高斯过程——笔记与推导

顾锴

2020 年 5 月 4 日

1 引入

1.1 线性模型

考虑一个线性模型

$$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{w}, \quad y = f(x) + \varepsilon \tag{1}$$

假定线性拟合的参数$\mathbf{w}$满足先验分布

$$\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \Sigma_p) \tag{2}$$

而ε则是高斯型噪声，

$$\varepsilon \sim \mathcal{N}(0, \sigma_n^2) \tag{3}$$

则似然函数是高斯函数。对于一组输入X和对应的输出$\mathbf{y}$，似然函数为

$$p(\mathbf{y}|X, \mathbf{w}) = \prod_{i=0}^{n} p(y_i|\mathbf{x}_i, \mathbf{w}) = \prod_{i=0}^{n} \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left(-\frac{(y_i - \mathbf{x}_i^\top \mathbf{w})^2}{2\sigma_n^2}\right) \sim \mathcal{N}(X^\top \mathbf{w}, \sigma_n^2 I) \tag{4}$$

根据贝叶斯定理，

$$p(\mathbf{w}|X, \mathbf{y}) = \frac{p(\mathbf{y}|X, \mathbf{w})p(\mathbf{w})}{\int \mathrm{d}\mathbf{w}\, p(\mathbf{y}|X, \mathbf{w})p(\mathbf{w})} \tag{5}$$

分母是常数，而分子是两个高斯函数，乘积仍然为高斯，因此整体应该仍然为高斯函数，故而可以暂时忽略可以合并入归一化因子的分母，即

$$p(\mathbf{w}|X, \mathbf{y}) \propto \exp\left[-\frac{1}{2\sigma_n^2}(\mathbf{y} - X^\top \mathbf{w})^\top (\mathbf{y} - X^\top \mathbf{w})\right] \exp\left(-\frac{1}{2}\mathbf{w}^\top \Sigma_p^{-1} \mathbf{w}\right) \tag{6}$$

$$\propto \exp\left[-\frac{1}{2}(\mathbf{w} - \boldsymbol{\mu}_w)^\top \Sigma_w^{-1} (\mathbf{w} - \boldsymbol{\mu}_w)\right] \tag{7}$$

上下对比，可得

$$\begin{cases} \Sigma_w^{-1} = \frac{1}{\sigma_n^2} XX^\top + \Sigma_p^{-1} \\ \boldsymbol{\mu}_w = \frac{1}{\sigma_n^2} \Sigma_w X\mathbf{y} \end{cases} \tag{8}$$

则$\mathbf{w}$的后验概率满足

$$p(\mathbf{w}|X,\mathbf{y}) \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{w}}, \Sigma_p) = \mathcal{N}\left(\frac{1}{\sigma_n^2}\left(\frac{XX^\top}{\sigma_n^2} + \Sigma_p^{-1}\right)^{-1} X\mathbf{y}, \left(\frac{XX^\top}{\sigma_n^2} + \Sigma_p^{-1}\right)^{-1}\right) \tag{9}$$

对于一组新的输入$\mathbf{x}_*$，可以以此预测其输出，表现为边缘分布的形式

$$p(f_*|\mathbf{x}_*, X, \mathbf{y}) = \int \mathrm{d}\mathbf{w}\, p(f_*|\mathbf{x}_*, \mathbf{w}) p(\mathbf{w}|X,\mathbf{y}) \tag{10}$$

$$\sim \mathcal{N}(\mathbf{x}_*^\top \boldsymbol{\mu}_w, \mathbf{x}_*^\top \Sigma_w \mathbf{x}_*) \tag{11}$$

$$= \mathcal{N}\left(\frac{1}{\sigma_n^2}\mathbf{x}_*^\top \left(\frac{XX^\top}{\sigma_n^2} + \Sigma_p^{-1}\right)^{-1} X\mathbf{y}, \mathbf{x}_*^\top \left(\frac{XX^\top}{\sigma_n^2} + \Sigma_p^{-1}\right)^{-1} \mathbf{x}_*\right) \tag{12}$$

1.2 拓展的线性形式

如果现在的函数形式为

$$f(\mathbf{x}) = \boldsymbol{\phi}(\mathbf{x})^\top \mathbf{w} \tag{13}$$

例如，对于某个一维的输入，将其映射到多项式空间

$$x \mapsto \boldsymbol{\phi}(\boldsymbol{x}) = (1, x, x^2, x^3, \dots)^\top \tag{14}$$

进行多项式拟合，就可以利用上述形式。

在这种情况下，上一节的推导可以全部重复使用，只需要把所有的$\mathbf{x}$用$\boldsymbol{\phi}(\mathbf{x})$代替即可。令$\Phi = \boldsymbol{\phi}(X)$和$\boldsymbol{\phi}_* = \boldsymbol{\phi}(\mathbf{x}_*)$，有

$$p(f_*|\mathbf{x}_*, X, \mathbf{y}) \sim \mathcal{N}\left(\frac{1}{\sigma_n^2}\boldsymbol{\phi}_*^\top \left(\frac{\Phi\Phi^\top}{\sigma_n^2} + \Sigma_p^{-1}\right)^{-1} \Phi\mathbf{y}, \boldsymbol{\phi}_*^\top \left(\frac{\Phi\Phi^\top}{\sigma_n^2} + \Sigma_p^{-1}\right)^{-1} \boldsymbol{\phi}_*\right) \tag{15}$$

上式可以重新写作

$$\begin{aligned} p(f_*|\mathbf{x}_*, X, \mathbf{y}) \sim \mathcal{N}(&k(\mathbf{x}_*, X)(k(X,X) + \sigma_n^2 I)^{-1}\mathbf{y}, \\ &k(\mathbf{x}_*, \mathbf{x}_*) - k(\mathbf{x}_*, X)(k(X,X) + \sigma_n^2 I)^{-1} k(X, \mathbf{x}_*)) \end{aligned} \tag{16}$$

其中

$$k(X, X')_{ij} = \boldsymbol{\phi}(X_{\cdot,i})^\top \Sigma_p \boldsymbol{\phi}(X'_{\cdot,j}) \Leftrightarrow k(X, X') = \boldsymbol{\phi}(X)^T \Sigma_p \boldsymbol{\phi}(X') \tag{17}$$

证明是平凡的。到此，即有了高斯过程的雏形。

1.3 定义

高斯过程，是**一系列随机变量的集合，其中任意个的组合都满足联合高斯分布**。它可以由均值函数（一般设为零）和协方差函数联合定义。

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')) \tag{18}$$

$$m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})] \tag{19}$$

$$k(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))] \tag{20}$$

可以考虑成时空上连续的一个变量，其每一点并非确定的函数值，而是一个高斯分布；任意数量的点可以组成一个联合高斯分布。

例如，贝叶斯线性回归模型$f(\mathbf{x}) = \boldsymbol{\phi}(\mathbf{x})^\top \mathbf{w}$中，基于先验分布$\mathbf{w} \sim \mathcal{N}(0, \Sigma_p)$，

$$\mathbb{E}[f(\mathbf{x})] = \boldsymbol{\phi}(\mathbf{x})^\top \mathbb{E}[\mathbf{w}] = 0 \tag{21}$$

$$\mathbb{E}[f(\mathbf{x} f(\mathbf{x}')] = \boldsymbol{\phi}(\mathbf{x})^\top \mathbb{E}[\mathbf{w}\mathbf{w}^\top] \boldsymbol{\phi}(\mathbf{x}) = \boldsymbol{\phi}(\mathbf{x})^\top \Sigma_p \boldsymbol{\phi}(\mathbf{x}) \tag{22}$$

关于协方差函数和均值函数的选取会在后续有更多的讨论。一个常用的协方差函数为平方指数函数(Squared Exponential, SE)或称为（最常用的）径向基函数(Radial Basis Function, RBF)，其形式为

$$k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left(-\frac{1}{2l^2}|\mathbf{x} - \mathbf{x}'|^2\right) + \sigma_n^2 \delta(\mathbf{x} - \mathbf{x}') \tag{23}$$

这里的特征长度尺度l，信号方差σ_f和噪声方差σ_n都是可变的，它们被称为**超参数**。一般而言，协方差函数中的超参数需要进行优化以使得拟合结果与真实情况尽可能接近，否则则会引入过多的白噪声。

对于输入数据无噪声的情况，即$\mathbf{y} \sim \mathcal{N}(0, K(X, X))$，

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K(X, X) & K(X, X_*) \\ K(X_*, X) & K(X_*, X_*) \end{bmatrix}\right) \tag{24}$$

$$\Rightarrow p(\mathbf{f}_* | X_*, X, \mathbf{y}) \sim \mathcal{N}(K(X_*, X)K(X, X)^{-1}\mathbf{y}, K(X_*, X_*) - K(X_*, X)K(X, X)^{-1}K(X, X_*)) \tag{25}$$

