基于高斯过程的回归分析

孙望涛

清华大学电子工程系

2020 年 1 月

摘　要

本文主要探讨了回归任务中的几种方法：最小二乘法，贝叶斯回归，广义线性回归与高斯过程回归，着重研究了高斯过程回归的核函数以及模型选择方法，对学生成绩预测的模型给出了有效的均值函数与协方差函数（核函数），在理论推导的基础上实现了上述算法应用于学生成绩预测的模型，根据仿真结果比较了各算法的性能，分析其特点与使用范围。

关键词：回归分析，线性模型，贝叶斯，高斯过程回归，核函数，超参数

Abstract

This paper mainly discusses several methods in the regression task: least squares method, Bayesian regression, generalized linear regression and Gaussian process regression. It focuses on the kernel function and model selection method of Gaussian process regression, and gives the model of student performance prediction. An effective mean function and covariance function (kernel function) were developed. Based on the theoretical derivation, the above-mentioned algorithm was applied to the model of student performance prediction. The performance of each algorithm was compared based on the simulation results, and its characteristics and application range were analyzed.

Keywords: Regression analysis, linear model, Bayesian, Gaussian process regression, kernel function, hyperparameter

1.介绍

回归分析是统计学、信号处理、机器学习等多领域中的基础研究问题之一。回归分析研究的是变量与变量间的关系，其中一个变量称为自变量$x \in S \subset R^d$，另一个变量称为因变量$y \in R$。假设两者存在如下的关系$y = f(x) + e$ 其中，e 为表示误差的随机变量，f(x)称为回归函数（或预测函数、拟合函数）。给定一组观测 D={(xi,yi)|i= 1,2, ..., n}，D 常称为训练数据，回归分析希望能就此找出在某个准则下最好的回归函数 f(*)。传统的回归分析包括了两个层次的问题，一是确定合适的回归函数形式；二是在给定回归函数形式下，依据训练数据求出具体的回归函数。

线性回归是假设观测量与目标量是线性关系的一种模型，即$y = f(x) = x'w$，其中$w = [w_1\ w_2 \cdots w_d]'$为 d 维的权重矢量。在给定此线性模型的条件下，可以通过训练集确定均方误差最小的 w，从而得到具体的回归函数形式。具体求解方法：最小二乘法和概率求解（贝叶斯线性回归）将在后文介绍。

但绝大多数的数据无法用简单的线性关系来描述，因此我们引入基函数$\phi(x): R^d \to R^N$对观测量 x 进行升维，再令$y = f(x) = \phi(x)'w$，确定 w 从而得到具体的回归函数。这里$\phi(x)$可以是多项式函数$\phi(x) = [x'\ x'^2\ x'^3 \cdots x'^N]$，也可以是三角函数$\phi_i(x) = \cos(w_i a_i' x)$等等。这称为广义线性回归模型。

高斯过程回归假设$f(x)$为一高斯过程（被高斯噪声污染后仍为高斯过程），其由均值函数$u(x)$与协方差函数$K(x,x')$完全确定。于是，对此随机过程进行采样，得到的训练集 f 与测试集未知的 f*应服从联合高斯分布，即可通过已有数据点(x,f)推测 f*的边缘分布。

2.模型

本文主要研究以下问题：

通过电子系学生 8 门前置课（微积分、线性代数、复变函数与数理方程……）的成绩预测随机过程课程的成绩。即希望获得函数$y = f(x)$，其中$y \in R, x \in R^8$。每门课的成绩均用百分制表示，使用均方误差

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(y_i - y_i^*)^2$$

来评估预测的准确程度。

3.基本理论

（1）最小二乘法(Least square)

设$f(x)=x'w$去估计y，定义均方误差$J \triangleq \|Y-X'W\|$，对 w 求导可得

$$\frac{\partial J}{\partial W}=X(Y-X'W)=0$$

$$\widehat{W}=(XX')^{-1}XY$$

即$W=\widehat{W}$时均方误差最小。

（2）贝叶斯线性回归(Bayesian linear regression)

