# 基于回归分析的成绩预测算法研究

高枫 2015011208
gaof15@mails.tsinghua.edu.cn

## 摘要

本文基于高斯过程回归（Gauss Process Regression, 以下简称 GPR）技术，横向比较了传统线性回归方法与不同核函数下的 GPR 方法在随机过程成绩的预测问题上的应用性能，此外，还将最大似然的 GPR 方法、神经网络方法以及 RJMCMC-RBF 方法进行了比较，从而对于随机过程成绩的预测问题进行了全面的方法介绍，并在给定的数据集上取得了较好的结果。

***Keyword**: Gauss Process Regression, Squared Exponential, Periodic, Rational Quadratic*

## I. 概述

回归分析是统计学中十分重要的一种数据分析方法，随着大数据的兴起，其在大数据科学中也体现出十分重要的作用。回归分析在数据分析中是一种预测性的建模技术，它研究因变量和自变量之间的关系，通常使用于预测分析，时间序列模型以及发现变量之间的因果关系，因此回归方法的研究具有很重要的实际意义。

回归的最早形式是最小二乘法，由 1805 年的勒让德 (Legendre)[1]，和 1809 年的高斯 (Gauss) 出版 [2]。勒让德和高斯都将该方法应用于从天文观测中确定关于太阳的物体的轨道（主要是彗星，但后来是新发现的小行星）的问题。后来又由于高斯过程具有广泛的应用，人们也开始将其与回归问题相结合，从而有了高斯回归方法，[3][4] 对高斯回归方法进行了综合性的介绍，而 [5] 则介绍了 GPR 中常用的几种核函数。另外，回归问题其实也就是函数的拟合问题，因此近年来大热的神经网络方法 [6] 和传统的 RBF 网络 [7] 也自然地可以应用其中。本文将基于学生成绩数据，比较上述几种方法在实际问题上的应用效果，并最后得出较好的随机过程成绩的预测结果。

本文的结构安排如下，$II$ 中将介绍传统的线性回归模型及相关实现方法，$III$ 中将介绍基于高斯过程的非线性回归方法及常用的核函数，$IV$ 中将介绍基于反向传播方法的神经网络模型，$V$ 中将介绍使用 RJMCMC 方法的 RBF 模型选择技术，然后将在 $VI$ 中介绍上述回归方法在给定数据集上的实验结果，并进行分析比较，最后会在 $VII$ 中进行总结。

## II. 线性回归模型

回归分析研究的是变量与变量间的关系，记其中一个变量称为自变量 $x \in S \subset \mathbb{R}^d$，另一个变量称为因变量 $y \in \mathbb{R}$，假设两者存在如下关系

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

其中 $e$ 为表示误差的随机变量，$f(x)$ 称为回归函数。回归分析即希望找出在某准则下最好的回归函数。广义线性模型是一类应用广泛的回归模型，它假定回归函数为基函数的线性组合，即

$$y = \sum_{i=1}^{M} \omega_i \phi_i(x) + e$$

其中 $\phi_i(x)$ 称为基函数。若 $\phi_i(x) = x_i$，则变为标准线性模型。

### A. 最小二乘法

取 $\phi_i(x) = x_i$，即使用标准线性回归模型，则

$$y = \sum_{i=1}^{M} \omega_i x_i \Rightarrow Y = X\omega$$

其中 X 的维数为 $N \times d$，$\omega$ 的维度为 $d \times 1$，N 为样本数，d 为样本维度。

利用最小二乘法 (Least Squares, LS)，求解即伪逆

$$\omega = (X^T X)^{-1} X^T Y$$

### B. 贝叶斯线性回归

贝叶斯线性回归 (Bayesian Regression, BR) 也是求解线性回归模型的回归方法，

$$y = f(x) + e = x\omega + e \Rightarrow Y = X\omega + E$$

在贝叶斯回归中，先验假设 $e \sim \mathcal{N}(0, \sigma^2), \omega \sim \mathcal{N}(0, \Sigma_p)$，则经过增广

$$\begin{pmatrix} Y \\ \omega \end{pmatrix} = \begin{pmatrix} I & X \\ 0 & I \end{pmatrix} \begin{pmatrix} E \\ \omega \end{pmatrix}$$

从而

$$\begin{pmatrix} Y \\ \omega \end{pmatrix} \sim \mathcal{N}\left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 I + X\Sigma_p X^T & X\Sigma_p \\ \Sigma_p X^T & \Sigma_p \end{pmatrix} \right\}$$

