置为标准正态分布，并且将稀疏高斯过程非线性转换的输出输送到 Softmax 函数中得到输出概率。

在文献[4]基础上，本文将高斯过程非线性转换拓展到深度高斯过程（Deep Gaussian Precess，DGP）非线性转换，从而得到针对小批量无标签数据分布估计的生成式深度高斯过程模型。

1 相关工作

对于多元类别数据分布估计，文献[5]提出利用线性高斯模型（Linear Gaussian Model，LGM）在隐空间中建立标准正态分布先验分布，然后进行线性转换，并输出到 Softmax 似然函数

基金项目： 国家自然科学基金面上项目“多元复杂时空数据建模与监控方法研究”（71672109）。

作者简介： 刘姝君（1994—），女，硕士，主研方向为智能故障诊断、贝叶斯机器学习；李艳婷，副教授、博士。

收稿日期： 2017-12-12 **修回日期：** 2018-01-30 **E-mail：** liushujun_uestc@163.com

中。针对线性模型表达能力的不足,文献[4]提出一种类别隐高斯过程(Categorical Latent Gaussian Process,CLGP)模型。高斯过程模型[6-7]是一种灵活的非参数贝叶斯模型,应用于机器人[8-9]、地理统计[10-11]、数值运算[12]、激励感知[13-14]以及参数优化问题[15]等领域。由于高斯过程的非线性,CLGP 模型较线性版本的 LGM 更加灵活,能够较好地学习多模态类别数据的特征。然而相对于深度高斯过程模型,单层高斯过程模型仍然存在学习能力较弱的缺陷。对于大规模类别数据的表示学习,深度神经网络如自编码[16]或长短时记忆取得较好的效果。但这些模型在小批量类别性数据集上表现并不好[17],因为深度神经网络包含大量参数,需要大量数据做参数优化。而高斯过程模型由于其完全贝叶斯的处理方式,因此不易过拟合,能够有效处理小批量类别数据集,同时能够提供预测结果的不确定度。本文所提出的类别隐深度高斯过程(Categorical Latent Deep Gaussian Process,CLDGP)模型可以看成深度高斯过程[18]的无监督学习模型版本,也可以看成类似于深度高斯过程模型对于单层高斯过程模型拓展的对于类别隐高斯过程模型的延伸。

在组合核函数、深度核函数高斯过程模型[19-21]及深度高斯过程模型的组合型架构中,基于过程组合的深度高斯过程模型[22]通常采用近似推断的方法,如变分推断[25-26]、期望传播来进行参数优化。由于基础均值场变分推断法[27-28]中变量独立性假设过强、变分参数线性增长的缺陷,利用变分自编码的方法对隐层变量后验分布用认知模型构建[26],采用核函数稀疏谱的方法做稀疏化处理[28]等方法相继提出。文献[29]提出新的基于伪点利用准确模型作为变分后验分布的推断方法,并且结合蒙特卡洛随机采样与随机梯度下降法将双重随机变分推断应用到深度高斯过程模型中。本文将利用蒙特卡洛采样的变分推断法应用到无监督的类别隐深度高斯过程模型中,同时采用单层高斯过程模型中对核函数噪声参数的融合处理的方法减少变分参数数量。

关于各种基本模型与本文模型之间的关系结构如图 1 所示,其中,箭头旁标注表示 4 个不同方向的模型变化,横向表示线性模型至非线性模型的变化,同理纵向 1 层至 2 层,2 层至 3 层以及前向的含义。

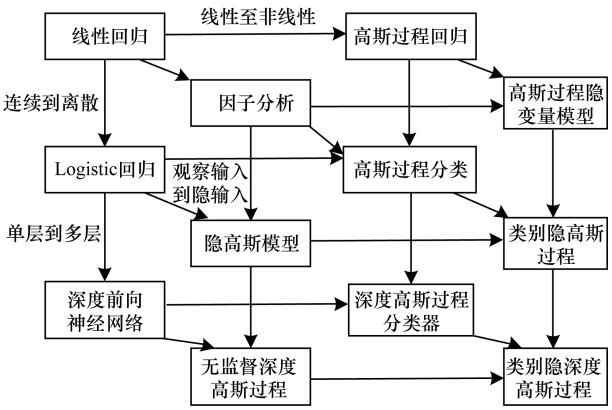

图 1　各模型与本文模型之间的关系结构

2　小样本多元类别数据的深度高斯过程模型

2.1　深度高斯过程模型

问题目标为在给定多元类别数据集 Y 的情况下对数据集进行分布估计,即给出数据集中不同样本数据的分布概率 $p(y)$ 进行数据集分布估计。主要思路通过建立隐输入及深度高斯过程变换中间层及 Softmax 输出层,并对模型进行变分推断参数优化,得到隐变量分布及中间变换层参数,从而得到数据分布生成过程。

假设存在数据集 $\boldsymbol{Y}$ 包含 N 个观察数据以及 D 个类别变量,即第 n 个观察数据 $\boldsymbol{y}_n=(y_{n1},y_{n2},\cdots,y_{nD})$,$n\in\{1,2,\cdots,N\}$。第 n 个观察数据中的第 d 个类别变量 y_{nd} 是一个取值范围为 0 到 K_d 的类别变量。为了表示方便,现假设所有的类别变量都有相同的 K_d,即 $K_d\equiv K,\forall\, d=1,2,\cdots,D$。

