An Introduction to

VARIATIONAL GAUSSIAN PROCESSES

Torsten Scholak
twitter @tscholak
github tscholak
blog tscholak.github.io
August 18, 2016

What Is Bayesian Inference?

1. **Observe** the phenomenon, gather $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^N$, $\mathbf{x}_i \in \mathbb{R}^D$.
2. **Build** a model, $p(\mathbf{X}, \mathbf{z})$ with latent variables $z_1, \ldots, z_d$.
3. **Infer** the posterior, $p(\mathbf{z}|\mathbf{X}) = p(\mathbf{X}, \mathbf{z}) / \int p(\mathbf{X}, \mathbf{z}) \, \mathrm{d}\mathbf{z}$, in order t **reason** about the phenomenon.
4. **Criticize** the model, revise it (↦2), or collect additional data (↦1).
5. **Apply** the model, i.e. calculate integrals over $p(\mathbf{z}|\mathbf{X})$: expectations of $f(\mathbf{z})$, posterior predictive $p(\mathbf{x}^*|\mathbf{X})$, etc.

Why Is Bayesian Inference Hard?

Most posteriors $p(\mathbf{z}|\mathbf{X})$ are not analytically tractable.

A possible solution is numerical estimation via MCMC, e.g. via:

- Metropolis Hastings Sampling,
- Gibbs Sampling,
- Hamiltonian Monte Carlo Sampling,
- No-U-Turn Hamiltonian Monte Carlo Sampling (NUTS).

NUTS is close to exact, but also slow and sequential.

What Is Variational Inference (VI)?

VI is a class of algorithms which cast posterior inference as optimization:

3. ~~**Infer**~~ **Approximate** the posterior, $p(\mathbf{z}|\mathbf{X})$:
    a. **Build** a variational model, $q(\mathbf{z};\lambda)$, over $\mathbf{z}$ with parameters λ.
    b. **Match** $q(\mathbf{z};\lambda)$ to $p(\mathbf{z}|\mathbf{X})$ by optimizing over λ,

$$\lambda^* = \operatorname{argmin}_\lambda \operatorname{divergence}\left(p(\mathbf{z}|\mathbf{X}), q(\mathbf{z};\lambda)\right).$$

    c. **Use** $q(\mathbf{z};\lambda^*)$ instead of $p(\mathbf{z}|\mathbf{X})$.
    d. **Criticize** the variational model, revise it (↦a).

Effectively, VI is an additional layer of approximation that facilitates convenient model iteration.

Matching And Optimizing

The **Kullback-Leibler divergence** from q to p is a good measure for closeness between p and q,

$$\mathrm{KL}\left(q(\mathbf{z};\lambda)\|p(\mathbf{z}|\mathbf{X})\right) \triangleq \mathbb{E}_{q(\mathbf{z};\lambda)}\left[\log \frac{q(\mathbf{z};\lambda)}{p(\mathbf{z}|\mathbf{X})}\right].$$

Minimization of this with respect to λ is intractable, though, because it directly depends on $p(\mathbf{z}|\mathbf{X})$.
Maximize the Evidence Lower BOund (ELBO) instead,

$$\begin{aligned}\mathcal{L}(\lambda) &\triangleq \log p(\mathbf{X}) - \mathrm{KL}\left(q(\mathbf{z};\lambda)\|p(\mathbf{z}|\mathbf{X})\right) \\ &= \mathbb{E}_{q(\mathbf{z};\lambda)}[\log p(\mathbf{X},\mathbf{z})] - \mathbb{E}_{q(\mathbf{z};\lambda)}[\log q(\mathbf{z};\lambda)].\end{aligned}$$

