Variational Learning for Sparse GP

Background

Given observation vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, we want to find the latent function value $\boldsymbol{f}$. Our goal is the posterior distribution $p(\boldsymbol{f}|\boldsymbol{y}) \propto (\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f})$ through maximizing the log marginal likelihood $\log p(\boldsymbol{y}) = \log[N(\boldsymbol{y}|\mathbf{0}, \sigma^2 I + K_{nn})]$, which uses $O(n^3)$ time.

Principles behind Variaitional Inference in Sparse GP

Besides latent variable $\boldsymbol{f}$, let us introduce two new latent variables $\boldsymbol{f}_m$ and $\boldsymbol{z}$. $\boldsymbol{f}_m$, named inducing variables, are function values evaluated at pseudo-inputs X_m, which live in the same space as training inputs $\boldsymbol{x}$ but are independent from them. For latent variable $\boldsymbol{z}$, it is any set of true function values.

Remember that we want posterior $p(\boldsymbol{f}|\boldsymbol{y})$, we can describe it by predictive distribution $p(\boldsymbol{z}|\boldsymbol{y})$

$$
\begin{aligned}
p(\boldsymbol{z}|\boldsymbol{y}) &= \iint p(\boldsymbol{z}, \boldsymbol{f}, \boldsymbol{f}_m|\boldsymbol{y})\, d\boldsymbol{f}\, d\boldsymbol{f}_m \\
&= \iint p(\boldsymbol{z}|\boldsymbol{f}, \boldsymbol{f}_m, \boldsymbol{y})p(\boldsymbol{f}, \boldsymbol{f}_m|\boldsymbol{y})\, d\boldsymbol{f}\, d\boldsymbol{f}_m \\
&= \iint p(\boldsymbol{z}|\boldsymbol{f}, \boldsymbol{f}_m)p(\boldsymbol{f}, \boldsymbol{f}_m|\boldsymbol{y})\, d\boldsymbol{f}\, d\boldsymbol{f}_m
\end{aligned}
$$

Now we want to use $q(\boldsymbol{z})$ to approximate $p(\boldsymbol{z}|\boldsymbol{y})$, i.e., $q(\boldsymbol{z}) \approx p(\boldsymbol{z}|\boldsymbol{y})$:

If we can approximate the posterior distribution $p(\boldsymbol{f}, \boldsymbol{f}_m|\boldsymbol{y})$ by variational distribution $q(\boldsymbol{f}, \boldsymbol{fm})$ and the inducing variables $\boldsymbol{f}_m$ are good enough to be sufficient statistics for $\boldsymbol{f}$, i.e., $q(\boldsymbol{f}, \boldsymbol{f}_m) \approx p(\boldsymbol{f}, \boldsymbol{f}_m|\boldsymbol{y})$, $p(\boldsymbol{z}|\boldsymbol{f}, \boldsymbol{f}_m) = p(\boldsymbol{z}|\boldsymbol{f}_m)$

Thus

$$
\begin{aligned}
p(\boldsymbol{z}|\boldsymbol{y}) \approx q(\boldsymbol{z}) &= \iint p(\boldsymbol{z}|\boldsymbol{f}_m)q(\boldsymbol{f}, \boldsymbol{f}_m)\, d\boldsymbol{f}\, d\boldsymbol{f}_m \\
&= \iint p(\boldsymbol{z}|\boldsymbol{f}_m)p(\boldsymbol{f}|\boldsymbol{f}_m)\phi(\boldsymbol{f}_m)\, d\boldsymbol{f}\, d\boldsymbol{f}_m \\
&= \int p(\boldsymbol{z}|\boldsymbol{f}_m)\{\int p(\boldsymbol{f}|\boldsymbol{f}_m)\, d\boldsymbol{f}\}\phi(\boldsymbol{f}_m)\, d\boldsymbol{f}_m \\
&= \int p(\boldsymbol{z}|\boldsymbol{f}_m)\phi(\boldsymbol{f}_m)\, d\boldsymbol{f}_m \\
&= \int q(\boldsymbol{z}, \boldsymbol{f}_m)\, d\boldsymbol{f}_m
\end{aligned}
$$

where $\phi(\boldsymbol{f}_m)$ is free variation Gaussian distribution for $\boldsymbol{f}_m$ with mean μ and covariance A. We can obtain the mean and covariance of this approximate posterior $q(\boldsymbol{z})$:

$$m(\boldsymbol{x}) = K_{\boldsymbol{x}m}K_{mm}^{-1}\mu \text{equation}(1)$$

$$k(\boldsymbol{x}, \boldsymbol{x\prime}) = k(\boldsymbol{x}, \boldsymbol{x\prime}) - K_{\boldsymbol{x}m}K_{mm}^{-1}K_{m\boldsymbol{x\prime}} + K_{\boldsymbol{x}m}K_{mm}^{-1}AK_{mm}^{-1}K_{m\boldsymbol{x\prime}}$$

The above defines the general form of sparse posterior GP, which is computed in $O(nm^2)$

Now our task is to find the optimal variational distribution $q(\boldsymbol{f}, \boldsymbol{f}_m) \approx p(\boldsymbol{f}, \boldsymbol{f}_m|\boldsymbol{y})$
In the theory of Variational Inference, the optimal variational distribution is always derived from a restricted distribution family, which has a factorization property. Here, the variational distribution $q(\boldsymbol{f}, \boldsymbol{f}_m)$ must satisfy the factorization: $q(\boldsymbol{f}, \boldsymbol{f}_m) = p(\boldsymbol{f}|\boldsymbol{f}_m)\phi(\boldsymbol{f}_m)$

