高斯过程回归方法及其应用

作者: XX

2018.11.1

目 录

摘　要

本文简单介绍了高斯过程回归方法的理论基础, 包括高斯分布, 多维高斯分布, 高斯过程, 以及如何应用高斯过程解决回归问题.

关键词：高斯过程回归; 预测

1 高斯过程回归理论

本章首先介绍随机过程与高斯分布, 特别是重点讨论了多维高斯分布的性质, 这是高斯过程应用于回归的理论基础. 然后介绍高斯过程和高斯过程回归模型的建立. 最后是模型的求解, 简单讨论了核函数选择和模型超参数的求解问题.

1.1 高斯过程回归模型

高斯过程回归模型 (Gaussian process regression, GPR) 是基于贝叶斯理论的非参数模型 (Non-pararmetric Model). 引入核函数方法使得模型能够对复杂的非线性数据进行拟合. 高斯过程回归建立的模型实际上是整个后验分布, 输出结果包括均值与方差, 因此还能够简单计算出置信区间, 评估预测结果. 要了解 GPR 其中的一个关键基础是高斯分布.

1.1.1 高斯分布及其性质

1.1.1.1 高斯分布

定义 1.1 (高斯分布)**.** 一个连续型随机变量 X 其概率密度函数 (probability density function, pdf) 若为

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\qquad x\in\mathbb{R}. \tag{1.1}$$

其中 $\mu\in\mathbb{R}$, $\sigma^2\geqslant 0$, 则称 X 服从均值为 μ 方差为 σ^2 的高斯分布 (Gaussian distribution) 或正态分布 (normal distribution). 记作 $X\sim\mathcal{N}(\mu,\sigma^2)$

随机变量的概念很容易推广到多维随机变量 (multivariate random variables) 或称为随机向量 (random vectors). 如有向量 $Y=(Y_1,\ldots,Y_d)^T$, 其每一个分量都是定义在同一个概率空间下的随机变量, 此时 Y 即是一个随机向量. 随机向量 Y 的一个样本 y 应有 $y\in\mathbb{R}^d$. 随机向量各个分量是定义在相同概率空间下的, 这使得我们能够研究各分量随机变量间的关系.

我们可以将与随机变量相关的一些概念也推广到随机向量, 假设 $Y=(Y_1,\ldots,Y_d)^T$ 与 $Z=(Z_1,\ldots,Z_n)^T$ 是两个分别为 d 维与 n 维的随机向量, 我们可以定义随机向量的均值、方差以及两随机向量之间的协方差

$$\mathrm{E}(Y)=(\mathrm{E}(Y_1),\ldots,\mathrm{E}(Y_d))^T. \tag{1.2}$$

$$\mathrm{var}(Y)=\mathrm{E}(YY^T)-\mathrm{E}(Y)\mathrm{E}(Y^T)=\mathrm{E}[(Y-\mathrm{E}(Y))(Y-\mathrm{E}(Y))^T]. \tag{1.3}$$

$$\mathrm{cov}(Y,Z)=\mathrm{E}(YZ^T)-\mathrm{E}(Y)\mathrm{E}(Z^T). \tag{1.4}$$

1.1.1.2 多维高斯分布

服从多维高斯分布 (multivariate Gaussian distribution) 的随机向量是我们需要重点讨论的对象.

定义 1.2 (多维高斯分布)**.** 若有随机向量 $Y = (Y_1, \ldots, Y_d)^T$, 其分量的任意线性组合而成的随机变量都服从于高斯分布, 即

$$\forall \alpha \in \mathbb{R}^d,\ \alpha_1 Y_1 + \cdots + \alpha_d Y_d = \alpha^T Y \sim \mathcal{N}. \tag{1.5}$$

此时称随机向量 Y 服从于 d 维高斯分布, Y 也被称为高斯随机向量.

d 维高斯分布的概率密度函数由下式给出.

$$f_Y(y) = \frac{1}{|2\pi\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(y-\mu)^T\Sigma^{-1}(y-\mu)\right) \qquad y \in \mathbb{R}^d. \tag{1.6}$$

为了表达上的简便有时也简记为

$$f(y) = \mathcal{N}(\mu, \Sigma) \tag{1.7}$$

上式中 μ 是 d 维的均值向量 (mean vector), 其每个分量 $\mu_i = \mathrm{E}(Y_i)$, 而 Σ 是一个 $d \times d$ 的矩阵, 其 i 行 j 列的分量 $\Sigma_{i,j} = \mathrm{cov}(Y_i, Y_j)$, 矩阵 Σ 被称为协方差矩阵 (covariance matrix).

就像高斯分布能被一组 μ 与 σ^2 唯一确定一样, 多维高斯分布的特性也能由均值向量 μ 与协方差矩阵 Σ 唯一确定. 因此服从多维高斯分布的随机向量可记为 $Y \sim \mathcal{N}(\mu, \Sigma)$.

由协方差矩阵定义可以得出协方差矩阵总是对称半正定 (symmetric and positive semi-definite) 的 (证明见附录A.1), 即有

$$\Sigma_{i,j} = \Sigma_{j,i}. \tag{1.8}$$

$$\forall \alpha \in \mathbb{R}^d,\ \alpha^T\Sigma\alpha \geqslant 0. \tag{1.9}$$

事实上, 只要是满足上面两条性质的矩阵都可以视为是协方差矩阵. 下面介绍多维高斯分布两个重要的性质, 特别是条件分布的性质在后面有重要应用.

1.1.1.3 高斯分布性质

性质 1.1. 对于高斯向量来说, 各分量之间不相关等价于各分量之间独立.

这是高斯分布独有的特性, 在非高斯分布的一般条件下变量间独立可推出不相关, 但不相关往往推不出变量间独立. 这个性质利用独立的定义也容易得到证明.

对于一个各分量之间独立的高斯向量 Y, 易知其协方差矩阵是一个对角阵, 且对角线元素即各分量的方差.

$$\Sigma = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_d^2 \end{pmatrix} \tag{1.10}$$

假设 Y 服从协方差矩阵为 Σ, 均值向量为 μ 的高斯分布, 则 Y 的概率密度函数可拆解为

$$\begin{aligned} f(y) &= \frac{1}{|2\pi\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(y-\mu)^T\Sigma^{-1}(y-\mu)\right) \\ &= \frac{1}{\sqrt{2\pi\sigma_1^2}\times\cdots\times\sqrt{2\pi\sigma_d^2}} \exp\left(-\sum_{i=1}^{d}\frac{(y_i-\mu_i)^2}{2\sigma_i^2}\right) \\ &= \prod_{i=1}^{d} f(y_i). \end{aligned} \tag{1.11}$$

上式中 $f(y_i)$ 即 Y_i 的概率密度函数, 由独立定义可得 Y 各分量间独立.

为了研究高斯分布的条件分布, 我们把高斯随机向量划分为两部分, $Y = \binom{Y_1}{Y_2}$, Y_1 与 Y_2 都可能为高斯随机向量或只是单变量 (univariate). 同时也对应将均值向量与协方差矩阵分块

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}. \tag{1.12}$$

$$\Sigma = \begin{pmatrix} \Sigma_{1,1} & \Sigma_{1,2} \\ \Sigma_{2,1} & \Sigma_{2,2} \end{pmatrix}. \tag{1.13}$$

在上述设定下, 有以下条件分布性质.