每一个类别变量 y_{nd} 的类别分布概率由一个 Softmax 函数给出,Softmax 函数的输入为权重 $\boldsymbol{f}_{nd}^{L}=(0,f_{nd1}^{L},\cdots,f_{ndK}^{L})$,每一个权重 $f_{ndk}^{L}(k\in\{1,2,\cdots,K\})$ 为一个 L 层深度高斯过程的第 L 层输出。深度高斯过程模型的输入是一个 Q 维且服从标准正态分布的隐变量 $\boldsymbol{x}_n=(x_{n1},x_{n2},\cdots,x_{nQ})\in\mathbb{R}^{Q}$,为表示方便,本文将输入 $\boldsymbol{x}$ 用深度高斯过程模型第 0 层 $\boldsymbol{F}^{0}$ 来表示,$\boldsymbol{F}_n^{0}=(f_{n1}^{0},f_{n2}^{0},\cdots,f_{nQ}^{0})$,具体模型结构如式(1)~式(6)所示。

$$f_{nq}^{0}\overset{iid}{\sim}N(0,\sigma_x^2),q\in\{1,2,\cdots,Q\} \tag{1}$$

$$\mathcal{F}^{1}\sim GP(\boldsymbol{\mu}^{1},K(\cdot)^{1}),\boldsymbol{F}^{1}=\mathcal{F}^{1}(\boldsymbol{F}^{0}),\boldsymbol{U}^{1}=\mathcal{F}^{1}(\boldsymbol{Z}^{0}),\ \boldsymbol{F}^{1}\in\mathbb{R}^{N\times D^{1}},\boldsymbol{Z}^{0}\in\mathbb{R}^{M\times Q},\boldsymbol{U}^{1}\in\mathbb{R}^{M\times D^{1}} \tag{2}$$

$$\mathcal{F}^{l}\sim GP(\boldsymbol{\mu}^{l},K(\cdot)l),\boldsymbol{F}^{l}=\mathcal{F}^{l}(\boldsymbol{F}^{l-1}),\boldsymbol{U}^{l}=\mathcal{F}^{l}(\boldsymbol{Z}^{l-1}),\ \boldsymbol{F}^{l}\in\mathbb{R}^{N\times D^{l}},\boldsymbol{Z}^{l-1}\in\mathbb{R}^{M\times D^{l-1}},\boldsymbol{U}^{l}\in\mathbb{R}^{M\times D^{l}} \tag{3}$$

$$\mathcal{F}^{L-1}\sim GP(\boldsymbol{\mu}^{L-1},K(\cdot)L-1),\boldsymbol{F}^{L-1}=\mathcal{F}^{L-1}(\boldsymbol{F}^{L-2}),\boldsymbol{U}^{L-1}=\mathcal{F}^{L-1}(\boldsymbol{Z}^{l-2}),\ \boldsymbol{F}^{L-1}\in\mathbb{R}^{N\times D^{L-1}},\boldsymbol{Z}^{L-1}\in\mathbb{R}^{M\times D^{L-2}},\ \boldsymbol{U}^{L-1}\in\mathbb{R}^{M\times D^{L-1}} \tag{4}$$

$$\mathcal{F}_{dk}^{L}\sim GP(\boldsymbol{\mu}^{L},K(\cdot)_{d}^{L}),\boldsymbol{f}_{dk}^{L}=\mathcal{F}_{dk}^{L}(\boldsymbol{F}^{L-1}),\boldsymbol{u}_{dk}^{L}=\mathcal{F}_{dk}^{L}(\boldsymbol{z}^{L-1}),$$
$$\boldsymbol{F}^{L}\in\mathbb{R}^{N\times D\times K},\boldsymbol{z}^{L-1}\in\mathbb{R}^{M\times D^{L-1}},\boldsymbol{U}^{L}\in\mathbb{R}^{M\times D\times K}\quad(5)$$
$$\boldsymbol{y}_{d}\sim Softmax(\boldsymbol{F}_{d}^{L}),Y\in\mathbb{R}^{N\times D\times K}\quad(6)$$

在上述结构中，$\mathcal{F}^{l}$ 为第 l 层 GP 非线性转换函数，$\boldsymbol{\mu}^{l}$、$K(\cdot)^{l}$ 为第 l 层 GP 的均值与核函数。同时，为减小高斯过程模型的计算成本，本文采用文献[30]提出的稀疏伪点法，$\boldsymbol{Z}^{l-1}$ 为第 l 层的伪点输入，$\boldsymbol{U}^{l}$ 为第 l 层的伪点输出。对于第 l 层，由于输出为 d 维，因此有 d 个 GP 模型。模型的概率图模型结构如图 2所示。

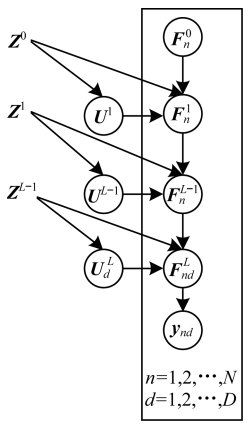

图 2 CLDGP 概率图模型

由于输出数据 $\boldsymbol{Y}$ 的分布取决于第 L 层隐函数变量 $\boldsymbol{F}^{L}$。同时根据单层高斯过程隐变量模型的推断方法，单层隐变量 $\boldsymbol{F}^{L}$ 的分布完全取决于该层伪点 $\boldsymbol{U}^{l}$ 的分布。根据 $\boldsymbol{F}^{L}$ 基于 $\boldsymbol{U}^{l}$ 的条件分布以及 $\boldsymbol{U}^{l}$ 的分布可以得到每一层 $\boldsymbol{F}^{L}$ 与 $\boldsymbol{U}^{l}$ 的联合分布，从而可以得到 DGP 模型的联合分布为：

$$p(\boldsymbol{Y},\{\boldsymbol{F}^{l},\boldsymbol{U}^{l}\}_{l=1}^{L},\boldsymbol{X})=$$
$$p(\boldsymbol{Y}|\boldsymbol{F}^{L})\{\prod_{l=1}^{L}p(\boldsymbol{F}^{l}|\boldsymbol{U}^{l};\boldsymbol{F}^{l-1},\boldsymbol{Z}^{l-1})\cdot$$
$$p(\boldsymbol{U}^{l};\boldsymbol{Z}^{l-1})\}p(\boldsymbol{F}^{0})\quad(7)$$

2.2 CLDGP 模型核函数处理

在 DGP 模型中，通常对每一层 GP 输出专门做参数化处理，本文不对这些变量做参数化处理。根据文献[29]方法，对于每一层的核函数 $K(\cdot)^{l}$，将噪声吸收进核函数中得到噪声化的核函数为：

$$K^{l}=K_{\text{noise}}^{l}(\boldsymbol{x}_{i},\boldsymbol{x}_{j})=k(\boldsymbol{x}_{i},\boldsymbol{x}_{j})+\sigma_{l}^{2}\delta_{ij}\quad(8)$$