则

$$\omega | Y \sim \mathcal{N}(\mu_\omega, \Sigma_\omega)$$

其中

$$\mu_\omega = \Sigma_p X^T (\sigma^2 I + X\Sigma_p X^T)^{-1} Y$$

$$\Sigma_\omega = \Sigma_p - \Sigma_p X^T (\sigma^2 I + X\Sigma_p X^T)^{-1} X\Sigma_p$$

因此给定一个新的数据 $x^*$，

$$x^*\omega \sim \mathcal{N}(x^*\mu_\omega, x^*\Sigma_\omega x^{*T})$$

### C. 广义线性回归

在广义线性回归 (Generalized Linear Regression, GLR) 中，取基函数为多项式函数，则原标准线性回归就转变为了实际上的非线性回归。这里取 $\phi_i(x) = ax_i + bx_i^2 + cx_i^3$。又因为其中的超参数 a,b,c 是线性组合系数，实际上可以看做是对增广后的 X 矩阵进行标准线性回归，使用上述介绍的最小二乘回归或贝叶斯线性回归即可。

$$X' = [X, X^{(2)}, X^{(3)}]$$

## III. 高斯过程回归模型

与传统线性模型中需要确定模型参数不同，GPR 模型把回归问题中的 $y$ 看成是服从高斯分布的，因此回归过程就变为了求解 $y$ 的后验概率分布，从而在新给定的 $x^*$ 上也能得到较好的回归结果。

设有核函数 $k(x_1, x_2)$，则在选定核函数和数据集后，有

$$K = \begin{bmatrix} k(x_1, x_1) & k(x_1, x_2) & \dots & k(x_1, x_n) \\ k(x_2, x_1) & k(x_2, x_2) & \dots & k(x_2, x_n) \\ \vdots & \vdots & \ddots & \vdots \\ k(x_n, x_1) & k(x_n, x_2) & \dots & k(x_n, x_n) \end{bmatrix}$$

$$K_* = [k(x^*, x_1) k(x^*, x_2) \dots k(x^*, x_n)]$$

$$K_{**} = k(x^*, x^*)$$

其中，常用的核函数有 SE 核,PER 核和 RQ 核，具体表达式如下表所示：

| Kernel | Function |
|---|---|
| SE(Squared Exponential) | $\sigma^2 exp(-\frac{(x_1-x_2)^2}{2l^2})$ |
| PER(Periodic) | $\sigma^2 exp(-\frac{2\sin^2(\pi(x_1-x_2)/p)}{l^2})$ |
| RQ(Rational Quadratic) | $\sigma^2(1+\frac{(x_1-x_2)^2}{2\alpha l^2})^{-\alpha}$ |

Table 1: Some Common Kernel Function

则根据 [4]，若给定核函数参数，y 的增广向量满足

$$\begin{pmatrix} Y \\ y^* \end{pmatrix} \sim \mathcal{N}\left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} K & K_*^T \\ K_* & K_{**} \end{pmatrix} \right\}$$

从而得到 $y^*$ 的后验分布概率为

$$y^* | Y \sim \mathcal{N}(K_* K^{-1} Y, K_{**} - K_* K^{-1} K_*^T)$$

故而原回归问题可以转化为在给定数据集下对核函数进行优化，考虑对数似然值

$$\log p(Y|X,\Theta) = -\frac{1}{2}Y^T K_y^{-1} Y - \frac{1}{2}\log|K_y| - \frac{n}{2}\log 2\pi$$

其中 $K_y = K + \sigma_n^2 I$。使用梯度下降法进行优化，可以推导出其导函数为

$$\frac{\partial}{\partial\theta_j}\log p(Y|X,\Theta) = \frac{1}{2}tr((\alpha\alpha^T - K^{-1})\frac{\partial K}{\partial\theta_j}),\ \alpha = K^{-1}Y$$

由此，即可利用梯度下降算法等优化方法对核函数进行优化，使 Y 的后验分布概率达到最大值，确定其核函数的超参数，从而能够在给定的 $x^*$ 上利用最大似然的后验分布求出预测的 $y^*$。故而称为使用最大似然估计的高斯过程回归 (MLE-GPR)。

# IV. 神经网络模型

神经网络模型近年来被广泛应用于语音、图像、文本处理等各个领域，而其本质其实就是函数的拟合、回归，因此我想到将其应用于这次的成绩预测问题上，将往年的学生成绩作为输入，其随机过程成绩作为输出，使用反向传播算法进行训练。网络结构示意图如下