$\mathbf{f}_*$可根据后验概率抽样产生。

对于输入数据存在观测噪声的情况，即$\mathbf{y} \sim \mathcal{N}(0, K(X, X) + \sigma_n^2 I)$，只需要将上述内容中的$K(X, X)$用$K(X, X) + \sigma_n^2 I$代替即可，后验概率为

$$p(\mathbf{f}_* | X_*, X, \mathbf{y}) \sim \mathcal{N}(K(X_*, X)[K(X, X) + \sigma_n^2 I]^{-1}\mathbf{y},$$

$$K(X_*, X_*) - K(X_*, X)[K(X, X) + \sigma_n^2 I]^{-1} K(X, X_*)) \tag{26}$$

可以注意到均值是已知输出的线性叠加，所以高斯过程是**线性预测器**，对于有无噪声的情况皆然。

此外需要引入边缘似然函数$p(\mathbf{y}|X)$，这里是指对函数值的边缘化

$$p(\mathbf{y}|X) = \int \mathrm{d}\mathbf{f}\, p(\mathbf{y}|\mathbf{f}, X) p(\mathbf{f}|X) \sim \mathcal{N}(\mathbf{0}, K(X, X) + \sigma_n^2 I) \tag{27}$$

这里先验概率$p(\mathbf{f}|X) \sim \mathcal{N}(\mathbf{0}, K(X, X))$而似然$p(\mathbf{y}|\mathbf{f}, X) \sim \mathcal{N}(\mathbf{f}, \sigma_n^2 I)$。

1.4 光滑

高斯过程是一种**线性平滑器**。

对于训练集的预测均值，记$K = K(X, X)$，则有

$$\bar{\mathbf{f}} = K(K + \sigma_n^2 I)^{-1} \mathbf{y} \tag{28}$$

其中，如果将K用其本征态展开（K是实对称矩阵，必然可对角化）$K = \sum_i \lambda_1 \mathbf{u}_i \mathbf{u}_i^\top$，则

$$\bar{\mathbf{f}} = \sum_i \frac{\lambda_i (\mathbf{u}_i^\top \mathbf{y})}{\lambda_i + \sigma_n^2} \mathbf{u}_i \tag{29}$$

对于绝大多数情况而言，对应较大本征值的本征向量变化较慢，因此可以将$\mathbf{y}$中高频变化的部分平滑化。有效参数数或者平滑器的自由度定义为

$$\mathrm{tr}\big(K(K + \sigma_n^2 I)^{-1}\big) = \sum_i \frac{\lambda_i}{\lambda_i + \sigma_n^2} \tag{30}$$

如果我们定义权重函数$\mathbf{h}(\mathbf{x}_*)$，

$$\mathbf{h}(\mathbf{x}_*) = (K(X, X) + \sigma_n^2 I)^{-1} k(\mathbf{x}_*) \tag{31}$$

则

$$\bar{f}(\mathbf{x}_*) = \mathbf{h}(\mathbf{x}_*)^\top \mathbf{y} \tag{32}$$

即为已知输出的线性叠加，且权重函数不依赖于已知输出$\mathbf{y}$。理想的权重函数被称为**等价核**。

对于核平滑器，一种方法是Nadaraya-Watson的核加权平均，

$$\bar{f}(\mathbf{x}_*) = \frac{\sum_i k(\mathbf{x}_*, \mathbf{x}_i) y_i}{\sum_i k(\mathbf{x}_*, \mathbf{x}_i)} \tag{33}$$

这样权重随距离平滑衰减，拟合结果也会比较平滑。

1.5 均值函数

一般来说均值函数会设为零，但这不是必须的。如果我们指定一个均值函数，即

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')) \tag{34}$$

则后验分布的均值变为

$$\bar{\mathbf{f}}_* = \mathbf{m}(X_*) + K(X_*, X)(K(X, X) + \sigma_n^2 I)^{-1}(\mathbf{y} - \mathbf{m}(X)) \tag{35}$$

而协方差不变。

很多时候很难指定一个特定的均值函数，但是可以考虑用某种函数基展开。考虑以下情况

$$g(\mathbf{x}) = f(\mathbf{x}) + \mathbf{h}(\mathbf{x})^\top \boldsymbol{\beta}, \quad f(\mathbf{x}) \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}')) \tag{36}$$

其中$\boldsymbol{\beta}$是额外引入的参数，$\mathbf{h}(\mathbf{x})$是一系列预设好的基函数（例如$\mathbf{h}(\mathbf{x}) = (1, x, x^2, \dots)^\top$），而$f(\mathbf{x})$是一个均值为零的高斯过程。如果假定$\boldsymbol{\beta}$满足先验分布$\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{b}, B)$，则有

$$g(\mathbf{x}) \sim \mathcal{GP}(\mathbf{h}(\mathbf{x})^\top \mathbf{b}, k(\mathbf{x}, \mathbf{x}') + h(\mathbf{x})^\top B h(\mathbf{x}')) \tag{37}$$

对应地可以导出

$$p(\mathbf{g}_* | X_*, X, \mathbf{y}) \sim \mathcal{N}(\bar{\mathbf{g}}_*, \mathbb{V}(\mathbf{g}_*)) \tag{38}$$

其中如果令$H = H(X)$，$H_* = H(X_*)$和$K_y = K(X, X) + \sigma_n^2 I$，

$$\begin{aligned}
\bar{\mathbf{g}}_* =& H_*^\top \mathbf{b} + [K(X_*, X) + H_*^\top B H](K_y + H^\top B H)^{-1}(\mathbf{y} - H^\top \mathbf{b}]) && (39)\\
=& H_*^\top \bar{\boldsymbol{\beta}} + K(X_*, X) K_y^{-1}(\mathbf{y} - H^\top \bar{\boldsymbol{\beta}}) && (40)\\
=& \bar{\mathbf{f}}_* + R^\top \bar{\boldsymbol{\beta}} && (41)
\end{aligned}$$

这里$\bar{\boldsymbol{\beta}} = (B^{-1} + H K_y^{-1} H^\top)^{-1}(H K_y^{-1} \mathbf{y} + B^{-1}\mathbf{b})$，是参数的均值，可以看成先验分布与数据的折衷（类似于后验分布）。$R = H_* - H K_y^{-1} K(X, X_*)$，可以认为是余项矩阵。而$\bar{\mathbf{f}}_* = K(X_*, X) K_y^{-1} \mathbf{y}$，是原高斯过程的预测均值。证明过程是平凡的，其含义为预测均值是线性均值输出加上高斯模型从残差中预测的值。协方差是

$$\begin{aligned}
\mathbb{V}[\mathbf{g}_*] =& K(X_*, X_*) + H_*^\top B H_* \\
& - [K(X_*, X) + H_*^\top B H](K_y + H^\top B H)^{-1}[K(X, X_*) + H^\top B H_*] && (42)\\
=& \mathbb{V}[\mathbf{f}_*] + R^\top (B^{-1} + H K_y^{-1} H^\top)^{-1} R && (43)
\end{aligned}$$

这里$\mathbb{V}[\mathbf{f}_*] = K(X_*, X_*) - K(X_*, X)K_y^{-1}K(X, X_*)$。证明也是类似的，可以看成通常协方差项和新的非负贡献之和。而且，如果$B^{-1} \to 0$，即先验分布是均匀的，则$\bar{\boldsymbol{\beta}}$以及由此导致的$\mathbf{g}_*$和$\mathbb{V}[\mathbf{g}_*]$都不依赖于$\mathbf{b}$。

我们可以类似地写出边缘似然分布，

$$p(\mathbf{y}|X, \mathbf{b}, B) \sim \mathcal{N}(H^T\mathbf{b}, K_y + H^\top BH) \tag{44}$$

当先验分布均匀化的情况下，即$B^{-1} \to 0$，则

$$p(\mathbf{y}|X) \sim \mathcal{N}(\mathbf{0}, [K_y^{-1} + K_y^{-1}H^\top (HK_y^{-1}H^\top)^{-1}HK_y^{-1}]^{-1}) \tag{45}$$

2 协方差函数

协方差函数在高斯过程中非常重要，因为其定义了高斯过程中数据集之间的接近度或相似度。一般的以$\mathbf{x}$和$\mathbf{x}'$为变量的函数并不能作为协方差函数。常见的协方差函数如下表

协方差函数	表达式	是否平稳	是否非简并
常数	σ_0^2	是	
线性	$\sum_{d=1}^{D} \sigma_d^2 x_d x'_d$		
多项式	$(\mathbf{x} \cdot \mathbf{x}' + \sigma_0^2)^p$		
平方指数	$\exp\left(-\frac{r^2}{2l^2}\right)$	是	是
马特恩	$\frac{1}{2^{\nu-1}\Gamma(\nu)}\left(\frac{\sqrt{2\nu}}{l}r\right)^{\nu} K_\nu\left(\frac{\sqrt{2\nu}}{l}r\right)$	是	是
指数	$\exp\left(-\frac{r}{l}\right)$	是	是
γ指数	$\exp\left[-\left(\frac{r}{l}\right)^\gamma\right]$	是	是
有理二次	$\left(1 + \frac{r^2}{2\alpha l^2}\right)^{-\alpha}$	是	是
神经网络	$\sin^{-1}\left(\frac{2\tilde{\mathbf{x}}^\top \Sigma \tilde{\mathbf{x}}'}{\sqrt{(1+2\tilde{\mathbf{x}}^\top \Sigma \tilde{\mathbf{x}})(1+2\tilde{\mathbf{x}}'^\top \Sigma \tilde{\mathbf{x}}')}}\right)$		是

