Latent Variable Augmentation in Bayesian Inference

Applications for Gaussian Processes

vorgelegt von
Dipl.-Ing.
Théo Galy-Fajou
geb. in Castres

von der Fakultät IV - Elektrotechnik und Informatik
der Technischen Universität Berlin
zur Erlangung des akademischen Grades

Doktor der Naturwissenschaften
-Dr.-Ing.-

genehmigte Dissertation

Promotionsausschuss:
Vorsitzender: Prof. A
Gutachter: Prof. Manfred Opper
Gutachterin: Prof. C
Gutachter: Prof. D
Tag der wissenschaftlichen Aussprache: XX. xxxx 2021

Berlin 2021

Zusammenfassung

Hier kommt der deutsche Abstrakt rein... ÜÖ sind ok.

Abstract

Put your abstract here...

Dedicated to ...

Acknowledgements

I would like to acknowledge the thousands of individuals who have coded for open-source projects for free. It is due to their efforts that scientific work with powerful tools is possible.

Table of Contents

List of Figures

List of Tables

Abbreviations

This document is incomplete. The external file associated with the glossary 'abbreviations' (which should be called `thesis.gls-abr`) hasn't been created.

Check the contents of the file `thesis.glo-abr`. If it's empty, that means you haven't indexed any of your entries in this glossary (using commands like `\gls` or `\glsadd`) so this list can't be generated. If the file isn't empty, the document build process hasn't been completed.

Try one of the following:

- Add `automake` to your package option list when you load `glossaries-extra.sty`. For example:
  `\usepackage[automake]{glossaries-extra}`
- Run the external (Lua) application:
  `makeglossaries-lite.lua "thesis"`
- Run the external (Perl) application:
  `makeglossaries "thesis"`

Then rerun LaTeX on this document.

This message will be removed once the problem has been fixed.

Symbols

This document is incomplete. The external file associated with the glossary 'symbolslist' (which should be called `thesis.syi`) hasn't been created.

Check the contents of the file `thesis.syg`. If it's empty, that means you haven't indexed any of your entries in this glossary (using commands like `\gls` or `\glsadd`) so this list can't be generated. If the file isn't empty, the document build process hasn't been completed.

Try one of the following:

- Add `automake` to your package option list when you load `glossaries-extra.sty`. For example:
  `\usepackage[automake]{glossaries-extra}`
- Run the external (Lua) application:
  `makeglossaries-lite.lua "thesis"`
- Run the external (Perl) application:
  `makeglossaries "thesis"`

Then rerun LaTeX on this document.

This message will be removed once the problem has been fixed.

1 Introduction

1.1 Following Bayes

- Bayes is awesome

1.2 The use of Gaussian Processes

- All these things you can do with Gaussian processes

1.3 The underestimated importance of representation

- Different representation lead to very different results, efficiency etc

2

Background

2.1 Probabilistic Bayesian Modeling

The Bayes' theorem is one of the simplest theorem in probabilites and its demonstration holds in one line, its implications are however more complex.

Let's give the very general modeling setting. We have a set of observed variables $\boldsymbol{x}$ and a set of latent (unobserved) variables $\boldsymbol{\theta}$. Given a prior distribution on $\boldsymbol{\theta}$, $p(\boldsymbol{\theta})$, and a likelihood function $p(\boldsymbol{x} \mid \boldsymbol{\theta})$ we are interested in the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{x})$ which is given by:

$$p(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{x})} = \frac{p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}} \tag{2.1}$$

The posterior is of interest for making prediction on previously unseen data. Let's take the simple example of logistic regression: Given some input $\boldsymbol{X} \in \mathbb{R}^D$ and binary label $y \in \{0, 1\}$ we model the generative model as:

$$y \sim \text{Bernoulli}\left(\sigma(\boldsymbol{\theta}^\top \boldsymbol{X})\right), \tag{2.2}$$

where $\boldsymbol{\theta} \in \mathbb{R}^D$ and σ is the logistic function $\sigma(x) = \frac{1}{1+\exp(-x)}$. We put a simple isotropic Normal prior on $\boldsymbol{\theta}$: $p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}|0, I_D)$ and use following likelihood function: $p(y_i|\boldsymbol{\theta}, \boldsymbol{X}_i) = \sigma\left(2(y_i - 1)\boldsymbol{\theta}^\top \boldsymbol{X}_i\right)$. Assuming that we now know the posterior $p(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{X})$ we can make predictions for new data using the following:

$$p(y^*|\boldsymbol{X}^*, \boldsymbol{y}, \boldsymbol{X}) = \int p(y^*, \boldsymbol{\theta}|\boldsymbol{X}^*\boldsymbol{y}, \boldsymbol{X})d\boldsymbol{\theta} = \int p(y^* = 1|\boldsymbol{\theta}, \boldsymbol{X}^*)p(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{X})d\boldsymbol{\theta} \tag{2.3}$$

2.2 Gaussian Processes

GP! (**GP!**) are a class of non-parametric models to approximate functions. By definition, a **GP!** is a stochastic process where the joint distribution on any collection of variables X_t follows a (multivariate) Gaussian distribution. This Gaussian nature is what make them so attractive since operations on Gaussian variables tend to be easier and many calculus have

closed-form solutions. The Gaussian distribution is to statistics what the harmonic oscillator is to physics. Although, **GP!** are defined to be a non-parametric model, one still needs to define how the covariance between each variable of the process is defined. One of the most popular interpretation of **GP!** is as a prior on functions in the **RKHS!** (**RKHS!**). In practice the **RKHS!** is infinite-dimensional, to be able to perform any computation one needs to project it into a finite-dimensional space.