性质 1.2. 已知 Y_2 的条件下 Y_1 的条件分布也是高斯分布. 若 Y_1 为随机向量, 则该条件分布是多维高斯分布, 且该多维高斯分布的均值向量与协方差函数由下式给出 (证明见附录A.2)

$$Y_1|Y_2 \sim \mathcal{N}(\mu_c, \Sigma_c).$$

$$\mu_c = \mu_1 + \Sigma_{1,2}\Sigma_{2,2}^{-1}(Y_2 - \mu_2). \tag{1.14}$$

$$\Sigma_c = \Sigma_{1,1} - \Sigma_{1,2}\Sigma_{2,2}^{-1}\Sigma_{2,1}. \tag{1.15}$$

1.1.2 高斯过程

多维随机向量的概念可以推广到随机过程 (stochastic processes), 可认为是由多维随机变量推广到了 “无限维”. 此时, 取这个 “无限维” 随机向量的有限个分量构成的随机向量将服从于某种分布, 根据服从的分布定义了不同的随机过程, 这里特别重要的是高斯过程.

1.1.2.1 高斯过程

定义 1.3. (高斯过程) 若一个定义在域 $D \in \mathbb{R}^d$ 上的随机过程 Z, 对于 $\forall n \in \mathbb{N}$, $\forall x_i \in D$, $(Z(x_1), \ldots, Z(x_n))$ 是一个高斯随机向量, 则 Z 是高斯过程 (Gaussian Process, GP).

与高斯分布类似, 高斯过程的性质可以由定义在 D 上的均值函数 (mean function) 与定义在 $D \times D$ 上的协方差函数 (covariance function) 完全确定. 通常用 $m(x)$ 与 $k(x, x')$ 来表示均值函数与协方差函数.

$$m(x) = \mathrm{E}(Z(x)) \tag{1.16}$$

$$k(x, x') = \mathrm{cov}(Z(x), Z(x')) \tag{1.17}$$

于是高斯过程 Z 可简记为

$$Z \sim \mathcal{GP}\left(m(x), k(x, x')\right)$$

1.1.2.2 均值函数与协方差函数

均值函数与协方差函数是高斯过程两个关键要素, 一组均值与协方差函数就唯一确定了一个高斯分布. 常见的均值函数有常数均值函数与线性均值函数等, 其表达式如下[1]

$$\text{常数均值函数 } m(x) = c \qquad c \in \mathbb{R}. \tag{1.18}$$

$$\text{线性均值函数 } m(x) = \alpha^T x \qquad \alpha \in \mathbb{R}^d. \tag{1.19}$$

然而在实际应用中总是先将数据预处理成是 0 均值的, 即使得 $m(x) = 0$. 这样不仅在理论推导时带来便利, 在实际运算中也提高了效率.

对于高斯过程的一个样本 $(Z(x_1), \ldots, Z(x_n))$ 是服从多维高斯分布的, 这个高斯分布的协方差矩阵 K 由协方差函数计算得到

$$K = \begin{pmatrix} k(x_1, x_1) & k(x_1, x_2) & \cdots & k(x_1, x_n) \\ k(x_2, x_1) & k(x_2, x_2) & \cdots & k(x_2, x_n) \\ \vdots & \vdots & \ddots & \vdots \\ k(x_n, x_1) & k(x_n, x_2) & \cdots & k(x_n, x_n) \end{pmatrix} \tag{1.20}$$

由于协方差矩阵总是对称半正定的 (见附录A.1), 可以以此来定义半正定函数 (positive semi-definite functions), 协方差函数是半正定函数, 事实上任何对称半正定函数都可以视为是协方差函数.

定义 1.4. (半正定函数) 若定义在 $D \times D$ 的函数 $k(\cdot, \cdot)$, 对于 $\forall n \in \mathbb{N}$, $\forall \{x_1, \cdots, x_n\} \in D^n$, $\forall \alpha \in \mathbb{R}^n$, 有

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j k(x_i, x_j) \geqslant 0$$

则 $k(\cdot, \cdot)$ 为半正定函数.

协方差函数也被称为核函数 (kernel function), 以下用统一采用核函数这一说法. 核函数形式繁多也有多种分类方法, 常见核函数如下

表 1　常见核函数

核函数	表达式	参数
$constant$	σ_f^2	σ_f^2
$linear$	$\sum_{i=1}^{d} \sigma_i^2 x_i x_i'$	$\{\sigma_1^2, \ldots, \sigma_d^2\}$
$polynomial$	$(xx' + \sigma_f^2)^p$	σ_f^2
$squared\ exponential$	$\sigma_f^2 \text{exp}\left(-\frac{\|x-x'\|^2}{2l^2}\right)$	σ_f^2, l
$Matérn$	$\frac{2^{1-v}}{\Gamma(v)}\left(\frac{\sqrt{2v}r}{l}\right)^v K_v\left(\frac{\sqrt{2v}r}{l}\right)$	v, l
$exponential$	$\sigma_f^2 \text{exp}\left(-\frac{\|x-x'\|}{l}\right)$	σ_f^2, l
$rational\ quadratic$	$\sigma_f^2\left(1+\frac{\|x-x'\|^2}{2\alpha l^2}\right)^{-\alpha}$	σ_f^2, α, l

核函数中的一大类又被称之为静态核函数 (stationary kernel), 如表1中的 SE 核 (squared exponential) 与 RQ 核 (rational quadratic) 等, 特点是与输入的 x 与 x' 无关而与 $\|x - x'\|$ 有关, 这些核函数在应用上更为广泛. 另外, 通常一种核函数有两个常用版本即 iso 形式 (isotropic) 与 ard 形式 (automatic relevance determination), 两者的差异是输入 x 的各个维度是否使用相同的长度尺度 (length scale), iso 形式的函数使用相同长度尺度, 而 ard 使用不同的尺度. 如表1中的线性核 (linear kernel) 是其 ard 版本, 其 iso 版本为 $\sigma^2 \sum_{i=1}^{d} x_i x_i'$, 此时核函数参数只有 σ^2, 可见 ard 形式的核函数参数较 iso 版本多. 表1中其余除常数核外的核函数皆为 iso 形式, 同时也各有 ard 形式的核函数.

1.1.3　高斯过程回归

下面进入正题, 即如何运用上述高斯过程理论解决回归问题. Rasmussen 在所著的《Gaussian processes in machine learning》一书中提到了两种介绍高斯过程回归的两种思路, 即借助于线性回归的权重空间的观点 (weight-space view) 和直接的函数空间的观点 (function-space view).

1.1.3.1　权重空间观点解释

权重空间观点的解释需要借助于贝叶斯线性模型 (Bayesian linear model), 贝叶斯线性回归模型是选择模型为线性模型的前提下, 假设权向量 $\boldsymbol{w}$ 的服从先验分布是高斯分布, 再结合贝叶斯理论来解决回归问题. 一个典型的贝叶斯线性模型假设如下 (只先考虑基本线性模型, 即输入不经过基函数 $\boldsymbol{\phi}(\cdot)$ 映射)

$$f(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{w}, \qquad y = f(\boldsymbol{x}) + \varepsilon. \tag{1.21}$$

同时假设噪声 ε 服从于独立的均值为 0, 方差为 σ_n^2 的高斯分布, 即 $\varepsilon \sim \mathcal{N}(0, \sigma_n^2)$. 在此假设下, 我们可以得到在已知 $\boldsymbol{x}$ 与 $\boldsymbol{w}$ 的条件下 y 的条件分布是高斯分布, 即

$$y|\boldsymbol{x}, \boldsymbol{w} \sim \mathcal{N}(\boldsymbol{x}^T\boldsymbol{w}, \sigma_n^2). \tag{1.22}$$

