Sparse Variational Gaussian Processes

目　录

第 1 章 Gaussian Distribution

本章主要回顾多元高斯分布及相应边缘概率和条件概率形式

1.1 联合概率

定义 $\boldsymbol{x} \in \mathbb{R}^d$ 为随机向量，其对应的均值为 $\boldsymbol{\mu}$，对应的协方差矩阵为 Σ，协方差矩阵衡量了随机变量两两间的相关程度。则维度为 d 的随机向量相应的高斯分布概率密度函数如下：

$$p(\boldsymbol{x}|\boldsymbol{\mu}, \Sigma) = \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) \tag{1.1}$$

1.2 边缘概率

定义随机向量 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 服从如下联合高斯分布：

$$\begin{bmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\boldsymbol{\mu_x}\\ \boldsymbol{\mu_y}\end{bmatrix}, \begin{bmatrix}\boldsymbol{A} & \boldsymbol{C}\\ \boldsymbol{C}^\top & \boldsymbol{B}\end{bmatrix}\right) \tag{1.2}$$

则随机向量 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 分别服从如下边缘分布:

$$p(\boldsymbol{x}) \sim \mathcal{N}(\boldsymbol{\mu_x}, \boldsymbol{A}) \tag{1.3}$$

$$p(\boldsymbol{y}) \sim \mathcal{N}(\boldsymbol{\mu_y}, \boldsymbol{B}) \tag{1.4}$$

1.3 条件概率

定义随机向量 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 服从如下联合高斯分布：

$$\begin{bmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\boldsymbol{\mu_x}\\ \boldsymbol{\mu_y}\end{bmatrix}, \begin{bmatrix}\boldsymbol{A} & \boldsymbol{C}\\ \boldsymbol{C}^\top & \boldsymbol{B}\end{bmatrix}\right) \tag{1.5}$$

则给定随机向量 $\boldsymbol{y}$ 取值时，随机向量 $\boldsymbol{x}$ 对应的条件概率服从如下高斯分布：

$$p(\boldsymbol{x}|\boldsymbol{y}) = \mathcal{N}(\mu_{\boldsymbol{x}|\boldsymbol{y}}, \Sigma_{\boldsymbol{x}|\boldsymbol{y}}) \tag{1.6}$$

其中：

$$\boldsymbol{\mu_{x|y}} = \boldsymbol{\mu_x} + \boldsymbol{C}\boldsymbol{B}^{-1}(\boldsymbol{y}-\boldsymbol{\mu_y}) \tag{1.7}$$

$$\Sigma_{\boldsymbol{x}|\boldsymbol{y}} = \boldsymbol{A} - \boldsymbol{C}\boldsymbol{B}^{-1}\boldsymbol{C}^\top \tag{1.8}$$

第 2 章 Gaussian Processes

本章涉及高斯过程先验分布、似然分布及后验分布的推理，最后给出高斯过程如何应用于预测

2.1 先验分布

高斯过程是关于函数值随机变量 $f(\boldsymbol{x})$ 的分布，分布的定义由均值函数 $m(\boldsymbol{x})$ 和核函数 $k(\boldsymbol{x},\boldsymbol{x}')$ 给定，其中 $\boldsymbol{x}$ 和 $\boldsymbol{x}'$ 为随机变量 $f(\boldsymbol{x})$ 及 $f(\boldsymbol{x}')$ 相应的输入。因此对于随机变量 $\boldsymbol{x} \in \mathbb{R}^d$ 的任意子集构成的矩阵 $\boldsymbol{X} \in \mathbb{R}^{n\times d}$，其对应的随机向量 $f(\boldsymbol{X})$ 服从如下多元高斯分布：

$$f(\boldsymbol{X}) \sim \mathcal{N}(m(\boldsymbol{X}), k(\boldsymbol{X},\boldsymbol{X})) \tag{2.1}$$

其中均值向量为 $\boldsymbol{\mu} = m(\boldsymbol{X})$，协方差矩阵 $\Sigma = k(\boldsymbol{X},\boldsymbol{X})$。公式2.1为高斯过程先验 (GP Prior)，应用中一般令均值函数 $m(\boldsymbol{x}) = 0$，协方差矩阵 Σ 由核函数 $k(\boldsymbol{x},\boldsymbol{x}')$ 给定。在高斯过程中称 $f(\boldsymbol{X})$ 服从的多元高斯分布为高斯过程先验，高斯过程先验即为我们假设随机变量 $f(\boldsymbol{X})$ 服从的联合分布。

2.2 核函数

核函数有不同的形式，分别对应的不同的随机变量 $f(\boldsymbol{x})$ 服从的先验分布。这里仅介绍在 EE 问题应用中涉及的径向基核函数 (RBF)，函数定义如下：

$$k_{\text{RBF}}(\boldsymbol{x}_{i,:},\boldsymbol{x}_{j,:}) = \sigma^2 \exp\left(-\frac{1}{2\ell^2}\sum_{q=1}^{d}\left(\boldsymbol{x}_{i,q} - \boldsymbol{x}_{j,q}\right)^2\right) \tag{2.2}$$

其中 σ^2 为方差，ℓ 为长度缩放系数。

2.3 似然分布

高斯过程先验仅涉及随机变量 $\boldsymbol{x}$，通过似然 (Likelihood) 可以将 GP Prior 中的随机变量 $\boldsymbol{x}$ 和相应的观察值 y 联系在一起，给定先验和似然后，即可计算后验，有了后验后，即可用于预测。似然为给定随机变量 $f(\boldsymbol{x})$ 后，观察到 y 的概率。假定给定随机变量 $\boldsymbol{x} \in \mathbb{R}^d$ 的任意子集构成的矩阵 $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ 后，随机向量 $y(\boldsymbol{X})$ 在给定 $f(\boldsymbol{X})$ 时，服从如下多元高斯分布：

$$p(y(\boldsymbol{X})|f(\boldsymbol{X})) = \mathcal{N}(f(\boldsymbol{X}), \eta^2\boldsymbol{I}_n) \tag{2.3}$$

上述形式即为似然函数，其均值为 $f(\boldsymbol{X})$，协方差为 $\eta^2\boldsymbol{I}_n$，其中 η^2 为模型参数，称为高斯噪声，$\boldsymbol{I}_n$ 为 $n\times n$ 的单位矩阵。

2.4 边缘分布

似然函数2.3等价定义为为如下高斯线性变换：

$$y(\boldsymbol{X}) = \boldsymbol{I}_n f(\boldsymbol{X}) + \epsilon \qquad \text{其中：} \ \epsilon \sim \mathcal{N}(\boldsymbol{0}, \eta^2 \boldsymbol{I}_n) \tag{2.4}$$

通过上述形式，可以应用高斯线性变换规则，推理出随机变量 $y(\boldsymbol{X})$ 的边缘分布。给定随机向量 $\boldsymbol{a}$ 服从如下多元高斯分布

$$\boldsymbol{a} \sim \mathcal{N}(\boldsymbol{\mu_a}, \Sigma_{\boldsymbol{a}}) \tag{2.5}$$

随机向量 $\boldsymbol{b}$ 为随机向量 $\boldsymbol{a}$ 的线性变换：

$$\boldsymbol{b} = \boldsymbol{A}\boldsymbol{a} + \eta \qquad \text{其中：} \ \eta \sim \mathcal{N}(0, \Sigma_\eta) \tag{2.6}$$

则随机向量 $\boldsymbol{b}$ 服从如下多元高斯分布：

$$\boldsymbol{b} \sim \mathcal{N}(\boldsymbol{A}\mu_{\boldsymbol{a}}, \boldsymbol{A}\Sigma_{\boldsymbol{a}}\boldsymbol{A}^\top + \Sigma_\eta) \tag{2.7}$$

给定上述线性变换规则后，由于 $f(\boldsymbol{X})$ 服从分布2.1，则 $y(\boldsymbol{X})$ 服从如下分布：

$$y(\boldsymbol{X}) \sim \mathcal{N}(m(\boldsymbol{X}), k(\boldsymbol{X}, \boldsymbol{X}) + \eta^2 \boldsymbol{I}_n) \tag{2.8}$$