表 1: 常见协方差函数。表中$r = |\mathbf{x} - \mathbf{x}'|$。

2.1 准备工作

常见的协方差函数有

- 平稳协方差函数 只依赖于$\mathbf{x} - \mathbf{x}'$的函数，它们具有平移不变性。进一步，如果函数只依赖于$|\mathbf{x} - \mathbf{x}'|$，则称其为各向同性的，也称为径向基函数(RBF)，例如平方指数

函数SE。在随机过程理论中，具有恒定均值且协方差函数对平移不变的过程称为宽平稳过程。如果一个过程的所有有限维分布都具有平移不变性，则该过程是严平稳的。

- 点积协方差函数 只依赖于$\mathbf{x}\cdot\mathbf{x}'$的函数，例如$k(\mathbf{x},\mathbf{x}')=\mathbf{x}\cdot\mathbf{x}'+\sigma_0^2$，这可以由满足$\mathcal{N}(0,1)$的先验系数和$\mathcal{N}(0,\sigma_0^2)$的先验偏差线性回归得到。再比如非齐次多项式核函数$k(\mathbf{X},\mathbf{x}')=(\sigma_0^2+\mathbf{x}\cdots\mathbf{x}')^p,\quad p\in\mathbb{Z}^+$。

更一般的称呼为**核函数**，这一术语来自积分算子理论。

对于一组输入$\{\mathbf{x}_i|i=1,\ldots,n\}$，定义**格拉姆矩阵**为$K_{ij}=k(\mathbf{x}_i,\mathbf{x}_j)$。如果$k$是协方差函数则其必为半正定函数，对应的格拉姆矩阵$K$称为**协方差矩阵**，且必为半正定矩阵。

如果存在，一维随机过程的特征长度尺度可以由u级上交次数决定。由零均值平稳且几乎肯定是连续的高斯过程得出的单位间隔上的u级的数量为

$$\mathbb{E}[N_u]=\frac{1}{2\pi}\sqrt{\frac{-k''(0)}{k(0)}}\exp\left(-\frac{u^2}{2k(0)}\right) \tag{46}$$

接下来考虑随机过程的连续性和可微性。

对于定点$\mathbf{x}_*\in\mathbb{R}^D$，如果有一个序列$\mathbf{x}_1,\mathbf{x}_2,\ldots$满足$\lim_{k\to\infty}|\mathbf{x}_k-\mathbf{x}_*|=0$，则如果满足$\lim_{k\to\infty}\mathbb{E}[|f(\mathbf{x}_k)-f(\mathbf{x}_*)|^2]=0$，随机过程$f(\mathbf{x})$在$\mathbf{x}_*$点均方连续。如果$f(\mathbf{x})$在$\forall\mathbf{x}_*\in A\subseteq\mathbb{R}^D$都连续，则称$f(\mathbf{x})$在区域$A$上均方连续。如果一个随机场的协方差函数$k(\mathbf{x},\mathbf{x}')$在$\mathbf{x}=\mathbf{x}'=\mathbf{x}_*$处连续，则该随机场在该点均方连续；对于平稳协方差函数，只需要检验$k(\mathbf{0})$的随机性即可。均方连续性并不意味着样本函数的连续性。

随机过程的均方偏导数定义为

$$\frac{\partial f(\mathbf{x})}{\partial x_i}=\lim_{h\to0}\frac{f(\mathbf{x}+h\mathbf{e}_i)-f(\mathbf{x})}{h} \tag{47}$$

这里要求极限存在，且极限是均方极限，$\mathbf{e}_i$是该方向上的单位向量。$\frac{\partial f(\mathbf{x})}{\partial x_i}$的协方差函数为$\frac{\partial^2 k(\mathbf{x},\mathbf{x}')}{\partial x_i\partial x_i'}$。

2.2 协方差函数举例

2.2.1 平稳协方差函数

平稳协方差函数是$\boldsymbol{\tau}=\mathbf{x}-\mathbf{x}'$的函数，常写作$k(\boldsymbol{\tau})$。本节中考虑的很多随机过程是复值的，相应的均值为零的过程的协方差函数为$k(\mathbf{x},\mathbf{x}')=\mathbb{E}[f(\mathbf{x})f^*(\mathbf{x}')]$，即需要取复共轭。

定理 2.1 (Bochner定理) 在$\mathbb{R}^D$上，如果对于某种正有限测度μ，当且仅当

$$k(\boldsymbol{\tau}) = \int_{\mathbb{R}^D} \mathrm{d}\mu(\mathbf{s}) \exp(2\pi \mathrm{i}\mathbf{s}\cdot\boldsymbol{\tau}) \tag{48}$$

则复值函数k是宽平稳均方连续复值随机过程的协方差函数

。如果μ具有某种密度$S(\mathbf{s})$则其为k的**谱密度**或**功率谱**。由此可以引出维纳-辛钦定理：宽平稳随机过程的功率谱密度是其自相关函数的傅里叶变换。

$$k(\boldsymbol{\tau}) = \int_{\mathbb{R}^D} \mathrm{d}\mathbf{s}\, S(\mathbf{s}) \exp(2\pi \mathrm{i}\mathbf{s}\cdot\boldsymbol{\tau}), \quad S(\mathbf{s}) = \int_{\mathbb{R}^D} \mathrm{d}\boldsymbol{\tau}\, k(\boldsymbol{\tau}) \exp(-2\pi \mathrm{i}\mathbf{s}\cdot\boldsymbol{\tau}) \tag{49}$$

如果协方差函数是各向同性的（即为$r = |\boldsymbol{\tau}|$的函数），则可以进行极坐标变换，

$$k(r) = \frac{2}{r^{\frac{D}{2}-1}} \int_0^\infty \mathrm{d}s\, S(s) J_{\frac{D}{2}-1}(2\pi r s) s^{\frac{D}{2}} \tag{50}$$

$$S(s) = \frac{2}{s^{\frac{D}{2}-1}} \int_0^\infty \mathrm{d}r\, k(r) J_{\frac{D}{2}-1}(2\pi r s) r^{\frac{D}{2}} \tag{51}$$

其中$J_{D/2-1}$是$D/2-1$阶贝塞尔函数。可以看出，函数显式依赖于维度。

以下是一些常用的各向同性协方差函数，并已归一化满足$k(0) = 1$，实际使用时乘上σ_f^2即可给出所需的协方差函数。

平方指数(SE)协方差函数：

$$k_{\mathrm{SE}}(r) = \exp\left(-\frac{r^2}{2l^2}\right) \tag{52}$$

其中l被称为特征长度尺度。由上文可得一维的SE过程的零级上交次数的期望值为$(2\pi l)^{-1}$。由于该函数无限可微，因此具有该协方差函数的高斯过程具有无穷阶均方导数，并因此非常平滑。其对应的谱密度为$S(s) = (2\pi l^2)^{\frac{D}{2}} \exp(-2\pi^2 l^2 s^2)$。该函数可能是机器学习领域使用最广泛的核函数。由于$[k(r)]^t$对于所有$t > 0$都是一个可用的核函数，故而其被称为无限可分的。

马特恩协方差函数：

$$k_{\mathrm{Matern}}(r) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu} r}{l}\right)^\nu K_\nu \left(\frac{\sqrt{2\nu} r}{l}\right) \tag{53}$$

其中ν和l是参数，Γ是Γ函数，而K_ν是ν阶的第二类修正贝塞尔函数。对应的D维谱密度为

$$S(s) = \frac{2^D \pi^{\frac{D}{2}} \Gamma\left(\nu + \frac{D}{2}\right)(2\nu)^\nu}{\Gamma(\nu) l^{2\nu}} \left(\frac{2\nu}{l^2} + 4\pi^2 s^2\right)^{-\left(\nu+\frac{D}{2}\right)} \tag{54}$$

当$\nu \to \infty$时其退化为SE协方差函数。当且仅当$\nu > k$时，这个随机过程$f(\mathbf{x})$是k阶均方可导的。特别地，如果ν是半整数，则其形式可以写作

$$k_{\nu=p+\frac{1}{2}}(r) = \exp\left(-\frac{\sqrt{2\nu}r}{l}\right)\frac{\Gamma(p+1)}{\Gamma(2p+1)}\sum_{i=0}^{p}\frac{(p+i)!}{i!(p-i)!}\left(\frac{\sqrt{8\nu}r}{l}\right)^{p-i} \tag{55}$$

