Stochastic Deep Gaussian Processes over Graphs

Naiqi Li[1] Wenjie Li[2] Jifeng Sun[1] Yinghua Gao[2] Yong Jiang[1,3], Shu-Tao Xia[2,3]

[1]Tsinghua-Berkeley Shenzhen Institute, Tsinghua University [2]Shenzhen International Graduate School, Tsinghua University

[3]PCL Research Center of Networks and Communications, Peng Cheng Laboratory

{ lnq18, liwj20, sjf19, yh-gao18}@mails.tsinghua.edu.cn {jiangy, xiast}@sz.tsinghua.edu.cn

http://github.com/naiqili/DGPG

Introduction

In this paper we propose Stochastic Deep Gaussian Processes over Graphs (DGPG), which are deep Gaussian models that learn the mappings between input and output signals in graph domains.
We summarize the contributions as follows:

- Present a deep Gaussian model (DGPG) that utilize graph information
- Propose a sampling method for optimizing the evidence lower bound
- Prove that the sampling variances of DGPG are strictly smaller than without using graph knowledge
- Show the superior performance of DGPG on various graph domains
- DGPG can model predictive uncertainties, and the ARD kernel allows it to automatically learn which neighbors are important for the prediction

Problem Statement

- Graph is defined as $G = \langle V, E \rangle$, where V (E) is the set of nodes (edges)
- Input signal: ψ: $V \to R^{d_{in}}$; output signal: ϕ: $V \to R^{d_{out}}$
- Our goal is to learn a function h: $\psi \mapsto \phi$ that maps the input signals to the output signals

Background

Sparse Gaussian Processes

Sparse Gaussian Processes (SGPs) [1] is proposed to assuage the $O(n^3)$ training complexity of traditional Gaussian processes. It introduces M inducing inputs $\mathbf{Z} = (\boldsymbol{z}_1, \dots, \boldsymbol{z}_m)^\top$,and the corresponding function values are $\mathbf{u} = f(\mathbf{Z})$. A variational distribution $q(\mathbf{f}, \mathbf{u})$ is used to approximate the posterior $p(\mathbf{f}, \mathbf{u} \mid \mathbf{y})$ .Parameters are inferred by optimizing the evidence lower bound (ELBO), defined as $L_{SGP} = E_{q(\mathbf{f},\mathbf{u})}[\log \frac{p(\mathbf{y},\mathbf{f},\mathbf{u})}{q(\mathbf{f},\mathbf{u})}]$.

Doubly Stochastic Deep Gaussian Processes

Stochastic Deep Gaussian Process (DGP) [2] stacks SGPs to form a hierarchical deep structure. A DGP model with L layers has corresponding ELBO as $L_{DGP} = \sum_{i=1}^{N} E_{q(\mathbf{f}_i^L)}[\log p(\mathbf{y}_i|\mathbf{f}_i^L)] - \sum_{l=1}^{L} KL[q(\mathbf{U}^l)||p(\mathbf{U}^l; \mathbf{Z}^{l-1})]$.

Methodology

Stochastic Deep Gaussian Processes over Graphs

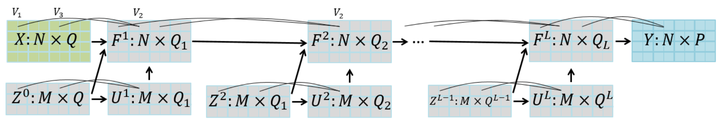

Figure 1. Factorization of the joint distribution.

The joint distribution of the proposed DGPG model is defined as:

$$p\left(\mathbf{Y}, \{\mathbf{F}^{l,k}, \mathbf{U}^{l,k}\}_{l,k}\right) = \underbrace{\prod_{n=1}^{N}\prod_{k=1}^{K} p\left(\mathbf{y}_n^k|\mathbf{F}_n^{L,k}\right)}_{\text{likelihood}} \cdot \underbrace{\prod_{l=1}^{L} p\left(\mathbf{F}^{l,k}|\mathbf{U}^{l,k}; \mathbf{F}^{l-1,pa(k)}, \mathbf{Z}^{l-1,pa(k)}\right) p\left(\mathbf{U}^{l,k}; \mathbf{Z}^{l-1,pa(k)}\right)}_{\text{GP prior}}$$

- $\mathbf{X}$, $\mathbf{Y} \in R^{N \times Kd}$ is the matrix representation of the input/output signal, where N is the number of training instances, K is the number of nodes, and d is the feature dimension per node
- $\mathbf{Z}^l$:inducing points for layer l; $\mathbf{U}^l$:function value of $\mathbf{Z}^l$
- $\mathbf{M}_i^{l,k}$ denotes the ith row, lth layer and kth node of the matrix $\mathbf{M}$

Recursive Sampling Scheme

The ELBO of DGPG is:

$$\mathcal{L}_{DGPG} = \sum_{n=1}^{N}\sum_{k=1}^{K} \mathbb{E}_{q(\mathbf{F}_n^{L,k})}\left[\log p\left(\mathbf{y}_n^k|\mathbf{F}_n^{L,k}\right)\right] - \sum_{l=1}^{L}\sum_{k=1}^{K} \mathrm{KL}\left[q\left(\mathbf{U}^{l,k}\right) \| p\left(\mathbf{U}^{l,k}; \mathbf{Z}^{l-1,pa(k)}\right)\right]$$