2.5 后验分布

给定高斯过程先验分布和边缘分布后，可以求得任意输入 $\boldsymbol{x'}$ 对应的随机变量 $f(\boldsymbol{x'})$ 的后验分布 $p(f(\boldsymbol{x'})|y(\boldsymbol{X}))$。为了求解该分布，由于随机变量 $f(\boldsymbol{x'})$ 和随机向量 $y(\boldsymbol{X})$ 均服从高斯分布，由高斯分布性质可知，其联合分布依然为高斯分布，形式如下：

$$\begin{bmatrix} f(\boldsymbol{x'}) \\ y(\boldsymbol{X}) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m(\boldsymbol{x'}) \\ m(\boldsymbol{X}) \end{bmatrix}, \begin{bmatrix} k(\boldsymbol{x'}, \boldsymbol{x'}) & \operatorname{cov}(f(\boldsymbol{x'}), y(\boldsymbol{X})) \\ \operatorname{cov}(f(\boldsymbol{x'}), y(\boldsymbol{X}))^\top & k(\boldsymbol{X}, \boldsymbol{X}) + \eta^2 \boldsymbol{I}_n \end{bmatrix} \right)$$

其中

$$\begin{aligned} \operatorname{cov}(f(\boldsymbol{x'}), y(\boldsymbol{X})) &= \mathbb{E}[(f(\boldsymbol{x'}) - m(\boldsymbol{x'}))(y(\boldsymbol{X}) - m(\boldsymbol{X}))] \\ &= \mathbb{E}[(f(\boldsymbol{x'}) - m(\boldsymbol{x'}))(f(\boldsymbol{X}) + \epsilon - m(\boldsymbol{X}))] \\ &= \mathbb{E}[(f(\boldsymbol{x'}) - m(\boldsymbol{x'}))(f(\boldsymbol{X}) - m(\boldsymbol{X})) + (f(\boldsymbol{x'}) - m(\boldsymbol{x'}))\epsilon] \\ &= \operatorname{cov}(f(\boldsymbol{x'}), f(\boldsymbol{X})) + \mathbb{E}[(f(\boldsymbol{x'}) - m(\boldsymbol{x'}))\epsilon] \\ &= k(\boldsymbol{x'}, \boldsymbol{X}) + 0 \end{aligned}$$

因此有：

$$\begin{bmatrix} f(\boldsymbol{x}') \\ y(\boldsymbol{X}) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m(\boldsymbol{x}') \\ m(\boldsymbol{X}) \end{bmatrix}, \begin{bmatrix} k(\boldsymbol{x}', \boldsymbol{x}') & k(\boldsymbol{x}', \boldsymbol{X}) \\ k(\boldsymbol{x}', \boldsymbol{X})^\top & k(\boldsymbol{X}, \boldsymbol{X}) + \eta^2 \boldsymbol{I}_n \end{bmatrix} \right)$$

结合上一章的条件概率公式可知

$$p(f(\boldsymbol{x}')|y(\boldsymbol{X})) = \mathcal{N}(\mu', \sigma^{2'})$$

其中：

$$\mu' = m(\boldsymbol{x}') + k(\boldsymbol{x}', \boldsymbol{X})(k(\boldsymbol{X}, \boldsymbol{X}) + \eta^2 \boldsymbol{I}_n)^{-1}(y(\boldsymbol{X}) - m(\boldsymbol{X})) \tag{2.9}$$

$$\sigma^{2'} = k(\boldsymbol{x}', \boldsymbol{x}') - k(\boldsymbol{x}', \boldsymbol{X})(k(\boldsymbol{X}, \boldsymbol{X}) + \eta^2 \boldsymbol{I}_n)^{-1} k(\boldsymbol{x}', \boldsymbol{X})^\top \tag{2.10}$$

2.6 预测应用

后验分布 $p(f(\boldsymbol{x}')|y(\boldsymbol{X}))$ 为给定 $y(\boldsymbol{X})$ 条件下，隐藏随机变量 $f(\boldsymbol{x}')$ 的概率分布，由于观测随机变量来自于引入高斯噪声后的隐藏随机变量。因此给定 $y(\boldsymbol{X})$ 条件时，观测随机变量 $y(\boldsymbol{x}')$ 满足如下线性变换：

$$y(\boldsymbol{x}')|y(\boldsymbol{X}) = f(\boldsymbol{x}')|y(\boldsymbol{X}) + \epsilon' \qquad \text{其中：} \ \epsilon' \sim \mathcal{N}(0, \eta^2) \tag{2.11}$$

因此有：

$$p(y(\boldsymbol{x}')|y(\boldsymbol{X})) = \mathcal{N}(\mu', \sigma^{2'} + \eta^2) \tag{2.12}$$

第 3 章 Variational GPs

前面在介绍高斯过程时，假定似然函数服从高斯分布。本章我们来看，当似然函数不服从高斯分布时，如何求解后验分布。

3.1 问题引出

假设我们要解决的是二分类问题，贝叶斯推断中总是包含先验和似然这两项，因此我们分别来看。

3.1.1 先验

对于二分类问题来说，先验项依然为随机变量 $f(\boldsymbol{X})$ 的联合分布，假设其服从如下高斯分布：

$$f(\boldsymbol{X}) \sim \mathcal{N}(m(\boldsymbol{X}), k(\boldsymbol{X}, \boldsymbol{X})) \tag{3.1}$$

其中 $m(\boldsymbol{X})$ 为均值，一般令均值为 0，$k(\boldsymbol{X}, \boldsymbol{X})$ 为协方差，可以采用 RBF 核函数，因此其概率密度函数如下所示：

$$p(\boldsymbol{f}; \ell, \sigma^2) = \frac{1}{\sqrt{(2\pi)^n \det(\boldsymbol{K})}} \exp\left(-\frac{1}{2}\boldsymbol{f}^\top \boldsymbol{K}^{-1} \boldsymbol{f}\right) \tag{3.2}$$

上述建模过程中未考虑测试噪声，和上一章类似，引入高斯噪声后：

$$f(\boldsymbol{X}) \sim \mathcal{N}(m(\boldsymbol{X}), k(\boldsymbol{X}, \boldsymbol{X}) + \eta^2 \boldsymbol{I}_n) \tag{3.3}$$

因此引入高斯噪声后，可以用 $k(\boldsymbol{X}, \boldsymbol{X}) + \eta^2 \boldsymbol{I}_n$ 替代 $k(\boldsymbol{X}, \boldsymbol{X})$ 用于有噪声建模。

3.1.2 似然

定义随机变量 y 对应二分类问题中观察到的 label，f 为随机变量 y 对应的潜在随机变量。由于二分类问题服从伯努利分布，其对应似然函数形式如下：

$$p(y_i|f_i) = g(f_i)^{y_i}(1 - g(f_i))^{(1-y_i)} \tag{3.4}$$

其中：

$$g(f_i) = \frac{e^{f_i}}{1 + e^{f_i}} \tag{3.5}$$

对于整个训练集来说有：

$$p(\boldsymbol{y}|\boldsymbol{f}) = \prod_{i=1}^{n} p(y_i|f_i) \tag{3.6}$$

3.1.3 后验

给定先验和似然后，由贝叶斯法则，可以知：

$$p(\boldsymbol{f}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f})}{\int p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f})d\boldsymbol{f}} \tag{3.7}$$

求解后验分布需要对分母的积分进行简化，求得其相应的解析表达式。由于先验为高斯分布，而似然为伯努利分布，伯努利分布非高斯分布的共轭分布，因此无法求得上述积分形式的解析表达式。