$\nu = 1/2$的情况过于粗糙，$\nu = 3/2$和$\nu = 5/2$的情况在机器学习中常用

$$k_{\nu=\frac{3}{2}}(r) = \left(1+\frac{\sqrt{3}r}{l}\right)\exp\left(-\frac{\sqrt{3}r}{l}\right) \tag{56}$$

$$k_{\nu=\frac{5}{2}}(r) = \left(1+\frac{\sqrt{5}r}{l}+\frac{5r^2}{3l^2}\right)\exp\left(-\frac{\sqrt{5}r}{l}\right) \tag{57}$$

而对于$\nu \geqslant 7/2$，由于高阶导数存在性较低，很难区别这些不同的ν（甚至区分有限和无穷的ν，即马特恩协方差函数和光滑的平方指数协方差函数）。

奥恩斯坦-乌伦贝克过程：即一维的具有$\nu = \frac{1}{2}$的马特恩协方差函数$k(r) = \exp(-r/l)$的随机过程，它均方连续但均方不可导。这个过程描述了一个有摩擦的布朗粒子。更一般地，对于一维情况，半整数$\nu + 1/2 = p$给出了连续时间自回归(p)高斯过程的一种特殊形式。

γ指数协方差函数：

$$k(r) = \exp\left[-\left(\frac{r}{l}\right)^{\gamma}\right] \tag{58}$$

它包括指数型和SE型。它与马特恩型的参数数目一样，但是更不灵活，这是因为除了$\gamma = 2$（即SE型）外对应的过程都均方不可导。

有理二次(Rational Quadratic, RQ)协方差函数：

$$k_{\mathrm{RQ}}(r) = \left(1+\frac{r^2}{2\alpha l^2}\right)^{-\alpha} \tag{59}$$

当$\alpha, l > 0$时，可以视为具有不同特征长度尺度的SE协方差函数的比例混合（无穷求和）。如果令$\tau = l^{-2}$而假定Γ分布$p(\tau|\alpha,\beta) \propto \tau^{\alpha-1}\exp(-\alpha\tau/\beta)$，则

$$k_{\mathrm{RQ}}(r) = \int \mathrm{d}\tau\, p(\tau|\alpha,\beta)k_{\mathrm{SE}}(r|\tau) \tag{60}$$

$$\propto \int \mathrm{d}\tau\, \tau^{\alpha-1}\exp\left(-\frac{\alpha\tau}{\beta}\right)\exp\left(-\frac{\tau r^2}{2}\right) \propto \left(1+\frac{r^2}{2\alpha l^2}\right)^{-\alpha} \tag{61}$$

其中令$\beta^{-1} = l^2$。当$\alpha \to \infty$时RQ协方差函数化为SE协方差函数，对应的特征长度尺度为l。这种方法可以有效地定义任意维的协方差函数，即对于一个基于特征长度尺度的分

布$p(l)$，

$$k(r) = \int \mathrm{d}l \exp\left(-\frac{r^2}{2l^2}\right) p(l) \tag{62}$$

紧凑形式的分段多项式(Piecewise Polynomial, PP)协方差函数：这里的紧凑形式是指超过一定距离则函数值为零，因此构建出的协方差矩阵较为稀疏，在计算上可能有一定优势，但是其正定性则是问题，经常只能在最大维数D的情况下保证。例如，如果令$j = \lfloor D/2 \rfloor + q + 1$，其中$q$是参数，则有以下例子

$$k_{\mathrm{pp}D,0}(r) = (1-r)_+^j \tag{63}$$

$$k_{\mathrm{pp}D,1}(r) = (1-r)_+^{j+1}[(j+1)r + 1] \tag{64}$$

$$k_{\mathrm{pp}D,2}(r) = \frac{1}{3}(1-r)_+^{j+2}[(j+3)(j+1)r^2 + 3(j+2)r + 3] \tag{65}$$

$$k_{\mathrm{pp}D,3}(r) = \frac{1}{15}(1-r)_+^{j+3}[(j+5)(j+3)(j+1)r^3 + 3(2j^2+12j+15)r^2 + 15(j+3)r + 15] \tag{66}$$

这些协方差函数是$2q$阶连续可微的，因此对应的随机过程是q阶均方可微的。

上述的这些协方差函数单调递减且恒正，但对于协方差函数而言这并不必要，例如满足$\nu \geqslant (D-2)/2$和$\alpha > 0$的平稳协方差函数$k(r) = c(\alpha r)^{-\nu} J_\nu(\alpha r)$，它是阻尼振荡的。

也存在各向异性的平稳协方差函数，只需要定义$r^2(\mathbf{x}, \mathbf{x}') = \mathbf{x}^\top M \mathbf{x}'$，其中$M$是半正定的。如果$M$是对角的，则它用于在不同维度上有不同特征长度尺度的情况；可以使用低秩M来实现从输入空间到低维特征空间的线性降维步骤；更一般地，可以定义

$$M = \Lambda\Lambda^\top + \Psi \tag{67}$$

其中Λ是$D \times k$型矩阵，定义了k个高相关的方向，而Ψ则是正的对角矩阵，表达一般的轴向相关性，因此M具有因子分析的形式，而适当选择k可以在灵活性和所需参数数量之间进行良好地折衷。

此外，还可以从一般的平稳协方差函数出发在周期域上构建平稳协方差函数，如对于一个一般的平稳协方差函数$k(r)$，$k_{\mathbb{T}}(r) = \sum_{m \in \mathbb{Z}} k(r + ml)$就是以$l$为周期的周期函数。

2.2.2 点积协方差函数

一个点积协方差函数的形式为$k(\mathbf{x}, \mathbf{x}') = \sigma_0^2 + \mathbf{x} \cdot \mathbf{x}'$，来源于线性回归。这里，如果$\sigma_0 = 0$则其被称为齐次线性核，否则是非齐次的。当然可以将其拓展为$k(\mathbf{x}, \mathbf{x}') = \sigma_0^2 + \mathbf{x}^\top \Sigma_p \mathbf{x}'$，其中$\Sigma_p$是一般的协方差矩阵，或者$k(\mathbf{x}, \mathbf{x}') = (\sigma_0^2 + \mathbf{x}^\top \Sigma_p \mathbf{x}')^p$，其中$p$是一个正整数。但是

对于多项式我们可以更显式地构建特征空间，这里以齐次为例，推广到非齐次的情况是平凡的

$$k(\mathbf{x},\mathbf{x}')=(\mathbf{x}\cdot\mathbf{x}')^p=\sum_{i=0}^{p}\sum_{d_i=1}^{D}(x_{d_i}x'_{d_i})=\sum_{d_1\dots d_p=1}^{D}(x_{d_1}\cdots x_{d_p})(x'_{d_1}\cdots)x'_{d_p} \tag{68}$$

但是实际上并不需要D^p项，很多是重复的，例如$x_1x_2=x_2x_1$，因此定义D维矢量$\mathbf{m}$，分量m_d为下标d出现的次数，且满足$\sum_{i=1}^{D}m_i=p$。对应于$\mathbf{m}$的特征映射为

$$\phi_{\mathbf{m}}(\mathbf{x})=\sqrt{\frac{p!}{m_1!\cdots m_D!}}x_1^{m_1}\cdots x_D^{m_D} \tag{69}$$

对于二维二次的情况，$\boldsymbol{\phi}(\mathbf{x})=(x_1^2,x_2^2,\sqrt{2}x_1x_2)^\top$。系数来源于简并度，所以可以将协方差函数写作$k(\mathbf{x},\mathbf{x}')=\boldsymbol{\phi}(\mathbf{x})\cdot\boldsymbol{\phi}(\mathbf{x}')$。

对于回归问题而言多项式核并不是一个好选项，因为在$|\mathbf{x}|>1$的情况下先验方差随着$\mathbf{x}$的增长而快速增长。但是在高维分类问题中很有效。

2.2.3 其它非平稳协方差函数

首先考虑一个单隐藏层神经网络模型，该层具有N_H个节点。整体的结果是偏差b与隐藏层结果的线性叠加

$$f(\mathbf{x})=b+\sum_{j=1}^{N_H}v_jh(\mathbf{x};\mathbf{u}_j) \tag{70}$$

这里v_j是隐藏层到输出层的参数，而$\mathbf{u}_j$则是输入层到隐藏层的参数，h是隐藏层的转换函数（例如，$h(x)=\tanh(\mathbf{x}\cdot\mathbf{u})$）。由于一般所用的转换函数有界，因此该分布的所有矩都有界，在$N_H\to\infty$的情况下收敛于高斯过程，于是可以得到协方差函数$\mathbb{E}[h(\mathbf{x};\mathbf{u})h(\mathbf{x}';\mathbf{u})]$。例如如果令$h(\mathbf{x};\mathbf{u})=\mathrm{erf}(u_0+\sum_{j=0}^{D}x_ju_j)$，且$\mathbf{u}\sim\mathcal{N}(\mathbf{0},\Sigma)$，则有

$$k_{\mathrm{NN}}(\mathbf{x},\mathbf{x}')=\frac{2}{\pi}\sin^{-1}\left(\frac{2\tilde{\mathbf{x}}^\top\Sigma\tilde{\mathbf{x}}'}{\sqrt{(1+2\tilde{\mathbf{x}}^\top\Sigma\tilde{\mathbf{x}})(1+2\tilde{\mathbf{x}}'^\top\Sigma\tilde{\mathbf{x}}')}}\right) \tag{71}$$