1） 设$Y=f(X)+E=X'W+E$，这里 E 为高斯噪声矢量，由于各学生的考试成绩彼此独立，因此 E 服从 n 维联合高斯分布，且协方差矩阵应为对角阵

$$E\sim N(0,\sigma^2 I)$$

因此

$$Y=X'W+E\sim N(X'W,\sigma^2 I)$$

W 是各门前置课对随机过程成绩的权重系数，我们先验的认为他服从 8 维的零均值联合高斯分布，即

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

这里的Σ_p表示其先验的协方差矩阵

由贝叶斯公式

$$\text{后验}=\frac{\text{似然}*\text{先验}}{\text{边缘似然}}$$

得到

$$p(W|Y,X)=\frac{p(Y|W,X)*p(W)}{p(Y|X)}$$

又分母与权重 W 无关，故

$$p(W|Y,X)\propto exp\left(-\frac{1}{2\sigma^2}(Y-X'W)'(Y-X'W)\right)exp\left(-\frac{1}{2}W'\Sigma_p{}^{-1}W\right)$$

$$\propto exp\left(-\frac{1}{2}(W-\bar{W})'(\frac{1}{\sigma^2}\mathrm{XX}'+\Sigma_p{}^{-1})(W-\bar{W})\right)$$

其中

$$\bar{W}=\sigma^{-2}(\sigma^{-2}XX'+\Sigma_p{}^{-1})^{-1}XY$$

即 W 的后验也为一高斯分布

$$p(W|Y,X)\sim N(\bar{W}=\sigma^{-2}\left(\sigma^{-2}XX'+\Sigma_p{}^{-1}\right)^{-1}XY,\qquad(\sigma^{-2}XX'+\Sigma_p{}^{-1})^{-1})$$

至此我们得到了 W 的后验分布，可以用此来预测测试集上的 f*

$$E\big(f(X_*)\big) = E(X_*'W) = X_*'E(W) = X_*'\sigma^{-2}(\sigma^{-2}XX' + \Sigma_p{}^{-1})^{-1}XY$$

2）方法 1 通过训练集所有的数据一次性得到了 W 的后验分布，我们考虑能否通过迭代的方式不断修正我们对 W 的认知

已知矢量$(W\ E)'$服从联合分布

$$\begin{pmatrix} W \\ E \end{pmatrix} \sim N\left(\begin{pmatrix} EW \\ 0 \end{pmatrix}, \begin{pmatrix} CovW & \\ & \sigma^2 I \end{pmatrix}\right)$$

又

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

故

$$\begin{pmatrix} W \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} EW \\ X'EW \end{pmatrix}, \begin{pmatrix} CovW & Cov(W)X \\ X'CovW & X'Cov(W)X + \sigma^2 I \end{pmatrix}\right)$$

从这个联合分布中抽取 W 的后验概率密度，由多维高斯分布的性质知其也为高斯分布，均值与协方差分别为

$$E(W|Y,X) = EW + Cov(W)X(X'Cov(W)X + \sigma^2 I)^{-1}(Y - X'EW)$$

$$Cov(W|Y,X) = CovW - Cov(W)X(X'Cov(W)X + \sigma^2 I)^{-1}XCovW$$

我们通过对一组数据(X,Y)的观测更新了我们对 W 的认识（均值，协方差），于是我们设定 W 的先验为

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

便可通过所有的输入对 W 进行更新，最终得到训练后的 W 的分布。这被称为增量贝叶斯学习(Incremental Bayesian learning)，相比于直接的贝叶斯线性回归，此方法优势在于可以在线学习，无需等待所有样本采集完毕，而是每输入一个样本便更新一次模型。

（3）高斯过程回归(Gaussian process regression)