Conventional Variational Modeling

Two conflicting demands:
i. Make q **simpler** than p, e.g. choose a factorized multivariate (mean-field) normal distribution,

$$q(\mathbf{z}; \lambda) = \prod_{i=1}^{d} \mathcal{N}\left(z_i; \mu_i, \sigma_i^2\right).$$

ii. Make q more **expressive** so that it can give good results, e.g. choose a full-rank multivariate normal distribution,

$$q(\mathbf{z}; \lambda) = \mathcal{N}\left(\mathbf{z}; \mu, \mathbf{\Sigma}\right).$$

Hierarchical Variational Modeling

iii. Use a mean-field distribution, $\prod_i q(z_i|\lambda_i)$, but softly constrai it by putting a prior $q(\lambda;\theta)$ on it,

$$q_{\mathrm{HVM}}(\mathbf{z};\theta) = \int \left[\prod_{i=1}^{d} q(z_i|\lambda_i)\right] q(\lambda;\theta)\,\mathrm{d}\lambda\,.$$

Hierarchy captures dependencies between latent variables, $\mathbf{z}$.
More computationally tractable than a variational model with full dependence structure.
Expressiveness is determined by the complexity of $q(\lambda;\theta)$.

Variational Gaussian Processes I

Let the mean-field parameters be $\lambda_i = f_i(\xi) \in \mathbb{R}$, $i = 1, \ldots, d$, where:

- The latent input $\xi \in \mathbb{R}^c$ is normally distributed, $\xi \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.
- The functions $f_i : \mathbb{R}^c \to \mathbb{R}$ are distributed according to a **Gaussian process** (GP),

$$f_i \sim \mathcal{GP}\left(\mathbf{0}, \mathbf{K}\right) | \, \theta,$$

  conditioned on a fake data set, θ, which is not modeled.

Then we draw mean-field samples, i.e. approximate posterior samples $\mathbf{z} \in \mathrm{supp}(p)$, conditioned on the output of the GP draw.

Variational Gaussian Processes II

$$q_{\mathrm{VGP}}(\mathbf{z};\theta) = \iint \left[\prod_{i=1}^{d} q(z_i|f_i(\xi))\right] \times \left[\prod_{i=1}^{d} \mathcal{GP}\left(f_i;\mathbf{0},\mathbf{K}\right)|\,\theta\right] \mathcal{N}(\xi;\mathbf{0},\mathbf{I})\,\mathrm{d}\mathbf{f}\,\mathrm{d}\xi\,.$$

Gaussian Processes

A Gaussian process is a **generalization** of the Gaussian probability distribution, $\mathcal{N}$.
Given data $\theta = \{(\mathbf{s}_j, \mathbf{t}_j)\}_j = \{\mathbf{S}, \mathbf{T}\}$, with inputs $\mathbf{s}_j \in \mathbb{R}^c$ and output $\mathbf{t}_j \in \mathbb{R}^d$,

$$p(\mathbf{f}|\theta) = \prod_{i=1}^{d} \mathcal{GP}\left(f_i; \mathbf{0}, \mathbf{K}\right) | \, \theta$$

forms a **distribution over functions** $\mathbf{f} : \mathbb{R}^c \to \mathbb{R}^d$ which interpolate between input-output pairs in θ.
$\mathbf{K}$ is the covariance matrix or **kernel** of the GP.

Gaussian Process Kernels

The standard choice is the **automatic relevance determination** kernel,

$$\mathbf{K}(\mathbf{S}, \mathbf{S}')_{jj'} = \eta^2 \exp\left[-\sum_{l=1}^{c} \rho_l^2 \left(s_{jl} - s'_{j'l}\right)^2\right] + \delta_{\mathbf{s}_j \mathbf{s}'_{j'}} \sigma^2,$$

where $\delta_{\mathbf{s}_j \mathbf{s}'_{j'}}$ is meant with respect to the identity of the points.
The more similar $\mathbf{s}_j$ and $\mathbf{s}'_{j'}$, the more similar $\mathbf{f}(\mathbf{s}_j)$ and $\mathbf{f}(\mathbf{s}'_{j'})$.
The larger ρ_l, the larger the weight on dimension l.
η is the scale of the outputs $\mathbf{T}$.
σ is the scale of the noise in $\mathbf{T}$.