I omit detailed derivations for the following equation:

$$\log p(\boldsymbol{y}) = \underbrace{\int q(\boldsymbol{f}, \boldsymbol{f}_m) \log \frac{p(\boldsymbol{f}, \boldsymbol{f}_m, \boldsymbol{y})}{q(\boldsymbol{f}, \boldsymbol{f}_m)}\, d\boldsymbol{f}\, d\boldsymbol{f}_m}_{ELBO} + KL(q(\boldsymbol{f}, \boldsymbol{f}_m)||p(\boldsymbol{f}, \boldsymbol{f}_m|y))$$

To get the optimal $q(\boldsymbol{f}, \boldsymbol{f}_m)$, we will minimize the KL divergence, $KL(q(\boldsymbol{f}, \boldsymbol{f}_m)||p(\boldsymbol{f}, \boldsymbol{f}_m|y))$, which is equivalent to maximize the evidence lower bound (ELBO), $\int q(\boldsymbol{f}, \boldsymbol{f}_m) \log \frac{p(\boldsymbol{f}, \boldsymbol{f}_m, \boldsymbol{y})}{q(\boldsymbol{f}, \boldsymbol{f}_m)}\, d\boldsymbol{f}\, d\boldsymbol{f}_m$. Now our objective function is ELBO, and our optimized variables are $\underbrace{X_m, \sigma^2, \theta}_{\text{variational parameters}}$ and $\underbrace{\phi(\boldsymbol{f}_m)}_{\text{free variational distribution}}$.

Therefore, I denote ELBO with $F_v(X_m, \sigma^2, \theta, \phi)$

We optimize $F_v(X_m, \sigma^2, \theta, \phi)$ with respect to $\phi(\boldsymbol{f}_m)$, and we get $\phi^*(\boldsymbol{f}_m) = N(\boldsymbol{f}_m|\mu, A)$, where $\mu = \sigma^{-2} K_{mm} \Sigma K_{mn} \boldsymbol{y}$, $A = K_{mm} \Sigma K_{mm}$, $\Sigma = (K_{mm} + \sigma^{-2} K_{mn} K_{nm})^{-1}$. Substitute μ and Σ into equation (1) and (2), then we can get the approximate posterior mean and variance.

Now with fixed optimal $\phi^*(\boldsymbol{f}_m)$, our objective function becomes

$$F_v = F_v(X_m, \sigma^2, \theta) = \log[N(\boldsymbol{y}|0, \sigma^2 I + Q_{nn})] - \frac{1}{2\sigma^2} \underbrace{Tr(K_{nn} - Q_{nn})}_{\text{total variance of } p(\boldsymbol{f}|\boldsymbol{f}_m)}$$

where $Q_{nn} = K_{nm} K_{mm}^{-1} K_{mn}$, and $K_{nn} - Q_{nn} = Var[\boldsymbol{f}|\boldsymbol{f}_m]$

So far, our goal is to maximize the $F_v = F_v(X_m, \sigma^2, \theta)$ with respect to variaitonal parameters (X_m, σ^2, θ). In standard variational inference, we implement the optimization with gradient descent method as the following algorithm shows:

Algorithm 1 standard variational inference

Initialize inducing inputs X_m
repeat
  $\theta = argmax F_v$, fixed σ^2, X_m
  $\sigma^2 = \frac{1}{n} \int \phi^*(\boldsymbol{f}_m) \parallel \boldsymbol{y} - K_{nm} K_{mm}^{-1} \boldsymbol{f}_m \parallel^2\, d\boldsymbol{f}_m + \frac{1}{n} Tr(K_{nn} - Q_{nn})$, fixed θ, X_m
  $X_m = argmax F_v$, fixed θ, σ^2
until Convergence

However, the standard gradient descent method will be difficult to implement; we can instead use gready selection method. Gready selection method results in a suboptimal solution, and it endures an easier algorithm. We have n training inputs, and basically, m inducing inputs will be selected among them. We start with with an empty inducing set $m = \emptyset$ and a remaining set $n - m = \{1, \cdots, n\}$. At each iteration, we add a training point $j \in J \subset n - m$, where J is a randomly chosen working set with the size of what we choose, into the inducing set that maximizes the selection criterion Δ_j. For me, I personly would like to choose the trace $Tr(K_{nn} - Q_{nn})$ to be the selection criterion, because it represents the total variance of the conditional prior $p(\boldsymbol{f}|\boldsymbol{f}_m)$. Thus smaller $Tr(K_{nn} - Q_{nn})$ means that the inducing variables $\boldsymbol{f}_m$ are more likely to contain more information of $\boldsymbol{f}$.

Here is the algorithm using greedy selection method.

Algorithm 2 variational inference using greedy selection method

Start with with an empty inducing set $m = \emptyset$ and a remaining set $n - m = \{1, \cdots, n\}$

repeat

(1). Add a training point $j \in J \subset n - m$, where J is a randomly chosen working set with the size of what we choose, into the inducing set that minimize $Tr(K_{nn} - Q_{nn})$, fixed θ, σ^2

(2). $\theta = argmax F_v$, fixed σ^2, X_m

(3). $\sigma^2 = \frac{1}{n} \int \phi^*(\boldsymbol{f}_m) \parallel \boldsymbol{y} - K_{nm} K_{mm}^{-1} \boldsymbol{f}_m \parallel^2 \ d\boldsymbol{f}_m + \frac{1}{n} Tr(K_{nn} - Q_{nn})$, fixed θ, X_m

until Convergence

This summary is my lastest understading after I read the Variational Inference Chapter in Pattern Recognition and Machine Learning, and the 2013 paper Stochastic Variational Inference. Thus the algorithm part of this summary is a little different from what I did in my previous code for the GSM kernel with variational inference. I will try to adapt my code to this summary, and to see whether I will obtain new improvements in my work.