我们从函数空间的观点推导高斯过程回归的方法。运用基函数的方法，将 x 投影到特征空间

$$f(x) = \phi(x)'W$$

这里认为 W 先验服从联合高斯分布

$$W \sim N(\overline{W}, \Sigma_p)$$

于是可得

$$E\big(f(x)\big) = \phi(x)'EW = \phi(x)'\overline{W}$$

$$E\big(f(x)f(y)\big) = \phi(x)'E(WW')\phi(y) = \phi(x)'\Sigma_p\phi(y)$$

我们令

$$u(x) \triangleq E\big(f(x)\big)$$

$$K(x,y) \triangleq E(f(x)f(y))$$

则 f(x)是以$u(x)$为均值函数，以$K(x,y)$为协方差函数的高斯过程，记作：

$$f(x) \sim GP(u(x),K(x,y))$$

考虑一个带噪声的回归模型

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

令训练集为

$$Y = f(X) + E$$

其中

$$E \sim N(0,\sigma^2 I)$$

测试集为 Z,希望预测的值

$$f_* = f(Z)$$

由上，我们可以知到训练集与测试集服从联合高斯分布

$$\begin{pmatrix} Y \\ f_* \end{pmatrix} \sim N(\begin{pmatrix} u(X) \\ u(Z) \end{pmatrix}, \begin{pmatrix} K(X,X)+\sigma^2 I & K(X,Z) \\ K(Z,X) & K(Z,Z) \end{pmatrix})$$

从而得到

$$f_* \sim N(\hat{u},\hat{\Sigma})$$

其中

$$\hat{u} = K(Z,X)(K(X,X)+\sigma^2 I)^{-1}\big(Y-u(X)\big)+u(Z)$$

$$\hat{\Sigma} = K(Z,Z) - K(Z,X)(K(X,X)+\sigma^2 I)^{-1}K(X,Z)$$

现在，我们虽然确定了$K(X,X)$的形式，但其内部的超参数还未被确定，这些超参数密切影响着我们模型对数据的贴合程度，考虑给定一组超参数和输入值条件下的边缘似然

$$p(Y|X,\theta) = \int p(y|f,X,\theta)p(f|X,\theta)df$$

在先验和似然都服从高斯分布的前提下

$$p(Y|X,\theta) = \int N(0,K)N(f,\sigma^2 I)df = N(0,K_\theta + \sigma^2 I)$$

求出对数边缘似然得

$$logp(Y|X,\theta) = -\frac{1}{2}Y'(K_\theta+\sigma^2 I)Y - \frac{1}{2}ln(\det(K_\theta+\sigma^2 I)) - \frac{n}{2}log2\pi$$

最大似然等价于负对数边缘似然最小

$$nlml = -logp(Y|X,\theta) = \frac{1}{2}Y'(K_\theta+\sigma^2 I)Y + \frac{1}{2}ln(\det(K_\theta+\sigma^2 I)) + \frac{n}{2}log2\pi$$

我们将其作为目标函数，使其达到最小时的超参数θ就满足了模型最大似然。

4.算法设计与分析

算法设计与实现中，将训练集的 80%用于训练，20%用于交叉验证，通过验证

集的均方误差评估各算法的预测性能。

（1）最小二乘法

直接通过$\widehat{W} = (XX')^{-1}XY$计算即可，但注意应将 X 升维为$\binom{X}{1}$，使得求出的 W 的最后一个元素能表示线性模型的偏置 b，均方误差更小。以下使用线性模型回归的算法均做相同处理。

（2）贝叶斯线性回归

1）直接用 W 后验分布的均值

$$\bar{W} = \sigma^{-2}(\sigma^{-2}XX' + {\Sigma_p}^{-1})^{-1}XY$$

作为模型中 W 的值，以此估计测试集的目标量。注意Σ_p是 W 的先验分布的协方差矩阵，需选择为半正定矩阵，由于目标量 Y 是标量，故σ^{-2}也选择标量即可。

2）同 1）选择$CovW$，EW 和σ^{-2}的先验后，对训练集进行循环，每次抽取一条学生的数据，用此更新 W 的数字特征

$$E(W|Y,X) = EW + Cov(W)X(X'Cov(W)X + \sigma^2 I)^{-1}(Y - X'EW)$$
$$Cov(W|Y,X) = CovW - Cov(W)X(X'Cov(W)X + \sigma^2 I)^{-1}XCovW$$

实现中发现，虽然增量贝叶斯学习可以重复利用数据，但在此线性模型中对训练集循环一遍后参数已收敛。

3）对数据投影到高维后，采用增量贝叶斯学习的方式进行回归，但实验中发现，使用大多数多项式基函数只能使训练集的 MSE 下降，而验证集的 MSE 上升，即发生了过拟合。