其中，δ_{ij} 为克罗内克 δ 函数，σ_{l}^{2} 是层与层之间的噪声方差。当噪声项被参数化时，变分变量上的变分分布为因子化的高斯分布，需要 $2N(D^{1}+D^{2}+\cdots+D^{L-1})$ 个变分参数（D^{l} 表示第 l 层的输出维度），在特定形式的核函数下关于对数边际似然可以得到可处理的变分下界。单独参数化噪声项的变分模型的另一个问题是输出的密度是带有相互独立高斯分布输入的单层高斯过程，会使得变分后验分布丢失所有层与层之间的相关性，因此不能表达整个模型的复杂性，可能会低估方差。本文将噪声项吸收到核函数中，采用保留真实模型的完整条件性结构方式，尽管关于边际似然损失了处理性功能，但是可以通过采样方式进行变分下界的估计。

2.3 基于变分推断的深度高斯过程模型参数优化

变分推断是一种通过优化来近似概率密度的机器学习方法。变分推断的基本思想是假设一个简化的分布族，然后找到分布族中与目标分布最为接近的分布。近似分布与目标分布的接近度用 Kullback-Leibler 距离来度量，即通过优化的方法近似给定观察变量条件下的隐变量的条件密度，通过自由变分参数来参数化隐变量的密度族，转变而成的优化问题将会寻找到分布族的具体变分参数，使得条件分布与目标分布的 KL 距离最小，所拟合的变分分布再被用作准确条件分布的近似来进行模型的预测。变分推断的综述性介绍见文献[31]。

DGP 推断的难点在于层内以及层级之间存在着复杂的相关性。根据文献[29]，本文利用稀疏变分推断的方法来简化层级内的相关性，但保留层级变量之间的相关性，所得到的变分下界并不能得到可处理结果，通过利用一元高斯分布得到变分下界的无偏采样样本。本文采用一个带有 3 个性质的后验分布，具体分析如下：

1）每一层隐函数变量 $\boldsymbol{F}^{l}$ 后验分布基于 $\boldsymbol{U}^{l}$ 保留了准确的模型形式，即没有做任何对先验分布的近似假设处理。

2）假设 $\boldsymbol{U}=(\boldsymbol{U}^{1},\boldsymbol{U}^{l},\cdots,\boldsymbol{U}^{L})$ 的分布在层级以及每一维度之间因子化，所得到的变分后验分布为：

$$q(\{\boldsymbol{F}^{l},\boldsymbol{U}^{l}\}_{l=1}^{L},\boldsymbol{F}^{0})=$$
$$\{\prod_{l=1}^{L}p(\boldsymbol{F}^{l}|\boldsymbol{U}^{l};\boldsymbol{F}^{l-1},\boldsymbol{Z}^{l-1})q(\boldsymbol{U}^{l})\}q(\boldsymbol{F}^{0})\quad(9)$$

3）$q(\boldsymbol{U}^{l})$ 设置为均值为 $\boldsymbol{m}^{l}$、方差为 $\boldsymbol{S}^{l}$ 的高斯分布。与单层稀疏高斯过程回归模型类似，能够在每一层中边际化伪点。边际化以后，得到针对每一层隐函数的变分后验分布为：

$$q(\{\boldsymbol{F}^{l}\}_{l=1}^{L})=\prod_{l=1}^{L}q(\boldsymbol{F}^{l}|\boldsymbol{m}^{l},\boldsymbol{S}^{l};\boldsymbol{F}^{l-1},\boldsymbol{Z}^{l-1})=$$
$$\prod_{l=1}^{L}N(\boldsymbol{F}^{l}|\tilde{\boldsymbol{\mu}}^{l},\tilde{\boldsymbol{\Sigma}}^{l})\quad(10)$$

以单层 GP 为例，$(\boldsymbol{F}^{l},\boldsymbol{U}^{l})$ 的联合变分分布为：

$$q(\boldsymbol{F}^{l},\boldsymbol{U}^{l})=p(\boldsymbol{F}^{l}|\boldsymbol{U}^{l};\boldsymbol{F}^{l-1},\boldsymbol{Z}^{l-1})q(\boldsymbol{U}^{l})\quad(11)$$

根据关于伪点的变分分布假设，$q(\boldsymbol{U}^{l})=N(\boldsymbol{U}^{l}|\boldsymbol{m}^{l},\boldsymbol{S}^{l})$ 可以得到隐函数变分后验分布为：

$$q(\boldsymbol{F}^{l}|\boldsymbol{m}^{l},\boldsymbol{S}^{l};\boldsymbol{F}^{l-1},\boldsymbol{Z}^{l-1})=$$
$$\int p(\boldsymbol{F}^{l}|\boldsymbol{U}^{l};\boldsymbol{F}^{l-1},\boldsymbol{Z}^{l-1})q(\boldsymbol{U}^{l})\mathrm{d}\boldsymbol{U}^{l}=$$
$$N(\boldsymbol{F}^{l}|\tilde{\boldsymbol{\mu}}^{l},\tilde{\boldsymbol{\Sigma}}^{l})\quad(12)$$

式(12)可以通过条件高斯分布等式求出结果。

其中,式(12)中的$\tilde{\boldsymbol{\mu}}^{l}$、$\tilde{\boldsymbol{\Sigma}}^{l}$分别如式(13)、式(14)所示。

$$[\tilde{\boldsymbol{\mu}}^{l}]_i = m^{l}(\boldsymbol{f}_i^{l}) + \alpha(\boldsymbol{f}_i^{l})^{\mathrm{T}}(\boldsymbol{m}^{l} - m^{l}(\boldsymbol{Z}^{l-1})) \tag{13}$$

$$[\tilde{\boldsymbol{\Sigma}}^{l}]_{ij} = \Sigma(\boldsymbol{f}_i^{l},\boldsymbol{f}_j^{l}) - \alpha(\boldsymbol{f}_i^{l})^{\mathrm{T}}(K(\boldsymbol{Z}^{l-1},\boldsymbol{Z}^{l-1}) - \boldsymbol{S}^{l})\alpha(\boldsymbol{f}_j^{l})^{\mathrm{T}} \tag{14}$$