为了求解 $\boldsymbol{w}$ 进行 n 次采样得到训练数据集 $\boldsymbol{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)$ 与 $\boldsymbol{y} = (y_1, \cdots, y_n)$, 利用最大似然估计法 (maximum likelihood estimation, MLE) 估计 $\boldsymbol{w}$. 可写出似然函数

$$\begin{aligned} p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}) &= \prod_{i=1}^{n} p(y_i|\boldsymbol{x}_i, \boldsymbol{w}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left(-\frac{(y_i - \boldsymbol{x}_i^T\boldsymbol{w})^2}{2\sigma_n^2}\right) \\ &= \frac{1}{(2\pi\sigma_n^2)^{n/2}} \exp\left(-\frac{1}{2\sigma_n^2}(\boldsymbol{y} - \boldsymbol{X}^T\boldsymbol{w})^T(\boldsymbol{y} - \boldsymbol{X}^T\boldsymbol{w})\right) \\ &= \mathcal{N}(\boldsymbol{X}^T\boldsymbol{w}, \sigma_n^2\boldsymbol{I}). \end{aligned} \tag{1.23}$$

上式成立需满足两个条件, 一是各次采样值之间独立同分布, 二是各次采样噪声 ε 也是独立同分布的, 这些条件包含在模型假设中. 为了利用贝叶斯理论得到 $\boldsymbol{w}$ 后验分布, 我们需要先假设 $\boldsymbol{w}$ 的先验分布

$$\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, \Sigma_p). \tag{1.24}$$

然后利用贝叶斯公式计算 $\boldsymbol{w}$ 的后验概率 $p(\boldsymbol{w}|\boldsymbol{y}, \boldsymbol{X})$.

$$p(\boldsymbol{w}|\boldsymbol{X}, \boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w})p(\boldsymbol{w})}{p(\boldsymbol{y}|\boldsymbol{X})}. \tag{1.25}$$

上式中的分母是与 $\boldsymbol{w}$ 无关的, 在此视为常数因子. 将式1.23, 式1.24代入式1.25进行整理 (详细推导过程见附录A.3).

$$\begin{aligned} p(\boldsymbol{w}|\boldsymbol{X}, \boldsymbol{y}) &\propto p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w})p(\boldsymbol{w}) \\ &\propto \exp\left(-\frac{1}{2}(\boldsymbol{w} - \bar{\boldsymbol{w}})^T(\frac{1}{\sigma_n^2}\boldsymbol{X}\boldsymbol{X}^T + \Sigma_p^{-1})(\boldsymbol{w} - \bar{\boldsymbol{w}})\right). \end{aligned} \tag{1.26}$$

若记 $A = \sigma_n^{-2}\boldsymbol{X}\boldsymbol{X}^T + \Sigma_p^{-1}$, 则上式中 $\bar{\boldsymbol{w}} = \sigma_n^{-2}A^{-1}\boldsymbol{X}\boldsymbol{y}$, 于是我们得到 $\boldsymbol{w}|\boldsymbol{X}, \boldsymbol{y}$ 也是服从高斯分布的, 即

$$\boldsymbol{w}|\boldsymbol{X}, \boldsymbol{y} \sim \mathcal{N}(\bar{\boldsymbol{w}}, A^{-1}). \tag{1.27}$$

有了 $\boldsymbol{w}|\boldsymbol{X}, \boldsymbol{y}$ 的后验分布, 如果有新的样本 $\boldsymbol{x}_*$, 记 $f_* = f(\boldsymbol{x}_*)$, 于是可以求得 $f_*|\boldsymbol{x}_*, \boldsymbol{X}, \boldsymbol{y}$ 的预测分布 (predictive distribution) 以实现预测 (详细推导过程见附录A.4).

$$
\begin{aligned}
p(f_*|\boldsymbol{x}_*, \boldsymbol{X}, \boldsymbol{y}) &= \int p(f_*|\boldsymbol{x}_*, \boldsymbol{w}) p(\boldsymbol{w}|\boldsymbol{X}, \boldsymbol{y}) \mathrm{d}\boldsymbol{w} \\
&= \mathcal{N}(\frac{1}{\sigma_n^2} \boldsymbol{x}_*^T A^{-1} \boldsymbol{X} \boldsymbol{y}, \boldsymbol{x}_*^T A^{-1} \boldsymbol{x}_*)
\end{aligned}
\tag{1.28}
$$

可见 $f_*|\boldsymbol{x}_*, \boldsymbol{X}, \boldsymbol{y}$ 的预测分布依然是服从高斯分布的.

如果在线性回归基本形式的基础上运用基函数 $\boldsymbol{\phi}(\cdot)$ 将输入 $\boldsymbol{x}$ 向高维映射, 这样就可能将线性回归方法用于非线性数据的建模. 此时模型为

$$
f(\boldsymbol{x}) = \boldsymbol{\phi}(\boldsymbol{x}) \boldsymbol{w} \tag{1.29}
$$

经过基函数的映射, 在预测分布的推导上没有什么不同, 只是将 $\boldsymbol{x}$ 用 $\boldsymbol{\phi}(\boldsymbol{x})$ 替代. 并记输入数据集 $\boldsymbol{X}$ 经过基函数映射结果为 $\boldsymbol{\Phi}$,$A = \sigma_n^{-2} \boldsymbol{\Phi} \boldsymbol{\Phi}^T + \Sigma_p^{-1}$.

$$
f_*|\boldsymbol{x}_*, \boldsymbol{X}, \boldsymbol{y} \sim \mathcal{N}\left(\frac{1}{\sigma_n^2} \boldsymbol{\phi}(\boldsymbol{x}_*)^T A^{-1} \boldsymbol{\Phi} \boldsymbol{y}, \boldsymbol{\phi}(\boldsymbol{x}_*)^T A^{-1} \boldsymbol{\phi}(\boldsymbol{x}_*)\right) \tag{1.30}
$$

为了与函数空间观点做对比, 对上式进行变形 (见附录A.5), 并记 $K = \boldsymbol{\Phi}^T \Sigma_p \boldsymbol{\Phi}$, $\boldsymbol{\phi}(\boldsymbol{x}_*) = \boldsymbol{\phi}_*$, 得

$$
\begin{aligned}
f_*|\boldsymbol{x}_*, \boldsymbol{X}, \boldsymbol{y} \sim \mathcal{N}\Big(& \boldsymbol{\phi}_*^T \Sigma_p \boldsymbol{\Phi} (K + \sigma_n^2 I)^{-1} \boldsymbol{y}, \\
& \boldsymbol{\phi}_*^T \Sigma_p \boldsymbol{\phi}_* - \boldsymbol{\phi}_*^T \Sigma_p \boldsymbol{\Phi} (K + \sigma_n^2 I)^{-1} \boldsymbol{\Phi}^T \Sigma_p \boldsymbol{\phi}_*\Big)
\end{aligned}
\tag{1.31}
$$

上式中各部分总是以 $\boldsymbol{\phi}_*^T \Sigma_p \boldsymbol{\phi}_*$, $\boldsymbol{\phi}_*^T \Sigma_p \boldsymbol{\Phi}$ 或 $\boldsymbol{\Phi}^T \Sigma_p \boldsymbol{\phi}_*$ 的内积形式出现. 由于 Σ_p 是协方差矩阵, 其为对称半正定矩阵, 我们总能找到矩阵 $\Sigma_p^{1/2}$, 使得 $\Sigma_p^{1/2} (\Sigma_p^{1/2})^T = \Sigma_p$(待证明, 可以用特征值分解来说明). 于是我们可以定义一个函数

$$
\begin{aligned}
k(\boldsymbol{x}, \boldsymbol{x}') &= \boldsymbol{\phi}(\boldsymbol{x})^T \Sigma_p \boldsymbol{\phi}(\boldsymbol{x}') \\
&= \boldsymbol{\phi}(\boldsymbol{x})^T \Sigma_p^{1/2} (\Sigma_p^{1/2})^T \boldsymbol{\phi}(\boldsymbol{x}') \\
&= [(\Sigma_p^{1/2})^T \boldsymbol{\phi}(\boldsymbol{x})]^T (\Sigma_p^{1/2})^T \boldsymbol{\phi}(\boldsymbol{x}') \\
&= \boldsymbol{\psi}(\boldsymbol{x})^T \boldsymbol{\psi}(\boldsymbol{x}')
\end{aligned}
\tag{1.32}
$$

上式在函数空间的观点下就被称之为协方差函数或核函数, 从中也可以看到核函数与基函数之间的联系, 两种观点由式1.32连接.