Gaussian Process Prediction

The distribution of the function's value at a finite number of test inputs, $\mathbf{S}^*$, is a multivariate normal distribution,

$$\mathbf{T}^*_i \,|\, \mathbf{S}, \mathbf{T}, \mathbf{S}^* \sim \mathcal{N}\big(\mathbf{K}(\mathbf{S}^*, \mathbf{S})\,\mathbf{K}(\mathbf{S}, \mathbf{S})^{-1}\,\mathbf{T}_i, \\ \mathbf{K}(\mathbf{S}^*, \mathbf{S}^*) - \mathbf{K}(\mathbf{S}^*, \mathbf{S})\,\mathbf{K}(\mathbf{S}, \mathbf{S})^{-1}\,\mathbf{K}(\mathbf{S}, \mathbf{S}^*)\big),$$

where $\mathbf{T}_i$ ($\mathbf{T}^*_i$) are (test) outputs for the i-th output dimension.

Gaussian Process Joint Distribution

The joint distribution of the observed target values and the functic
values at the test locations can be written as:

$$\begin{pmatrix}\mathbf{T}_i\\\mathbf{T}_i^*\end{pmatrix}\Bigg|\,\mathbf{S},\mathbf{S}^*\sim\mathcal{N}\left[\mathbf{0},\begin{pmatrix}\mathbf{K}(\mathbf{S},\mathbf{S})&\mathbf{K}(\mathbf{S},\mathbf{S}^*)\\\mathbf{K}(\mathbf{S}^*,\mathbf{S})&\mathbf{K}(\mathbf{S}^*,\mathbf{S}^*)\end{pmatrix}\right].$$

Demonstration

What Is The Target Function in The VGP?

Just for now, assume that the variational likelihood $q(\mathbf{z}|\mathbf{f}(\xi))$ is a point mass distribution,

$$q(\mathbf{z}|\mathbf{f}(\xi)) = \delta(\mathbf{z} - \mathbf{f}(\xi))\,.$$

Then, with the GP, we want to approximate a function, $\mathbf{f}^*$, that, wher applied to draws $\xi \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, produces samples $\mathbf{z} = \mathbf{f}^*(\xi)$ that are distributed as the posterior $p(\mathbf{z}|\mathbf{X})$, i.e. effectively

$$p(\mathbf{z}|\mathbf{X}) = \int_{\mathbb{R}^c} \delta(\mathbf{z} - \mathbf{f}^*(\xi))\,\mathcal{N}(\xi; \mathbf{0}, \mathbf{I})\,\mathrm{d}\xi.$$

Explicit Construction of f* I

1. Integrate on both sides:

$$p(\mathbf{z}'|\mathbf{X}) = \int_{\mathbb{R}^c} \delta(\mathbf{z}' - \mathbf{f}^*(\xi'))\, \mathcal{N}(\xi'; \mathbf{0}, \mathbf{I})\, \mathrm{d}\xi',$$

$$\int_{\{z'_i \leq f^*_i(\xi)\}_i} p(\mathbf{z}'|\mathbf{X})\, \mathrm{d}\mathbf{z}' = \int_{\mathbb{R}^c} \int_{\{z'_i \leq f^*_i(\xi)\}_i} \delta(\mathbf{z}' - \mathbf{f}^*(\xi'))\, \mathrm{d}\mathbf{z}'\, \mathcal{N}(\xi'; \mathbf{0}, \mathbf{I})\, \mathrm{d}\xi'.$$

2. The LHS is, by definition, the posterior cumulative density function, $P(\mathbf{z}|\mathbf{X}) \triangleq \mathbb{P}(\mathbf{z}' \leq \mathbf{z}|\mathbf{X})$, evaluated at $\mathbf{z} = \mathbf{f}^*(\xi)$.
3. The inner integral over $\mathbf{z}'$ on the RHS reduces to the Heaviside function, Θ, evaluated at $\mathbf{f}^*(\xi) - \mathbf{f}^*(\xi')$.

