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 la bite et les couilles

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.

Abbreviations

$\mathcal{GP}$ Gaussian Process

RKHS Reproducing Kernel Hilbert Space

$\mathcal{GP}$**s** Gaussian Processes

MCMC Markov Chain Monte Carlo

VI Variational Inference

VFE Variational Free Energy

ELBO Evidence Lower BOund

KL Kullback-Leibler

MF Mean-Field

CAVI Coordinate Ascent Variational Inference

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

Machine learning has become a wide field of research with a variety of sub-fields, each dedicated to solve various problems in different ways. One field in particular, usually called probabilistic machine learning aims at representing the statistical side of the different models.

1.1 Following Bayes

- Bayes is awesome

In this thesis we are going to follow Bayesian principles

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:

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

The last term of the equation involves the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{X})$. To solve this integral, we must either be able to know the posterior in closed form and solve the integral numerically, or be able to sample from it and compute this integral with Monte-Carlo integration. Computing the posterior (2.1) in closed-form involves computing the integral $\int p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$, which is intractable for most non-trivial models. In Section 2.3, we mention methods which

help solving this issue by introducing approximations or ways to sample directly from the posterior.

2.2 Gaussian Processes

Gaussian Process ($\mathcal{GP}$) are a class of non-parametric models to approximate functions. By definition, a $\mathcal{GP}$ 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, Gaussian Processes ($\mathcal{GP}$**s**) 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 $\mathcal{GP}$ is as a prior on functions in the Reproducing Kernel Hilbert Space (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. Considering a function f we wish to approximate with a $\mathcal{GP}$, we need some data $\boldsymbol{X}$ to evaluate f on. We then consider the finite-dimensional vector $\boldsymbol{f}$ where $f_i = f(X_i)$.

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 $x \in \mathbb{R}^D$, $x^\top K x > 0$.

2.2.1 Gaussian Process Regression

We now have a prior on the realisation of the function $\boldsymbol{f}$ on some data $\boldsymbol{X}$, $p(\boldsymbol{f}) = \mathcal{N}(\boldsymbol{f}|\boldsymbol{\mu}_0, \boldsymbol{K})$. We can now add information about some noisy observations $\boldsymbol{y}$ we got for $\boldsymbol{X}$:

$$y_i = f(X_i) + \epsilon_i, \tag{2.4}$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. This leads to the likelihood $p(y_i|f_i) = \mathcal{N}(y_i|f_i, \sigma^2)$. Fortunately, multiplying Gaussian probability distributions together lead to another Gaussian distribution function. The posterior for $\boldsymbol{f}$ is given by $p(\boldsymbol{f}|\boldsymbol{y}) = \mathcal{N}(\boldsymbol{f}|\boldsymbol{y}, \boldsymbol{K} + \sigma^2 I)$. The prediction of f^* on a new point $\boldsymbol{x}^*$ can be done by computing:

$$p(f^*|\boldsymbol{x}^*, \boldsymbol{X}, \boldsymbol{y}) = \int p(f^*|\boldsymbol{f}, \boldsymbol{x}^*) p(\boldsymbol{f}|\boldsymbol{X}, \boldsymbol{y}) d\boldsymbol{f}. \tag{2.5}$$

This integral is analytically solvable and results in $p(f^*|\boldsymbol{x}^*, \boldsymbol{X}, \boldsymbol{y}) = \mathcal{N}(f^*|m^*, s^*)$ where $m^* = K_{\boldsymbol{x}^*,\boldsymbol{X}}(K_{\boldsymbol{X},\boldsymbol{X}} + \sigma^2 I)^{-1} y$ and $s^* = K_{\boldsymbol{x}^*,\boldsymbol{x}^*} - K_{\boldsymbol{x}^*,\boldsymbol{X}} \left(K_{\boldsymbol{X},\boldsymbol{X}} + \sigma^2 I\right)^{-1} K_{\boldsymbol{X},\boldsymbol{x}^*}$.

2.2.2 Non-Conjugate Gaussian Processes

A Gaussian prior is only conjugate[1] to a Gaussian likelihood Therefore $\mathcal{GP}$**s** only give a Gaussian posterior with a Gaussian likelihood, for all other cases we talk about *Non-Conjugate Gaussian Processes*.

Since the posterior is not analytically tractable, one has to resort to some of the methods presented in Section 2.3.

[1] A prior is said conjugate to a given likelihood when the resulting posterior is of the same family of the prior.

2.2.3 Sparse Gaussian Processes

One of the largest issue of $\mathcal{GP}$**s**, regardless of if they are conjugate or not is the scalability with the number of observed samples.

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 $\mathcal{GP}$**s**. I will especially focus on Markov Chain Monte Carlo (MCMC) methods, where a chain of variable $\boldsymbol{\theta}^t$ is created with a Markovian assumption ($\boldsymbol{\theta}^t$ depends only of $\boldsymbol{\theta}^{t-1}$) and where the stationary distribution of $\boldsymbol{\theta}^t$ is the same as the target distribution (in our case the posterior $p(\boldsymbol{\theta}|\boldsymbol{x})$.