（3）高斯过程回归

首先我们需要给出 f(x)服从的高斯过程的均值函数与协方差函数。考虑到学生成绩的相对恒定，随机过程的成绩均值应与前八门课的平均成绩相当，故$E(f(x)) \approx mean(x)$，又考虑到随机过程的**难度较大**，成绩可能相对偏低，故通过测试集估计了前八门课的平均成绩与随机过程成绩的偏差 udelta

```
udelta = mean(mean(train_set(:,1:8),2)-train_set(:,9));
```

进而给出 f(x)的均值函数表示

$$E\big(f(x)\big) = u(x) = mean(x) - udelta$$

在许多文献中提到，实际应用中将均值函数直接取为 0 是一种常用的处理方法，因为均值的不确定性可以包含在协方差中；但此模型在上述分析中对均值有比较好的估计，经实验也发现使用上述均值函数比零均值均方误差更小，因此在这里不使用零均值的处理方法。

协方差函数，即核函数采用 SE 核，其含有三个超参数,sf,l, σ^2

$$Cov\big(f(x),f(x')\big)=K(x,x')=sf^2\exp\left(-\frac{1}{2l^2}(x-x')^2\right)+\sigma^2\delta_{xx'}$$

为了找到最贴合模型的超参数的值，我们需要进行超参学习，损失函数为负对数边缘似然(negative log marginal likelihood)

$$loss=\frac{1}{2}Y'(K_{XX}+\sigma^2I)Y+\frac{1}{2}ln(\det(K_{XX}+\sigma^2I))$$

实现中调用了 matlab 优化工具箱中的 fminunc 函数

```
para = fminunc(@calculate_loss,[sf,L,CovE]);
```

但是此时的目标函数并非凸函数，fminunc 结束迭代时返回的参数并不一定在可行域上最优，因此考虑直接在较大范围内随机生成多个初始参数，进行迭代，最后筛选出最优的参数。

最后根据公式

$$\hat{u}=K(Z,X)(K(X,X)+\sigma^2I)^{-1}\big(Y-u(X)\big)+u(Z)$$

便可对验证集和测试集的数据进行预测。

对上述算法进行实验，最小二乘、普通贝叶斯、增量贝叶斯、广义线性回归、高斯回归分别对应后缀 A、B1、B2、C、D；另外发现 matlab 有自带的高斯回归函数 fitrgp，用其作为参照，后缀 m。每组算法输出其在训练集和验证集上的均方误差（以验证集上的均方误差为主要指标衡量算法预测的准确度）。

工作区

名称	值
MSE_train_A	99.0861
MSE_train_B1	99.0896
MSE_train_B2	99.2472
MSE_train_C	95.7525
MSE_train_D	117.9927
MSE_train_m	85.5160
MSE_valid_A	143.4396
MSE_valid_B1	143.4922
MSE_valid_B2	143.9873
MSE_valid_C	139.5976
MSE_valid_D	109.0965
MSE_valid_m	125.3164

图 4-1：上述各算法在 2010 年数据集下的均方误差

可以看到，最小二乘与贝叶斯回归的均方误差几乎一致，这是由于在大量训练数据的冲洗下模型的先验完全消失，采用线性模型的贝叶斯回归的结果收敛到最小二乘所致。广义线性回归仅当选择$\phi(x)=x^2$时误差有小幅下降，但性能都

```
runtime of ols
历时 0.022520 秒。
runtime of Bayes
历时 0.010044 秒。
runtime of Incremental Bayes
历时 0.013218 秒。
runtime of Generalized Bayes
历时 0.676778 秒。
runtime of Gaussian process regression
历时 384.948136 秒。
runtime of fitrgp
历时 0.501809 秒。
```

