The Variational Gaussian Process

Juho Lee

References

- D. J. Rezende and S. Mohamed. Variational inference with normalizing flows. ICML 2015.
- R. Ranganath, D. Tran and D. M. Blei. Hierarchical variational models. ICML 2016.
- D. Tran, R. Ranganath and D. M. Blei. The Variational Gaussian process. ICLR 2016.

Mean-field variational inference (MFVI)

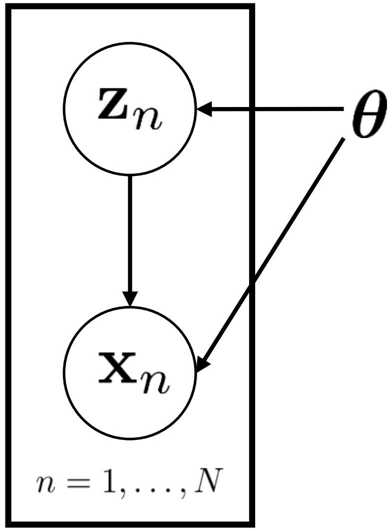
- Goal is to estimate $p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta})$, but intractable
- Posit a variational distribution $q(\mathbf{z};\boldsymbol{\lambda})$,

$$q(\mathbf{z};\boldsymbol{\lambda}) = \prod_{k=1}^{K} q(z_k;\lambda_k)$$

- Minimize $\mathrm{KL}[q(\mathbf{z};\boldsymbol{\phi})||p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta})]$ w.r.t. variational parameter $\boldsymbol{\lambda}$, equivalent to maximize evidence lower-bound (ELBO):

$$\mathcal{L}(\boldsymbol{\lambda}) = \mathbb{E}_{q(\mathbf{z};\boldsymbol{\lambda})}[\log p(\mathbf{x},\mathbf{z};\boldsymbol{\theta}) - \log q(\mathbf{z};\boldsymbol{\lambda})]$$

- Underestimates the posterior covariance, thus poor predictive performance



Variational Inference with normalizaing flows

- Start from a simple variational distribution, possibly a mean-field family

$$q_0(\mathbf{z}[0]) = \prod_{k=1}^{K} q_0(z_k[0])$$

- Transform variables with invertible and differentiable mappings

$$\mathbf{z} = \mathbf{z}[T] = f_T \circ f_{T-1} \circ \cdots \circ f_1(\mathbf{z}[0])$$

$$q(\mathbf{z}; \boldsymbol{\lambda}) = q_0(\mathbf{z}[0]) \prod_{t=1}^{T} \left| \det \frac{df_t}{d\mathbf{z}[t-1]} \right|^{-1}$$

- Can optimize $\boldsymbol{\lambda}$ with stochastic gradient descent

Hierarchical variational models

- Variational distribution is also a hierarchical Bayesian model

Variational
hyperparameter

$$q(\mathbf{z};\boldsymbol{\nu}) = \int \underbrace{\boxed{q(\mathbf{z}|\boldsymbol{\lambda})}}_{\text{variational likelihood}} \underbrace{\boxed{q(\boldsymbol{\lambda};\boldsymbol{\nu})}}_{\text{variational prior}} d\boldsymbol{\lambda} = \int \left[\prod_{k=1}^{K} q(z_k|\boldsymbol{\lambda})\right] q(\boldsymbol{\lambda};\boldsymbol{\nu}) d\boldsymbol{\lambda}.$$

- Can approximate complex posterior distributions without underestimating dependencies between latent variables
- Can we optimize ELBO w.r.t. $\boldsymbol{\nu}$? - no, in general we cannot evaluate $q(\mathbf{z};\boldsymbol{\nu})$

$$\mathcal{L}(\boldsymbol{\nu}) = \mathbb{E}_{q(\mathbf{z};\boldsymbol{\nu})}[\log p(\mathbf{x},\mathbf{z};\boldsymbol{\theta}) - \log q(\mathbf{z};\boldsymbol{\nu})]$$