2.3.2 Variational Inference

Variational Inference (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:

$$\boldsymbol{\varphi}^* = \arg_{\boldsymbol{\varphi}} \min \mathrm{KL}\left(q_{\boldsymbol{\varphi}}(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\boldsymbol{x})\right), \tag{2.6}$$

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

$$\mathrm{KL}\left(q(x)||p(x)\right) = \int q(x) \log \frac{q(x)}{p(x)} dx \tag{2.7}$$

The objective of equation (2.6) 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 Variational Free Energy (VFE) (or its negative counterpart the Evidence Lower BOund (ELBO)):

$$\begin{aligned}
\mathrm{KL}\left(q_{\boldsymbol{\varphi}}(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\boldsymbol{x})\right) &= \int q_{\boldsymbol{\varphi}}(\boldsymbol{\theta}) \left(\log q_{\boldsymbol{\varphi}}(\boldsymbol{\theta}) - \log p(\boldsymbol{\theta}|\boldsymbol{x})\right) d\boldsymbol{\theta} && (2.8)\\
&= \int q_{\boldsymbol{\varphi}}(\boldsymbol{\theta}) \left(\log q_{\boldsymbol{\varphi}}(\boldsymbol{\theta}) - \log p(\boldsymbol{\theta}, \boldsymbol{x}) - \log p(\boldsymbol{x})\right) d\boldsymbol{\theta} && (2.9)\\
&= \underbrace{-\log p(\boldsymbol{x})}_{\leq 0} + \int q_{\boldsymbol{\varphi}}(\boldsymbol{\theta}) \left(\log q_{\boldsymbol{\varphi}}(\boldsymbol{\theta}) - \log p(\boldsymbol{x}|\boldsymbol{\theta}) - \log p(\boldsymbol{\theta})\right) d\boldsymbol{\theta} && (2.10)\\
&\leq -\mathbb{E}_{q_{\boldsymbol{\varphi}}}\left[\log p(\boldsymbol{x}|\theta)\right] + \mathrm{KL}\left(q_{\boldsymbol{\varphi}}(\boldsymbol{\theta})||p(\boldsymbol{\theta})\right) = \mathcal{F}(\boldsymbol{\varphi}) && (2.11)
\end{aligned}$$

By minimizing the VFE:$\mathcal{F}(\boldsymbol{\varphi})$ instead of the Kullback-Leibler (KL) divergence, we expect to find a solution close to the optimum of the problem stated in (2.6). A standard way is to perform gradient descent on the variational parameters $\boldsymbol{\varphi}$

$$\boldsymbol{\varphi}^{t+1} = \boldsymbol{\varphi}^t - \epsilon \nabla_{\boldsymbol{\varphi}} \mathcal{F}(\boldsymbol{\varphi}^t). \tag{2.12}$$

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 of the most important addition to the VI method is the Mean-Field (MF) approximation. MF is the assumption that the variational distribution $q(\boldsymbol{\theta})$ assumes every compoment of $\boldsymbol{\theta}$ to be independent from each other. This way we can write

$$q_{\boldsymbol{\varphi}}^{MF}(\boldsymbol{\theta}) = \prod_{i=1}^{D} q_{\boldsymbol{\varphi}_i}(\theta_i) \tag{2.13}$$

A more general method is also to consider blocks of variables instead.

Following the MF approach, it is sometimes possible to find the optimal parameters $\boldsymbol{\varphi}^*$ in closed-form. By solving:

$$\nabla_{\varphi_i} \mathcal{F}(\boldsymbol{\varphi})|_{\varphi_i = \varphi_i^*} = 0, \tag{2.14}$$

for each variable φ_i we can find a local optima, which with additional assumptions, can prove to be the local minima we are looking for. The advantage of this method is that one can also perform it independently for each latent variable θ_i. Concretely the updates are of the form:

$$q_{\varphi_i}^*(\theta_i) \propto \exp\left(\mathbb{E}_{q_{\boldsymbol{\varphi}}(\boldsymbol{\theta}_{/i})}\left[\log p\left(\theta_i | \boldsymbol{\theta}_{/i}, \boldsymbol{x}\right)\right]\right) \tag{2.15}$$

where $\boldsymbol{\theta}_{/i}$ represent the collection of variables $\boldsymbol{\theta}_{/i} = \{\theta_j | j \neq i\}$. When working with distribution coming from exponential families, it is straightforward to get the optimal variational parameters φ_i. By updating the parameters one after another we get a Coordinate Ascent Variational Inference (CAVI) scheme[2]. Effectively, one update each variational parameter φ_i by its optimum given the rest of the variational parameters $\boldsymbol{\varphi}_{/i}$ via closed-form functions:

$$\varphi_i^{t+1} = f_i\left(\boldsymbol{\varphi}_{1:(i-1)}^{t+1}, \boldsymbol{\varphi}_{(i+1):D}^{t}\right). \tag{2.16}$$

The order of the updates do not matter as long as the variational parameters $\boldsymbol{\varphi}$ are initialized in their domain.

[2]The word ascent is used since the scheme was originally derived using the negative VFE or ELBO.

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