[Paper review 36]

Gaussian Processes for Big Data

(Hensman, et al., 2013)

[Contents]

1. Abstract

introduce **SVI** (Stochastic Variational Inference) for **GP** (Gaussian Process) models

- enable GP models to be **scalable**
- show that GPs can be variationally decomposed, to be dependent on a set of **globally relevant inducing variables**

2. Introduction

GP, used for regression, classification, unsupervised learning..

drawback : complexity of $O(n^3)$

To deal with this problem, various approximate techniques have been proposed

- 1) partition data set into separate groups
- 2) low rank approximation to the covariance matrix (complexity of $O(nm^2)$)
- 3) (by this paper)

  "recent advances in VI can be combined with the idea of INDUCING VARIABLES to develop a practical algorithm for fitting GPs using SVI"

3. Sparse GPs Revisited

inducing variables of Titsias (2009)

notation

- $\mathbf{y}$: data vector

  (consists of y_i, which are noisy observation of the function $f(\mathbf{x_i})$)

  (independent Gaussian, with precision β)
- $\mathbf{X}=\{\mathbf{x_i}\}_{i=1}^n$: all the datapoints
- **INDUCING VARIABLES** : $\mathbf{u}$ values of the function f at the points $mathbfZ=\{\mathbf{z}_i\}_{i=1}^m$

$$
\begin{aligned}
p(\mathbf{y} \mid \mathbf{f}) &= \mathcal{N}\left(\mathbf{y} \mid \mathbf{f}, \beta^{-1}\mathbf{I}\right) \\
p(\mathbf{f} \mid \mathbf{u}) &= \mathcal{N}\left(\mathbf{f} \mid \mathbf{K}_{nm}\mathbf{K}_{mm}^{-1}\mathbf{u}, \widetilde{\mathbf{K}}\right). \\
p(\mathbf{u}) &= \mathcal{N}\left(\mathbf{u} \mid \mathbf{0}, \mathbf{K}_{mm}\right)
\end{aligned}
$$

- $\mathbf{K}_{mm}$: covariance function evaluated between **all the inducing points**
- $\mathbf{K}_{nm}$: covariance function between **all inducing points** and **training points**
- $\tilde{\mathrm{K}}=\mathrm{K}_{nn}-\mathbf{K}_{nm}\mathbf{K}_{mm}^{-1}\mathbf{K}_{mn}$

Apply Jensen's inequality on the conditional probability $p(\mathbf{y} \mid \mathbf{u})$

$$
\begin{aligned}
\log p(\mathbf{y} \mid \mathbf{u}) &= \log\langle p(\mathbf{y} \mid \mathbf{f})\rangle_{p(\mathbf{f}\mid\mathbf{u})} \\
&\geq \langle\log p(\mathbf{y} \mid \mathbf{f})\rangle_{p(\mathbf{f}\mid\mathbf{u})} \triangleq \mathcal{L}_1
\end{aligned}.
$$

- $\langle\cdot\rangle_{p(x)}$: expectation under $p(x)$.
- $\log\langle p(\mathbf{y} \mid \mathbf{f})\rangle_{p(\mathbf{f}\mid\mathbf{u})}$: computed by $\mathcal{O}\left(n^3\right)$
- $\langle\log p(\mathbf{y} \mid \mathbf{f})\rangle_{p(\mathbf{f}\mid\mathbf{u})} \triangleq \mathcal{L}_1$: computed by $\mathcal{O}\left(m^3\right)$.

Interpretation : belows are equivalent

- $\mathbf{u}=\mathbf{f}$
- $\mathbf{K_{mm}}=\mathbf{K_{mn}}=\mathbf{K_{nm}}$
- $m=n$ inducing variables, and they are placed at training data locations
- no computational / storage advantage

when $p(\mathbf{y} \mid \mathbf{f})$ factorizes across the data, $p(\mathbf{y} \mid \mathbf{f})=\prod_{i=1}^n p\left(y_i \mid f_i\right)$.

$\rightarrow$ then $\exp(\mathcal{L}_1)=\prod_{i=1}^n \mathcal{N}\left(y_i \mid \mu_i, \beta^{-1}\right)\exp\left(-\frac{1}{2}\beta\tilde{k}_{i,i}\right)$.

- $\mu=\mathbf{K}_{nm}\mathbf{K}_{mm}^{-1}\mathbf{u}$.
- $\tilde{k}_{i,i}$: i th diagonal element of $\tilde{\mathbf{K}}$.