3.2 变分近似

虽然无法直接求得后验分布 $p(\boldsymbol{f}|\boldsymbol{y})$ 的解析表达式，但可以通过另外一个分布 $q(\boldsymbol{f})$ 来近似后验分布，希望该分布尽量和后验分布一致，称该近似分布为变分分布。假设变分分布服从均值为 $\boldsymbol{v}$，协方差矩阵为 $\boldsymbol{S}$ 的高斯分布，则有：

$$q(f) = \mathcal{N}(\boldsymbol{v}, \boldsymbol{S}) \tag{3.8}$$

给定变分分布后，可以通过 KL 距离衡量变分分布和后验分布近似程度：

$$\begin{aligned}
KL(q(\boldsymbol{f})||p(\boldsymbol{f}|\boldsymbol{y})) &= \mathbb{E}_{\boldsymbol{f}\sim q}\left[\log\left(\frac{q(\boldsymbol{f})}{p(\boldsymbol{f}|\boldsymbol{y})}\right)\right] \\
&= \mathbb{E}_{\boldsymbol{f}\sim q}\left[\log(q(\boldsymbol{f}))\right] - \mathbb{E}_{\boldsymbol{f}\sim q}\left[\log(p(\boldsymbol{f}|\boldsymbol{y}))\right] \\
&= \mathbb{E}_{\boldsymbol{f}\sim q}\left[\log(q(\boldsymbol{f}))\right] - \mathbb{E}_{\boldsymbol{f}\sim q}\left[\log\left(\frac{p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f})}{p(\boldsymbol{y})}\right)\right] \\
&= \mathbb{E}_{\boldsymbol{f}\sim q}\left[\log(q(\boldsymbol{f}))\right] - \mathbb{E}_{\boldsymbol{f}\sim q}\left[\log\left(p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f})\right)\right] + \mathbb{E}_{\boldsymbol{f}\sim q}\left[\log p(\boldsymbol{y})\right] \\
&= \mathbb{E}_{\boldsymbol{f}\sim q}\left[\log(q(\boldsymbol{f}))\right] - \mathbb{E}_{\boldsymbol{f}\sim q}\left[\log\left(p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f})\right)\right] + \log p(\boldsymbol{y})
\end{aligned} \tag{3.9}$$

令

$$\text{ELBO} = -(\mathbb{E}_{\boldsymbol{f}\sim q}\left[\log(q(\boldsymbol{f}))\right] - \mathbb{E}_{\boldsymbol{f}\sim q}\left[\log\left(p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f})\right)\right]) \tag{3.10}$$

因此有：

$$KL(q(\boldsymbol{f})||p(\boldsymbol{f}|\boldsymbol{y})) = \log p(\boldsymbol{y}) - \text{ELBO} \tag{3.11}$$

有上述形式可知，由于 $KL(q(\boldsymbol{f})||p(\boldsymbol{f}|\boldsymbol{y}))$ 大于等于 0，$\log p(\boldsymbol{y})$ 和参数无关，因此要最小化变分分布和后验分布的 KL 距离，可以最大化 ELBO。

3.3 ELBO 解析式

由上一小节可知，ELBO 公式形式如下：

$$\begin{aligned}
\text{ELBO} &= -(\mathbb{E}_{\boldsymbol{f}\sim q}\left[\log(q(\boldsymbol{f}))\right] - \mathbb{E}_{\boldsymbol{f}\sim q}\left[\log\left(p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f})\right)\right]) \\
&= -\mathbb{E}_{\boldsymbol{f}\sim q}[\log(q(\boldsymbol{f}))] + \mathbb{E}_{\boldsymbol{f}\sim q}[\log(p(\boldsymbol{y}|\boldsymbol{f}))] + \mathbb{E}_{\boldsymbol{f}\sim q}[\log(p(\boldsymbol{f}))] \\
&= \mathbb{E}_{\boldsymbol{f}\sim q}[\log(p(\boldsymbol{y}|\boldsymbol{f}))] - \mathbb{E}_{\boldsymbol{f}\sim q}\left[\frac{q(\boldsymbol{f})}{p(\boldsymbol{f})}\right] \\
&= \mathbb{E}_{\boldsymbol{f}\sim q}[\log(p(\boldsymbol{y}|\boldsymbol{f}))] - KL(q(\boldsymbol{f})||p(\boldsymbol{f})) \\
&= \int \log(p(\boldsymbol{y}|\boldsymbol{f}))q(\boldsymbol{f})d\boldsymbol{f} - KL(q(\boldsymbol{f})||p(\boldsymbol{f}))
\end{aligned} \tag{3.12}$$

可见 ELBO 形式中的似然项和 KL 距离项均为积分形式，求得上述两项的解析式后，即可通过梯度上升方法求解使得 ELBO 取得最大值的参数。

3.3.1 似然解析式

给定 ELBO 形式后，我们首先看如何求得似然项的解析式

$$\begin{aligned}
\int \log(p(\boldsymbol{y}|\boldsymbol{f}))q(\boldsymbol{f})d\boldsymbol{f} &= \int \log\left(\prod_{i=1}^{n} p(y_i|f_i)\right) q(\boldsymbol{f})d\boldsymbol{f} \\
&= \int \left(\sum_{i=1}^{n} \log\left(p(y_i|f_i)\right)\right) q(\boldsymbol{f})d\boldsymbol{f} \\
&= \sum_{i=1}^{n} \left(\int \log(p(y_i|f_i))q(\boldsymbol{f})d\boldsymbol{f}\right)
\end{aligned} \tag{3.13}$$

对上述求和形式中每一项的积分形式进行化简：

$$\begin{aligned}
\int \log(p(y_i|f_i))q(\boldsymbol{f})d\boldsymbol{f} &= \int \cdots \int \log(p(y_i|f_i))q([f_1,\ldots,f_n]^\top)df_1\ldots df_n \\
&= \iint \log(p(y_i|f_i))q([\boldsymbol{f}_{-i},f_i]^\top)d\boldsymbol{f}_{-i}df_i \\
&= \int \log(p(y_i|f_i)) \left(\int q([\boldsymbol{f}_{-i},f_i]^\top)d\boldsymbol{f}_{-i}\right) df_i \\
&= \int \log(p(y_i|f_i))q(f_i)df_i
\end{aligned} \tag{3.14}$$

上述公式中 $\boldsymbol{f}_{-i}$ 表示向量 $\boldsymbol{f}$ 中除 f_i 的所有元素。由多元高斯分布的性质可知，变分分布 $q(\boldsymbol{f})$ 的边缘分布为：

$$\begin{aligned}
q(f_i) &= \mathcal{N}(f_i;\mu_i,\Sigma_{i,i}^2) \\
&= \frac{1}{\sqrt{2\pi\Sigma_{i,i}^2}} \exp\left(-\frac{(f_i-\mu_i)^2}{2\Sigma_{i,i}^2}\right)
\end{aligned} \tag{3.15}$$

因此有：

$$\int \log(p(y_i|f_i))q(f_i)df_i = \int \log(p(y_i|f_i))\frac{1}{\sqrt{2\pi\Sigma_{i,i}^2}}\exp\left(-\frac{(f_i-\mu_i)^2}{2\Sigma_{i,i}^2}\right)df_i \tag{3.16}$$

可以通过 Gauss-Hermit quadrature 求得上述积分形式的数值近似解析式，其定义如下：

$$\int_{-\infty}^{+\infty} e^{-x^2}h(x)dx \approx \sum_{i=1}^{n} w_i h(x_i) \tag{3.17}$$

为了应用上述形式求解似然项的积分近似解析式，定义随机变量 f_i 满足如下形式：

$$f_i = \Sigma_{i,i}u + \mu_i \tag{3.18}$$

由高斯线性变换可知，随机变量 u 服从均值为 0，方差为 1 的高斯分布，具体推理如下：

$$\begin{aligned} f_i &\sim \mathcal{N}(\mu_i + \Sigma_{i,i}\cdot 0, \Sigma_{i,i}1\Sigma_{i,i}) \\ &\sim \mathcal{N}(\mu_i, \Sigma_{i,i}^2) \end{aligned} \tag{3.19}$$