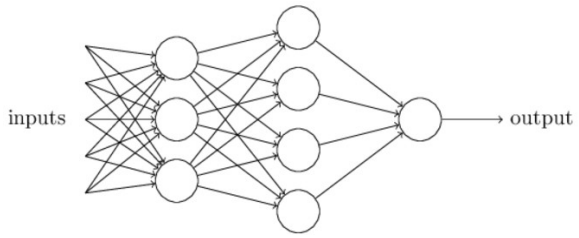


Figure 1: 2-layer Neural Network

因为 Gauss 模型属于相对简单的函数模型，故而这里仅选用了两层（单隐藏层-输出层）的神经网络进行测试，其中隐藏层节点数设置为 25，使用 ReLU 函数作为激活函数，各层之间的均为全连接方式。

# V. RBF 网络模型

这里采用了与第一次课程设计相同的 RJMCMCSA 算法，下面仅给出其实现算法，具体说明参考 [8]

**Algorithm 1** Reversible Jump Simulated Annealing

**Initialization**:

Set $(k^{(0)}, \theta^{(0)}) \in \Theta$

**Iteration i**:

a. Sampling u U[0,1] and set the temperature with a cooling schedule

b. if $u \leq b_{k^{(i)}}$

do "birth" move

else if $u \leq b_{k^{(i)}} + d_{k^{(i)}}$

do "death" move

else if $u \leq b_{k^{(i)}} + d_{k^{(i)}} + s_{k^{(i)}}$

do "split" move

else if $u \leq b_{k^{(i)}} + d_{k^{(i)}} + s_{k^{(i)}} + m_{k^{(i)}}$

do "merge" move

else

update the RBF centres

c. Perform an MH step with the annealed acceptance ratio.

d. i = i + 1

* 几种 Move 的具体实现可参考代码文件或 [7]

# VI. 相关试验

为了验证上述各方法的实际应用性能，在学生的成绩数据集上进行测试，将每年的成绩数据分成十份，其中九份作为训练集，一份作为验证集，如此进行十轮循环，完成交叉验证过程。将各方法在验证集上的 loss 曲线绘制如 Figure2-6 所示。