1.1.3.2 函数空间观点解释

函数空间观点的解释将很大程度依赖于高斯分布的性质. 假设 f 是定义在域 D 上的高斯分布, 并假设其均值为 0, 协方差函数为 $k(\boldsymbol{x}, \boldsymbol{x}')$, 即

$$
f \sim \mathcal{GP}\left(0, k(\boldsymbol{x}, \boldsymbol{x}')\right). \tag{1.33}
$$

经过采样得到训练数据集 $\{(\boldsymbol{x}_i, f_i)|i=1,\ldots,n\}$, 记 $\boldsymbol{f}=f(\boldsymbol{X})$, 同样 $\{(\boldsymbol{X}_*, \boldsymbol{f}_*)\}$ 为测试数据集. 先讨论无噪声 (noise-free) 的情况, 即认为观测值没有噪声叠加在上面, 此时的模型为

$$y=f(\boldsymbol{x}). \tag{1.34}$$

由高斯过程的定义可知, $(\boldsymbol{f}, \boldsymbol{f}_*)$ 是联合分布是多维高斯分布, 同时将协方差矩阵分块处理

$$\begin{bmatrix}\boldsymbol{f}\\ \boldsymbol{f}_*\end{bmatrix}\sim\mathcal{N}\left(0,\ \begin{bmatrix}K(\boldsymbol{X},\boldsymbol{X}) & K(\boldsymbol{X},\boldsymbol{X}_*)\\ K(\boldsymbol{X}_*,\boldsymbol{X}) & K(\boldsymbol{X}_*,\boldsymbol{X}_*)\end{bmatrix}\right). \tag{1.35}$$

利用高斯分布条件分布的性质和式1.14与式1.15易得

$$\begin{aligned}\boldsymbol{f}_*|\boldsymbol{X}_*,\boldsymbol{X},\boldsymbol{f}\sim\mathcal{N}(&K\big(\boldsymbol{X}_*,\boldsymbol{X}\big)K(\boldsymbol{X},\boldsymbol{X})^{-1}\boldsymbol{f},\\ &K(\boldsymbol{X}_*,\boldsymbol{X}_*)-K(\boldsymbol{X}_*,\boldsymbol{X})K(\boldsymbol{X},\boldsymbol{X})^{-1}K(\boldsymbol{X},\boldsymbol{X}_*)\big).\end{aligned} \tag{1.36}$$

若考虑噪声, 此时 y 与 f 不再相同, 模型为

$$y=f(\boldsymbol{x})+\varepsilon. \tag{1.37}$$

假设各次采样噪声 $\varepsilon\sim\mathcal{N}(0,\sigma_n^2)$. 在计算协方差矩阵的每一项时需要考虑噪声方差.

$$\mathrm{cov}(y_p,y_q)=k(\boldsymbol{x}_p,\boldsymbol{x}_q)+\sigma_n^2\delta_{pq}. \tag{1.38}$$

其中 δ_{pq} 为克罗内克函数 (Kronecker delta function), 其形式为

$$\delta_{pq}=\begin{cases}1 & p=q,\\ 0 & p\neq q.\end{cases} \tag{1.39}$$

或从协方差的角度看

$$\mathrm{cov}(\boldsymbol{y})=K(\boldsymbol{X},\boldsymbol{X})+\sigma_n^2 I. \tag{1.40}$$

故式1.35改写为

$$\begin{bmatrix}\boldsymbol{y}\\ \boldsymbol{f}_*\end{bmatrix}\sim\mathcal{N}\left(0,\ \begin{bmatrix}K(\boldsymbol{X},\boldsymbol{X})+\sigma_n^2 I & K(\boldsymbol{X},\boldsymbol{X}_*)\\ K(\boldsymbol{X}_*,\boldsymbol{X}) & K(\boldsymbol{X}_*,\boldsymbol{X}_*)\end{bmatrix}\right). \tag{1.41}$$

于是含噪声情况下高斯过程回归的预测分布为

$$\boldsymbol{f}_*|\boldsymbol{X}_*,\boldsymbol{X},\boldsymbol{y}\sim\mathcal{N}(\bar{\boldsymbol{f}}_*,\ \mathrm{cov}(\boldsymbol{f}_*)) \tag{1.42}$$

$$\bar{\boldsymbol{f}}_* = K\big(\boldsymbol{X}_*, \boldsymbol{X}\big)[K(\boldsymbol{X}, \boldsymbol{X}) + \sigma_n^2 I]^{-1}\boldsymbol{y} \tag{1.43}$$

$$\text{cov}(\boldsymbol{f}_*) = K(\boldsymbol{X}_*, \boldsymbol{X}_*) - K(\boldsymbol{X}_*, \boldsymbol{X})[K(\boldsymbol{X}, \boldsymbol{X}) + \sigma_n^2 I]^{-1}K(\boldsymbol{X}, \boldsymbol{X}_*). \tag{1.44}$$

对比发现式1.42, 1.43, 1.44与式1.31本质上是一致的. 从两种观点都可以推出预测分布的形式.

1.2 高斯过程回归模型的求解

1.2.1 核函数选择

除了表1中的单个核函数外, 通过核函数的组合可以生成更多的核函数以建立更复杂的回归模型. 事实上简单的组合方式如两核函数相加, 或两核函数做乘积等, 结果依然满足核函数的性质 (半正定), 于是就可以从已有的函数构造出新的核函数[2]. 在一般的回归问题中, SE 核被证明是拟合数据最有效的, 因此往往被作为首选.

1.2.2 超参数估计

选定核函数后所要做的就是用训练数据来估计得到核函数中的超参数 (hyperparameters), 这些参数确定后, 模型才最终确定, 然后用于预测或回归.

估计核函数参数的方法常用的有两种, 一是最大似然估计 (Maximum likelihood estimation, MLE), 另一种是基于交叉验证的误差最小化.

1.2.2.1 最大似然估计

最大似然估计本质上是使用对数损失函数 (logarithmic loss function) 的经验风险最小化 (empirical risk minimization, ERM) 方法.