其中$\tilde{\mathbf{x}}=(1,\mathbf{x})$。而简单的$k(\mathbf{x},\mathbf{x}')=\tanh(a+b\mathbf{x}\cdot\mathbf{x}')$则不正定。这就是一个神经网络协方差函数

再比如$h(\mathbf{x};\mathbf{u})=\exp\left(-|\mathbf{x}-\mathbf{u}|^2/2\sigma_g^2\right)$，且$\mathbf{u}\sim\mathcal{N}(\mathbf{0},\sigma_u^2I)$，则对应的神经网络结果为

$$k_{\mathrm{G}}(\mathbf{x},\mathbf{x}')=\left(\frac{\sigma_e}{\sigma_u}\right)^2\exp\left(-\frac{\mathbf{x}^\top\mathbf{x}+\mathbf{x}'^\top\mathbf{x}'}{2\sigma_m^2}-\frac{|\mathbf{x}-\mathbf{x}'|^2}{2\sigma_s^2}\right) \tag{72}$$

其中$1/\sigma_e^2 = 2/\sigma_g^2 + 1/\sigma_u^2$，$\sigma_m^2 = 2\sigma_u^2 + \sigma_g^2$，而$\sigma_s^2 = 2\sigma_g^2 + \sigma_g^4/sigma_u^2$。这是一个非平稳协方差函数，在$\sigma_u^2$有限的时候它是由高斯函数调制的SE协方差，但如果$\sigma_u^2 \to \infty$则其化为SE协方差函数。

另一些方法包括引入一个任意的非线性函数$\mathbf{u}(\mathbf{x})$，然后使用一个$\mathbf{u}$空间的平稳协方差函数，其中包括将一个一维变量x映射到二维$\mathbf{u}(x) = (\cos(x), \sin(x))$产生周期性。如果使用SE协方差函数，则有

$$k(x, x') = \exp\left(-\frac{2\sin^2\left(\frac{x-x'}{2}\right)}{l^2}\right) \tag{73}$$

另一类协方差函数的特征长度尺度是$\mathbf{x}$的函数。例如，如果所有的$l_i(\mathbf{x}) > 0$，

$$k(\mathbf{x}, \mathbf{x}') = \sqrt{\prod_{i=0}^{D} \frac{2l_i(\mathbf{x})l_i(\mathbf{x}')}{l_i^2(\mathbf{x}) + l_i^2(\mathbf{x}')}} \exp\left(-\sum_{i=0}^{D} \frac{(x_i - x_i')^2}{l_i^2(\mathbf{x}) + l_i^2(\mathbf{x}')}\right) \tag{74}$$

它满足$k(\mathbf{x}, \mathbf{x}) = 1$。在此基础上进行推广，如果令$k_{\mathrm{S}}$是任意维欧几里得空间上的各向同性的平稳协方差函数，而$\Sigma(\mathbf{x})$是正定矩阵函数，令$\Sigma(\mathbf{x}_i) = \Sigma_i$，有非平稳协方差函数

$$k_{\mathrm{NS}}(\mathbf{x}_i, \mathbf{x}_j) = 2^{\frac{D}{2}} |\Sigma_i|^{\frac{1}{4}} |\Sigma_j|^{\frac{1}{4}} |\Sigma_i + \Sigma_j|^{-\frac{1}{2}} k_{\mathrm{S}}(\sqrt{Q_{ij}}) \tag{75}$$

其中

$$Q_{ij} = (\mathbf{x}_i - \mathbf{x}_j)^\top \left(\frac{\Sigma_i + \Sigma_j}{2}\right)^{-1} (\mathbf{x}_i - \mathbf{x}_j) \tag{76}$$

此外还有维纳过程，$k(x, x') = \min(x, x')$，也是非平稳过程。

2.2.4 从旧函数创建新函数

两个核函数的和与积都是核函数，这里的和与积包括不同空间的直和与张量积。

也可以用卷积构建。例如对于映射$g(\mathbf{x}) = \int \mathrm{d}\mathbf{z}\, h(\mathbf{x}, \mathbf{z}) f(\mathbf{z})$，其中$h(\mathbf{x}, \mathbf{z})$是任意的固定内核，则$\mathrm{cov}(g(\mathbf{x}), g(\mathbf{x}')) = \int \mathrm{d}\mathbf{z}\, \mathrm{d}\mathbf{z}'\, h(\mathbf{x}, \mathbf{z}) k(\mathbf{z}, \mathbf{z}') h(\mathbf{x}', \mathbf{z}')$也是协方差函数。

还有核函数的归一化，即

$$\tilde{k}(\mathbf{x}, \mathbf{x}') = \frac{k(\mathbf{x}, \mathbf{x}')}{\sqrt{k(\mathbf{x}, \mathbf{x}) k(\mathbf{x}', \mathbf{x}')}} \tag{77}$$

这可以保证$k(\mathbf{x}, \mathbf{x}) = 1$。

2.3 核函数的本征函数分析

高斯过程可以看成是有无限基函数的贝叶斯线性回归。其中一种基函数是核函数的本征函数，即满足

$$\int \mathrm{d}\mu(\mathbf{x})\, k(\mathbf{x},\mathbf{x}')\phi(\mathbf{x}) = \lambda\phi(\mathbf{x}') \tag{78}$$

的函数ϕ是核k在μ测度下具有本征值λ的本征函数。常见测度为$\mathbb{R}^D$的紧子集$\mathcal{C}$上的勒贝格测度，或者具有某种密度（即$\mathrm{d}\mu(\mathbf{x}) = p(\mathbf{x})\,\mathrm{d}\mathbf{x}$）。本征函数的下标根据本征值大小排序，且本征函数正交归一，一般数目是无限的。

定理 2.2 (Mercer定理) 如果$(\mathcal{X},\mu)$是有限测度空间，核函数$k \in L_\infty(\mathcal{X}^2,\mu^2)$满足映射$T_k : L_2(\mathcal{X},\mu) \to L_2(\mathcal{X},\mu)$正定（即对任意$f \in L_2(\mathcal{X},\mu)$都有$\int \mathrm{d}\mu(\mathbf{x})\,\mathrm{d}\mu(\mathbf{x}')\, f(\mathbf{x})k(\mathbf{x},\mathbf{x}')f(\mathbf{x}') \geqslant 0$。如果$\phi_i \in L_2(\mathcal{X},\mu)$是$T_k$归一化的本征值为$\lambda_i > 0$的本征函数，则

1) $\{\lambda_i\}_{i=1}^{\infty}$数项级数绝对收敛

2)

$$k(\mathbf{x},\mathbf{x}') = \sum_{i=0}^{\infty} \lambda_i \phi_i^*(\mathbf{x})\phi_i(\mathbf{x}') \tag{79}$$

对μ^2几乎处处成立，其中该求和序列对μ^2几乎处处绝对且一致收敛。

这和厄米矩阵对角化是类似的。

定义 2.1 简并核函数是只有有限个非零本征值的核函数。

简并核函数的秩有限。例如，特征空间中的一个N维线性回归模型的核函数是简并的，最多只有N个非零特征值。

Mercer定理是对有限测度μ来说的。如果换成勒贝格测度，考虑任一平稳协方差函数的Bochner定理，

$$k(\mathbf{x}-\mathbf{x}') = \int_{\mathbb{R}^D} \mathrm{d}\mu(\mathbf{s}) \exp(2\pi \mathrm{i}\mathbf{s}\cdot(\mathbf{x}-\mathbf{x}')) = \int_{\mathbb{R}^D} \mathrm{d}\mu(\mathbf{s}) \exp(2\pi\mathrm{i}\mathbf{s}\cdot\mathbf{x})[\exp(2\pi\mathrm{i}\mathbf{s}\cdot\mathbf{x}')]^* \tag{80}$$

其中复指数$\exp(2\pi\mathrm{i}\mathbf{s}\cdot\mathbf{x})$是对于勒贝格测度的平稳核函数的本征函数。上式和Mercer定理是类似的，只不过求和换成了积分。

本征值的衰减速率提供了核函数平滑性的重要信息。例如，对于一维在$[0,1]$上均匀的测度μ，r次均方可导的过程渐进地有$\lambda_i \propto i^{-(2r+2)}$——更粗糙的过程在高频成分更多，因此衰减更慢。马特恩协方差函数的功率谱也有类似的现象。

2.3.1 一个解析的例子

如果密度函数是高斯$p(x) \sim \mathcal{N}(x|0, \sigma^2)$，核函数是SE，本征函数和本征值有解析解：

$$\lambda_k = \sqrt{\frac{2a}{A}} B^k \tag{81}$$

$$\phi_k(x) = \exp\bigl(-(c-a)x^2\bigr) H_k(\sqrt{2c}x) \tag{82}$$

其中H_k是k阶厄米多项式。式中的常数为

$$a^{-1} = 4\sigma^2, \quad b^{-1} = 2l^2, \quad c = \sqrt{a^2 + 2ab}, \quad A = a + b + c, \quad B = \frac{b}{A} \tag{83}$$