不如 matlab 自带的 fitrgp。本文实现的高斯过程回归在大范围大量选取初始参数进行优化后能得到优于 fitrgp 的回归结果，但耗时较长。

图 4-2：上述各算法在 2010 年数据集下的耗时

贝叶斯回归相对于最小二乘有一定速度上的优势，且增量贝叶斯学习的方法能够实时更新模型，相对较为实用。

考虑到单纯的 SE 核并不能很好的刻画该模型（两名学生 8 门前置课的成绩之间的相关程度），因此考虑修改核函数，增加超参数的量。我们令第一部分是一个 SE 核，表示两个同学随机过程成绩的相关性随他们前八门课成绩矢量的欧式距离增加而指数平方衰减；第二部分是一个 RQ 核，同样表示随机过程成绩的相关性随他们前八门课成绩矢量的欧式距离增加而幂指数衰减；第三部分为一个 SE 核，不同的是用变换过的成绩矢量代替了原成绩矢量，变换为对每门成绩乘以其序号再平方，这表示序号越大（即越靠后，离随机过程课程学习时间越近）的课程成绩与随机过程成绩相关性越大；第四项为噪声项。实现的核如下：

```
normd = sum((x1-x2).^2);
trend = sum(((x1.*(1:size(x1,2))).^2-(x2.*(1:size(x2,2))).^2).^2);
k1 = para(1)^2*exp(-normd/(2*para(2)^2));
k2 = para(3)^2*(1+normd/(2*para(4)^2))^(-para(5));
k3 = para(6)^2*exp(-trend/(2*para(7)^2));
k4 = para(8)^2*flag;
cov = (k1+k2+k3+k4)/1000000;
```

图 4-3：修改后的核函数 Kernel1

使用该核函数运行回归程序结果如下

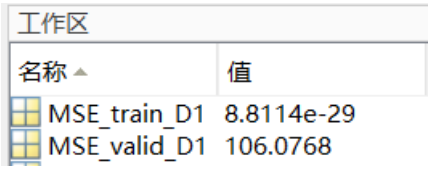
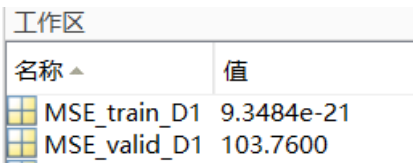


图 4-4：Kernel1 的回归结果

可以看到训练的模型几乎已经完全贴合训练集，对验证集的误差也有所降低。本文接下来又尝试了两种核构造方式，回归的结果如下：

Kernel2:将 1*8 成绩矢量相邻相加得到 4 个标量，用 2 组 4 个标量对应的差构造 4 个 SE 核，其和加上噪声作为核的输出。

Kernel3:直接用 2 组 8 个标量对应的差构造 8 个 SE 核，其和加上噪声作为核的输出。



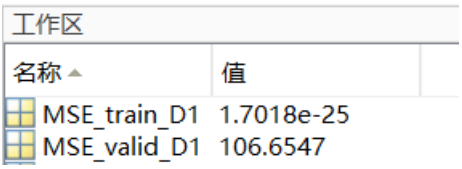


图 4-4：Kernel2/3 的回归结果

限于时间原因，本文没有尝试更多形式的核（核的组合）。针对具体问题，越是能反映变量之间相关程度的核，其回归结果贴合模型的程度越高。

5.结论

最小二乘法算法实现较为简单，但对高维复杂且线性关系不明显的模型拟合能力一般；贝叶斯回归在先验的基础上通过观测数据调整模型，但在先验信息不足的情况下会收敛到最小二乘的结果，增量式贝叶斯能在线学习，对数据的依赖性降低；高斯过程回归能拟合复杂的非线性模型，但比较依赖核的选择，另外由于超参学习是个非凸的优化问题，会给训练模型增加较大的计算量。

在核的选择方法上，应着重考虑如何反映变量之间的相关程度，通过基本的核（SE、RQ、PER、LIN）的组合生成更能反应模型特性的核。本问题中，同时应考虑到高维变量各项与目标量之间关系的强弱不同而针对性的设计核函数。

6.致谢

感谢父母为我提供舒适的宅居环境

感谢余书涵同学给予我极大精神上的关怀。

7.参考文献

[1]C. E. Rasmussen and C. K. I. Williams,Gaussian Processes for Machine Learning. MITPress, 2006.

[2]D. Duvenaud, J. R. Lloyd, R. Grosse, J. B. Tenenbaum, and Z. Ghahramani, “StructureDiscovery in Nonparametric Regression through Compositional Kernel Search,”Proceed-ings of the 30th International Conference on Machine Learning, 2013.

[3] Chen Z. Gaussian process regression methods and extensions for stock market prediction[D]. Department of Mathematics, 2017.

[4]知乎专栏：高斯世界下的 machine learning