式(13)、式(14)中$\alpha(\boldsymbol{f}_i^{l})$展开的表达式为:

$$\alpha(\boldsymbol{f}_i^{l}) = K(\boldsymbol{f}_i^{l},\boldsymbol{Z}^{l-1})K(\boldsymbol{Z}^{l-1},\boldsymbol{Z}^{l-1})^{-1} \tag{15}$$

根据式(15)的结果,对于每一层 GP 中的第 i 个变量的边际分布,有:

$$q(\boldsymbol{f}_i^{l}|\boldsymbol{m}^{l},\boldsymbol{S}^{l};\boldsymbol{F}^{l-1},\boldsymbol{Z}^{l-1}) = q(\boldsymbol{f}_i^{l}|\boldsymbol{m}^{l},\boldsymbol{S}^{l};\boldsymbol{f}_i^{l-1},\boldsymbol{Z}^{l-1}) = N(\boldsymbol{f}_i^{l}|[\tilde{\boldsymbol{\mu}}^{l}]_i,[\tilde{\boldsymbol{\Sigma}}^{l}]_{ij}) \tag{16}$$

含有 $\boldsymbol{d}$ 元变量的最后一层即 L 层伪点变分分布假设为:

$$q(\boldsymbol{U}^{L}) = \prod_{d=1}^{D}\prod_{k=1}^{K} N(\boldsymbol{U}_{dk}^{L}|\boldsymbol{m}_{dk}^{L},\boldsymbol{S}_d^{L}) \tag{17}$$

同理与前 $\boldsymbol{L}-1$ 层的分布推导,可得 $\boldsymbol{L}$ 层隐函数变分分布为:

$$q(\boldsymbol{f}_{id}^{L}|\boldsymbol{m}^{L-1},\boldsymbol{S}^{L-1};\boldsymbol{F}^{L-1},\boldsymbol{Z}^{l-1}) = \prod_{k=1}^{K} N(f_{idk}^{L}|[\tilde{\boldsymbol{\mu}}_d^{L}]_i,[\tilde{\boldsymbol{\Sigma}}_d^{L}]_{ij}) \tag{18}$$

其中:

$$[\tilde{\boldsymbol{\mu}}_d^{L}]_i = \boldsymbol{m}_d^{L}(\boldsymbol{f}_i^{L-1}) + \alpha(\boldsymbol{f}_i^{L-1})^{\mathrm{T}}(\boldsymbol{m}_d^{L} - \boldsymbol{m}_d^{L}(\boldsymbol{Z}^{L-1})) \tag{19}$$

$$\alpha(\boldsymbol{f}_{id}^{L}) = K_d^{L}(\boldsymbol{f}_{id}^{L},\boldsymbol{Z}^{L-1})K_d^{L}(\boldsymbol{Z}^{L-1},\boldsymbol{Z}^{L-1})^{-1} \tag{20}$$

$$[\tilde{\boldsymbol{\Sigma}}_d^{L}]_{ij} = K_d^{L}(\boldsymbol{f}_i^{L-1},\boldsymbol{f}_j^{L-1}) - \alpha(\boldsymbol{f}_{id}^{L-1})^{\mathrm{T}} \cdot (K_d^{L}(\boldsymbol{Z}^{L-1},\boldsymbol{Z}^{L-1}) - \boldsymbol{S}^{L})\alpha(\boldsymbol{f}_{jd}^{L})^{\mathrm{T}} \tag{21}$$

Softmax 输出可表示为:

$$y_{id} \sim \mathrm{Softmax}(\boldsymbol{f}_{id}^{L}) \tag{22}$$

$$\mathrm{Softmax}(y=k;\boldsymbol{F}^{L}) = \mathrm{Categorical}\left(\frac{\exp(\boldsymbol{f}_k^{L})}{1+\sum_{k'=1}^{K}\exp(\boldsymbol{f}_{k'}^{L})}\right) \tag{23}$$

根据式(7)定义的联合分布函数,对输入隐变量、各层函数隐变量以及各层伪点变量积分取对数可以得到数据的对数似然函数 $\ln p(\boldsymbol{Y})$,再依据变分推断的基本思路,结合式(9)~式(23)定义的变分分布,利用 Jensen 不等式可以得到对数似然函数的变分下界 L_{CLDGP}为:

$$\begin{aligned}\ln p(Y) &= \ln\int p(Y|\boldsymbol{F}^{L})\left\{\prod_{l=1}^{L}p(\boldsymbol{F}^{l}|\boldsymbol{U}^{l};\boldsymbol{F}^{l-1},\boldsymbol{Z}^{l-1})\cdot p(\boldsymbol{U}^{l};\boldsymbol{Z}^{l-1})\right\}\cdot p(\boldsymbol{F}^{0})\cdot \mathrm{d}\boldsymbol{F}^{0}\mathrm{d}\{\boldsymbol{F}^{l},\boldsymbol{U}^{l}\}_{l=1}^{L} \geqslant \\ &\int q(\{\boldsymbol{F}^{l},\boldsymbol{U}^{l}\}_{l=1}^{L},\boldsymbol{F}^{0})\ln\frac{p(Y|\boldsymbol{F}^{L})\left\{\prod_{l=1}^{L}p(\boldsymbol{F}^{l}|\boldsymbol{U}^{l};\boldsymbol{F}^{l-1},\boldsymbol{Z}^{l-1})p(\boldsymbol{U}^{l};\boldsymbol{Z}^{l-1})\right\}p(\boldsymbol{F}^{0})}{q(\{\boldsymbol{F}^{l},\boldsymbol{U}^{l}\}_{l=1}^{L},\boldsymbol{F}^{0})} \\ &p(\boldsymbol{F}^{0})\mathrm{d}\boldsymbol{F}^{0}\mathrm{d}\{\boldsymbol{F}^{l},\boldsymbol{U}^{l}\}_{l=1}^{L} = L_{\mathrm{CLDGP}}\end{aligned} \tag{24}$$

