基于高斯过程的回归分析

目录

摘要

本次项目主要完成了基础部分四个模块以及提高部分两个模块。

基础部分包含标准线性回归模型的非概率最小二乘求解，标准线性回归模型的贝叶斯回归模型求解，引入基函数构成广义线性回归模型并分别依据最小二乘和贝叶斯回归模型求解，以平方指数函数为核函数完成基于高斯回归过程的分析。在此部分重点探讨了核函数超参数的处理以及超参数初值的选择问题。其中超参数的处理使用了 RW 书中指出的最大边际似然法，并结合梯度下降法进行迭代收敛。再观察多组超参数收敛规律的同时，我也探讨了超参数迭代之间的联系与规律。

提高部分包含对 David Duvenaud [1]论文中所提到的 GPML 库函数的探索以及对所提供数据进行筛选去噪的探索。其中 GPML 库函数提供了多种核函数，因此在此部分重点探索了核函数的选择问题。对于数据重组问题，我采用了对比数据之间的相关性，选择相关性最高的四组数据进行分析。

相关研究

Carl Edward Rasmussen 与 Christopher K. I. Williams[2]在机器学习的高斯过程一书中给出了高斯过程的原理以及高斯过程的分析方法，本次项目主要依赖其中第五章关于超参数的模型选择和自适应的内容。

David Duvenaud[1]等在论文中提出了核函数结构以及组合核函数对于高斯问题的求解的影响，本次提高部分依照改论文提供的 GPML 库函数进行了探索。

模型

整体模型

本次大作业的主题是基于高斯过程的回归分析。回归分析研究的是变量与变量间的关系，假设两个变量存在一下关系：

$$y = f(X) + e$$

其中e表示误差，$f(X)$为回归函数，给定一组观测D，回归分析希望找出最佳回归函数f。回归问题包含两个层次的问题，一是确定合适的回归函数，二是在给定回归函数形式的情况下，依据训练数据求出具体回归函数。

标准线性模型

$$y = f(X) = x^T w$$

由于均方误差

$$J(w) = \frac{1}{n}\sum_{\{i=1\}}^{\{n\}} |y_i - f(x_i)|^2$$

为了最小化均方误差，有$J(w)$对w的偏导数为 0，求得：

$$W = (XX^T)^{-1}(XY)$$

标准线性模型（概率）

$$y = f(X) = x^T w$$

对于w有初值假设：

$$w \sim N(0, \Sigma_p)$$

由于为高斯过程，有：

$$p(y|X,w) = \prod_{\{i=1\}}^{\{n\}} p(y_i|x_i,w) = \prod_{\{i=1\}}^{\{n\}} \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left(-\frac{\left(y_i - x_i^T w\right)^2}{2\sigma_n^2}\right)$$

化简后得：

$$p(y|X,w) = \frac{1}{(2\pi\sigma_n^2)^{\frac{n}{2}}} \exp\left(-\frac{1}{2\sigma_n^2}|y - X^T w|^2\right) = N(X^T w, \sigma_n^2 I)$$

由贝叶斯原理知：

$$p(w|X,y) = \frac{p(y|X,w)p(w)}{p(y|X)}$$

最终可得：

$$p(w|X,y) \sim N\left(\frac{1}{\sigma_n^2} A^{-1} X y, A^{-1}\right)$$

其中：

$$A = \sigma_n^{-2} XX^T + \Sigma_p^{-1}$$

最大化后验分布（MAP）的优势在于最大化利用给定的训练数据，而且将各 种统计因素考虑在内。

广义线性回归模型

$$y = f(X) = \Phi(x)^T w$$

即将标准线性模型替换为依据基函数的线性模型，其中基函数可以不是线性函数，这样就方便引入非线性关系。具体分析方法同标准线性模型

核函数模型

在函数空间分析问题：

$$y = f(X) = \Phi(x)^T w$$

$$p(f|x,w,y) \sim N\left(\frac{1}{\sigma_n^2} \Phi(x)^T A^{-1} \Phi y, \Phi(x)^T A^{-1} \Phi(x)\right)$$

其中

$$A = \sigma_n^{-2} \Phi\Phi^T + \Sigma_p^{-1}$$

针对高斯过程有：

$$f(x) \sim GP\big(m(x), k(x,x')\big)$$
$$p(f|X,w,x_*) \sim N\big(f', cov(f)\big)$$