给定 f_i 的变换形式后，似然项的积分形式可以简化为如下形式：

$$\begin{aligned} &\int \log(p(y_i|f_i))q(f_i)df_i \\ &= \int \log(p(y_i|f_i))\frac{1}{\sqrt{2\pi\Sigma_{i,i}^2}}\exp\left(-\frac{(f_i-\mu_i)^2}{2\Sigma_{i,i}^2}\right)df_i \\ &= \int \log(p(y_i|\Sigma_{i,i}u+\mu_i))\frac{1}{\sqrt{2\pi\Sigma_{i,i}^2}}\exp\left(-\frac{(\Sigma_{i,i}u+\mu_i-\mu_i)^2}{2\Sigma_{i,i}^2}\right)d(\Sigma_{i,i}u+\mu_i) \\ &= \int \log(p(y_i|\Sigma_{i,i}u+\mu_i))\frac{1}{\sqrt{2\pi\Sigma_{i,i}^2}}\exp\left(-\frac{1}{2}u^2\right)\cdot\Sigma_{i,i}du \\ &= \int \log(p(y_i|\Sigma_{i,i}u+\mu_i))\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}u^2\right)du \end{aligned} \tag{3.20}$$

令 $t = 1/\sqrt{2}u$，则 $u = \sqrt{2}t$，因此有：

$$\begin{aligned} &\int \log(p(y_i|\Sigma_{i,i}u+\mu_i))\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}u^2\right)\cdot du \\ &= \int \log(p(y_i|\sqrt{2}\Sigma_{i,i}t+\mu_i))\frac{1}{\sqrt{2\pi}}\exp\left(-t^2\right)\cdot\sqrt{2}dt \\ &= \int \log(p(y_i|\sqrt{2}\Sigma_{i,i}t+\mu_i))\frac{1}{\sqrt{\pi}}\exp\left(-t^2\right)dt \end{aligned} \tag{3.21}$$

因此令

$$h(t) = \frac{1}{\sqrt{\pi}}\log(p(y_i|\sqrt{2}\Sigma_{i,i}t+\mu_i)) \tag{3.22}$$

其中 $p(y_i|\sqrt{2}\Sigma_{i,i}t+\mu_i)$ 服从伯努利分布，因此有：

$$\begin{aligned}&\int \log(p(y_i|\sqrt{2}\Sigma_{i,i}t+\mu_i))\frac{1}{\sqrt{\pi}}\exp(-t^2)dt\\&\approx\sum_{j=1}^{m}w_j\cdot\frac{1}{\sqrt{\pi}}\log\left(\left(\frac{e^{(\sqrt{2}\Sigma_{i,i}t_j+\mu_i)}}{1+e^{(\sqrt{2}\Sigma_{i,i}t_j+\mu_i)}}\right)^{y_i}\left(1-\frac{e^{(\sqrt{2}\Sigma_{i,i}t_j+\mu_i)}}{1+e^{(\sqrt{2}\Sigma_{i,i}t_j+\mu_i)}}\right)^{(1-y_i)}\right)\end{aligned}\tag{3.23}$$

上述形式即为似然项积分形式对应的近似解析式

3.3.2 KL 距离解析式

由上一小节可知，KL 距离项 $KL(q(\boldsymbol{f})||p(\boldsymbol{f}))$，其中 $q(\boldsymbol{f})$ 和 $p(\boldsymbol{f})$ 分别服从如下分布：

$$\begin{aligned}q(\boldsymbol{f})=N(\boldsymbol{v},\boldsymbol{S})&=\frac{1}{\sqrt{(2\pi)^d|\boldsymbol{S}|}}\exp\left(-\frac{1}{2}(\boldsymbol{f}-\boldsymbol{v})^\top\boldsymbol{S}^{-1}(\boldsymbol{f}-\boldsymbol{v})\right)\\p(\boldsymbol{f})=N(\boldsymbol{m},\boldsymbol{K})&=\frac{1}{\sqrt{(2\pi)^d|\boldsymbol{K}|}}\exp\left(-\frac{1}{2}(\boldsymbol{f}-\boldsymbol{m})^\top\boldsymbol{K}^{-1}(\boldsymbol{f}-\boldsymbol{m})\right)\end{aligned}$$

因此有:

$$\begin{aligned}&KL(q(\boldsymbol{f})||p(\boldsymbol{f}))\\&=\int q(\boldsymbol{f})\log\left(\frac{q(\boldsymbol{f})}{p(\boldsymbol{f})}\right)d\boldsymbol{f}\\&=\int q(\boldsymbol{f})\left(\log(q(\boldsymbol{f}))-\log(p(\boldsymbol{f}))\right)d\boldsymbol{f}\end{aligned}$$

其中:

$$\begin{aligned}\log(q(\boldsymbol{f}))&=\log\left(\frac{1}{\sqrt{(2\pi)^d|\boldsymbol{S}|}}\exp\left(-\frac{1}{2}(\boldsymbol{f}-\boldsymbol{v})^\top\boldsymbol{S}^{-1}(\boldsymbol{f}-\boldsymbol{v})\right)\right)\\&=\log\left(\frac{1}{\sqrt{(2\pi)^d|\boldsymbol{S}|}}\right)-\frac{1}{2}(\boldsymbol{f}-\boldsymbol{v})^\top\boldsymbol{S}^{-1}(\boldsymbol{f}-\boldsymbol{v})\\&=-\frac{1}{2}\log\left((2\pi)^d\right)-\frac{1}{2}\log(|\boldsymbol{S}|)-\frac{1}{2}(\boldsymbol{f}-\boldsymbol{v})^\top\boldsymbol{S}^{-1}(\boldsymbol{f}-\boldsymbol{v})\end{aligned}$$

$$\begin{aligned}\log(p(\boldsymbol{f}))&=\log\left(\frac{1}{\sqrt{(2\pi)^d|\boldsymbol{K}|}}\exp\left(-\frac{1}{2}(\boldsymbol{f}-\boldsymbol{m})^\top\boldsymbol{K}^{-1}(\boldsymbol{f}-\boldsymbol{m})\right)\right)\\&=\log\left(\frac{1}{\sqrt{(2\pi)^d|\boldsymbol{K}|}}\right)-\frac{1}{2}(\boldsymbol{f}-\boldsymbol{m})^\top\boldsymbol{K}^{-1}(\boldsymbol{f}-\boldsymbol{m})\\&=-\frac{1}{2}\log\left((2\pi)^d\right)-\frac{1}{2}\log(|\boldsymbol{K}|)-\frac{1}{2}(\boldsymbol{f}-\boldsymbol{m})^\top\boldsymbol{K}^{-1}(\boldsymbol{f}-\boldsymbol{m})\end{aligned}$$

结合上式可知：

$$\begin{aligned}&\log(q(\boldsymbol{f}))-\log(p(\boldsymbol{f}))\\&=\frac{1}{2}\log\left(\frac{|\boldsymbol{K}|}{|\boldsymbol{S}|}\right)-\frac{1}{2}(\boldsymbol{f}-\boldsymbol{v})^{\top}\boldsymbol{S}^{-1}(\boldsymbol{f}-\boldsymbol{v})+\frac{1}{2}(\boldsymbol{f}-\boldsymbol{m})^{\top}\boldsymbol{K}^{-1}(\boldsymbol{f}-\boldsymbol{m})\end{aligned}$$

结合 Matrix Cookbook 第 8.2 节的性质：假设 $\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{m},\Sigma)$，则：

$$\mathbb{E}\left[(\boldsymbol{x}-\boldsymbol{m}')^{\top}\boldsymbol{A}(\boldsymbol{x}-\boldsymbol{m}')\right]=(\boldsymbol{m}-\boldsymbol{m}')^{\top}\boldsymbol{A}(\boldsymbol{m}-\boldsymbol{m}')+\mathrm{Tr}(\boldsymbol{A}\Sigma)\tag{3.24}$$