对多维的情况，如果σ^2和l^2对所有维度一致，则本征值$(2a/A)^{D/2}B^k$的简并度为$\binom{k+D-1}{D-1} = \mathcal{O}(k^{D-1})$。考虑到$\sum_{j=0}^{k}\binom{j+D-1}{D-1} = \binom{k+D}{D}$，第$\binom{k+D}{D}$个本征值为$(2a/A)^{D/2}B^k$，这可以体现本征值的衰减速率。

2.3.2 本征函数的数值近似

如果$\mathrm{d}\mu(\mathbf{x}) = p(\mathbf{x})\,\mathrm{d}x$，则可以做如下近似：

$$\lambda_i \phi_i(\mathbf{x}') = \int \mathrm{d}\mathbf{x}\, p(\mathbf{x}) k(\mathbf{x}, \mathbf{x}') \phi_i(\mathbf{x}) \approx \frac{1}{n} \sum_{l=1}^{n} k(\mathbf{x}_l, \mathbf{x}') \phi_i(\mathbf{x}_l) \tag{84}$$

其中$\mathbf{x}_l$是从$p(\mathbf{x})$采样得到。联立$\mathbf{x}' = \mathbf{x}_l$可得矩阵本征值问题$K\mathbf{u}_i = \lambda_i^{\mathrm{mat}}\mathbf{u}_i$，其中$K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$，$\lambda_i^{\mathrm{mat}}$是矩阵的第$i$个本征值。考虑归一化，$\phi_i(\mathbf{x}_j) \sim \sqrt{n}(\mathbf{u}_i)_j$，于是$\lambda_i \approx \lambda_i^{\mathrm{mat}}/n$，且当$n \to \infty$时收敛。利用主成分分析，在特征空间中对任意$l \in [1, n]$，$\mathbb{E}_n[\frac{1}{n}\sum_{i=1}^{l}\lambda_i^{\mathrm{mat}}] \geqslant \sum_{i=1}^{l}\lambda_i$，且$\mathbb{E}_n[\frac{1}{n}\sum_{i=l+1}^{n}\lambda_i^{\mathrm{mat}}] \leqslant \sum_{i=l+1}^{n}\lambda_i$，其中$\mathbb{E}_n$是从$p(\mathbf{x})$中采样$n$个点的期望值。

利用Nyström方法，

$$\phi_i(\mathbf{x}') \approx \frac{\sqrt{n}}{\lambda_i^{\mathrm{mat}}} \mathbf{k}(\mathbf{x}')^\top \mathbf{u}_i \tag{85}$$

其中$vbk(\mathbf{x}') = (k(\mathbf{x}_1, \mathbf{x}'), \ldots, k(\mathbf{x}_n, \mathbf{x}'))$。

2.4 非矢量输入型核函数

常见的非矢量输入包括字符串（如DNA序列），树（语言学分析）和图（化学分子）。我们需要从输入对象中提取一些特征，并根据这些特征构建预测变量。

2.4.1 字符串核函数

如果$\mathcal{A}$是有限的字符组成的字母表，xy是字符串x和y的连接，$|x|$是x的长度，如果$x = usv$则s是x的子串。如果$\phi_s(x)$是子串s在x中出现的次数，则两个字符串之间的核函数可以定义为

$$k(x, x') = \sum_{s \in \mathcal{A}^*} w_s \phi_s(x) \phi_s(x') \tag{86}$$

其中w_s是某种非负权重，可以有以下选项：

- $w_s = \lambda^{|s|}$且$\lambda \in [0, 1]$，这样可以使较短的字符串权重较大。
- 对所有$|s| > 1$有$w_s = 0$，即字符袋核函数，特征矢量是每个$\mathcal{A}$中字符在字符串中出现的次数。
- 在文本分析中，我们不妨考虑单词出现的频率。如果我们要求s以空格为边界，则将获得“单词袋”表示，它对于文档分类和检索任务却可能出奇地有效。也可以为不同的单词设置不同的权重，使用在信息检索区域开发的“词频-逆向文件频率”(Term Frequency - Inversed Document Frequency, TF-IDF)加权方案。
- 如果只考虑长度为k的子串，那可以获得k谱核函数。

有些使用后缀树的方法计算的核函数与$|x| + |x'|$成正比，但需要对权重做出一定的限制。上述方法也可以进一步拓展，如将子串拓展到不一定连续的子序列，或者最多有m处不匹配的x和x'子串s和s'进行匹配的核函数。字符串内核的思想可以轻松地扩展到树，例如通过查看子树的匹配。

2.4.2 费歇尔核函数

费歇尔核函数可以通过使用一个生成模型$p(x|\boldsymbol{\theta})$并计算特征向量（有时也称之为分数向量）$\phi_{\boldsymbol{\theta}}(\mathbf{x}) = \nabla_{\boldsymbol{\theta}} \log p(x|\boldsymbol{\theta})$来提取输入是任意长度的结构化对象（如字符串）的特征。例如对于字符串的马尔可夫模型，记x_k为字符串x中第k位的字符，$\boldsymbol{\theta} = (\boldsymbol{\pi}, A)$，$(\boldsymbol{\pi})_j$为$x_1$是字母表$\mathcal{A}$中第$j$个字符的概率，而$A$是$|\mathcal{A}| \times |\mathcal{A}|$的随机矩阵且$a_{jk} = p(x_{i+1} = k|x_i = j)$，则$p(x|\boldsymbol{\theta}) = p(x_1|\boldsymbol{\pi}) \prod_{i=1}^{|x|-1} p(x_{i+1}|x_i, A)$，如此即可计算给定$x$的分数向量。其它的生成模型包括由前$k$个符号预测$x_i$的$k$阶马尔可夫模型，其中$k-1$阶模型的分数向量与$k$谱核函数的特征有相似之处。也可以选择隐马尔可夫模型作为生成模型。从各向同性的高斯函数出发也可以构建线性核函数。

基于分数向量的核函数可以为$k(x,x') = \phi_{\boldsymbol{\theta}}^{\top}(x)M^{-1}\phi_{\boldsymbol{\theta}}(x')$，其中$M$严格正定。信息几何学中广泛研究了$p(x|\boldsymbol{\theta})$对$\boldsymbol{\theta}$的依赖，并证明了$\log p(x|\boldsymbol{\theta})$是黎曼流形，其度规张量是费歇尔信息矩阵$F$的逆，

$$F = \mathbb{E}_x[\phi_{\boldsymbol{\theta}(x)}\phi_{\boldsymbol{\theta}}^{\top}(x)] \tag{87}$$

如果令$M = F$则得到的就是费歇尔核函数，如果F很难计算则可以令$M = I$。使用费歇尔信息矩阵的优点是，它使流形上的弧长对于$\boldsymbol{\theta}$的重新参数化而言不变。也可以是SE型核函数如$k(x,x') = \exp\left(-\alpha|\phi_{\boldsymbol{\theta}}(x) - \phi_{\boldsymbol{\theta}}(x')|^2\right)$，其中$\alpha > 0$。

3 模型选择与超参数适应

3.1 模型选择问题

模型选择涵盖离散选择和协方差函数的连续（超）参数设置。实际上，模型选择既可以帮助完善模型的预测，又可以为用户提供有关数据属性的有价值的解释，例如非平稳协方差函数可能优于平稳协方差函数。协方差函数及其参数的选择称为高斯过程的训练。例如对于SE核函数，

$$k(\mathbf{x}_p, \mathbf{x}_q) = \sigma_f^2 \exp\left(-\frac{1}{2}(\mathbf{x}_p - \mathbf{x}_q)^{\top} M (\mathbf{x}_p - \mathbf{x}_q)\right) + \sigma_n^2 \delta_{pq} \tag{88}$$

包含所有超参数的矢量为$\boldsymbol{\theta} = (\{M\}, \sigma_f^2, \sigma_n^2)^{\top}$，其中$\{M\}$是对称矩阵$M$的所有参数，一般有以下选项

$$M_1 = l^{-2}I, \quad M_2 = \mathrm{diag}(\boldsymbol{l})^{-2}, \quad M_3 = \Lambda\Lambda^{\top} + \mathrm{diag}(\boldsymbol{l})^{-2} \tag{89}$$

其中$\boldsymbol{l}$是正值矢量，而Λ是$D \times k (k < D)$矩阵。对于许多协方差函数，解释超参数的含义在尝试理解数据时非常重要。例如M_2中的$l_1, \ldots, l_D$是对应维度的特征长度尺度，大致来说是需要在输入空间中移动（沿特定轴）多远才能使函数值不相关。这样的协方差函数实现了自动相关性确定(Auto Relevance Determination, ARD)，因为长度尺度的倒数确定了输入的相关性：如果长度尺度具有非常大的值，则协方差将几乎独立于该输入，有效地将其从推理中删除。M_3的参数化称为因子分析距离，这是由于与（无监督）因子分析模型的类比，该模型试图通过低秩加对角线分解来解释数据。对于高维数据集，Γ的k列可以标识输入空间中具有特别高“相关性”的几个方向，并且它们的长度给出了这些方向的特征长度尺度的倒数。即使在一族协方差函数中也有很大的变化范围，而我们需要基于一组训练数据来推断协方差函数的形式和参数，或者等效地推断数据的关系。