在高斯过程的模型下, 设模型参数向量 $\boldsymbol{\theta}$, 假设各次采样值 $\boldsymbol{y} \in \mathbb{R}^d$ 独立同分布于同一个高斯分布. 记 $\Sigma_{xx} = K(\boldsymbol{X}, \boldsymbol{X})$, $\Sigma_{yy} = \text{cov}(\boldsymbol{y})$.

$$\begin{aligned} p(\boldsymbol{y}; \boldsymbol{\theta}) &= \frac{1}{\sqrt{(2\pi)^d |\Sigma_{yy}|}} \text{exp}\left(-\frac{1}{2}\boldsymbol{y}^T \Sigma_{yy}^{-1} \boldsymbol{y}\right) \\ &= \frac{1}{\sqrt{(2\pi)^d |\Sigma_{yy}|}} \text{exp}\left(-\frac{1}{2}\boldsymbol{y}^T (\Sigma_{xx} + \sigma_n^2 I)^{-1} \boldsymbol{y}\right). \end{aligned} \tag{1.45}$$

n 次采样可写出似然函数

$$L(\boldsymbol{Y}; \boldsymbol{\theta}) = \prod_{i=1}^{n} p(\boldsymbol{y}_i; \boldsymbol{\theta}) = \left(\frac{1}{\sqrt{(2\pi)^d |\Sigma_{yy}|}}\right)^n \text{exp}\left(-\frac{1}{2}\sum_{i=1}^{n} \boldsymbol{y}_i^T \Sigma_{yy}^{-1} \boldsymbol{y}_i\right). \tag{1.46}$$

$$\log(L(\boldsymbol{Y}; \boldsymbol{\theta})) = -\frac{dn}{2}\log(2\pi) - \frac{n}{2}\log(|\Sigma_{yy}|) - \frac{1}{2}\sum_{i=1}^{n} \boldsymbol{y}_i^T \Sigma_{yy}^{-1} \boldsymbol{y}_i. \tag{1.47}$$

为求解

$$\max_{\boldsymbol{\theta}} \ L(\boldsymbol{Y};\boldsymbol{\theta}). \tag{1.48}$$

简化为计算：

$$\min_{\boldsymbol{\theta}} \ -2\log(L(\boldsymbol{Y};\boldsymbol{\theta})) = \min_{\boldsymbol{\theta}} \ dn\log(2\pi) + n\log(|\Sigma_{yy}|) + \sum_{i=1}^{n}(\boldsymbol{y}_i^T\Sigma_{yy}^{-1}\boldsymbol{y}_i). \tag{1.49}$$

$$\boldsymbol{\theta} = \text{argmin}\ dn\log(2\pi) + n\log(|\Sigma_{yy}|) + \sum_{i=1}^{n}(\boldsymbol{y}_i^T\Sigma_{yy}^{-1}\boldsymbol{y}_i). \tag{1.50}$$

1.2.2.2 基于交叉验证的误差最小化

最大似然估计虽然简单, 但也有不小的问题, 即得到的参数容易发生过拟合, 也就是模型的泛化能力较差, 对新的输入不能很好得预测. 交叉验证是一种常用的估算模型测试误差的方法, 可以用来选择模型. 用测试误差评定模型的优劣往往更合理. 交叉验证通常使用 k 折 (k-fold) 来处理训练数据, 即把训练数据分成 k 份, 每次取其中一份作为测试集, 其余 k-1 份作为训练集, 于是每一次都能计算出预测误差 (如均方误差 MSE), 轮流作测试集后将求平均误差作为对模型测试误差的估计. 求解参数的过程即使得测试误差最小.

$$MSE_{CV}(\boldsymbol{\theta}) = \frac{1}{k}\sum_{i=1}^{k} MSE_i. \tag{1.51}$$

$$MSE_i = \frac{1}{m_i}\sum_{j=1}^{m_i} \|\boldsymbol{y}_{ij} - \hat{\boldsymbol{y}}_{ij}\|^2. \tag{1.52}$$

上式中 m_i 为第 i 组作为测试集时包含的样本数. 求解优化问题

$$\min_{\boldsymbol{\theta}} MSE_{CV}(\boldsymbol{\theta}). \tag{1.53}$$

$$\boldsymbol{\theta} = \text{argmin}\ MSE_{CV}. \tag{1.54}$$

然而在试验中发现, 基于交叉验证的估计方法所建模型的预测效果并没有明显地好于使用最大似然估计, 甚至有时稍差, 而且计算量更大, 故接下来的试验都使用 MLE 求解模型.

参考文献

[1] Carl Edward Rasmussen, Hannes Nickisch. The gpml toolbox version 4.1[EB/OL]. 2017.

[2] C. E. Rasmussen, C. K. I. Williams. Gaussian processes in machine learning[M]. Cambridge: the MIT Press, 2006.

附录 A

A.1 协方差矩阵性质

协方差矩阵总是对称半正定的. 说明其对称是容易的, 即

$$\Sigma_{i,j} = \text{cov}(Y_i, Y_j) = \text{cov}(Y_j, Y_i) = \Sigma_{j,i}. \tag{A.1}$$

下面证明协方差矩阵是半正定的.

证明 设 X 为一服从均值向量为 μ, 协方差矩阵为 Σ 的 d 维高斯随机向量,$\mu = \text{E}(X)$,$\Sigma = \text{var}(X)$, 由协方差矩阵的定义可得

$$\Sigma = \text{var}(X) = \text{E}[(X-\mu)(X-\mu)^T]. \tag{A.2}$$

设 $\forall \alpha \in \mathbb{R}^d$, 有

$$\alpha^T \Sigma \alpha = \alpha^T \text{E}[(X-\mu)(X-\mu)^T]\alpha = \text{E}[\alpha^T (X-\mu)(X-\mu)^T \alpha]. \tag{A.3}$$

令 $Y = \alpha^T(X-\mu) = (X-\mu)^T\alpha$, $X-\mu$ 是服从于中心化后 (centered) 的多维高斯分布, 由多维高斯分布的定义可知 $Y \sim \mathcal{N}$, 且易得 $\text{E}(Y) = 0$, 于是

$$\alpha^T \Sigma \alpha = \text{E}(Y^2) = \text{var}(Y) \geqslant 0. \tag{A.4}$$

由矩阵半正定的定义可得 Σ 为半正定矩阵. [证毕]

A.2 高斯条件分布

为证明式1.14与式1.15, 需要先介绍分块矩阵求逆定理 (inversion of a partitioned matrix)[2]. 设矩阵 A 为分块矩阵, 其逆阵为 A^{-1}

$$A = \begin{pmatrix} P & Q \\ R & S \end{pmatrix}, \qquad A^{-1} = \begin{pmatrix} \tilde{P} & \tilde{Q} \\ \tilde{R} & \tilde{S} \end{pmatrix}. \tag{A.5}$$

其中 P 与 $\tilde{P}$ 为 $n_1 \times n_1$ 的方阵, S 与 $\tilde{S}$ 为 $n_2 \times n_2$ 的方阵. 在已知 P,Q,R,S 的条件下, 可由下式计算得到 $\tilde{P}$,$\tilde{Q}$,$\tilde{R}$,$\tilde{S}$.

$$\begin{cases} \tilde{P} = N. \\ \tilde{Q} = -NQS^{-1}. \\ \tilde{R} = -S^{-1}RN. \\ \tilde{S} = S^{-1} + S^{-1}RNQS^{-1}. \end{cases} \tag{A.6}$$

其中 $N = (P - QS^{-1}R)^{-1}$. 下面在基于性质 2.2 的设定下完成对式1.14与式1.15的证明.