因此有：

$$\begin{aligned}&KL(q(\boldsymbol{f})||p(\boldsymbol{f}))\\&=\int q(\boldsymbol{f})\left(\frac{1}{2}\log\left(\frac{|\boldsymbol{K}|}{|\boldsymbol{S}|}\right)-\frac{1}{2}(\boldsymbol{f}-\boldsymbol{v})^{\top}\boldsymbol{S}^{-1}(\boldsymbol{f}-\boldsymbol{v})+\frac{1}{2}(\boldsymbol{f}-\boldsymbol{m})^{\top}\boldsymbol{K}^{-1}(\boldsymbol{f}-\boldsymbol{m})\right)d\boldsymbol{f}\\&=\frac{1}{2}\log\left(\frac{|\boldsymbol{K}|}{|\boldsymbol{S}|}\right)+\mathbb{E}_{\boldsymbol{f}\sim q}\left[-\frac{1}{2}(\boldsymbol{f}-\boldsymbol{v})^{\top}\boldsymbol{S}^{-1}(\boldsymbol{f}-\boldsymbol{v})+\frac{1}{2}(\boldsymbol{f}-\boldsymbol{m})^{\top}\boldsymbol{K}^{-1}(\boldsymbol{f}-\boldsymbol{m})\right]\\&=\frac{1}{2}\log\left(\frac{|\boldsymbol{K}|}{|\boldsymbol{S}|}\right)+\frac{1}{2}\left((\boldsymbol{v}-\boldsymbol{m})^{\top}\boldsymbol{K}^{-1}(\boldsymbol{v}-\boldsymbol{m})+\mathrm{Tr}(\boldsymbol{K}^{-1}\boldsymbol{S})\right)-\frac{1}{2}\mathrm{Tr}(\boldsymbol{S}^{-1}\boldsymbol{S})\\&=\frac{1}{2}\left(\log\left(\frac{|\boldsymbol{K}|}{|\boldsymbol{S}|}\right)+(\boldsymbol{v}-\boldsymbol{m})^{\top}\boldsymbol{K}^{-1}(\boldsymbol{v}-\boldsymbol{m})+\mathrm{Tr}(\boldsymbol{K}^{-1}\boldsymbol{S})-n\right)\end{aligned}\tag{3.25}$$

上式即为 KL 距离的解析式。

3.4 预测应用

求得后验分布 $p(\boldsymbol{f}(\boldsymbol{X})|\boldsymbol{y}(\boldsymbol{X}))$ 的变分分布 $q(\boldsymbol{f}(\boldsymbol{X}))$ 近似后，我们即可应用该分布求解函数在 $\boldsymbol{x}'$ 处的函数分布，令随机变量 f' 表示 $f(\boldsymbol{x}')$，随机向量 $\boldsymbol{f}$ 表示 $f(\boldsymbol{X})$，随机向量 $\boldsymbol{y}$ 表示 $y(\boldsymbol{X})$，则有：

$$\begin{aligned}p(f(\boldsymbol{x}')|y(\boldsymbol{X}))&=\int p(f(\boldsymbol{x}'),f(\boldsymbol{X})|y(\boldsymbol{X}))df(\boldsymbol{X})=\int p(f',\boldsymbol{f}|\boldsymbol{y})d\boldsymbol{f}\\&=\int p(f'|\boldsymbol{f},\boldsymbol{y})p(\boldsymbol{f}|\boldsymbol{y})d\boldsymbol{f}=\int p(f'|\boldsymbol{f})p(\boldsymbol{f}|\boldsymbol{y})d\boldsymbol{f}\\&=\int p(f'|\boldsymbol{f})q(\boldsymbol{f})d\boldsymbol{f}\end{aligned}$$

由多元高斯分布条件概率公式可知：

$$p(f'|\boldsymbol{f})=\mathcal{N}(\mu',\sigma^{2'})$$

其中：

$$\mu' = \boldsymbol{m} + k(\boldsymbol{x}', \boldsymbol{X})k(\boldsymbol{X}, \boldsymbol{X})^{-1}(\boldsymbol{f} - \boldsymbol{m})$$
$$\sigma^{2'} = k(\boldsymbol{x}', \boldsymbol{x}') - k(\boldsymbol{x}', \boldsymbol{X})k(\boldsymbol{X}, \boldsymbol{X})^{-1}k(\boldsymbol{x}', \boldsymbol{X})^{\top}$$

由于现实应用中一般 $\boldsymbol{m}$ 为 $\boldsymbol{0}$，因此有：

$$\mu' = \underbrace{k(\boldsymbol{x}', \boldsymbol{X})k(\boldsymbol{X}, \boldsymbol{X})^{-1}}_{\boldsymbol{a}^{\top}}\boldsymbol{f}$$
$$\sigma^{2'} = \underbrace{k(\boldsymbol{x}', \boldsymbol{x}') - k(\boldsymbol{x}', \boldsymbol{X})k(\boldsymbol{X}, \boldsymbol{X})^{-1}k(\boldsymbol{x}', \boldsymbol{X})^{\top}}_{b}$$
$$f'|\boldsymbol{f} \sim \mathcal{N}(\boldsymbol{a}^{\top}\boldsymbol{f}, b)$$

由于随机向量 $\boldsymbol{f}$ 服从变分分布 $\boldsymbol{f} \sim \mathcal{N}(\boldsymbol{v}, \boldsymbol{S})$，根据高斯线性变换可知：

$$f'|\boldsymbol{f} \sim \mathcal{N}(\boldsymbol{a}^{\top}\boldsymbol{v}, \boldsymbol{a}^{\top}\boldsymbol{S}\boldsymbol{a} + b) \tag{3.26}$$

因此有：

$$\begin{aligned} p(f'|\boldsymbol{f}) &= \mathcal{N}(\boldsymbol{a}^{\top}\boldsymbol{v}, \boldsymbol{a}^{\top}\boldsymbol{S}\boldsymbol{a} + b) \\ p(f(\boldsymbol{x}')|y(\boldsymbol{X})) &= \int p(f'|\boldsymbol{f})q(\boldsymbol{f})d\boldsymbol{f} \\ &= \mathcal{N}(\boldsymbol{a}^{\top}\boldsymbol{v}, \boldsymbol{a}^{\top}\boldsymbol{S}\boldsymbol{a} + b)\int q(\boldsymbol{f})d\boldsymbol{f} \\ &= \mathcal{N}(\boldsymbol{a}^{\top}\boldsymbol{v}, \boldsymbol{a}^{\top}\boldsymbol{S}\boldsymbol{a} + b) \end{aligned} \tag{3.27}$$

上式即为给定 $\boldsymbol{x}'$ 时其取值函数对应的分布，这里未考虑高斯噪声。

第 4 章 Sparse VGPs

从上一章节可知通过变分高斯过程，我们可以在预测阶段给定任意样本点时，给出其对应函数值随机变量服从的高斯分布密度函数。通过分析可见，在求解过程中需要构建训练数据集上所有样本点对应的协方差矩阵 $k(\boldsymbol{X},\boldsymbol{X})$ 的逆，该计算的时间复杂度为 $O(n^3)$，可见当 n 较大时，时间复杂度较高。针对上述问题稀疏变分高斯过程在前述方案基础上通过引入 $m(m \ll n)$ 个稀疏归纳点，有效解决了变分高斯过程存在的时间复杂度和空间复杂度较高的问题。

4.1 问题引出

由于协方差矩阵计算复杂度较高，针对该问题，我们定义 m 个稀疏点，所有稀疏点构成矩阵 $\boldsymbol{Z} = [z_1, \ldots, z_M]^\top \in \mathbb{R}^{m\times d}$，令随机向量 $f(\boldsymbol{Z})$ 服从如下高斯分布：

$$f(\boldsymbol{Z}) \sim \mathcal{N}(m(\boldsymbol{Z}), k(\boldsymbol{Z},\boldsymbol{Z})) \tag{4.1}$$

本节介绍引入 m 个稀疏点后，如何定义问题求解所需的高斯过程先验，似然及后验分布。上述建模过程中未考虑测试噪声，和前述章节类似，引入高斯噪声后：

$$f(\boldsymbol{Z}) \sim \mathcal{N}(m(\boldsymbol{Z}), k(\boldsymbol{Z},\boldsymbol{Z}) + \eta^2\boldsymbol{I}_n) \tag{4.2}$$