One resorts to kernel functions [**NEED TO CITE THIS**]. The kernel matrix K is defined by $K_{ij} = k(x_i, x_j)$. K is positive-definite, i.e. for $K \in \mathbb{R}^{D\times D}$, and $x \in \mathbb{R}^D$, $x^\top Kx > 0$.

2.3 Approximate Bayesian Inference

The posterior distribution in Eq.(2.1) cannot be computed in closed-form for non-trivial problems. To still be able to make predictions and render the model useful one can resort to different approximations. Out of a very large number of methods two of the most used are sampling and variational inference.

2.3.1 Sampling

When the posterior $p(\boldsymbol{\theta}|\boldsymbol{x})$ is not available in closed-form, it may be possible to draw samples from it. The set of methods is far too large to be even mentioned in this thesis, I will restrict the scope to methods tailored or adapted to **GPs!** (**GPs!**).

2.3.2 Variational Inference

VI! (**VI!**), sometimes called Variational Bayes, consists in approximating the posterior with another parametrized distribution. Given a family of distributions $\mathcal{Q}$, parametrized by parameters $\boldsymbol{\varphi}$ one aims to solve the following optimization problem:

$$\boldsymbol{\varphi}^* = \arg_{\boldsymbol{\varphi}} \min \mathrm{KL}\left(q_{\boldsymbol{\varphi}}(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\boldsymbol{x})\right), \tag{2.4}$$

where the KL (Kullback-Leibler) divergence is defined (for continuous distributions as:

$$\mathrm{KL}\left(q(x)||p(x)\right) = \int q(x) \log \frac{q(x)}{p(x)} dx \tag{2.5}$$

The objective of equation (2.4) is generally not tractable. Since computing $p(\boldsymbol{\theta}|\boldsymbol{x})$ involves the typically intractable normalization constant $p(\boldsymbol{x})$, one resort to a surrogate function, the **VFE!** (**VFE!**) (or its negative counterpart the **ELBO!** (**ELBO!**)):

$$\begin{aligned}
\mathrm{KL}\left(q_{\boldsymbol{\varphi}}(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\boldsymbol{x})\right) &= \int q_{\boldsymbol{\varphi}}(\boldsymbol{\theta}) \left(\log q_{\boldsymbol{\varphi}}(\boldsymbol{\theta}) - \log p(\boldsymbol{\theta}|\boldsymbol{x})\right) d\boldsymbol{\theta} && (2.6)\\
&= \int q_{\boldsymbol{\varphi}}(\boldsymbol{\theta}) \left(\log q_{\boldsymbol{\varphi}}(\boldsymbol{\theta}) - \log p(\boldsymbol{\theta}, \boldsymbol{x}) - \log p(\boldsymbol{x})\right) d\boldsymbol{\theta} && (2.7)\\
&= \underbrace{-\log p(\boldsymbol{x})}_{\leq 0} + \int q_{\boldsymbol{\varphi}}(\boldsymbol{\theta}) \left(\log q_{\boldsymbol{\varphi}}(\boldsymbol{\theta}) - \log p(\boldsymbol{x}|\boldsymbol{\theta}) - \log p(\boldsymbol{\theta})\right) d\boldsymbol{\theta} && (2.8)\\
&\leq -\mathbb{E}_{q_{\boldsymbol{\varphi}}}\left[\log p(\boldsymbol{x}|\theta)\right] + \mathrm{KL}\left(q_{\boldsymbol{\varphi}}(\boldsymbol{\theta})||p(\boldsymbol{\theta})\right) = \mathcal{F}(\boldsymbol{\varphi}) && (2.9)
\end{aligned}$$

By minimizing the **VFE!**:$\mathcal{F}(\boldsymbol{\varphi})$ instead of the **KL!** (**KL!**) divergence, we expect to find a solution close to the optimum of the problem stated in (2.4). A standard way is to perform gradient descent on the variational parameters $\boldsymbol{\varphi}$

$$\boldsymbol{\varphi}^{t+1} = \boldsymbol{\varphi}^{t} - \epsilon \nabla_{\boldsymbol{\varphi}} \mathcal{F}(\boldsymbol{\varphi}^{t}). \tag{2.10}$$

Computing the gradient $\nabla_{\boldsymbol{\varphi}} \mathcal{F}(\boldsymbol{\varphi})$ can be non-trivial but many methods were developed to tackle this problem.

[**INTRODUCE DIFFERENT METHODS HERE**].

One method which interest us is to find the optimal parameters $\boldsymbol{\varphi}^{*}$ in closed-form. By solving:

$$\nabla_{\boldsymbol{\varphi}} \mathcal{F}(\boldsymbol{\varphi})|_{\boldsymbol{\varphi}=\boldsymbol{\varphi}^{*}} = 0 \tag{2.11}$$

3

Efficient Gaussian Process Classification Using Polya-Gamma Data Augmentation

4

Multi-Class Gaussian Process Classification Made Conjugate: Efficient Inference via Data Augmentation

5

Automated Augmented Conjugate Inference for Non-conjugate Gaussian Process Models

6

Variational Gaussian Particle Flow

7

Discussion

Appendix A