将变量变分分布代入式(24),同时对表达式进行拆分可以得到 3 项表达式:第 1 项是关于条件似然函数的积分,第 2 项和第 3 项为关于输入隐变量与各层伪点变量的变分分布与真实分布的 KL 距离项。具体可表示为:

$$\begin{aligned}L_{\mathrm{CLDGP}} = &\int q(\{\boldsymbol{F}^{l},\boldsymbol{U}^{l}\}_{l=1}^{L},\boldsymbol{F}^{0})\ln p(\boldsymbol{Y}|\boldsymbol{F}^{L})\cdot \\ &\mathrm{d}\boldsymbol{F}^{0}d\{\boldsymbol{F}^{l},\boldsymbol{U}^{l}\}_{l=1}^{L} + \int q(\{\boldsymbol{F}^{l},\boldsymbol{U}^{l}\}_{l=1}^{L},\boldsymbol{F}^{0})\cdot \\ &\ln\frac{\prod_{l=1}^{L}p(\boldsymbol{U}^{l};\boldsymbol{Z}^{l-1})}{q(\{\boldsymbol{U}^{l}\}_{l=1}^{L})}\mathrm{d}\boldsymbol{F}^{0}\mathrm{d}\{\boldsymbol{F}^{l},\boldsymbol{U}^{l}\}_{l=1}^{L} + \\ &\int q(\boldsymbol{F}^{0})\ln\frac{p(\boldsymbol{F}^{0})}{q(\boldsymbol{F}^{0})}\mathrm{d}\boldsymbol{F}^{0} = \\ &\sum_{i=1}^{N}\sum_{d=1}^{D}\int q(\boldsymbol{f}_i^{L-1})q(\boldsymbol{U}_d^{L})p(\boldsymbol{f}_{id}^{L}|\boldsymbol{f}_i^{L-1},\boldsymbol{U}_d^{L})\cdot \\ &\ln p(\boldsymbol{y}_{id}|\boldsymbol{f}_{id}^{L})\mathrm{d}\boldsymbol{f}_i^{L-1}\mathrm{d}\boldsymbol{f}_{id}^{L}\mathrm{d}\boldsymbol{U}_d^{L} - \\ &KL(q(\boldsymbol{F}^{0})\,\|\,p(\boldsymbol{F}^{0})) - \sum_{i=1}^{L}KL(q(\boldsymbol{U}^{l})\,\|\,p(\boldsymbol{U}^{l}))\end{aligned} \tag{25}$$

对于输入隐变量的变分分布同样假设为相互独立的正态分布,有:

$$q(\boldsymbol{F}_0) = q(X) = \prod_{i=1}^{N}\prod_{q=1}^{Q} N(x_{iq}|\mu_{iq},\sigma_{iq}^{2}) \tag{26}$$

在得到变分参数的变分下界后,通常采用梯度下降法优化变分参数。为了以低方差蒙特卡洛法估计变分下界的梯度,本文采用再参数化技巧。再参数化使得随机性并不取决于求取梯度的参数,而是通过再参数化的形式引入另一个随机源。

对于输入隐变量,再参数化形式可表示为:

$$x_{iq} = \mu_{iq} + \sigma_{iq}\varepsilon_{iq}^{0},\ \varepsilon_{iq}^{0} \sim N(0,1) \tag{27}$$

变量 x_{iq}的随机性部分全部转移到服从标准正态分布的参数 ε_{iq}^{0}上。在变分后验分布的求取过程中,可以直接通过 ε_{iq}^{0}采样得到 x_{iq}的采样结果。同理,对于隐函数变量,由于第 l 层第 i 个变量的变分后验分布仅取决于第前 $l-1$ 层第 i 个变量的边际分布,即:

$$q(\boldsymbol{f}_i^{l}) = \int\prod_{j=1}^{l-1}q(\boldsymbol{f}_i^{j}|\boldsymbol{m}^{j},\boldsymbol{S}^{j};\boldsymbol{f}_i^{j-1},\boldsymbol{Z}^{j-1})\mathrm{d}\boldsymbol{f}_i^{j} \tag{28}$$

每一层隐函数变量引入随机参数 ε_i^{l},可以从第 0 层开始对隐函数变量逐层向前采样,有:

$$\hat{\boldsymbol{f}}_i^{l} = [\tilde{\boldsymbol{\mu}}^{l}]_i + \varepsilon_i^{l}\sqrt{[\tilde{\boldsymbol{\Sigma}}^{l}]_{ii}} \tag{29}$$

式(29)中的$[\tilde{\boldsymbol{\mu}}^{l}]_i$、$[\tilde{\boldsymbol{\Sigma}}^{l}]_{ii}$是关于第 $l-1$ 层输出的函数,具体表达式如式(13)、式(14)。对于最后一层,伪点再参数化方法类似,引入随机参数 $\boldsymbol{\varepsilon}_{dk}^{L(u)}$、$\boldsymbol{\varepsilon}_{idk}^{L(f)}$。

$$\boldsymbol{U}_{dk}^{L}=\boldsymbol{m}_{dk}^{L}+\boldsymbol{L}_{d}^{L}\boldsymbol{\varepsilon}_{dk}^{L(u)},\boldsymbol{\varepsilon}_{dk}^{L(u)}\sim N(0,\boldsymbol{I}_{M}) \quad (30)$$

其中，$\boldsymbol{L}_{d}^{L}$ 为 $\boldsymbol{S}_{d}^{L}$ 的 Cholesky 分解表达式 $\boldsymbol{S}_{d}^{L}=\boldsymbol{L}_{d}^{L}(\boldsymbol{L}_{d}^{L})^{\mathrm{T}}$

$$f_{idk}^{L}=[\tilde{\boldsymbol{\mu}}_{d}^{L}]_{i}+\sqrt{[\tilde{\boldsymbol{\Sigma}}_{d}^{L}]_{ii}}\varepsilon_{idk}^{L(f)},\varepsilon_{idk}^{L(f)}\sim N(0,1) \quad (31)$$