因此引入高斯噪声后，可以用 $k(\boldsymbol{Z},\boldsymbol{Z}) + \eta^2\boldsymbol{I}_n$ 替代 $k(\boldsymbol{Z},\boldsymbol{Z})$ 用于有噪声建模。

4.1.1 稀疏先验

假设训练数据和稀疏归纳点服从如下联合高斯分布：

$$\begin{bmatrix} f(\boldsymbol{X}) \\ f(\boldsymbol{Z}) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m(\boldsymbol{X}) \\ m(\boldsymbol{Z}) \end{bmatrix}, \begin{bmatrix} k(\boldsymbol{X},\boldsymbol{X}) & k(\boldsymbol{X},\boldsymbol{Z}) \\ k(\boldsymbol{X},\boldsymbol{Z})^\top & k(\boldsymbol{Z},\boldsymbol{Z}) \end{bmatrix}\right) \tag{4.3}$$

4.1.2 似然函数

和前序章节类似，通过似然建立先验和训练数据之间的关系，因此有：

$$p(y(\boldsymbol{X})|f(\boldsymbol{X}), f(\boldsymbol{Z})) = p(y(\boldsymbol{X})|f(\boldsymbol{X})) \tag{4.4}$$

该分布对应的具体形式和应用相关，对于二分类问题，其概率密度函数为伯努利分布，因此有：

$$p(y_i|f_i) = g(f_i)^{y_i}(1 - g(f_i))^{(1-y_i)} \tag{4.5}$$

其中：

$$g(f_i) = \frac{e^{f_i}}{1 + e^{f_i}} \tag{4.6}$$

对于整个训练集来说有：

$$p(y(\boldsymbol{X})|f(\boldsymbol{X})) = \prod_{i=1}^{n} p(y_i|f_i) \tag{4.7}$$

对于回归问题，一般采用高斯分布，同时假设观察值来自于引入高斯噪声后的真实值，因此有：

$$\begin{aligned} p(y(\boldsymbol{X})|f(\boldsymbol{X})) &= \mathcal{N}(f(\boldsymbol{X}), \eta^2 \boldsymbol{I}_n) \\ &= \mathcal{N}(m(\boldsymbol{X}), k(\boldsymbol{X}, \boldsymbol{X}) + \eta^2 \boldsymbol{I}_n) \end{aligned} \tag{4.8}$$

其中 η^2 为各向同性的噪声方差。

4.1.3 后验分布

后验分布为给定训练集后，求如下概率密度函数：

$$p(f(\boldsymbol{X}), f(\boldsymbol{Z})|y(\boldsymbol{X})) = \frac{p(y(\boldsymbol{X})|f(\boldsymbol{X}), f(\boldsymbol{Z}))p(f(\boldsymbol{X}), f(\boldsymbol{Z}))}{\iint p(y(\boldsymbol{X})|f(\boldsymbol{X}), f(\boldsymbol{Z}))p(f(\boldsymbol{X}), f(\boldsymbol{Z}))df(\boldsymbol{X})df(\boldsymbol{Z})} \tag{4.9}$$

令符号 $\boldsymbol{f}$ 表示随机变量 $f(\boldsymbol{X})$，$\boldsymbol{f}_Z$ 表示随机变量 $f(\boldsymbol{Z})$，$\boldsymbol{y}$ 表示随机变量 $y(\boldsymbol{X})$，因此有：

$$p(\boldsymbol{f}, \boldsymbol{f}_Z|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{f}, \boldsymbol{f}_Z)p(\boldsymbol{f}, \boldsymbol{f}_Z)}{\iint p(\boldsymbol{y}|\boldsymbol{f}, \boldsymbol{f}_Z)p(\boldsymbol{f}, \boldsymbol{f}_Z)d\boldsymbol{f}d\boldsymbol{f}_Z} \tag{4.10}$$

由于稀疏先验为高斯分布，若问题为分类问题，则似然函数非高斯先验。因此无法求得上述后验分布的解析式。

4.2 变分近似

由于无法直接求得后验分布，因此考虑采用变分分布 $q(\boldsymbol{f}, \boldsymbol{f}_Z$ 近似后验分布 $p(\boldsymbol{f}, \boldsymbol{f}_Z|\boldsymbol{y})$，并定义：

$$q(\boldsymbol{f}, \boldsymbol{f}_Z) = p(\boldsymbol{f}|\boldsymbol{f}_Z)q(\boldsymbol{f}_Z) \tag{4.11}$$

其中 $q(\boldsymbol{f}_Z)$ 服从如下高斯分布：

$$q(\boldsymbol{f}_Z) = \mathcal{N}(\boldsymbol{v}, \boldsymbol{S}) \tag{4.12}$$

由高斯过程先验结合多元高斯分布条件概率可知：

$$p(\boldsymbol{f}|\boldsymbol{f}_Z) = \mathcal{N}(\boldsymbol{\mu}, \Sigma) \tag{4.13}$$

其中：

$$\boldsymbol{\mu} = m(\boldsymbol{X}) + k(\boldsymbol{X}, \boldsymbol{Z})k(\boldsymbol{Z}, \boldsymbol{Z})^{-1}(\boldsymbol{f}_Z - m(\boldsymbol{Z})) \tag{4.14}$$

$$\Sigma = k(\boldsymbol{X}, \boldsymbol{X}) - k(\boldsymbol{X}, \boldsymbol{Z})k(\boldsymbol{Z}, \boldsymbol{Z})^{-1}k(\boldsymbol{X}, \boldsymbol{Z})^{\top} \tag{4.15}$$

给定变分分布后，可通过 KL 距离度量变分分布和后验分布的近似程度：

$$\begin{aligned}
&KL(q(\boldsymbol{f}, \boldsymbol{f}_Z)||p(\boldsymbol{f}, \boldsymbol{f}_Z|\boldsymbol{y})) \\
&= \iint q(\boldsymbol{f}, \boldsymbol{f}_Z) \log \frac{q(\boldsymbol{f}, \boldsymbol{f}_Z)}{p(\boldsymbol{f}, \boldsymbol{f}_Z|\boldsymbol{y})} d\boldsymbol{f} d\boldsymbol{f}_Z \\
&= \iint q(\boldsymbol{f}, \boldsymbol{f}_Z) \log \frac{q(\boldsymbol{f}, \boldsymbol{f}_Z)p(\boldsymbol{y})}{p(\boldsymbol{y}|\boldsymbol{f}, \boldsymbol{f}_Z)p(\boldsymbol{f}, \boldsymbol{f}_Z)} d\boldsymbol{f} d\boldsymbol{f}_Z \\
&= \iint q(\boldsymbol{f}, \boldsymbol{f}_Z) \log(p(\boldsymbol{y})) d\boldsymbol{f} d\boldsymbol{f}_Z + \iint q(\boldsymbol{f}, \boldsymbol{f}_Z) \log \frac{q(\boldsymbol{f}, \boldsymbol{f}_Z)}{p(\boldsymbol{y}|\boldsymbol{f}, \boldsymbol{f}_Z)p(\boldsymbol{f}, \boldsymbol{f}_Z)} d\boldsymbol{f} d\boldsymbol{f}_Z \\
&= \log(p(\boldsymbol{y})) - \iint q(\boldsymbol{f}, \boldsymbol{f}_Z) \log \frac{p(\boldsymbol{y}|\boldsymbol{f}, \boldsymbol{f}_Z)p(\boldsymbol{f}, \boldsymbol{f}_Z)}{q(\boldsymbol{f}, \boldsymbol{f}_Z)} d\boldsymbol{f} d\boldsymbol{f}_Z \\
&= \log(p(\boldsymbol{y})) - \iint q(\boldsymbol{f}, \boldsymbol{f}_Z) \log(p(\boldsymbol{y}|\boldsymbol{f}, \boldsymbol{f}_Z)) d\boldsymbol{f} d\boldsymbol{f}_Z - \iint q(\boldsymbol{f}, \boldsymbol{f}_Z) \log \frac{p(\boldsymbol{f}, \boldsymbol{f}_Z)}{q(\boldsymbol{f}, \boldsymbol{f}_Z)} d\boldsymbol{f} d\boldsymbol{f}_Z \\
&= \log(p(\boldsymbol{y})) - \iint q(\boldsymbol{f}, \boldsymbol{f}_Z) \log(p(\boldsymbol{y}|\boldsymbol{f})) d\boldsymbol{f} d\boldsymbol{f}_Z + KL(q(\boldsymbol{f}, \boldsymbol{f}_Z)||p(\boldsymbol{f}, \boldsymbol{f}_Z))
\end{aligned}$$