其中均值为：

$$f' = K(x_*, X)[K(X, X) + \sigma_n^2 I]^{-1} y$$

方差为：

$$cov(f) = K(x_*, x_*) - K(x_*, X)\,[K(X, X) + \sigma_n^2 I]^{-1} K(X, x_*)$$

其中X为训练数据x_*为测试数据。

方法分析

最小二乘法

现有模型：

$$y = x^T w$$

对于回归问题就是求解：

$$w = argmin(y - Xw)^T (y - Xw)$$

对w求导得到：

$$\frac{\partial E}{\partial w} = 2 * X^T (Xw - y)$$

展开得到：

$$w = (XX\text{^}T\ )\text{^}(-1)\ (XY)$$

贝叶斯概率

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

最大边际似然法

已知优化函数$L(\theta)$，在所有可能的 θ 取值中，寻找到一个值使得这个样本集的“可能性”最大化，就是使得样本集的可能性函数取得最大值。
由于$L(\theta)$与$logL(\theta)$的导数交与 1，因此可以通过估计$logL(\theta)$的极值点来估计$L(\theta)$的极值点。

梯度下降法

$$x^{(2)} = x^{(1)} - \alpha * \nabla f'(x^{(1)})$$

其中α为步长。

算法

标准线性模型最小二乘法

$Input\ \ X(train\ data)\ Y(train\ label)\ X'(test\ data)$

$Calculate\ \ W = (XX^T)^{-1}(XY)$

$Output\ \ Y' = X'^T W$

标准线性模型的概率角度分析

$Input\ \ X(train\ data)\ Y(train\ label)\ X'(test\ data)$

$Calculate\ \ \ A = \sigma_n^{-2}\ \Phi\Phi^T + \Sigma_p^{-1}$

$$p(w|X,y) \sim N\left(\frac{1}{\sigma_n^2} A^{-1} X y, A^{-1}\right)$$

$Output\ \ Y' = X'^T W$

广义线性回归

$Input\ \ X(train\ data)\ Y(train\ label)\ X'(test\ data)$

$Calculate\ \ W = (\Phi(X)\Phi(X)^T)^{-1}(\Phi(X)Y)$

$Output\ \ Y' = \Phi(X')\text{^}T\ \ W$

或

$Input\ \ X(train\ data)\ Y(train\ label)\ X'(test\ data)$

$Calculate\ \ \ A = \sigma_n^{-2}\ \Phi\Phi^T + \Sigma_p^{-1}$

$$p(w|X,y) \sim N\left(\frac{1}{\sigma_n^2} A^{-1} X y, A^{-1}\right)$$

$Output\ \ Y' = \Phi(X')\text{^}TW$

平方指数函数作为核函数

$Input\ \ X(train\ data)\ Y(train\ label)\ X'(test\ data)$

$Calculate\ \ E(Y') = K(x_*, X)[K(X,X) + \sigma_n^2 I]^{-1} Y$

$Calculate\ \ Cov(Y') = \ \ K(x_*, x_*) - K(x_*, X)\ [K(X,X) + \sigma_n^2 I]^{-1} K(X, x_*)$

$Output\ \ Y'$

算法分析

- 标准线性模型适用范围较窄，仅试用于线性的情况，对于非线性问题准确度不高。最小二乘法是从均值上误差最小进行的拟合，贝叶斯算法是从概率上进行拟合。整体复杂度均较低。
- 广义线性回归模型可引入非线性函数作为基函数，整体算法复杂度高于标准线性模型，

但是对于非线性函数的处理能力上升。

- 核函数法也是引入函数空间，这点和广义线性回归没有本质区别，但是核函数之间的组合以及核函数的选择十分丰富，因此大大增加了对于数据的处理能力，但是由于超参数的优化以及核函数的计算问题，算法复杂度均高于上述两种。

文件清单

文件名	内容	运行时间
Basic_f.m	基础部分四种方法的全部代码	38s
C_least_square.m	C 部分用最小二乘法得出的结果	0.1s
GPML.m	复现论文结果，调用 GPML 库函数	4s
Cor_select.m	利用数据相关度对数据处理再拟合	40s
GPR.m	包含 a,b,c,d,e1,e2,e3,optimal 的数据	/
报告.pdf	报告	/