Explicit Construction of f* II

4. The Heaviside function reduces the integration domain of the remaining integral over ξ' on the RHS:

$$P\left(\mathbf{f}^*(\xi)|\mathbf{X}\right) = \int_{\{\xi' : f_i^*(\xi') \leq f_i^*(\xi), i=1,\ldots,d\}} \mathcal{N}(\xi'; \mathbf{0}, \mathbf{I})\, \mathrm{d}\xi'.$$

5. At this point, I wish I had evidence that allowed me to replace the integration domain with $\{\xi' : \xi'_l \leq \xi_l, l = 1, \ldots, c\}$, because then the RHS would reduce to the standard multivariate normal cumulative distribution function, Φ, evaluated at ξ. We would then have $\mathbf{f}^*(\xi) = \mathbf{P}^{-1}(\Phi(\xi)) \triangleq \{\mathbf{z} : P(\mathbf{z}|\mathbf{X}) = \Phi(\xi)\}$.
But I don't have that evidence :(

VGP Approximation Theorem

Let $q_{\mathrm{VGP}}(\mathbf{z};\theta)$ denote the variational Gaussian process. Let $p(\mathbf{z}|\mathbf{X})$ be a posterior distribution with a finite number of latent variables and continuous quantile function (inverse cumulative distribution function), $\mathbf{P}^{-1}(\mathbf{z})$. Then there exists a sequence of parameters θ_m such that

$$\lim_{m\to\infty} \mathrm{KL}\left(q_{\mathrm{VGP}}(\mathbf{z};\theta_m)\|p(\mathbf{z}|\mathbf{X})\right) = 0.$$

Every posterior with strictly positive density $p(\mathbf{z}|\mathbf{X})$ can be represented by a VGP (because those always have continuous quantile functions).

Black Box Inference

Unfortunately, the ELBO,

$$\mathcal{L} = \mathbb{E}_{q_{\mathrm{VGP}}(\mathbf{z};\theta)}[\log p(\mathbf{X}, \mathbf{z})] - \mathbb{E}_{q_{\mathrm{VGP}}(\mathbf{z};\theta)}[\log q_{\mathrm{VGP}}(\mathbf{z};\theta)]\,,$$

is not analytically tractable because of the log-density $\log q_{\mathrm{VGP}}(\mathbf{z};\theta)$.
The paper derives a weaker lower bound for the log-evidence,

$$\begin{aligned}\tilde{\mathcal{L}} \triangleq\ & \mathbb{E}_{q_{\mathrm{VGP}}(\mathbf{z};\theta)}[\log p(\mathbf{X}|\mathbf{z})] \\ & - \mathbb{E}_{q_{\mathrm{VGP}}(\mathbf{z};\theta)}\left[\mathrm{KL}\big(q(\mathbf{z}|\mathbf{f}(\xi))\|p(\mathbf{z})\big) + \mathrm{KL}\big(q(\xi,\mathbf{f})\|r(\xi,\mathbf{f}|\mathbf{z})\big)\right]\,,\end{aligned}$$

and shows how to maximize that instead, where r is an auxiliary distribution...

• • •

Algorithm Complexity

$$\mathcal{O}(d + m^3 + LH^2),$$

where:

- d is the number of latent variables,
- m is the size of the fake data set θ,
- L is the number of layers of a neural network leveraged for optimization with
- H the average hidden layer size.

How Do We Use This?

We Don't (For Now)

@dustinvtran is Edward implementing the VGP from your ICLR 2016 paper? I looked through the code and couldn't find it anywhere.
— Torsten Scholak (@tscholak) August 12, 2016

@tscholak thanks for asking! its in a private branch at the moment, waiting for api changes to enable more expressive variational model
— Dustin Tran (@dustinvtran) August 12, 2016