令

$$\text{ELBO} = \iint q(\boldsymbol{f}, \boldsymbol{f}_Z) \log(p(\boldsymbol{y}|\boldsymbol{f})) d\boldsymbol{f} d\boldsymbol{f}_Z - KL(q(\boldsymbol{f}, \boldsymbol{f}_Z)||p(\boldsymbol{f}, \boldsymbol{f}_Z)) \tag{4.16}$$

因此有：

$$KL(q(\boldsymbol{f}, \boldsymbol{f}_Z)||p(\boldsymbol{f}, \boldsymbol{f}_Z|\boldsymbol{y})) = \log(p(\boldsymbol{y})) - \text{ELBO} \tag{4.17}$$

因此要最小化变分分布和后验分布的差异，需要最大化 ELBO。

4.3 ELBO 解析式

由 ELBO 解析式4.16可知，其包含两项分别为似然项和 KL 距离项，下面分别求取两项的解析式

4.3.1 似然解析式

似然项为二重积分形式，下面给出其解析式求解过程：

$$\begin{aligned}
&\iint q(\boldsymbol{f}, \boldsymbol{f}_Z) \log(p(\boldsymbol{y}|\boldsymbol{f})) d\boldsymbol{f} d\boldsymbol{f}_Z \\
&= \int \log(p(\boldsymbol{y}|\boldsymbol{f})) \left(\int q(\boldsymbol{f}, \boldsymbol{f}_Z) d\boldsymbol{f}_Z \right) d\boldsymbol{f} \\
&= \int \log(p(\boldsymbol{y}|\boldsymbol{f})) q(\boldsymbol{f}) d\boldsymbol{f}
\end{aligned} \tag{4.18}$$

为了求解上述积分形式的解析式，需要求得 $q(\boldsymbol{f})$ 和 $p(\boldsymbol{y}|\boldsymbol{f})$ 的解析式：

$$\begin{aligned} q(\boldsymbol{f}) &= \int q(\boldsymbol{f}, \boldsymbol{f}_Z) d\boldsymbol{f}_Z = \int p(\boldsymbol{f}|\boldsymbol{f}_Z) q(\boldsymbol{f}_Z) d\boldsymbol{f}_Z \\ &= \int \mathcal{N}(\boldsymbol{f}; \boldsymbol{\mu}, \Sigma) \cdot \mathcal{N}(\boldsymbol{f}_Z; \boldsymbol{v}, \boldsymbol{S}) d\boldsymbol{f}_Z \end{aligned} \tag{4.19}$$

其中：

$$\begin{aligned} \boldsymbol{\mu} &= m(\boldsymbol{X}) + k(\boldsymbol{X}, \boldsymbol{Z}) k(\boldsymbol{Z}, \boldsymbol{Z})^{-1} (\boldsymbol{f}_Z - m(\boldsymbol{Z})) \\ &= m(\boldsymbol{X}) + k(\boldsymbol{X}, \boldsymbol{Z}) k(\boldsymbol{Z}, \boldsymbol{Z})^{-1} \boldsymbol{f}_Z - k(\boldsymbol{X}, \boldsymbol{Z}) k(\boldsymbol{Z}, \boldsymbol{Z})^{-1} m(\boldsymbol{Z}) \end{aligned} \tag{4.20}$$

由于：

$$\boldsymbol{f}_Z \sim \mathcal{N}(\boldsymbol{v}, \boldsymbol{S}) \tag{4.21}$$

由高斯线性变换可知：

$$\begin{aligned} \mathcal{N}(\boldsymbol{f}; \boldsymbol{\mu}, \Sigma) &= \mathcal{N}(\boldsymbol{\mu}_{qf}, \Sigma_{qf}) \\ \boldsymbol{\mu}_{qf} &= m(\boldsymbol{X}) + k(\boldsymbol{X}, \boldsymbol{Z}) k(\boldsymbol{Z}, \boldsymbol{Z})^{-1} \boldsymbol{v} - k(\boldsymbol{X}, \boldsymbol{Z}) k(\boldsymbol{Z}, \boldsymbol{Z})^{-1} m(\boldsymbol{Z}) \\ &= m(\boldsymbol{X}) + k(\boldsymbol{X}, \boldsymbol{Z}) k(\boldsymbol{Z}, \boldsymbol{Z})^{-1} (\boldsymbol{v} - m(\boldsymbol{Z})) \\ \Sigma_{qf} &= k(\boldsymbol{X}, \boldsymbol{Z}) k(\boldsymbol{Z}, \boldsymbol{Z})^{-1} \boldsymbol{S} (k(\boldsymbol{X}, \boldsymbol{Z}) k(\boldsymbol{Z}, \boldsymbol{Z})^{-1})^{\top} + \Sigma \end{aligned} \tag{4.22}$$

其中 Σ 为公式4.15，上述形式不包含随机变量 $\boldsymbol{f}_Z$，因此有：

$$\begin{aligned} q(\boldsymbol{f}) &= \int \mathcal{N}(\boldsymbol{f}; \boldsymbol{\mu}, \Sigma) \cdot \mathcal{N}(\boldsymbol{f}_Z; \boldsymbol{v}, \boldsymbol{S}) d\boldsymbol{f}_Z \\ &= \mathcal{N}(\boldsymbol{f}; \boldsymbol{\mu}_{qf}, \Sigma_{qf}) \int \mathcal{N}(\boldsymbol{f}_Z; \boldsymbol{v}, \boldsymbol{S}) d\boldsymbol{f}_Z \\ &= \mathcal{N}(\boldsymbol{f}; \boldsymbol{\mu}_{qf}, \Sigma_{qf}) \end{aligned} \tag{4.23}$$

给定上述形式后，似然项为：

$$\begin{aligned} &\int \log(p(\boldsymbol{y}|\boldsymbol{f})) q(\boldsymbol{f}) d\boldsymbol{f} \\ &= \sum_{i=1}^{n} \left(\int \log(p(y_i|f_i)) q(\boldsymbol{f}_i) d\boldsymbol{f}_i \right) \end{aligned} \tag{4.24}$$

由于：

$$q(f_i) = \mathcal{N}(f_i; \mu_i, \Sigma_{ii}^2) \tag{4.25}$$

因此若求得上述一元高斯分布中的 μ_i 和 Σ_{ii}^2 则后面的近似公式和变分高斯过程中求对数似然的形式保持一致，因此这里重点给出 μ_i 和 Σ_{ii}^2 的公式，由 $q(\boldsymbol{f})$ 分布公式可以较容易

推断出 μ_i 和 Σ_{ii}^2 形式，并最终得出似然项的数值积分近似解析式。

4.3.2 KL 距离解析式

下面我们来看如何求得 KL 距离项解析式

$$\begin{aligned}
&KL(q(\boldsymbol{f},\boldsymbol{f}_Z)||p(\boldsymbol{f},\boldsymbol{f}_Z)) \\
&= \iint q(\boldsymbol{f},\boldsymbol{f}_Z)\log\left(\frac{q(\boldsymbol{f},\boldsymbol{f}_Z)}{p(\boldsymbol{f},\boldsymbol{f}_Z)}\right)d\boldsymbol{f}d\boldsymbol{f}_Z \\
&= \iint q(\boldsymbol{f},\boldsymbol{f}_Z)\log\left(\frac{p(\boldsymbol{f}|\boldsymbol{f}_Z)q(\boldsymbol{f}_Z)}{p(\boldsymbol{f}|\boldsymbol{f}_Z)p(\boldsymbol{f}_Z)}\right)d\boldsymbol{f}d\boldsymbol{f}_Z \\
&= \iint \log\left(\frac{q(\boldsymbol{f}_Z)}{p(\boldsymbol{f}_Z)}\right)q(\boldsymbol{f},\boldsymbol{f}_Z)d\boldsymbol{f}d\boldsymbol{f}_Z \\
&= \int \log\left(\frac{q(\boldsymbol{f}_Z)}{p(\boldsymbol{f}_Z)}\right)\left(\int q(\boldsymbol{f},\boldsymbol{f}_Z)d\boldsymbol{f}\right)d\boldsymbol{f}_Z \\
&= \int q(\boldsymbol{f}_Z)\log\left(\frac{q(\boldsymbol{f}_Z)}{p(\boldsymbol{f}_Z)}\right)d\boldsymbol{f}_Z
\end{aligned} \tag{4.26}$$