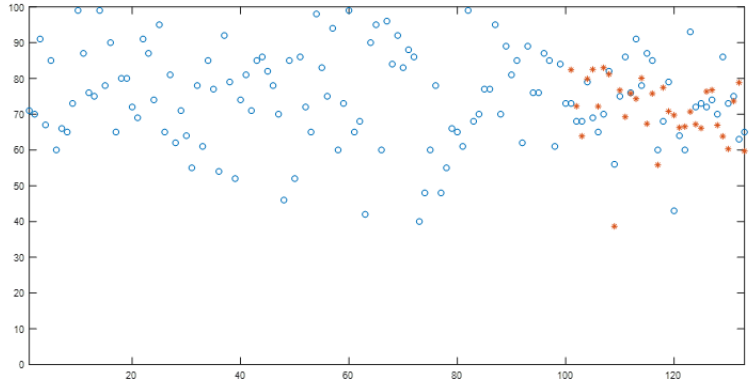
数值结果分析

对 2010 年数据作部分验证的 MSE

以下图中左端蓝色部分为训练数据，右端与红色部分重叠蓝色部分为测试数据，没有参与测试，红色部分为估计结果。

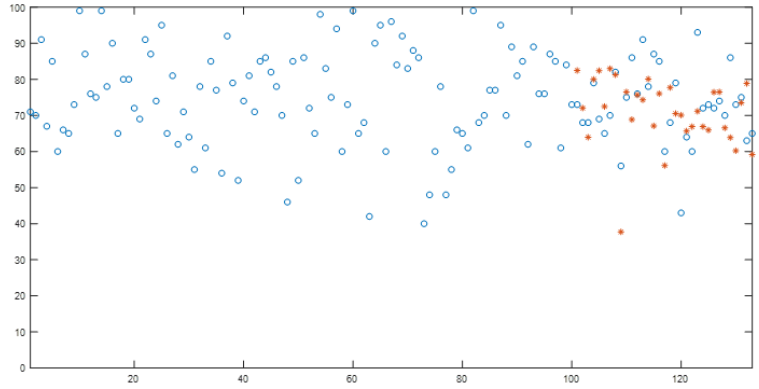
方法	MSE
标准线性模型最小二乘法	131.5543
标准线性模型概率估计法	133.7417
函数空间广义线性法（最小二乘）	139.8694
函数空间广义线性法（概率法）	128.1113
平方指数核函数法	110.9199
GPML 库函数法	107.9835
重组数据平方指数核函数法	90.7187

标准线性模型最小二乘法结果



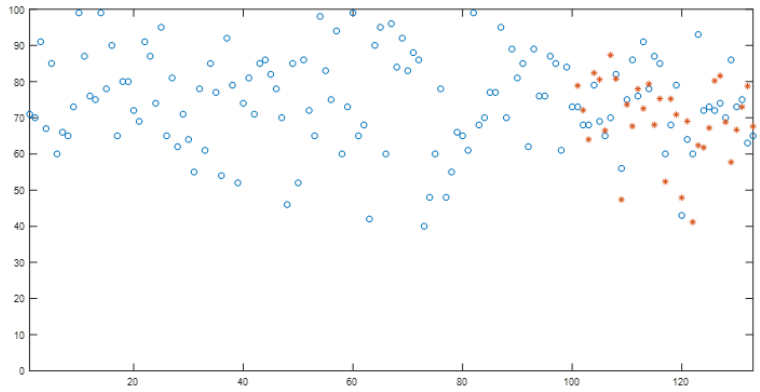
可以看到大部分点预测都比较准确，只有一个点预估过小，经过观察该点仅有一门课高于 80 粉，其余分数均较低，因此线性模型估计分数过低。