本文利用随机梯度下降法优化得到最优变分分布。带有噪声梯度的梯度下降法在给定的学习率下可以收敛到局部最优，但在实际中很难操作。学习率的初始值集合会在算法收敛处影响学习率，错误设置的初始值会导致偏离。基于上述原因，一些新的算法被提出来处理噪声梯度，如 AdaGrad[32]、RMSPROP[33]等优化算法，这些算法在梯度处理上均略有不同。文献[34]在不同的单元检测上对这些不同的优化技巧做对比，表明 RMSPROP 在多数检验集合上相较其他的优化方法都有较好的表现。因此，本文在变分下界的优化中选择 RMSPROP 算法。本文利用带有自动求导机制的 Tensorflow 平台来进行模型推断、参数优化以及模型预测。模型的推断流程如图 3 所示。

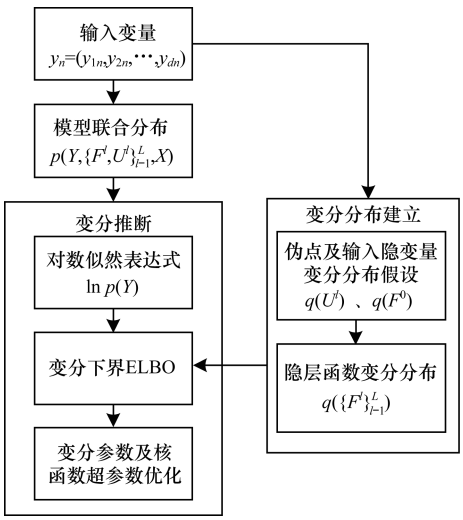

图 3 CLDGP 模型变分推断流程

2.4 分布估计整体步骤

分布估计过程的具体步骤如下：

步骤 1 建立训练集 $\boldsymbol{Y}$ 及检验集数据集 $\boldsymbol{Y}^{*}$。为与现有文献研究结果进行对比，采用将检验数据集 $\boldsymbol{Y}^{*}$ 随机抽取部分元素设为缺失值，即 $\boldsymbol{Y}^{*}=(\boldsymbol{Y}_{U}^{*},\boldsymbol{Y}_{O}^{*})$。其中，$\boldsymbol{Y}_{U}^{*}$ 为缺失部分，$\boldsymbol{Y}_{O}^{*}$ 为观测部分。

步骤 2 根据第 2.1 节和第 2.2 节所述建立模型，将部分观测点的输入隐变量与训练集数据结合，利用本文提供的变分推断架构训练。

步骤 3 对检验数据集中的缺失部分进行分布估计，得到 $P(\boldsymbol{Y}_{U}^{*}\mid\boldsymbol{Y}_{O}^{*},\boldsymbol{Y}^{*})$。

步骤 4 利用分布估计指标评估效果，进行性能对比。

3 实例结果与分析

为与 CLGP 模型进行性能比对，本文采用文献[4]所采用的部分代表性数据集以及 CLGP 模型的结果进行实验对比。

在实验中，LGM 模型、CLGP 模型、以及本文提出的 CLDGP 模型都以二维隐空间初始化，CLDGP 模型的层数选择为 3 层。隐输入的均值 μ_i 以标准正态分布进行随机初始化，所有层的伪输出 $\boldsymbol{m}_{dk}^{l}$ 以标准偏差为 10^{-2} 的正态分布随机初始化，这种初始化方法等价于对所有值使用一个均匀初始分布。

每一层隐函数的标准偏差初始化为 0.1，自动相关决策径向基函数协方差函数的尺度参数初始化为 0.1，变分分布优化迭代 500 次，每一次迭代优化所有参数变量但是保持伪点 $\boldsymbol{u}_{dk}^{l}$ 的变分参数固定，然后再优化伪点 $\boldsymbol{u}_{dk}^{l}$ 的变分参数且保持其他量固定不变。

本文模型可用于带有部分观察数据的半监督学习，部分观察点的隐部分通过训练集合来进行优化，然后预测缺失值。本文采用文献[4]和文献[5]相同的评估指标困惑度[35]作为评估模型性能的误差指标。困惑度是一种常用的用来评估样本数据概率分布估计效果的统计量，该统计量定义为负平均对数预测概率的指数值，即对于所预测的 d 元输出概率 $p(y_{nd}),n=1,2,\cdots,N,d=1,2,\cdots,D$ 的困惑度指标定义为 $2^{-\sum_{n=1}^{N}\frac{1}{N}\mathrm{lb}\,p(y_{nd})}$。例如，二元数据正确预测结果的误差为 0，随机猜测即对每一个数据点的每一种二元值预测概率 0.5 会使得困惑度误差为 1，而对任意正确类别值给予概率 0 的困惑度误差为 ∞。

本文实验均用 Python 语言，高斯过程模型建立在 Tensorflow[36]平台上的 GPflow[37]平台，模型参数优化辅助利用 Tensorflow 的自动求导方法。

3.1 小批量多元类别手写字母数据集

数据来源于文献[38]，其中，手写字母数据集包含 10 个手写数字（0～9）以及 26 个手写大写字母（A～Z），每一张图像分辨率为 20×16 像素，每一个类别包含 39 张图像。将图像重置为 10×8 像素可以得到 1 404（39×36）像素个 80 维的变量数据点。图 4所示为 10 个手写数字与部分手写字母的示例。

图 4 二元手写数字字幕数据集示例

本文将 36 个类别的图像分别划分为 30 个训练样本与 9 个检验样本。对于检验样本集合，随机移除 20% 的像素做模型检验，然后训练模型，评估预测误差。

图 5 所示为基于困惑度的检验误差在训练集以及检验集上的变化结果。CLGP 模型收敛速度比 CLDGP 模型快,CLGP 模型最终预测误差比 CLDGP 模型高,CLGP 模型的训练集检验集误差分别为 0.634、0.705,CLDGP 模型的训练集检验集误差分别为 0.571、0.658。图 6 所示为36 种类别数据点在优化以后二元隐空间上的投影,即将不同类别数据点以不同颜色形状区分投影到模型优化所得的二元输入隐变量空间中,不同类别数据区分度越大表示隐空间优化效果越好。从图 6 可以看出,CLDGP 模型对于不同类别的数据点区分度要大一些,即区别能力比 CLGP 模型好。