Hierarchical variational models

- To approximate the entropy, introduce another auxiliary distribution

$$
\begin{aligned}
\mathbb{E}_{q(\mathbf{z};\boldsymbol{\nu})}[\log q(\mathbf{z};\boldsymbol{\nu})] &\leq \mathbb{E}_{q(\mathbf{z};\boldsymbol{\nu})}[\log q(\mathbf{z};\boldsymbol{\nu}) + \mathrm{KL}\left[q(\boldsymbol{\lambda}|\mathbf{z};\boldsymbol{\nu})||r(\boldsymbol{\lambda}|\mathbf{z};\boldsymbol{\phi})\right]] \\
&= \mathbb{E}_{q(\mathbf{z},\boldsymbol{\lambda};\boldsymbol{\nu})}\left[\log q(\mathbf{z};\boldsymbol{\nu}) + \log q(\boldsymbol{\lambda}|\mathbf{z};\boldsymbol{\nu}) - \log r(\boldsymbol{\lambda}|\mathbf{z};\boldsymbol{\phi})\right] \\
&= \mathbb{E}_{q(\mathbf{z},\boldsymbol{\lambda};\boldsymbol{\nu})}\left[\log q(\mathbf{z}|\boldsymbol{\lambda}) + \log q(\boldsymbol{\lambda};\boldsymbol{\nu}) - \log r(\boldsymbol{\lambda}|\mathbf{z};\boldsymbol{\phi})\right]
\end{aligned}
$$

- Hence, we get the following lower-bound on ELBO:

$$
\mathcal{L}(\boldsymbol{\nu},\boldsymbol{\phi}) = \mathbb{E}_{q(\mathbf{z},\boldsymbol{\lambda};\boldsymbol{\nu})}\left[\log p(\mathbf{x},\mathbf{z};\boldsymbol{\theta}) - \sum_{k=1}^{K}\log q(z_k|\boldsymbol{\lambda}) - \log q(\boldsymbol{\lambda};\boldsymbol{\nu}) + \log r(\boldsymbol{\lambda}|\mathbf{z};\boldsymbol{\phi})\right]
$$

Hierarchical variational models

- Possible choices of variational prior $q(\boldsymbol{\lambda}; \boldsymbol{\nu})$:
  - Mixture of Gaussians

$$q(\boldsymbol{\lambda}; \boldsymbol{\nu}) = \sum_{\ell=1}^{L} \pi_\ell \mathcal{N}(\boldsymbol{\lambda}; \boldsymbol{\mu}_\ell, \boldsymbol{\Sigma}_\ell).$$

  - Normalizing flows:

$$q(\boldsymbol{\lambda}; \boldsymbol{\nu}) = q_0(\boldsymbol{\lambda}[0]) \prod_{t=1}^{T} \left| \det \frac{df_t}{d\boldsymbol{\lambda}[t-1]} \right|^{-1}$$

$$r(\boldsymbol{\lambda}|z; \boldsymbol{\phi}) = r_0(\boldsymbol{\lambda}[0]|\mathbf{z}) \prod_{t=1}^{T} \left| \det \frac{dg_t}{d\boldsymbol{\lambda}[t-1]} \right|^{-1}$$

Normalizing flows vs hierarchical variational models with normalizing flows

Normalizing flows

$$q(\mathbf{z};\boldsymbol{\lambda}) = q_0(\mathbf{z}[0]) \prod_{t=1}^{T} \left| \det \frac{df_t}{d\mathbf{z}[t-1]} \right|^{-1}$$

Hierchical variational model with normalizing flow prior

$$q(\boldsymbol{\lambda};\boldsymbol{\nu}) = q_0(\boldsymbol{\lambda}[0]) \prod_{t=1}^{T} \left| \det \frac{df_t}{d\boldsymbol{\lambda}[t-1]} \right|^{-1}$$

$$q(\mathbf{z}|\boldsymbol{\lambda}) = \prod_{k=1}^{K} q(z_k|\boldsymbol{\lambda})$$

Black box methods	Computation	Storage	Dependency	Class of models
BBVI (Ranganath et al., 2014)	$\mathcal{O}(d)$	$\mathcal{O}(d)$	✗	discrete/continuous
DSVI (Titsias and Lázaro-Gredilla, 2014)	$\mathcal{O}(d^2)$	$\mathcal{O}(d^2)$	✓	continuous-diff.
COPULA VI (Tran et al., 2015)	$\mathcal{O}(d^2)$	$\mathcal{O}(d^2)$	✓	discrete/continuous
MIXTURE (Jaakkola and Jordan, 1998)	$\mathcal{O}(Kd)$	$\mathcal{O}(Kd)$	✓	discrete/continuous
NF (Rezende and Mohamed, 2015)	$\mathcal{O}(Kd)$	$\mathcal{O}(Kd)$	✓	continuous-diff.
HVM w/ NF prior	$\mathcal{O}(Kd)$	$\mathcal{O}(Kd)$	✓	discrete/continuous

The variational Gaussian process

- Hierarchical variational model with Gaussian process variational prior

latent input　　Variational data

$$\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad \mathcal{D} = \{(\mathbf{s}_n, \mathbf{t}_n)\}_{n=1}^{N}$$

$$f_k \sim \mathrm{GP}(\mathbf{0}, \mathbf{K})|\mathcal{D}, \quad k = 1, \ldots, K.$$

$$q(\mathbf{z}|\{f_k\}_{k=1}^{K}, \boldsymbol{\xi}) = \prod_{k=1}^{K} q(z_k|f_k(\boldsymbol{\xi}))$$