标准线性模型概率估计结果

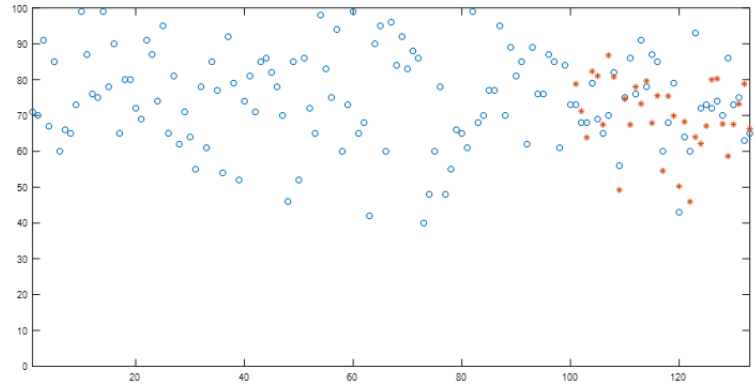


可以看到大部分点预测都比较准确，只有一个点预估过小，经过观察该点仅有一门课高于 80 粉，其余分数均较低，因此线性模型估计分数过低。

函数空间的广义线性法（最小二乘法）

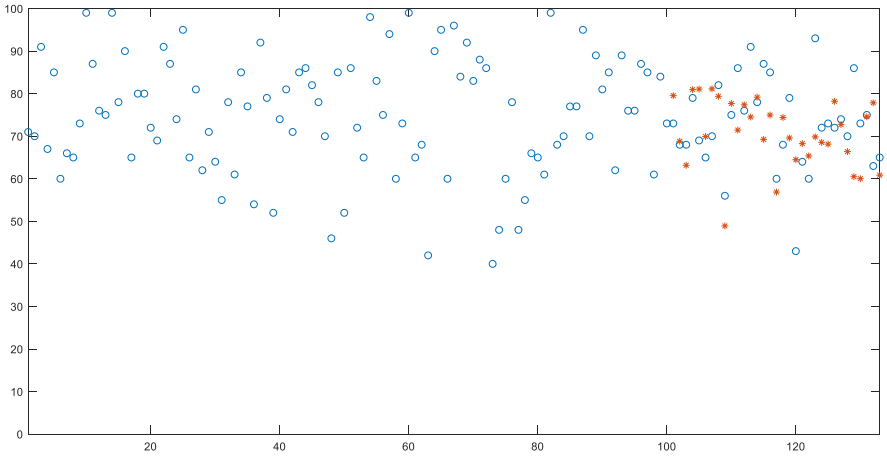


函数空间的广义线性法（概率法）

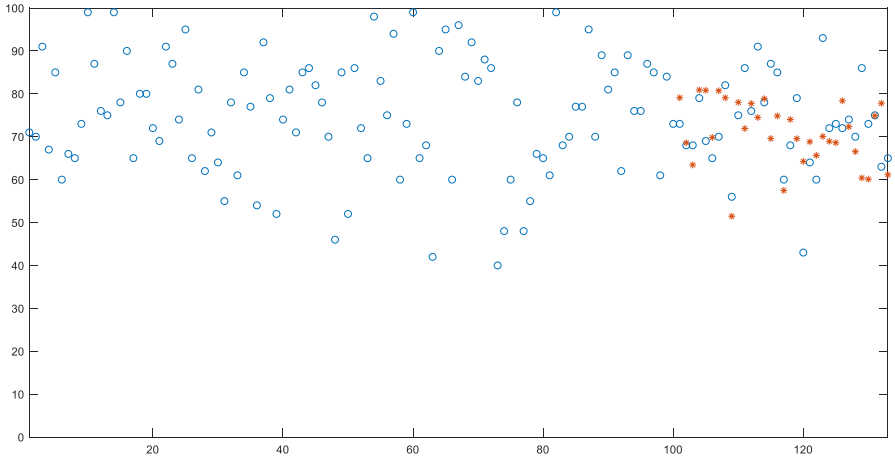


可以观察到对于该不合线性规律的点，引入非线性的基函数显然拟合程度更好，因此该点的误差减小。但是由于基函数增加了参数，因此在数据集不够大的情况下对于别的线性度较好的点的误差会增大。因此整体 MSE 变化不大。