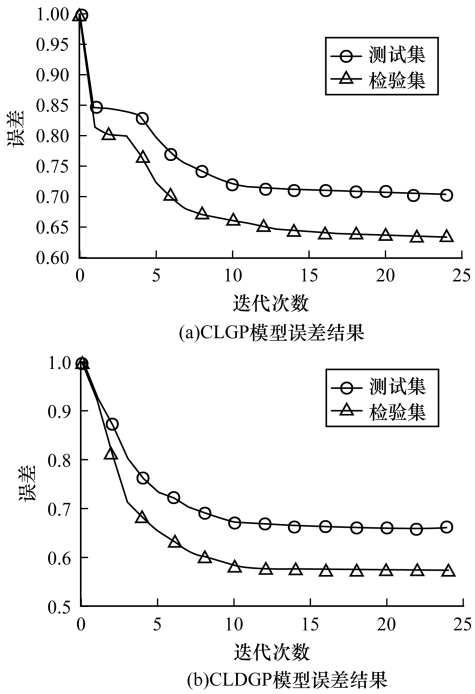

图 5　2 种模型误差在测试集和检验集上的对比结果

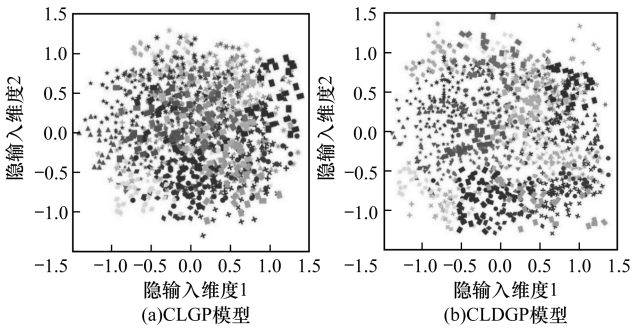

图 6　二元隐空间分类绘图

3.2　小批量多元类别医疗诊断数据集

威斯康星肺癌数据集[39]由 683 个数据点组成,带有 9 个取值在 1 ~ 10 的类别变量,以及一个取值 0、1 的二值类别变量,共 2×10^9 种可能的配置结果。本文利用 75% 数据集作为训练数据集,剩余作为检验数据集,实验重复 3 次,且对数据做 3 次随机划分。在检验集合中,随机移除 10 个类别变量中的 1 个,然后检验模型对数据的恢复能力。

本文将 CLGP、基准模型、LGM 模型以及 CLDGP 模型作对比。表 1 所示为 3 种模型在 3 种数据划分情况下的检验集数据误差结果。从表 1 可以看出,基准模型的检验集合困惑度最高,尤其在数据集合 3 上出现了较大的分布估计误差,CLGP 模型在 3 项划分数据集上较 LGM 模型均有所提升,CLDGP 模型在检验数据困惑度与估计结果方差上性能都比较好。

表 1　3 种模型在 3 种数据下检验集数据误差结果

划分集合	LGM 模型	CLGP 模型	CLDGP 模型
1	3.57 ±0.208	2.86 ±0.119	1.99 ±0.107
2	3.47 ±0.252	3.36 ±0.186	2.21 ±0.098
3	12.13 ±9.705	3.34 ±0.096	2.79 ±0.170

4　结束语

为提高类别隐高期模型的表达能力,本文通过深度高斯过程作非线性变换,建立生成式模型对多元类别数据的分布进行估计。利用小批量手写数字字母、医疗诊断数据集、类别隐高斯过程模型以及隐高斯模型的性能进行了对比,验证了模型的有效性。下一步将在参数化的过程中加入控制协变量,以提高参数估计稳定性,同时将神经网络与高斯过程模型相结合,以提高模型容量。

参考文献

[1] 王盛玉,曾碧卿,胡翩翩. 基于卷积神经网络参数优化的中文情感分析[J]. 计算机工程,2017,43(8):200-207,214.

[2] BENGIO Y, SCHWENK H, SENÉCAL J S, et al. Neural probabilistic language models[M]. In Innovations in Machine Learning. Berlin, Germany: Springer, 2006: 137-186.

[3] COLLOBERT R, WESTON J. A unified architecture for natural language processing: deep neural networks with multitask learning[C]//Proceedings of the 25th International Conference on Machine learning. New York, USA: ACM Press,2008:160-167.

[4] GAL Y, CHEN Y, ZOUBIN G. Latent Gaussian process for distribution estimation of multivariate categorical data [EB/OL]. [2017-11-18]. https://arxiv.org/pdf/1503.02182.pdf

[5] KHAN M E, MOHAMED S, MARLIN B R, et al. A stick-breaking likelihood for categorical data analysis with latent Gaussian models[EB/OL]. [2017-11-18]. https://www.shakirm.com/papers/catLGM-AIstats 20 12.pdf.

[6] 何志昆,刘光斌,赵曦晶,等. 高斯过程回归方法综述[J]. 控制与决策,2013,28(8):1121-1129,1137.

[7] RASMUSSEN C E, WILLIAM K I. Gaussian process for machine learning[EB/OL]. [2017-11-18]. http://www.gaussianprocess.org/gpml/.

[8] KO J, FOX D. GP-Bayes filters: Bayesian filtering using Gaussian process prediction and observation models[C]// Proceedings of IEEE/RSJ Intelligent Robots and Systems. Washington D.C., USA: IEEE Press, 2008: 3471-3476.

[9] DEISENROTH M P, RASMUSSEN C E. PILCO: a model-based and data-efficient approach to policy search[C]// Proceedings of the 28th International Conference on Machine Learning. [S.l.]: Omnipress, 2011: 465-472.