证明 不失一般性, 我们假定 Y 是经过中心化的, 即 $\mu = 0$. 对于高斯分布我们通常只需关注指数部分, 从多维高斯分布的 pdf 出发, 代入1.13我们得到

$$\begin{aligned} f_Y(y) &\propto \exp\left(-\frac{1}{2}y^T\Sigma^{-1}y\right) \\ &\propto \exp\left(-\begin{pmatrix} y_1^T & y_2^T \end{pmatrix}\begin{pmatrix} \Sigma_{1,1} & \Sigma_{1,2} \\ \Sigma_{2,1} & \Sigma_{2,2} \end{pmatrix}^{-1}\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}\right). \end{aligned} \tag{A.7}$$

下面求分块矩阵的逆, 令

$$\begin{pmatrix} \Sigma_{1,1} & \Sigma_{1,2} \\ \Sigma_{2,1} & \Sigma_{2,2} \end{pmatrix}^{-1} = \begin{pmatrix} A & B \\ C & D \end{pmatrix}. \tag{A.8}$$

由A.6计算得

$$\begin{cases} A = N. \\ B = -N\Sigma_{1,2}\Sigma_{2,2}^{-1}. \\ C = -\Sigma_{2,2}^{-1}\Sigma_{2,1}N. \\ D = \Sigma_{2,2}^{-1} + \Sigma_{2,2}^{-1}\Sigma_{2,1}N\Sigma_{1,2}\Sigma_{2,2}^{-1}. \end{cases} \tag{A.9}$$

上式中 $N = (\Sigma_{1,1} - \Sigma_{1,2}\Sigma_{2,2}^{-1}\Sigma_{2,1})^{-1}$, 由协方差矩阵性质 $\Sigma_{1,2} = \Sigma_{2,1}^T$.
将式A.9代入式A.7, 并进行整理

$$\begin{aligned} f_Y(y) &\propto \exp\left(-\begin{pmatrix} y_1^T & y_2^T \end{pmatrix}\begin{pmatrix} A & B \\ C & D \end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}\right) \\ &\propto \exp(-y_1^TAy_1 - y_2^TCy_1 - y_1^TBy_2 - y_2^TDy_2) \\ &\propto \exp(-y_1^T(\Sigma_{1,1} - \Sigma_{1,2}\Sigma_{2,2}^{-1}\Sigma 2,1)^{-1}y_1 \\ &\qquad + y_2^T\Sigma_{2,2}^{-1}\Sigma_{2,1}(\Sigma_{1,1} - \Sigma_{1,2}\Sigma_{2,2}^{-1}\Sigma_{2,1})^{-1}y_1 \\ &\qquad + y_1^T(\Sigma_{1,1} - \Sigma_{1,2}\Sigma_{2,2}^{-1}\Sigma_{2,1})^{-1}\Sigma_{1,2}\Sigma_{2,2}^{-1}y_2 \\ &\qquad - y_2^T\left[\Sigma_{2,2}^{-1} + \Sigma_{2,2}^{-1}\Sigma_{2,1}(\Sigma_{1,1} - \Sigma_{1,2}\Sigma_{2,2}^{-1}\Sigma_{2,1})^{-1}\Sigma_{1,2}\Sigma_{2,2}^{-1}\right]y_2) \\ &\propto \exp(-y_1^T(\Sigma_{1,1} - \Sigma_{1,2}\Sigma_{2,2}^{-1}\Sigma_{2,1})^{-1}(y_1 - \Sigma_{1,2}\Sigma_{2,2}^{-1}y_2) \\ &\qquad + y_2^T\Sigma_{2,2}^{-1}\Sigma_{2,1}(\Sigma_{1,1} - \Sigma_{1,2}\Sigma_{2,2}^{-1}\Sigma_{2,1})^{-1}(y_1 - \Sigma_{1,2}\Sigma_{2,2}^{-1}y_2)) \\ &\propto \exp(-(y_1^T - y_2^T\Sigma_{2,2}^{-1}\Sigma_{2,1})(\Sigma_{1,1} - \Sigma_{2,2}^{-1}\Sigma_{2,1})^{-1}(y_1 - \Sigma_{1,2}\Sigma_{2,2}^{-1}y_2)) \\ &\propto \exp(-(y_1 - \Sigma_{1,2}\Sigma_{2,2}^{-1}y_2)^T(\Sigma_{1,1} - \Sigma_{2,2}^{-1}\Sigma_{2,1})^{-1}(y_1 - \Sigma_{1,2}\Sigma_{2,2}^{-1}y_2)). \end{aligned} \tag{A.10}$$

在 y_2 已知的条件下, $\mu_c = \Sigma_{1,2}\Sigma_{2,2}^{-1}y_2$ 与 $\Sigma_c = \Sigma_{1,1} - \Sigma_{2,2}^{-1}\Sigma_{2,1}$ 是确定的, 故 $Y_1|Y_2$ 服从均值为 μ_c, 协方差矩阵为 Σ_c 的高斯分布, 即

$$Y_1|Y_2 \sim \mathcal{N}(\mu_c, \Sigma_c).$$

如果不考虑进行中心化, 则只需将式A.10中 y_1 与 y_2 分别用 $y_1-\mu_1$ 与 $y_2-\mu_2$ 替代, 即得到式1.14与式1.15结果.

[证毕]

A.3 贝叶斯线性模型 $\boldsymbol{w}$ 的后验分布推导

将式1.23, 式1.24代入式1.25进行整理

$$
\begin{aligned}
p(\boldsymbol{w}|\boldsymbol{X},\boldsymbol{y}) &\propto p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{w})p(\boldsymbol{w}) \\
&\propto \exp\left(-\frac{1}{2\sigma_n^2}(\boldsymbol{y}-\boldsymbol{X}^T\boldsymbol{w})^T(\boldsymbol{y}-\boldsymbol{X}^T\boldsymbol{w})\right)\exp\left(-\frac{1}{2}\boldsymbol{w}^T\Sigma_p^{-1}\boldsymbol{w}\right) \\
&\propto \exp\left(-\frac{1}{2}\left[\frac{1}{\sigma_n^2}(\boldsymbol{y}-\boldsymbol{X}^T\boldsymbol{w})^T(\boldsymbol{y}-\boldsymbol{X}^T\boldsymbol{w})-\boldsymbol{w}^T\Sigma_p^{-1}\boldsymbol{w}\right]\right) \\
&\propto \exp\left(-\frac{1}{2}\left[\frac{1}{\sigma_n^2}\boldsymbol{y}^T\boldsymbol{y}-\frac{1}{\sigma_n^2}\boldsymbol{y}^T\boldsymbol{X}^T\boldsymbol{w}-\frac{1}{\sigma_n^2}\boldsymbol{w}^T\boldsymbol{X}\boldsymbol{y}+\frac{1}{\sigma_n^2}\boldsymbol{w}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{w}+\boldsymbol{w}^T\Sigma_p^{-1}\boldsymbol{w}\right]\right) \\
&\propto \exp\left(-\frac{1}{2}\left[-\frac{2}{\sigma_n^2}\boldsymbol{w}^T\boldsymbol{X}\boldsymbol{y}+\frac{1}{\sigma_n^2}\boldsymbol{w}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{w}+\boldsymbol{w}^T\Sigma_p^{-1}\boldsymbol{w}\right]\right) \\
&\propto \exp\left(-\frac{1}{2}\left[-\frac{2}{\sigma_n^2}\boldsymbol{w}^T\boldsymbol{X}\boldsymbol{y}+\boldsymbol{w}^T(\frac{1}{\sigma_n^2}\boldsymbol{X}\boldsymbol{X}^T+\Sigma_p^{-1})\boldsymbol{w}\right]\right) \\
&\propto \exp\left(-\frac{1}{2}\left[-\frac{2}{\sigma_n^2}\boldsymbol{w}^T\boldsymbol{X}\boldsymbol{y}+\boldsymbol{w}^TA\boldsymbol{w}\right]\right).
\end{aligned}
\tag{A.11}
$$