平方指数核函数法

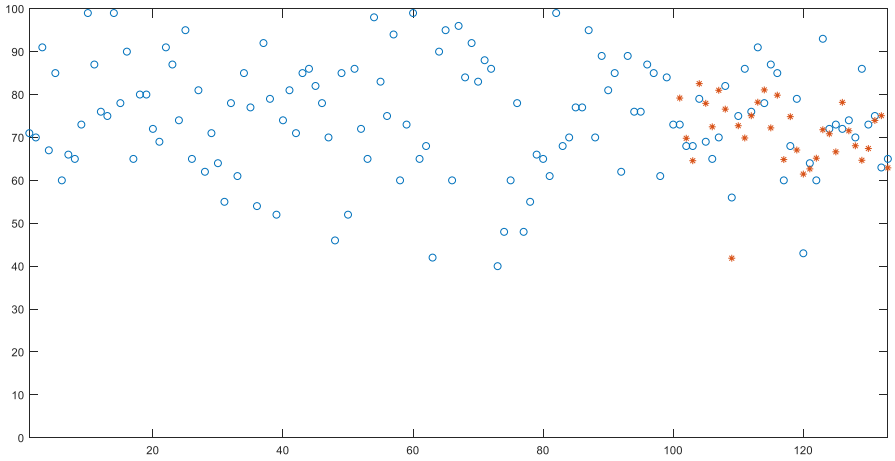


GPML 库函数法



可以看到 GPML 库函数法数据更加集中，整体拟合效果较线性函数好，和平方指数核函数类似。对于部分差异较大的点的拟合度也更好。

重组数据平方指数核函数法



可以观察到该方法数据结合了核函数方法较为集中的特点，又有部分极值类似线性法，最终得到了最小的误差。

核函数超参数初值分析

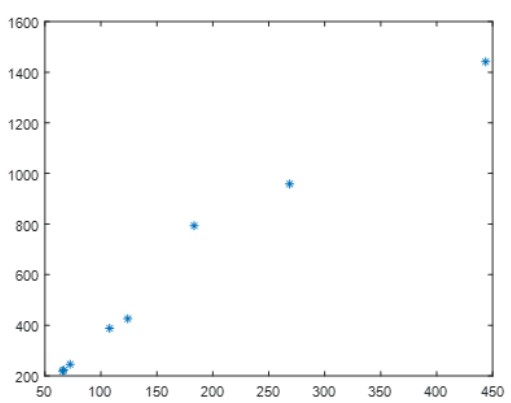
对于 2010 年的数据，我研究了不同初值超参数最终收敛点的规律，平方指数核函数有三个超参数σ_f, σ_n, l，为保证求导的方便，在代码中我用l代替了l^2

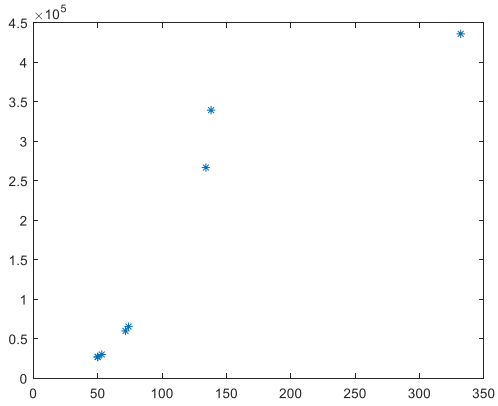
2010 年数据：

σ _n	10.5123	10.4312	10.3725	10.6148	10.5342	10.4978	10.4721	10.4508
σ _f	65.9	66.8	72.5	107.8	124.0	183.3	268.6	443.8
l^2	48133	49769	60614	150700	181760	629700	918440	2077200

2012 年数据：

σ _n	13.6123	13.5608	13.5890	13.6743	13.4976	13.6017	13.7102	13.5782
σ _f	73.9	50	134	138	332	71.5	53.1	49.9
l^2	65605	27511	266900	339530	436190	60083	30287	26974