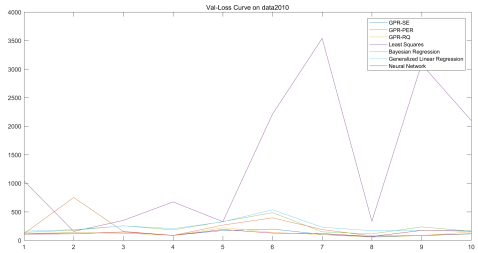


Figure 2: Val-Loss Curve on data2010

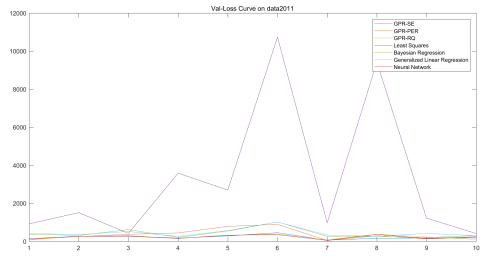


Figure 3: Val-Loss Curve on data2011

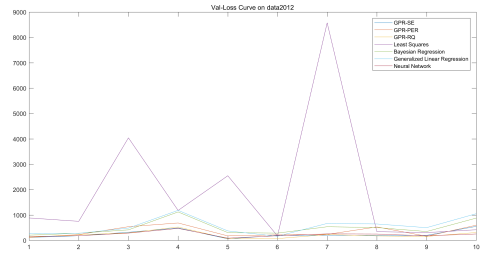


Figure 4: Val-Loss Curve on data2012

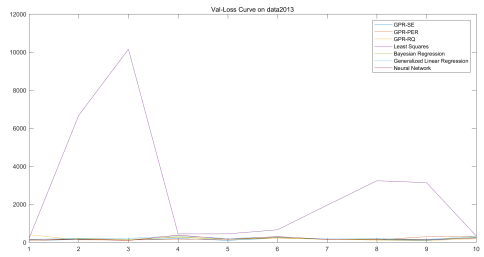


Figure 5: Val-Loss Curve on data2013

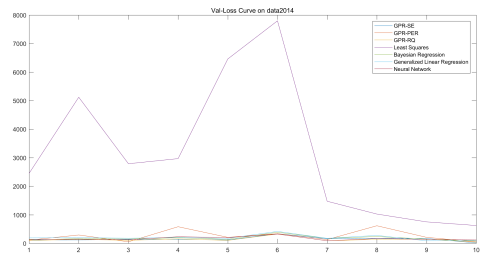


Figure 6: Val-Loss Curve on data2014

* 曲线图中并没有 RJMCMC-RBF 模型的数据，因为其损失值远高于其他方法，故而仅以表格的形式呈现。

为更清楚地看出各方法的相对变化趋势，去掉最小二乘方法的数据，重新绘制 loss 曲线如 Figure7-11 所示。

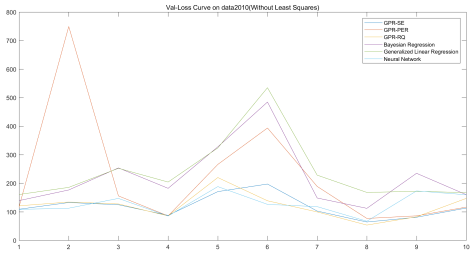


Figure 7: Val-Loss Curve on data2010 (Without LS)

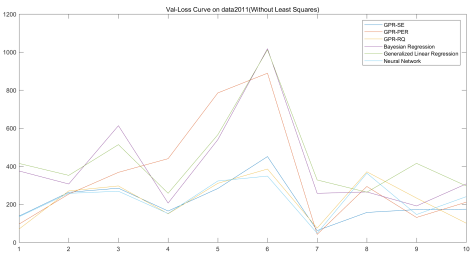


Figure 8: Val-Loss Curve on data2011 (Without LS)

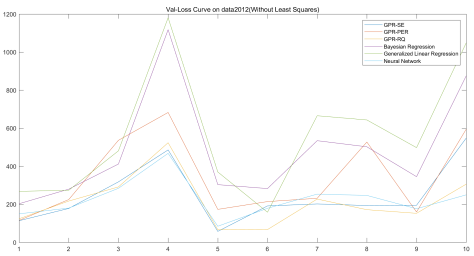


Figure 9: Val-Loss Curve on data2012 (Without LS)

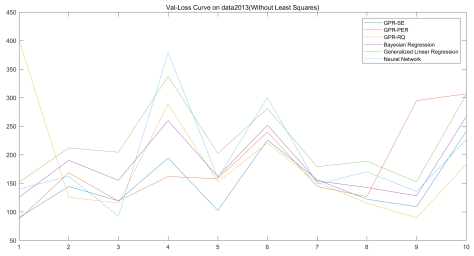


Figure 10: Val-Loss Curve on data2013 (Without LS)

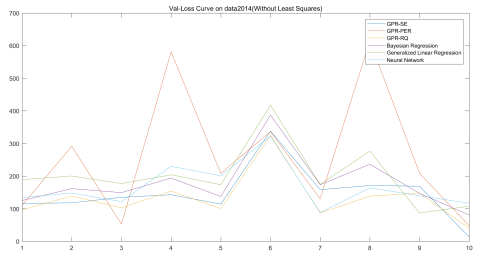


Figure 11: Val-Loss Curve on data2014 (Without LS)

从 Figure7-11 中可以看出，使用 PER 核的 GPR 方法、贝叶斯线性回归和广义线性回归方法的效果要差于其他三种方法，且波动性更大。为了更清楚地显示出不同数据集下不同方法的最优性能，将其数据整理成下表，其给出的 loss 均为交叉验证过程中的最小 loss。

| | 2010 | 2011 | 2012 | 2013 | 2014 |
|---|---|---|---|---|---|
| SE | 64.012 | 61.530 | **57.305** | 89.950 | **13.037** |
| PER | 76.594 | **41.519** | 115.074 | **86.711** | 47.591 |
| RQ | **53.354** | 69.524 | 65.969 | 89.631 | 39.944 |
| NN | 66.789 | 47.591 | 84.360 | 91.980 | 88.117 |
| LS | 155.739 | 405.661 | 170.480 | 201.055 | 622.867 |
| BR | 111.970 | 191.012 | 202.638 | 125.450 | 80.257 |
| GLR | 160.438 | 257.635 | 158.084 | 151.797 | 86.682 |
| RBF | 6179 | 9383 | 11703 | 17672 | 3247 |