其中 $A=\sigma^{-2}\boldsymbol{X}\boldsymbol{X}^T+\Sigma_p^{-1}$. 最后需要凑成多维高斯分布指数的形式, 可以使用待定系数法来求 $\bar{\boldsymbol{w}}$. 从目标形式出发

$$
\begin{aligned}
p(\boldsymbol{w}|\boldsymbol{X},\boldsymbol{y}) &\propto \exp\left(-\frac{1}{2}(\boldsymbol{w}-\bar{\boldsymbol{w}})^TA(\boldsymbol{w}-\bar{\boldsymbol{w}})\right) \\
&\propto \exp\left(-\frac{1}{2}(\boldsymbol{w}^TA\boldsymbol{w}-\boldsymbol{w}^TA\bar{\boldsymbol{w}}-\bar{\boldsymbol{w}}^TA\boldsymbol{w}+\bar{\boldsymbol{w}}^TA\bar{\boldsymbol{w}})\right) \\
&\propto \exp\left(-\frac{1}{2}(\boldsymbol{w}^TA\boldsymbol{w}-2\boldsymbol{w}^TA\bar{\boldsymbol{w}})\right).
\end{aligned}
\tag{A.12}
$$

上式推导中注意 $\bar{\boldsymbol{w}}$ 应为与 $\boldsymbol{w}$ 无关的常数. 对比式A.11与式A.12即可得

$$
\begin{aligned}
A\bar{\boldsymbol{w}} &= \frac{1}{\sigma_n^2}\boldsymbol{X}\boldsymbol{y} \\
\bar{\boldsymbol{w}} &= \frac{1}{\sigma_n^2}A^{-1}\boldsymbol{X}\boldsymbol{y}.
\end{aligned}
\tag{A.13}
$$

A.4 贝叶斯线性模型预测分布推导

为了得到 $f_*|\boldsymbol{x}_*,\boldsymbol{X},\boldsymbol{y}$ 的后验分布, 先求 $y_*|\boldsymbol{x}_*,\boldsymbol{X},\boldsymbol{y}$ 的后验分布, 然后利用 $y_*=f_*+\varepsilon_*$ 求得 $f_*|\boldsymbol{x}_*,\boldsymbol{X},\boldsymbol{y}$.

$$p(y_*|\boldsymbol{x}_*,\boldsymbol{X},\boldsymbol{y})=\int p(y_*|\boldsymbol{x}_*,\boldsymbol{w})p(\boldsymbol{w}|\boldsymbol{X},\boldsymbol{y})\mathrm{d}\boldsymbol{w}. \tag{A.14}$$

借助式1.22与式1.26, 有

$$\begin{aligned}
p(y_*|\boldsymbol{x}_*,\boldsymbol{X},\boldsymbol{y}) &\propto \int \exp\left(-\frac{(y_*-\boldsymbol{w}^T\boldsymbol{x}_*)^2}{2\sigma_n^2}\right)\exp\left(-\frac{1}{2}(\boldsymbol{w}-\bar{\boldsymbol{w}})^T A(\boldsymbol{w}-\bar{\boldsymbol{w}})\right)\mathrm{d}\boldsymbol{w}\\
&\propto \int \exp\left[-\frac{1}{2}\left(\frac{1}{\sigma_n^2}y_*^2-\frac{2}{\sigma_n^2}\boldsymbol{w}^T\boldsymbol{x}y_*+\frac{1}{\sigma_n^2}\boldsymbol{w}^T\boldsymbol{x}_*\boldsymbol{x}_*^T\boldsymbol{w}+\boldsymbol{w}^TA\boldsymbol{w}-2\boldsymbol{w}^TA\bar{\boldsymbol{w}}\right)\right]\mathrm{d}\boldsymbol{w}\\
&\propto \int \exp\left[-\frac{1}{2}\left(\boldsymbol{w}^T(\frac{1}{\sigma_n^2}\boldsymbol{x}_*\boldsymbol{x}_*^T+A)\boldsymbol{w}-2\boldsymbol{w}^T(\frac{1}{\sigma_n^2}\boldsymbol{x}_*y_*+A\bar{\boldsymbol{w}})+\frac{1}{\sigma_n^2}y_*^2\right)\right]\mathrm{d}\boldsymbol{w}\\
&\propto \int \exp\left[-\frac{1}{2}\left(\boldsymbol{w}^TL\boldsymbol{w}-2\boldsymbol{w}^T(\frac{1}{\sigma_n^2}\boldsymbol{x}_*y_*+A\bar{\boldsymbol{w}})+\frac{1}{\sigma_n^2}y_*^2\right)\right]\mathrm{d}\boldsymbol{w}.
\end{aligned} \tag{A.15}$$

上式中 $L=\sigma_n^{-2}\boldsymbol{x}_*\boldsymbol{x}_*^T+A$, 构造 $\boldsymbol{w}$ 服从均值为 $\boldsymbol{m}$, 协方差矩阵为 L 的高斯分布指数部分, 用待定系数法求出 $\boldsymbol{m}$.

$$\begin{aligned}
&(\boldsymbol{w}-\boldsymbol{m})^TL(\boldsymbol{w}-\boldsymbol{m})\\
=&\boldsymbol{w}^TL\boldsymbol{w}-2\boldsymbol{w}^TL\boldsymbol{m}+\boldsymbol{m}^TL\boldsymbol{m}
\end{aligned} \tag{A.16}$$

对比式A.15与式A.16, 可求出 $\boldsymbol{m}$.

$$\begin{aligned}
L\boldsymbol{m}&=\frac{1}{\sigma_n^2}\boldsymbol{x}_*y_*+A\bar{\boldsymbol{w}}\\
\boldsymbol{m}&=L^{-1}(\frac{1}{\sigma_n^2}\boldsymbol{x}_*y_*+A\bar{\boldsymbol{w}})
\end{aligned} \tag{A.17}$$

$\boldsymbol{m}$ 与 L 是独立于 $\boldsymbol{w}$ 的, 于是

$$\begin{aligned}
p(y_*|\boldsymbol{x}_*,\boldsymbol{X},\boldsymbol{y}) &\propto \int \exp\left(-\frac{1}{2}(\boldsymbol{w}-\boldsymbol{m})^TL(\boldsymbol{w}-\boldsymbol{m})\right)\exp\left[-\frac{1}{2}(\frac{1}{\sigma_n^2}y_*^2-\boldsymbol{m}^TL\boldsymbol{m})\right]\mathrm{d}\boldsymbol{w}\\
&\propto \exp\left[-\frac{1}{2}\left(\frac{1}{\sigma_n^2}y_*^2-\boldsymbol{m}^TL\boldsymbol{m}\right)\right]\int \exp\left(-\frac{1}{2}(\boldsymbol{w}-\boldsymbol{m})^TL(\boldsymbol{w}-\boldsymbol{m})\right)\mathrm{d}\boldsymbol{w}\\
&\propto \exp\left[-\frac{1}{2}\left(\frac{1}{\sigma_n^2}y_*^2-\boldsymbol{m}^TL\boldsymbol{m}\right)\right]
\end{aligned} \tag{A.18}$$

$$\begin{aligned}
\boldsymbol{m}^TL\boldsymbol{m}&=\left(\frac{1}{\sigma_n^2}\boldsymbol{x}_*y_*+A\bar{\boldsymbol{w}}\right)^TL^{-1}\cdot L\cdot L^{-1}\left(\frac{1}{\sigma_n^2}\boldsymbol{x}_*y_*+A\bar{\boldsymbol{w}}\right)\\
&=\left(\frac{1}{\sigma_n^4}\boldsymbol{x}_*^TL^{-1}\boldsymbol{x}_*\right)y_*^2+2\left(\frac{1}{\sigma_n^2}\boldsymbol{x}_*^TL^{-1}A\bar{\boldsymbol{w}}\right)y_*+\bar{\boldsymbol{w}}^TAL^{-1}A\bar{\boldsymbol{w}}
\end{aligned} \tag{A.19}$$