$\Rightarrow$

Kernel hyperparameters
+ variational data

$$q(\boldsymbol{\lambda}; \boxed{\boldsymbol{\nu}}) = \mathcal{N}(\boldsymbol{\xi}; \mathbf{0}, \mathbf{I}) \prod_{k=1}^{K} \mathrm{GP}(f_k; \mathbf{0}, \mathbf{K})|\mathcal{D}$$

$$q(\mathbf{z}|\boldsymbol{\lambda}) = \prod_{k=1}^{K} q(z_k|f_k(\boldsymbol{\xi}))$$

- Note that “variational data” is needed as anchor points to express complex nonlinear mappings

The variational Gaussian process

Theorem 1 (universal approximation) Let $q(\mathbf{z};\boldsymbol{\nu})$ denote the variational Gaussian process. Consider a posterior distribution $p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta})$ with finite number of latent variables and continuous quantile (inverse CDF) function. Then there exists a sequence of variational hyperparameters $(\boldsymbol{\nu}_t)$ such that

$$\lim_{t\to\infty} \mathrm{KL}[q(\mathbf{z};\boldsymbol{\nu}_t)||p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta})] = 0.$$

In other words, any posterior distribution under some condition can be exactly matched with sufficiently many variational data.

The variational Gaussian process

- Objective function: introduce auxiliary distribution $r(\boldsymbol{\xi}, f|\mathbf{z}; \boldsymbol{\phi})$

$$\mathcal{L}(\boldsymbol{\nu}, \boldsymbol{\phi}) = \mathbb{E}_{q(\mathbf{z}, \boldsymbol{\xi}, f; \boldsymbol{\nu})}\left[\log p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) - \sum_{k=1}^{K} \log q(\mathbf{z}|f_k(\boldsymbol{\xi})) - \log q(\boldsymbol{\xi}, f; \boldsymbol{\nu}) + r(\boldsymbol{\xi}, f|\mathbf{z}; \boldsymbol{\phi})\right]$$

- $r(\boldsymbol{\xi}, f; \boldsymbol{\phi})$ is specified as fully factorized Gaussian

$$r(\boldsymbol{\xi}, f; \boldsymbol{\phi}) = \mathcal{N}(\boldsymbol{\xi}; \boldsymbol{\mu}_0, \boldsymbol{\sigma}_0^2) \prod_{k=1}^{K} \mathcal{N}(f_k(\boldsymbol{\xi}_k); \mu_k, \sigma_k^2).$$

- Parameters can efficiently be optimized via stochastic gradient descent (with reparametrization trick)

The variational Gaussian process

- Inference network with reparametrization

$$\boldsymbol{\nu} = \mathrm{MLP}(\mathbf{x}), \quad \boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

Map input data to (variatioal data + kernel hyperparameters), draw latent input

$$f_k(\boldsymbol{\xi}) = \boldsymbol{\mu}_k(\boldsymbol{\nu}, \boldsymbol{\xi}) + \boldsymbol{\Sigma}_k^{-\frac{1}{2}}(\boldsymbol{\nu}, \boldsymbol{\xi})\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}).$$

Draw variational parameters with reparametrization

$$\mathbf{z} \sim q(\mathbf{z}|f_k(\boldsymbol{\xi}))$$

Draw latent variables (do reparametrization if possible)

$$\boldsymbol{\phi} = \mathrm{MLP}(\mathbf{z}, \mathbf{x})$$

Map latent variables + input data to auxiliary parameters

The variational Gaussian process

- Experiments on binarized MNIST dataset

Model	$-\log p(\mathbf{x})$	$\leq$
DLGM + VAE [1]		86.76
DLGM + HVI (8 leapfrog steps) [2]	85.51	88.30
DLGM + NF ($k = 80$) [3]		85.10
EoNADE-5 2hl (128 orderings) [4]	84.68	
DBN 2hl [5]	84.55	
DARN 1hl [6]	84.13	
Convolutional VAE + HVI [2]	81.94	83.49
DLGM 2hl + IWAE ($k = 50$) [1]		82.90
DRAW [7]		80.97
DLGM 1hl + VGP		84.79
DLGM 2hl + VGP		81.32
DRAW + VGP		**79.88**

The variational Gaussian process

- Experiments on Sketch dataset

Model	Epochs	$\leq -\log p(\mathbf{x})$
DRAW	100	526.8
	200	479.1
	300	464.5
DRAW + VGP	100	**460.1**
	200	**444.0**
	300	**423.9**