Bound of Titsias (2009)

- by marginalizing the inducing variables $\mathbf{u}$

$$\log p(\mathbf{y} \mid \mathbf{X}) = \log \int p(\mathbf{y} \mid \mathbf{u}) p(\mathbf{u}) \mathrm{d}\mathbf{u} \\ \geq \log \int \exp\{\mathcal{L}_1\} p(\mathbf{u}) \mathrm{d}\mathbf{u} \triangleq \mathcal{L}_2 .$$

- $\mathcal{L}_2 = \log \mathcal{N}\left(\mathbf{y} \mid \mathbf{0}, \mathbf{K}_{nm}\mathbf{K}_{mm}^{-1}\mathbf{K}_{mn} + \beta^{-1}\mathbf{I}\right) - \frac{1}{2}\beta \operatorname{tr}(\widetilde{\mathbf{K}})$.
  - $\mathbf{\Lambda} = \beta \mathbf{K}_{mm}^{-1}\mathbf{K}_{mn}\mathbf{K}_{nm}\mathbf{K}_{mm}^{-1} + \mathbf{K}_{mm}^{-1}$.
  - $\hat{\mathbf{u}} = \beta \mathbf{\Lambda}^{-1}\mathbf{K}_{mm}^{-1}\mathbf{K}_{mn}\mathbf{y}$.

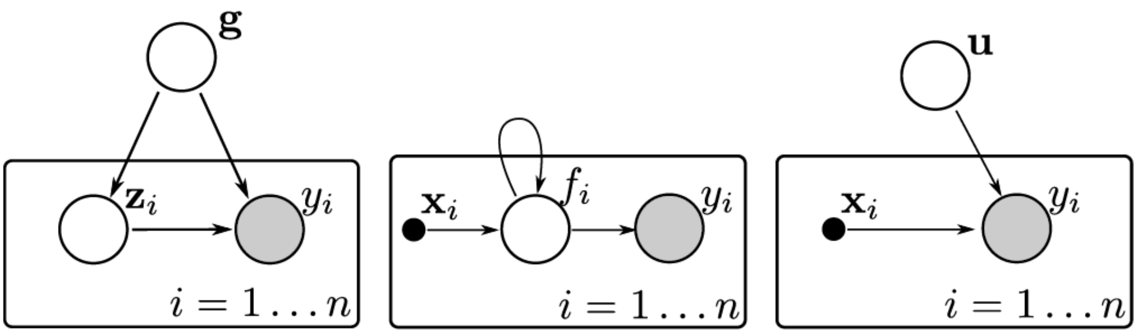
4. SVI for GPs

novelties of Titsias bound : rather than explicitly representing variational distribution for $q(\mathbf{u})$, these are collapsed!

but for SVI to work on GP, we need to "maintain an explicit representation of these inducing variables"

SVI (Stochastic Variational Inference)

- works on large dataset
- but can be only applied to (probabilistic) models
  - which have **global variables**
  - which factorizes in the observations and latent variables
- by introducing $\mathbf{u}$, we satisfies the condition!



(a) Requirements for SVI (b) Gaussian Process regression (c) Variational GP regression

- But in above we have found lower bound as below
  ($\log \int \exp\{\mathcal{L}_1\} p(\mathbf{u}) \mathrm{d}\mathbf{u} \triangleq \mathcal{L}_2$)

4-1. Global Variables

New lower bound : ($\mathcal{L}_2 \geq \mathcal{L}_3$,)

$\log p(\mathbf{y} \mid \mathbf{X}) \geq \langle \mathcal{L}_1 + \log p(\mathbf{u}) - \log q(\mathbf{u}) \rangle_{q(\mathbf{u})} \triangleq \mathcal{L}_3$.

Now, parameterize our variational distribution as $q(\mathbf{u}) = \mathcal{N}(\mathbf{u} \mid \mathbf{m}, \mathbf{S})$.

$$\mathcal{L}_3=\sum_{i=1}^{n}\left\{\log \mathcal{N}\left(y_i \mid \mathbf{k}_i^{\top} \mathbf{K}_{mm}^{-1} \mathbf{m}, \beta^{-1}\right)-\frac{1}{2} \beta \tilde{k}_{i,i}-\frac{1}{2} \operatorname{tr}\left(\mathbf{S} \boldsymbol{\Lambda}_i\right)\right\}-\mathrm{KL}(q(\mathbf{u}) \| p(\mathbf{u})) .$$

- $\mathbf{k}_i$: vector of the i^{th} column of $\mathbf{K}_{mn}$
- $\boldsymbol{\Lambda}_i=\beta \mathbf{K}_{mm}^{-1} \mathbf{k}_i \mathbf{k}_i^{\top} \mathbf{K}_{mm}^{-1}$.
- $\mathcal{L}_3$ can be written as sum of n terms!