把式A.19代入A.18, 注意 L,A,$\bar{\boldsymbol{w}}$ 都独立于 y_*, 整理得

$$
\begin{aligned}
p(y_*|\boldsymbol{x}_*, \boldsymbol{X}, \boldsymbol{y}) &\propto \exp\left[-\frac{1}{2}\left(\frac{1}{\sigma_n^2}y_*^2 - \boldsymbol{m}^T L \boldsymbol{m}\right)\right] \\
&\propto \exp\left[-\frac{1}{2}\left(\left(\frac{1}{\sigma_n^2} - \frac{1}{\sigma_n^4}\boldsymbol{x}_*^T L^{-1}\boldsymbol{x}_*\right)y_*^2 - 2\left(\frac{1}{\sigma_n^2}\boldsymbol{x}_*^T L^{-1} A\bar{\boldsymbol{w}}\right)y_*\right)\right]
\end{aligned} \tag{A.20}
$$

同样从目标出发

$$
\lambda(y_* - \bar{y}_*)^2 = \lambda y_*^2 - 2\lambda\bar{y}_* y_* + \lambda\bar{y}_*^2 \tag{A.21}
$$

对比式A.20括号内部分与式A.21得

$$
\lambda = \frac{1}{\sigma_n^2}(1 - \frac{1}{\sigma_n^2}\boldsymbol{x}_*^T L^{-1}\boldsymbol{x}_*). \tag{A.22}
$$

$$
\bar{y}_* = \frac{1}{\lambda}(\frac{1}{\sigma_n^2}\boldsymbol{x}_*^T L^{-1} A\bar{\boldsymbol{w}}). \tag{A.23}
$$

于是

$$
p(y_*|\boldsymbol{x}_*, \boldsymbol{X}, \boldsymbol{y}) \propto \exp\left(-\frac{1}{2}\lambda(y_* - \bar{y}_*)^2\right). \tag{A.24}
$$

下面计算 λ 与 $\bar{y}_*$ 以得到预测分布得均值与方差. 为简便起见, 首先计算 L^{-1}, $\boldsymbol{x}_*^T L^{-1}\boldsymbol{x}_*$, 以及 $\boldsymbol{x}_*^T L^{-1}$. 计算 L^{-1} 需要用到矩阵求逆引理 (matrix inversion lemma)[2]. 以下计算中记 $\alpha = \boldsymbol{x}_*^T A^{-1}\boldsymbol{x}$.

$$
\begin{aligned}
L^{-1} &= \left(A + \boldsymbol{x}_*\frac{1}{\sigma_n^2}\boldsymbol{x}_*^T\right)^{-1} \\
&= A^{-1} - A^{-1}\boldsymbol{x}_*(\sigma_n^2 + \boldsymbol{x}_*^T A^{-1}\boldsymbol{x}_*)^{-1}\boldsymbol{x}_*^T A^{-1} \\
&= A^{-1} - \frac{1}{\sigma_n^2 + \alpha}A^{-1}\boldsymbol{x}_*\boldsymbol{x}_*^T A^{-1}.
\end{aligned} \tag{A.25}
$$

$$
\boldsymbol{x}_*^T L^{-1}\boldsymbol{x}_* = \boldsymbol{x}_*^T A^{-1}\boldsymbol{x}_* - \frac{1}{\sigma_n^2 + \alpha}\boldsymbol{x}_*^T A^{-1}\boldsymbol{x}_*\boldsymbol{x}_*^T A^{-1}\boldsymbol{x}_* = \frac{\alpha\sigma_n^2}{\sigma_n^2 + \alpha}. \tag{A.26}
$$

$$
\boldsymbol{x}_*^T L^{-1} = \frac{\sigma_n^2}{\sigma_n^2 + \alpha}\boldsymbol{x}_*^T A^{-1}. \tag{A.27}
$$

故

$$
\lambda = \frac{1}{\sigma_n^2 + \alpha}. \tag{A.28}
$$

$$
\bar{y}_* = \frac{1}{\sigma_n^2}\boldsymbol{x}_n^T A^{-1}\boldsymbol{X}\boldsymbol{y}. \tag{A.29}
$$

$$y_*|\boldsymbol{x}_*, \boldsymbol{X}, \boldsymbol{y} \sim \mathcal{N}(\frac{1}{\sigma_n^2}\boldsymbol{x}_n^T A^{-1}\boldsymbol{X}\boldsymbol{y}, \frac{1}{\lambda} = \sigma_n^2 + \boldsymbol{x}_*^T A^{-1}\boldsymbol{x}). \tag{A.30}$$

由 $f_* = y_* - \varepsilon_*$

$$f_*|\boldsymbol{x}_*, \boldsymbol{X}, \boldsymbol{y} \sim \mathcal{N}(\frac{1}{\sigma_n^2}\boldsymbol{x}_n^T A^{-1}\boldsymbol{X}\boldsymbol{y}, \boldsymbol{x}_*^T A^{-1}\boldsymbol{x}). \tag{A.31}$$

A.5 贝叶斯线性模型预测分布变形

首先对式1.30均值进行变形.

$$\begin{aligned}
\boldsymbol{\phi}_*^T \left(\frac{1}{\sigma_n^2} A^{-1}\boldsymbol{\Phi}\right)\boldsymbol{y} &= \boldsymbol{\phi}_*^T \left[A^{-1} \cdot \frac{1}{\sigma_n^2}\boldsymbol{\Phi}(K + \sigma_n^2 I) \cdot (K + \sigma_n^2 I)^{-1}\right]\boldsymbol{y} \\
&= \boldsymbol{\phi}_*^T \left[A^{-1} \cdot A\Sigma_p\boldsymbol{\Phi} \cdot (K + \sigma_n^2 I)^{-1}\right]\boldsymbol{y} \\
&= \boldsymbol{\phi}_*^T \Sigma_p \boldsymbol{\Phi}(K + \sigma_n^2 I)^{-1}\boldsymbol{y}.
\end{aligned} \tag{A.32}$$

然后方差的变形需要先用矩阵求逆引理展开 A^{-1}.

$$\begin{aligned}
A^{-1} &= \left(\frac{1}{\sigma_n^2}\boldsymbol{\Phi}\boldsymbol{\Phi}^T + \Sigma_p^{-1}\right)^{-1} \\
&= \Sigma_p - \Sigma_p\boldsymbol{\Phi}(\boldsymbol{\Phi}^T\Sigma_p\boldsymbol{\Phi} + \sigma_n^2 I)^{-1}\boldsymbol{\Phi}^T\Sigma_p \\
&= \Sigma_p - \Sigma_p\boldsymbol{\Phi}(K + \sigma_n^2 I)^{-1}\boldsymbol{\Phi}^T\Sigma_p
\end{aligned} \tag{A.33}$$

故

$$\boldsymbol{\phi}_*^T A^{-1}\boldsymbol{\phi}_* = \boldsymbol{\phi}_*^T\Sigma_p\boldsymbol{\phi}_* - \boldsymbol{\phi}_*^T\Sigma_p\boldsymbol{\Phi}(K + \sigma_n^2 I)^{-1}\boldsymbol{\Phi}^T\Sigma_p\boldsymbol{\phi}_* \tag{A.34}$$