上式和上一章求取 KL 距离项的形式完全一致，因此其最终的解析式如下：

$$KL(q(\boldsymbol{f},\boldsymbol{f}_Z)||p(\boldsymbol{f},\boldsymbol{f}_Z)) = \frac{1}{2}\left(\log\left(\frac{|\boldsymbol{K}|}{|\boldsymbol{S}|}\right) + (\boldsymbol{v}-\boldsymbol{m})^\top\boldsymbol{K}^{-1}(\boldsymbol{v}-\boldsymbol{m}) + \mathrm{Tr}(\boldsymbol{K}^{-1}\boldsymbol{S}) - m\right) \tag{4.27}$$

4.4 预测应用

给定稀疏点对应的高斯过程先验及后验分布的变分近似后，我们即可以进行测试样本点对应的概率密度函数预测。

$$\begin{aligned}
p(\boldsymbol{f}'|\boldsymbol{y}) &= \iint p(\boldsymbol{f}',\boldsymbol{f},\boldsymbol{f}_Z|\boldsymbol{y})d\boldsymbol{f}d\boldsymbol{f}_Z \\
&= \iint p(\boldsymbol{f}'|\boldsymbol{f},\boldsymbol{f}_Z,\boldsymbol{y})p(\boldsymbol{f},\boldsymbol{f}_Z|\boldsymbol{y})d\boldsymbol{f}d\boldsymbol{f}_Z \\
&= \iint p(\boldsymbol{f}'|\boldsymbol{f},\boldsymbol{f}_Z)p(\boldsymbol{f},\boldsymbol{f}_Z|\boldsymbol{y})d\boldsymbol{f}d\boldsymbol{f}_Z \\
&= \iint p(\boldsymbol{f}'|\boldsymbol{f},\boldsymbol{f}_Z)q(\boldsymbol{f},\boldsymbol{f}_Z)d\boldsymbol{f}d\boldsymbol{f}_Z \\
&= \iint p(\boldsymbol{f}'|\boldsymbol{f},\boldsymbol{f}_Z)p(\boldsymbol{f}|\boldsymbol{f}_Z)q(\boldsymbol{f}_Z)d\boldsymbol{f}d\boldsymbol{f}_Z \\
&= \int\left(\int p(\boldsymbol{f}'|\boldsymbol{f},\boldsymbol{f}_Z)p(\boldsymbol{f}|\boldsymbol{f}_Z)d\boldsymbol{f}\right)q(\boldsymbol{f}_Z)d\boldsymbol{f}_Z \\
&= \int\left(\int p(\boldsymbol{f}',\boldsymbol{f}|\boldsymbol{f}_Z)d\boldsymbol{f}\right)q(\boldsymbol{f}_Z)d\boldsymbol{f}_Z \\
&= \int p(\boldsymbol{f}'|\boldsymbol{f}_Z)q(\boldsymbol{f}_Z)d\boldsymbol{f}_Z
\end{aligned} \tag{4.28}$$

从如上形式可见，最终进行预测时只和稀疏归纳点及变分分布有关，且有：

$$\begin{bmatrix} \boldsymbol{f}' \\ \boldsymbol{f}_Z \end{bmatrix} = \mathcal{N}\left(\begin{bmatrix} m(\boldsymbol{f}') \\ m(\boldsymbol{f}_Z) \end{bmatrix}, \begin{bmatrix} k(\boldsymbol{X}', \boldsymbol{X}') & k(\boldsymbol{X}', \boldsymbol{Z}) \\ k(\boldsymbol{X}', \boldsymbol{Z})^\top & k(\boldsymbol{Z}, \boldsymbol{Z}) \end{bmatrix} \right)$$

因此有：

$$p(\boldsymbol{f}'|\boldsymbol{f}_Z) = \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{f}'}, \Sigma_{\boldsymbol{f}'}) \tag{4.29}$$

其中：

$$\boldsymbol{\mu}_{\boldsymbol{f}'} = m(\boldsymbol{X}') + k(\boldsymbol{X}', \boldsymbol{Z})k(\boldsymbol{Z}, \boldsymbol{Z})^{-1}(\boldsymbol{f}_Z - m(\boldsymbol{Z})) \tag{4.30}$$

$$\Sigma_{\boldsymbol{f}'} = k(\boldsymbol{X}', \boldsymbol{X}') - k(\boldsymbol{X}', \boldsymbol{Z})k(\boldsymbol{Z}, \boldsymbol{Z})^{-1}k(\boldsymbol{X}', \boldsymbol{Z})^\top \tag{4.31}$$

由于 $\boldsymbol{f}_Z$ 服从高斯分布 $\mathcal{N}(\boldsymbol{v}, \boldsymbol{S})$，由高斯线性变换可知：

$$p(\boldsymbol{f}'|\boldsymbol{f}_Z) = \mathcal{N}(\boldsymbol{\mu}_p, \Sigma_p) \tag{4.32}$$

其中：

$$\boldsymbol{\mu}_p = m(\boldsymbol{X}') + k(\boldsymbol{X}', \boldsymbol{Z})k(\boldsymbol{Z}, \boldsymbol{Z})^{-1}(\boldsymbol{v} - m(\boldsymbol{Z})) \tag{4.33}$$

$$\begin{aligned} \Sigma_p &= k(\boldsymbol{X}', \boldsymbol{Z})k(\boldsymbol{Z}, \boldsymbol{Z})^{-1}\boldsymbol{S}(k(\boldsymbol{X}', \boldsymbol{Z})k(\boldsymbol{Z}, \boldsymbol{Z})^{-1})^\top + \Sigma_{\boldsymbol{f}'} \\ &= k(\boldsymbol{X}', \boldsymbol{Z})k(\boldsymbol{Z}, \boldsymbol{Z})^{-1}\boldsymbol{S}k(\boldsymbol{Z}, \boldsymbol{Z})^{-1}k(\boldsymbol{X}', \boldsymbol{Z})^\top + \Sigma_{\boldsymbol{f}'} \\ &= k(\boldsymbol{X}', \boldsymbol{X}') - k(\boldsymbol{X}', \boldsymbol{Z})k(\boldsymbol{Z}, \boldsymbol{Z})^{-1}(k(\boldsymbol{Z}, \boldsymbol{Z}) - \boldsymbol{S})k(\boldsymbol{Z}, \boldsymbol{Z})^{-1}k(\boldsymbol{X}', \boldsymbol{Z})^\top \end{aligned} \tag{4.34}$$

从上述可见 $p(\boldsymbol{f}'|\boldsymbol{f}_Z)$ 和随机向量 $\boldsymbol{f}_Z$ 无关，因此有：

$$\begin{aligned} p(\boldsymbol{f}'|\boldsymbol{y}) &= \int p(\boldsymbol{f}'|\boldsymbol{f}_Z)q(\boldsymbol{f}_Z)d\boldsymbol{f}_Z \\ &= p(\boldsymbol{f}'|\boldsymbol{f}_Z)\int q(\boldsymbol{f}_Z)d\boldsymbol{f}_Z = p(\boldsymbol{f}'|\boldsymbol{f}_Z) \end{aligned} \tag{4.35}$$

上式即为给定任意测试点时，预测的函数分布。