模型选择本质上是开放式的。即使对于平方指数协方差函数，也存在多种可能的距离度量，这代表了学习的可能性，但这需要一种系统实用的模型选择方法。简而言之，我们需要能够比较两种（或多种）方法不同，这些方法在特定参数的值或协方差函数的形式方面有所不同，或者将高斯过程模型与任何其他类型的模型进行比较。对此有三个通用原则：（1）给定数据后计算模型的概率；（2）估计泛化误差；（3）约束泛化误差。泛化误差表示与训练集的分布相同的未见测试示例的平均误差。注意到训练误差通常不能很好地替代泛化误差，因为模型可能拟合训练集中的噪声（过度拟合），导致训练误差较小，但泛化性能较差。

3.2 贝叶斯模型选择

常使用层次模型。底层是参数$\mathbf{w}$，如线性模型的参数或者神经网络模型中的权重。第二层是控制底层参数分布的超参数$\boldsymbol{\theta}$，如神经网络中的“权重衰减”项或岭回归中的“岭”项就是超参数。顶层是可能模型结构的一个（离散）集合$\mathcal{H}_i$。对于贝叶斯推断，一次发生在一个层次上。在底层，贝叶斯定理给出了对参数的后验分布

$$p(\mathbf{w}|\mathbf{y}, X, \boldsymbol{\theta}, \mathcal{H}_i) = \frac{p(\mathbf{y}|X, \mathbf{w}, \mathcal{H}_i)p(\mathbf{w}|\boldsymbol{\theta}, \mathcal{H}_i)}{p(\mathbf{y}|X, \boldsymbol{\theta}, \mathcal{H}_i)} \tag{90}$$

其中$p(\mathbf{y}|X, \mathbf{w}, \mathcal{H}_i)$是似然函数，而$p(\mathbf{w}|\boldsymbol{\theta}, \mathcal{H}_i)$是参数先验分布，是在已知数据之前对参数的信息编码而得概率分布。如果仅对参数具有模糊的先验信息，则用先验分布反映这一点。后验通过似然结合了先验信息和数据中的信息。分母

$$p(\mathbf{y}|X, \boldsymbol{\theta}, \mathcal{H}_i) = \int \mathrm{d}\mathbf{w}\, p(\mathbf{y}|X, \mathbf{w}, \mathcal{H}_i)p(\mathbf{w}|\boldsymbol{\theta}, \mathcal{H}_i) \tag{91}$$

是归一化常数，不依赖于参数，称为边缘似然函数。下一层可以给出类似的超参数后验分布

$$p(\boldsymbol{\theta}|\mathbf{y}, X, \mathcal{H}_i) = \frac{p(\mathbf{y}|X, \boldsymbol{\theta}, \mathcal{H}_i)p(\boldsymbol{\theta}|\mathcal{H}_i)}{p(\mathbf{y}|X, \mathcal{H}_i)} \tag{92}$$

其中$p(\boldsymbol{\theta}|\mathcal{H}_i)$是超先验分布（超参数的先验分布）。归一化常数为

$$p(\mathbf{y}|X, \mathcal{H}_i) = \int \mathrm{d}\boldsymbol{\theta}\, p(\mathbf{y}|X, \boldsymbol{\theta}, \mathcal{H}_i)p(\boldsymbol{\theta}|\mathcal{H}_i) \tag{93}$$

顶层是模型的后验分布

$$p(\mathcal{H}_i|\mathbf{y}, X) = \frac{p(\mathbf{y}|X, \mathcal{H}_i)p(\mathcal{H}_i)}{p(\mathbf{y}|X)} \tag{94}$$

其中$p(\mathbf{y}|X) = \sum_i p(\mathbf{y}|X, \mathcal{H}_i)p(\mathcal{H}_i)$。贝叶斯推断要求计算数个积分。根据模型的信息详细度，这些积分可能不易解析计算，通常，人们可能不得不求助于分析近似解或马尔可夫链

蒙特卡洛(Markov Chain Monte Carlo, MCMC)方法。在实践中尤其难计算超参数的归一化积分；作为一种近似，可能回避使用超参数后验而使用能最大化参数的边际似然的超参数，这种近似称为II型最大似然(Type II Maximum Likelihood, ML-II)。当然，在进行这种优化步骤　　时应格外小心，因为尤其是在有许多超参数的情况下可能过拟合。然后，可以使用最大值附近的局部展开来近似后验边缘似然的积分（拉普拉斯方法）。如果θ的后验分布比较尖锐则这种近似将是好的近似，而这种情况在超参数上比参数更常见。对模型的先验分布通常被认为是平坦的，因此先验分布中各模型权重大致一致，则模型的后验概率与超参数的边缘似然成正比。参数的边际似然是将贝叶斯推断方案与基于优化的其他方案区分开的主要原因。边际似然函数的特性是其会自动纳入模型拟合和模型复杂性之间的权衡。这就是为什么边际似然对解决模型选择问题有价值。

对于超参数的边际似然$p(\mathbf{y}|X, \mathcal{H}_i)$与可能的数据集$\mathbf{y}$之间的关系，简单模型（如线性回归）只能说明目标值可能集的有限范围故而分布更尖锐，而复杂模型（如大型神经网络）可以处理更广泛的数据集故而分布更平坦，且由于它们是概率分布因此有必要归一化，于是对于特定的一组数据集$\mathbf{y}_0$，某个中等复杂度的模型可能具有更大的边际似然而更值得采用。边际似然不只是偏向于最适合训练数据的模型，即奥卡姆剃刀原理，“如无必要，勿增实体”。数据拟合和模型复杂度之间的权衡是自动的，无需在外部设置参数即可解决此问题。不要将自动奥卡姆剃刀原理与贝叶斯方法中使用先验相混淆。即使先验条件在复杂性方面是“平坦”的，边际似然仍将倾向于支持能够解释数据的最简单模型。因此，可以使用边际似然来选择非常适合数据的模型复杂度。

模型的规格为模型结构以及参数的先验分布等。如果不清楚如何设置参数的先验分布则可将其视为超参数，并进行模型选择以确定如何设置它们。同时应强调的是，先验分布对应于关于数据的（概率）假设。如果先验与数据的分布大不相同，则推断仍将在先验分布蕴涵的假设下进行。为避免这种情况，应该小心不要使用太窄的先验分布，从而排除了对数据的合理解释。

3.3　交叉验证

用交叉验证(Cross-Validation, CV)进行模型选择的基本思路是将训练集划分为两个不相交集，其中一个作为训练集而另一个**验证集**用于检查模型的表现。在验证集上的表现可以用于代替泛化误差，并以此度量决定模型选择。实际上这样做的一个缺点是只用了训练集的一部分进行训练，且如果验证集过小则对模型表现的估计可能会有较大的波动。为此，一般都会进行k折交叉验证：将训练集等分为k份（k一般取3到10），每次用其中一份验证而剩下$k-1$份训练，如此重复k次。这样做的好处在于大部分数据都用于训练且所有

数据都经过验证，坏处是需要训练k次。极端的情况是$k=n$，被称为留一验证(Leave-One-Out Cross-Validation, LOO-CV)，一般来说计算量过大，但是在特定的情况（如GPR）下能简便计算。可以使用任何损失函数进行交叉验证，常用平方误差为回归损失函数，但对于诸如高斯过程之类的概率模型也会考虑使用负对数概率损失进行交叉验证。

3.4 高斯回归的模型选择

3.4.1 边缘似然函数

贝叶斯定理为贝叶斯推断提供了有说服力和一致的框架。不幸的是，对于大多数有趣的机器学习模型，所需的计算（在参数空间上的积分）在分析上是难以解决的，并且很难轻易得出良好的近似值。具有高斯噪声的高斯过程回归模型是一个罕见的例外：参数上的积分在分析上易于处理，同时模型非常灵活。高斯过程模型是非参数模型，模型的参数并不显而易见。通常将训练输入$\mathbf{f}$处的无噪声潜函数值视为参数，于是训练集越大则参数就越多。使用权重空间视图则可以等效地将参数视为使用基函数ϕ的线性模型的权重，该基函数可以选择为协方差函数的本征函数。当然这对于非简并协方差函数而言并不方便，因为它们具有无穷多的权重。

对于模型第一层（参数）的贝叶斯推断，回顾前文，有预测性分布

$$p(\mathbf{f}_*|X_*,X,\mathbf{y}) \sim \mathcal{N}(K(X_*,X)K_y^{-1}\mathbf{y}, \\ K(X_*,X_*)-K(X_*,X)K_y^{-1}K(X,X_*)) \tag{95}$$