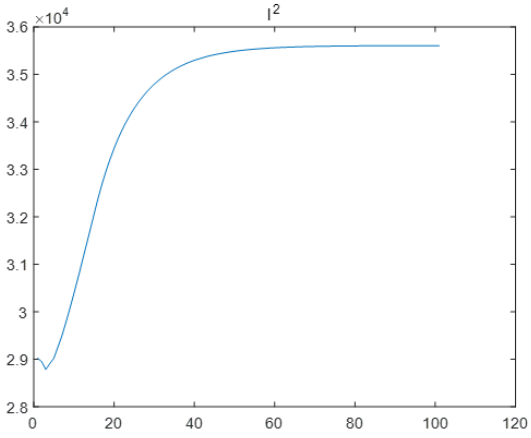


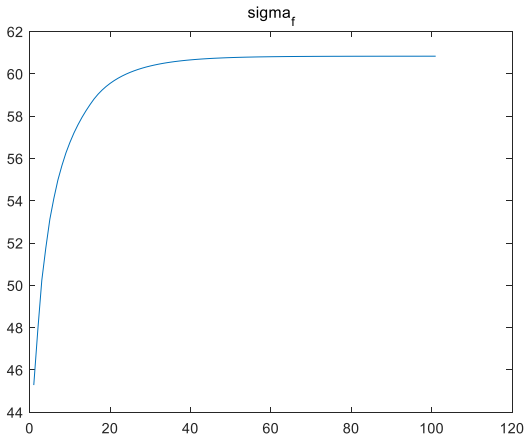


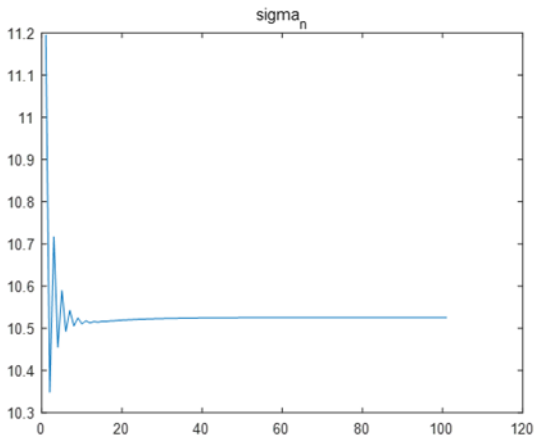
可以观察到σ_n作为数据的噪声分别稳定在 10.5 与 13.6 左右，因为它是由数据本身决定的，因此作为超参数是比较独立的。
对于σ_f与I，由于平方指数核函数的格式知σ_f增大函数值会减小，I增大函数值会变大。由于估计的数据大小不变，因此为保证最终结果不变，两者同时增大或减小，这是有核函数本身决定的性质。

核函数超参数收敛曲线分析

本次大作业，我采用了梯度下降法进行收敛分析。
开始由于核函数步长选取较小，因此造成了超参数初值对于最后结果影响过大的问题，随后考虑增加跳转步长，方便其收敛。
后续发现 l 的取值较大但是跳转能力及其有限，因此选择相对于其他超参数进一步扩大 l 跳转的步长（乘 5000）来避免由于初始值选取的范围有误造成的误差较大问题。再跳转了 15 次后，每增加一次跳转，步长减少为原来的 1/1.1 倍。以下是某一次的收敛曲线图，可以发现收敛速度较快且保证了跳转范围足够大。由此避免了初值的影响。







基于 GPML 库的核函数选择分析

MSE 来源：将 2010 年数据分为两部分（1~100 为训练数据 101~133 为测试数据）

核函数	平方指数	线性	有理二次	Matern
MSE	107.9835	106.4178	114.1263	108.3267

由于数据的相关度接近于线性，惊奇的发现线性核函数的误差居然是最小的，整体来看平方指数核函数 Matern 核函数和线性核函数相关度差别不大。
对于非线性数据的容忍度理论上来说平方指数函数，Matern 函数会优于剩下两种，但是对于线性函数，这几种函数的差别不大。
在训练集较小的情况下，由于超参数数量的影响，可能会出现线性函数由于其他函数的情况。

数据噪声分析

在该模型中，我从原有的 8 个数据中选择了 4 个噪声较小的进行拟合。判断依据为和结果的相关度，相关度较高的被认为噪声较小。
在对数据分析后得到了较小的数据的列分别为：

2	7	5	6
2	3	8	1
3	2	7	6
6	5	1	7
3	7	6	5
8	4	5	2
7	6	1	3

并由此得到了最小的 MSE，说明部分数据的噪声对结果起干扰作用。
因此数据集并不是越大越好，而是需要综合考虑噪声等因素。

结论

1. 本次数据线性度较高，因此以线性函数作为基函数或核函数得到的结果也较为优异。
2. 核函数选取的初值对于核函数整体表现会有较大影响，这时需要修改跳转步长，方便快速跳转。
3. 超参数由核函数决定，因此彼此之间依据核函数存在相关的联系。
4. 本次实验数据噪声较大，因此可以对部分数据进行去除，得到相对噪声较小的数据再进行分析。
5. GPML 库函数中的函数调用十分方便，算法也十分优异，在研究的过程中要学会用已有成果进行分析，不要拘泥于造轮子。

致谢

感谢王子奕同学与赵子嘉同学与我进行了关于超参数选择问题的交流。
感谢占文皓同学与我进行了 GPML 库函数的交流。

参考文献

[1] David Duvenaud, James Robert Lloyd, Roger Grosse, Joshua B. Tenenbaum, Zoubin

Ghahramani† .Structure Discovery in Nonparametric Regression through Compositional Kernel Search. Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013
[2] Carl Edward Rasmussen and Christopher K. I. Williams. Gaussian Processes for Machine Learning. Page[105~128]