Gradients of lower bound $\mathcal{L}_3$

- $\frac{\partial \mathcal{L}_3}{\partial \mathbf{m}}=\beta \mathbf{K}_{mm}^{-1} \mathbf{K}_{mn} \mathbf{y}-\boldsymbol{\Lambda} \mathbf{m}$
- $\frac{\partial \mathcal{L}_3}{\partial \mathbf{S}}=\frac{1}{2} \mathbf{S}^{-1}-\frac{1}{2} \boldsymbol{\Lambda}$
- setting the above to zero...
  - $\mathbf{S}=\boldsymbol{\Lambda}^{-1}, \mathbf{m}=\hat{\mathbf{u}}$.
  - This is when $\mathcal{L}_2=\mathcal{L}_3$

4-2. Natural Gradients

SVI works by taking steps in the direction of **approximate natural gradient** (= $\tilde{\mathrm{g}}(\theta)=G(\theta)^{-1} \frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}}$)

canonical and expectation parameters

- $\theta_1=\mathrm{S}^{-1}\ \mathrm{m}$

  $\theta_2=-\frac{1}{2}\ \mathrm{S}^{-1}$
- $\eta_1=\mathbf{m}$

  $\eta_2=\mathrm{mm}^{\top}+\mathrm{S}$

simplification of natural gradient : $\tilde{\mathrm{g}}(\boldsymbol{\theta})=G(\boldsymbol{\theta})^{-1} \frac{\partial \mathcal{L}_3}{\partial \boldsymbol{\theta}}=\frac{\partial \mathcal{L}_3}{\partial \boldsymbol{\eta}}$.

Therefore , by $\boldsymbol{\theta}_{(t+1)}=\boldsymbol{\theta}_{(t)}+\ell \frac{\mathrm{d} \mathcal{L}_3}{\mathrm{d} \eta}$,,

- $$\begin{aligned} \theta_{2(t+1)} & =-\frac{1}{2} \mathbf{S}_{(t+1)}^{-1} \\ & =-\frac{1}{2} \mathbf{S}_{(t)}^{-1}+\ell\left(-\frac{1}{2} \boldsymbol{\Lambda}+\frac{1}{2} \mathbf{S}_{(t)}^{-1}\right) \\ \boldsymbol{\theta}_{1(t+1)} & =\mathbf{S}_{(t+1)}^{-1} \mathbf{m}_{(t+1)} \\ & =\mathbf{S}_{(t)}^{-1} \mathbf{m}_{(t)}+\ell\left(\beta \mathbf{K}_{mm}^{-1} \mathbf{K}_{mn} \mathbf{y}-\mathbf{S}_{(t)}^{-1} \mathbf{m}_{(t)}\right) \end{aligned}$$.

4-3. Latent Variables

- enable **online learning** for GPR using SVI
- to perform SVI with latent variable, need **factorization** (like figure 1-(a))

$$
\begin{aligned}
\log p(\mathbf{y}) &= \log \int p(\mathbf{y} \mid \mathbf{X}) p(\mathbf{X}) \mathrm{d}\mathbf{X} \\
&\geq \int q(\mathbf{X}) \left\{ \mathcal{L}_3 + \log p(\mathbf{X}) - \log q(\mathbf{X}) \right\} \mathrm{d}\mathbf{X}
\end{aligned}.
$$

- $q(\mathbf{X}) = \prod_{i=1}^{n} q\left(\mathrm{x}_i\right)$.

Too perform SVI in this model, alternate between....

- (1) selecting a mini-batch of data
- (2) optimizing relevant variables of $q(\mathbf{X})$ (with $q(\mathbf{u})$ fixed)

  and updating $q(\mathbf{u})$ using the approximate natural gradient

4-4. Non-Gaussian likelihood

Advantage of using

$$\mathcal{L}_3 = \sum_{i=1}^{n} \left\{ \log \mathcal{N}\left(y_i \mid \mathbf{k}_i^{\top} \mathbf{K}_{mm}^{-1} \mathbf{m}, \beta^{-1}\right) - \frac{1}{2} \beta \tilde{k}_{i,i} - \frac{1}{2} \operatorname{tr}(\mathbf{S} \mathbf{\Lambda}_i) \right\} - \mathrm{KL}(q(\mathbf{u}) \| p(\mathbf{u}))$$

$\rightarrow$ enable inference with non-Gaussian likelihoods

ex) binary, probit likelihood ....

5. Discussion

method for inference in GP using SVI (enable scalability)

discuss the bound on $p(y \mid u)$ in detail

(becomes tight when $\mathbf{Z} = \mathbf{X}$)

complexity becomes $O(m^3)$