以及边缘似然

$$p(\mathbf{y}|X,\boldsymbol{\theta}) \sim \mathcal{N}(0,K_y) \tag{96}$$

$$\Rightarrow \log p(\mathbf{y}|X,\boldsymbol{\theta}) = -\frac{1}{2}\mathbf{y}^\top K_y^{-1}\mathbf{y} - \frac{1}{2}\log|K(X,X)| + \sigma_n^2 I - \frac{n}{2}\log 2\pi \tag{97}$$

其中$K_y = K(X,X)+\sigma_n^2 I$，取决于输入数据是否有噪声。这里的边缘似然是针对潜函数的边缘。如果从函数空间的角度，“边缘”和超参数的“超”都显得无法理解。上式中的三项，第一项是数据拟合，第二项是复杂度惩罚，第三项是归一化系数。随着长度尺度的增长，模型适应性降低导致数据拟合项单调递减而负的复杂度惩罚项单调递增，二者合力导致边缘分布的最大值在1附近，超过后模型适应性降低而使得边缘似然迅速递减，不足时则由于可以容纳数据但在数据点之外的地方误差线则会迅速增长而导致边缘似然函数的缓慢递减。此外，随着训练集的增长，特征长度尺度也会越来越明显，即边缘似然在特征长度尺度位置有一个显著的峰值。

为了求得使边缘似然最大的超参数，需要对其求偏导，

$$\frac{\partial}{\partial\theta_j}\log p(\mathbf{y}|X,\boldsymbol{\theta}) = \frac{1}{2}\mathbf{y}^\top K_y^{-1}\frac{\partial K_y}{\partial\theta_j}K_y^{-1}\mathbf{y} - \frac{1}{2}\operatorname{tr}\left(K_y^{-1}\frac{\partial K_y}{\partial\theta_j}\right) \tag{98}$$

$$= \frac{1}{2}\operatorname{tr}\left[(\boldsymbol{\alpha}\boldsymbol{\alpha}^T - K_y^{-1})\frac{\partial K_y}{\partial\theta_j}\right] \tag{99}$$

其中$\boldsymbol{\alpha} = K_y^{-1}\mathbf{y}$。这里的计算复杂度主要来源于求正定矩阵$K_y$的逆（求行列式的对数可以作为副产物产生），一般需要$\mathcal{O}(n^3)$的时间，之后对每个超参数则需要$\mathcal{O}(n^2)$的时间。

空间统计学中已长期研究通过优化边缘似然来估计超参数$\boldsymbol{\theta}$。一般认为随着训练集增大超参数估计也会更精确，但有必要对比定域渐近（在某个区域内观察越来越密集）与增长渐近（观察区域的大小随n增大）。在时间序列一般使用增长渐进，而在空间（和机器学习）设置中则常用定域渐近。

显然，当超参数与产生数据所用的参数一致时对数边缘似然取到极大值；在方差为真实方差（产生数据所用的方差）的情况下，当长度尺度达到一定长度后边缘似然基本与特征长度尺度无关，这是由于该模型将所有内容解释为噪声而无需信号协方差引起的。反之若长度尺度为真实值则当噪声方差小到一定程度时边缘似然也与噪声方差无关，这是由于模型能够在此较短的长度范围内精确内插数据而造成的。注意到尽管此超参数区域（即噪声方差小或长度尺度长的区域）中的模型也能准确解释所有数据点，但是根据边际似然判断它仍然是不好的。

无法保证边缘似然不会有多个局部最优。简单协方差函数的实践经验似乎表明，局部极值值不是破坏性的问题，但确实存在。实际上，每个局部最大值对应于数据的特定解释，如一个最优值对应于一个噪声较低的相对复杂的模型，而另一个最优值对应于一个具有更多噪声的简单得多的模型。如果数据点过少则模型就无法确定是这两种中的哪一种。对于较复杂的模型，边际似然的数值比简单模型高约60%。根据贝叶斯形式体系，一个人应该根据从后验概率得到的替代解释来权衡预测。在实践中，对于更大的数据集，人们经常发现一个局部最优比其他局部最优的概率高几个数量级，因此对所有解释做平均可能没必要，但注意不要以糟糕的局部最优而告终。此外，可能会有几个共享相同超参数的数据集，即多任务学习，此时可以简单地将各个问题的对数边际似然相加来进行超参数优化。

3.4.2 交叉验证

留出第i个训练样例的对数预测概率分布为

$$\log p(y_i|X,\mathbf{y}_{-i},\boldsymbol{\theta}) = -\frac{1}{2}\log\sigma_i^2 - \frac{(y_i-\mu_i)^2}{2\sigma_i^2} - \frac{1}{2}\log 2\pi \tag{100}$$

其中

$$\mu_i = K(\mathbf{x}_i, X_{-i})[K(X_{-i}, X_{-i}) + \sigma_n^2 I]^{-1}\mathbf{y}_{-i} \tag{101}$$

$$\sigma_i = K(\mathbf{x}_i, \mathbf{x}_i) - K(\mathbf{x}_i, X_{-i})[K(X_{-i}, X_{-i}) + \sigma_n^2 I]^{-1}K(X_{-i}, \mathbf{x}_i) \tag{102}$$

与上文是类似的。于是，留一验证预测概率的对数（有时也称为伪似然）为

$$L_{\text{LOO}} = \sum_{i=1}^{n} \log p(y_i|X, \mathbf{y}_{-i}, \boldsymbol{\theta}) \tag{103}$$

注意到在每轮LOO-CV过程中，固定超参数的高斯过程模型的贝叶斯推断实际上就是计算协方差矩阵的逆来求均值和方差（即不存在参数模型中的参数拟合）。由于计算过程非常相似（从完整的协方差矩阵中移除一行一列），这一计算可以通过分块矩阵求逆来完成。对于样条模型也是类似的。预测的均值和方差为

$$\mu_i = y_i - \frac{[K^{-1}\mathbf{y}]_i}{[K^{-1}]_{ii}}, \quad \sigma_i^2 = \frac{1}{[K^{-1}]_{ii}} \tag{104}$$

总计算量对于求K的逆是$\mathcal{O}(n^3)$，对于后续的LOO-CV是$\mathcal{O}(n^2)$，所以LOO-CV的计算量可以忽略。于是，我们可以根据伪似然对超参数的偏导进行优化来做模型选择。对于均值和方差，其对于超参数的导数为

$$\frac{\partial \mu_i}{\partial \theta_j} = \frac{[Z_j\boldsymbol{\alpha}]_i}{[K^{-1}]_{ii}} - \frac{\alpha_i[Z_jK^{-1}]_{ii}}{[K^{-1}]_{ii}^2}, \quad \frac{\partial \sigma_i^2}{\partial \theta_j} = \frac{[Z_jK^{-1}]_{ii}}{[K^{-1}]_{ii}^2} \tag{105}$$

其中$Z_j = K^{-1}\frac{\partial K}{\partial \theta_j}$而$\boldsymbol{\alpha} = K^{-1}\mathbf{y}$。于是有

$$\frac{\partial L_{\text{LOO}}}{\partial \theta_j} = \sum_{i=1}^{n}\left(\frac{\partial \log p(y_i|X, \mathbf{y}_{-i}, \boldsymbol{\theta})}{\partial \mu_i}\frac{\partial \mu_i}{\partial \theta_j} + \frac{\partial \log p(y_i|X, \mathbf{y}_{-i}, \boldsymbol{\theta})}{\partial \sigma_i^2}\frac{\partial \sigma_i^2}{\partial \theta_j}\right) \tag{106}$$

$$= \sum_{i=1}^{n}\frac{1}{[K^{-1}]_{ii}}\left[\alpha_i[Z_j\boldsymbol{\alpha}]_i - \left(1 + \frac{\alpha_i^2}{[K^{-1}]_{ii}}\right)\frac{[Z_jK^{-1}]_{ii}}{2}\right] \tag{107}$$

对于每个超参数的计算量都为$\mathcal{O}(n^3)$，负担比对边缘似然的类似计算更重。

这里将负对数验证密度作为损失函数，但还可以设想使用其他损失函数，如常用的平方误差损失函数。但是，此损失函数只是预测平均值的函数，而忽略了验证集方差。此外，由于均值预测与协方差的大小无关（即，将信号和噪声的协方差乘以任意正常数不会改变均值预测结果），因此基于平方误差（或仅取决于均值预测的任何其他）损失函数的LOO-CV会导致有一个自由度是不确定的，而显然完整的预测分布确实取决于协方差函数的尺度。此外，基于平方误差损失的导数计算具有与负对数验证密度损失相似的计算复杂度。总之，将基于平方误差损失的LOO-CV用于超参数选择似乎没有吸引力。

将LOO-CV方法的伪似然与前文的的边际似然进行比较，他们的计算量基本一致，于是在什么情况下某种方法可能更可取是一个有趣的问题。这个问题还没有进行太多的经验研究。但是要注意边际似然在给定模型假设的情况下给出了观测概率，而LOO-CV则不管是否有任何假设的条件下给出了预测概率的对数的估计，所以交叉验证在对抗模型的推测失误上可能是更鲁棒的。