[10] CRESSIE N, WIKLE K. Statistics for spatio-temporal data[EB/OL]. [2017-11-18]. https://www.wiley.com/en-us/Statistics+for+Spatio+Temporal+Data-p-9780471692744.

[11] 王鑫, 李红丽. 台风最大风速预测的高斯过程回归模型[J]. 计算机应用研究, 2015, 32(1): 59-62.

[12] BRIOL F X, OATES C J, GIROLAMI M, et al. Probabilistic integration: a role for statisticians in numerical analysis? [EB/OL]. [2017-11-18]. https://arxiv.org/pdf/1512.00933v5.pdf.

[13] GUESTRIN C, KRAUSE A, SINGH A P. Near-optimal sensor placements in Gaussian processes[C]//Proceedings of the 22nd International Conference on Machine Learning. New York, USA: ACM Press, 2005: 265-272.

[14] 孙晓燕, 陈姗姗, 巩敦卫, 等. 基于区间适应值交互式遗传算法的加权多输出高斯过程代理模型[J]. 自动化学报, 2014, 40(2): 172-184.

[15] SNOEK J, LAROCHELLE H, ADAMS R P. Practical bayesian optimization of machine learning algorithms[C]// Proceedings of the 25th International Conference on Neural Information Processing Systems [S.l.]: Curran Associates Inc., 2012: 2951-2959.

[16] REZENDE D J, MOHAMED S, WIERSTRA D. Stochastic back propagation and approximate inference in deep generative models[J]. Pattern Recognition and Machine Learning, 2014, 32(2): 1278-1286.

[17] GHAHRAMANI Z. Probabilistic machine learning and artificial intelligence[J]. Nature, 2015, 521: 452-459.

[18] DAMIANOU A C, LAWRENCE N D. Deep Gaussian processes[EB/OL]. [2017-11-18]. https://core.ac.uk/download/pdf/46564399.pdf.

[19] WILSON A G, HU Z, SALAKHUTDINOV R, et al. Deep kernel learning[EB/OL]. [2017-11-18]. https://arxiv.org/pdf/1511.02222.pdf.

[20] DURRANDE N, GINSBOURGER D, ROUSTANT O. Additive kernels for Gaussian process modeling[EB/OL]. [2017-11-18]. https://arxiv.org/pdf/1103.4023.pdf.

[21] DAVID D, JAMES R L, ROGER G, et al. Structure discovery in nonparametric regression through compositional kernel search[EB/OL]. [2017-11-18]. http://www.cs.toronto.edu/~rgrosse/icml2013-gp.pdf.

[22] HENSMAN J, LAWRENCE N D. Nested variational compression in deep Gaussian processes[EB/OL]. [2017-11-18]. https://arxiv.org/pdf/1412.1370.pdf.

[23] VAFA K. Training deep Gaussian processes with sampling[EB/OL]. [2017-11-18]. http://approximateinference.org/accepted/Vafa2016.pdf.

[24] WANG Y, BRUBAKER M, CHAIB-DRAA B, et al. Sequential inference for deep Gaussian process[EB/OL]. [2017-11-18]. http://proceedings.mlr.press/v51/wang16c.pdf.

[25] BUI T D, HERNÁNDEZ-LOBATO D, LI Y, et al. Deep Gaussian processes for regression using approximate expectation propagation[EB/OL]. [2017-11-18]. http://proceedings.mlr.press/v48/bui16.pdf.

[26] DAI Z, DAMIANOU A, GONZÁLEZ J, et al. Variational auto-encoded deep Gaussian processes[EB/OL]. [2017-11-18]. https://arxiv.org/pdf/1511.06455.pdf.

[27] DAMIANOU A D, TITSIAS M K, LAWRENCE N D. Variational Gaussian process dynamical systems[EB/OL]. [2017-11-18]. http://papers.nips.cc/paper/4330-variational-gaussian-process-dynamical-systems.pdf.

[28] CUTAJAR K, BONILLA E V, MICHIARDI P, et al. Practical learning of deep Gaussian processes via random fourier features[EB/OL]. [2017-11-18]. https://pdfs.semanticscholar.org/bafa/7e2d586e7bfe77d9a55ac1cff4eb2f6ff292.pdf.

[29] HUGH S, MARC D. Doubly stochastic variational inference for deep Gaussian processes[EB/OL]. [2017-11-18]. https://arxiv.org/pdf/1705.08933.pdf.

[30] TITSIAS M K. Variational learning of inducing variables in sparse Gaussian processes[EB/OL]. [2017-11-18]. http://proceedings.mlr.press/v5/titsias09a/titsias09a.pdf.

[31] BLEI D M, KUCUKELBIR A, MCAULIFFE J D. Variational inference: a review for statisticians[EB/OL]. [2017-11-18]. https://arxiv.org/pdf/1601.00670.pdf.

[32] JOHN D, ELAD H, YORAM S. Adaptive subgradient methods for online learning and stochastic optimization[J]. Journal of Machine Learning Research, 2011, 12: 2121-2159.

[33] TIELEMAN T, HINTON G. Lecture 6.5- rmsprop, COURSERA: neural networks for machine learning[EB/OL]. [2017-11-18]. http://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf.

[34] TOM S, IOANNIS A, DAVID S. Unit tests for stochastic optimization[EB/OL]. [2017-11-18]. https://arxiv.org/pdf/1312.6055.pdf.

[35] Perplexity[G/OL]. [2017-11-18]. https://en.wikipedia.org/wiki/Perplexity.

[36] Tensorflow[EB/OL]. [2017-11-18]. https://tensorflow.google.cn/.

[37] MATTHEWS A G D G, MARK V D W, NICKSON T, et al. GPflow: a gaussian process library using tensorflow[EB/OL]. [2017-11-18]. http://adsabs.harvard.edu/abs/2016arXiv161008733M.

[38] Handwritten digits[DB/OL]. [2017-11-18]. https://cs.nyu.edu/~roweis/data.html.

[39] MATJAZ Z, MILAN S. Breast cancer data set[DB/OL]. [2017-11-18]. http://archive.ics.uci.edu/ml/datasets/Breast+Cancer.

编辑 赵 辉