The key observation is that by using the reparameterize trick, this ELBO can be achieved by a recursive sampling scheme:

$$\hat{\mathbf{F}}_i^{l,k} = \mu_{\mathbf{m}^{l,k}, \mathbf{Z}^{l-1,pa(k)}}(\hat{\mathbf{F}}_i^{l-1,pa(k)}) + \boldsymbol{\epsilon}_i^{l,k} \odot \sqrt{\Sigma_{\mathbf{S}^{l,k}, \mathbf{Z}^{l-1,pa(k)}}(\hat{\mathbf{F}}_i^{l-1,pa(k)}, \hat{\mathbf{F}}_i^{l-1,pa(k)})}$$

Sampling is performed recursively, since the sampled function at layer l dependents only on its **parent nodes** at layer $l-1$.

Theoretical Analysis of Sampling Variances

The following theorem rigorously shows that under some technical conditions, the sampling variance of DGPG is strictly smaller than its counterpart without utilizing graph information.

Theorem 1. Denote the sampling variance of DGPG as $\tilde{\Sigma}_{ii}$, and the sampling variance of its counterpart without graph knowledge as Σ_{ii}. Under some mild conditions (satisfied by most nontrivial graphs) we have $\tilde{\Sigma}_{ii} < \Sigma_{ii}$.

Experimental Study

Precise Mean Prediction

DGPG can achieve good performance on challenging regression tasks like traffic flow prediction, it's competitive to sophisticated graph neural networks

Table 1. DGPG* utilizes validation data during training. Terms with underline denote best results. Terms with wavy underline indicate second best.

	T	Metrics	VAR	FC-LSTM	DCRNN	DGPG (1/2/3/4)	DGPG*
LA	15 min	MAE	4.42	3.44	2.77	3.06 / 3.04 / 3.02 / 3.02	3.00
		RMSE	7.89	6.30	5.38	5.40 / 5.35 / 5.32 / 5.32	5.31
		MAPE	10.2%	10.9%	7.3%	6.6% / 6.0% / 6.6% / 6.5%	6.5%
	30 min	MAE	5.41	3.77	3.15	3.57 / 3.42 / 3.42 / 3.39	3.39
		RMSE	9.13	7.23	6.45	6.37 / 6.16 / 6.16 / 6.12	6.13
		MAPE	12.7%	10.9	8.8%	7.5% / 7.3% / 7.3% / 7.2%	7.2%
	60 min	MAE	6.52	4.37	3.6	4.02 / 3.83 /3.00 / 3.80	3.8
		RMSE	10.11	8.69	7.59	7.12 / 6.93 / 6.94 / 6.94	6.85
		MAPE	15.8%	13.2%	10.5%	8.4% / 8.1% / 7.9% / 8.0%	5.0%

Accurate Uncertainty Estimation

DGPG can make near-ground uncertainty estimation.

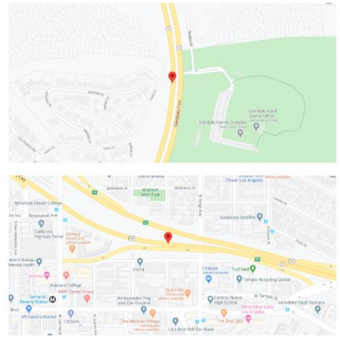

Table 2. Variance analysis. For Gaussian $(\mu - k\sigma, \mu + k\sigma)$ CI covers 68.3%/95.5%/99.7% of the density, so the test instances falling in the predictive intervals have similar ratio.

	T	$\pm 1\sigma$	$\pm 2\sigma$	$\pm 3\sigma$
LA	15 min	84.4%	94.7%	97.7%
	30 min	84.2%	94.5%	97.5%
	60 min	83.7%	94.3%	97.5%
BAYS	15 min	87.4%	95.6%	97.9%
	30 min	86.0%	94.5%	97.2%
	60 min	85.1%	94.1%	97.0%

Figure 2. The sensor with the least uncertainty locates at sparsely populated area with simple traffic condition; the sensor with the largest uncertainty locates at the business center where the traffic condition is very complex.

Automatic Relevance Discovery

DGPG can automatically discover which neighbors and nodes are more important for prediction.

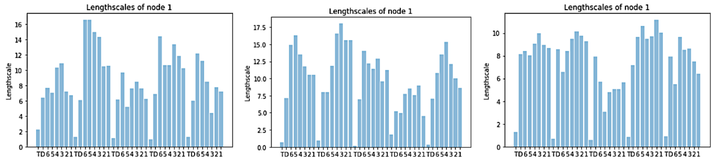

Figure 3. Lengthscales inferred by the ARD kernel in: (left) 15 min; (middle) 30 min; (right) 60 min. Each node has 5 neighbors and 8 features:'T' represents time in a day; 'D' represents day in a week; t={6,...,1} represent the traffic at 5t minutes before. DGPG can discover that time plays a significant role, while which day the record occurred is less relevant.

References

[1] M. Titsias, "Variational learning of inducing variables in sparse gaussian processes," in Artificial Intelligence and Statistics (AISTATS-2009), pp. 567–574, 2009.

[2] H. Salimbeni and M. Deisenroth, "Doubly stochastic variational inference for deep gaussianprocesses," in Advances in Neural Information Processing Systems (NIPS-2017), pp. 4588–4599,2017.