Table 2: Min Val-Loss on Different Methods

* 注：其中 SE、PER、RQ 为相应核函数下的 GPR 模型的数据，NN 对应神经网络模型的数据，LS 对应最小二乘法得到的数据，BR 和 GLR 则分别对应贝叶斯回归和广义线性回归，RBF 为使用 RJMCMCSA 方法的 RBF 模型

从 Table2 中可以更加清晰的看出，使用不同核函数下的 GPR 模型和神经网络模型得到的结果要明显好于其他方法，且 RBF 模型的结果要远差于其他方法。不过这种结果与数据的格式有很大的关系，这次实验所使用的学生成绩数据并不适合与 RBF 模型，而 GPR 模型则可以取得较好的应用性能。

若结合表格和 loss 变化曲线来看，不难发现，PER 核下取得的结果虽然随数据集波动较大，但是在某些时刻能够取得最优的预测结果，而 SE 核和 RQ 核相对稳定，且其 loss 始终处于较低值，神经网络模型则略差于不同核函数下的 GPR 模型。

# VII. 总结

本文先后介绍了最传统的线性回归模型、基于高斯过程的回归模型、神经网络模型和 RBF 网络模型，并在学生成绩数据集上进行了多次试验，从这些结果不难看出，线性回归模型，包括使用最小二乘法的线性模型和贝叶斯回归模型，在固定数据集上的结果相对稳定，但是在不同数据集上波动较大，而 GPR 模型在不同数据集上往往能够取得更好的结果，但是由于其为概率生成模型，在某一给定数据集上的结果具有一定的波动性，需要多次试验才能够找到更好的结果。此外，神经网络模型和 RBF 模型为在上述模型外进行的独立探索，不过由于 RBF 模型并不适合这次的数据集，其取得的效果很差。神经网络模型则不负所望，取得了仅次于 GPR 模型的结果，而且其预测值比较稳定，波动小于 GPR 模型。

# VIII. 致谢

在完成这次课程设计的过程中，我与全雨晗同学进行了多次讨论，这让我对回归分析的原理和实现技巧有了更加深刻的认识。

另外在阅读相关文献时，我曾多次参考欧智坚老师的课程讲义，这些知识对于我理解高斯过程回归的相关知识提供了极大的帮助，同时还要感谢老师和助教的辛苦付出，这次作业令我收获很多！

# IX. 文件说明

最终提交的.mat 文件中共有 8 个向量，分别为

a: 使用最小二乘方法的线性回归

b: 贝叶斯线性回归

c: 广义线性回归

d: 使用 SE 核的高斯过程回归

e1: 使用 PER 核的高斯过程回归

e2: 使用 RQ 核的高斯过程回归

e3: 神经网络模型

optimal: 选用 SE 核的 GPR 模型的结果作为最优结果

* 注：由于 RBF 模型的结果与其他结果相差很大，故不再提交其结果

# 参考文献


[1] A.M. Legendre. *Nouvelles méthodes pour la détermination des orbites des comètes*. Nineteenth Century Collections Online (NCCO): Science, Technology, and Medicine: 1780-1925. F. Didot, 1805.

[2] C.F. Gauss. *Theoria motus corporum coelestium in sectionibus conicis solem ambientium*. Carl Friedrich Gauss Werke. sumtibus F. Perthes et I. H. Besser, 1809.

[3] Carl Edward Rasmussen. Gaussian processes for machine learning. MIT Press, 2006.

[4] M.Ebden. Gaussian processes for regression: a quick introduction. 2008.

[5] David Duvenaud, James Lloyd, Roger Grosse, Joshua Tenenbaum, and Ghahramani Zoubin. Structure discovery in nonparametric regression through compositional kernel search. In *Proceedings of the 30th International Conference on Machine Learning*, pages 1166–1174. PMLR, 2013.

[6] Yann LeCun, Léon Bottou, Genevieve B. Orr, and Klaus-Robert Müller. Efficient backprop. In *Neural Networks: Tricks of the Trade, This Book is an Outgrowth of a 1996 NIPS Workshop*, pages 9–50, London, UK, UK, 1998. Springer-Verlag.

[7] Christophe Andrieu, Nando de Freitas, and Arnaud Doucet. Reversible jump MCMC simulated annealing for neural networks. *CoRR*, abs/1301.3833, 2013.

[8] Feng Gao. Automatic model selection techniques for RBF network. 2017.