高斯过程回归与股票预测

林宝诚 2019080030
linbc19@mails.tsinghua.edu.cn

摘要：本次实验利用典型的随机回归模型 – 高斯过程回归进行股票预测。首要目标是深入学习并总结高斯过程回归模型的思路，并通过动手实践的方式观察模型在股票价格预测上的性能，同时与其他文献中的模型进行对比分析。最后根据结果，提出一些潜在研究方向以进一步改善模型性能。

1. 问题背景

自从上世纪末，利用算法进行股票交易股市已经开始受到专业人士的关注，其热度和普遍性也蒸蒸日上。作为具有高度随机性和波动性的时序数据，对股票价格进行预测的研究工作特别有意思，一方面因其与金钱利益相关；另一方面则是股价的变化是个非常复杂的过程（是人类的决策行为互相交互后的产物），想要进行可信度高的预测是个富有挑战性的任务。对于股市变化是否可预测，至今学术界仍未达到共识，实验中会将股市价格变化视为可预测的过程。由于不具备丰富的金融领域专业知识，本次实验会以学习实践高斯过程回归为主，简单地利用“时间”作为输入特征进行模型训练和预测，而不会将重心放在研究影响股市走向的各个因素。当然，若想要设计出性能更好的模型，使用“时间”作为唯一的特征不太合理，这点在最后的部分会做些探讨。

2. 高斯过程回归理论介绍

初次接触到高斯过程回归时，很难直观地理解高斯过程与回归之间的联系，下面将从两个角度进行理解分析。

- 从回归模型出发（Weight-space View）

已知包含加性高斯噪声（$N(0,\sigma_n^2)$）的线性回归模型的形式如下：

$$y = \vec{x}^T\vec{w} + \varepsilon = f(\vec{x}) + \varepsilon$$

为了对上述模型引入随机性并采用贝叶斯的方法进行预测，令权重 $\vec{w}$ 的先验概率服从零均值高斯分布，即 $\vec{w}\sim N(0,\Sigma_p)$（使用高斯分布是因为其计算方便且在自然界中很常见）。假设X是所有输入特征 $\vec{x}$ 以列的形式排在一起的矩阵，$\vec{y}$为相应函数输出的向量，在获得训练数据（$X,\vec{y}$）后，可以利用贝叶斯定理和高斯分布的计算可得权重的后验概率：

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

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

其中，$A = \frac{1}{\sigma_n^2}XX^T + \Sigma_p^{-1}$，所有概率密度都有$X$作为条件，为了方便已省略。值得一提的是，上述条件下的权重均值也可以由 MAP 的方法导出。

有了权重的后验概率后，就能够对新的数据点$\overrightarrow{x_*}$进行预测：

$$p(f_*|\vec{y}) = \int p(f_*|\vec{w},\vec{y})p(\vec{w}|\vec{y})dw$$

$$f_*|\vec{y}\sim N(\frac{1}{\sigma_n^2}\overrightarrow{x_*}A^{-1}X\vec{y}, \overrightarrow{x_*}^T A^{-1}\overrightarrow{x_*})$$

其中，概率密度条件中省略$X, \overrightarrow{x_*}$

通常情况下，为了避免线性模型带来的局限性，会将输入特征向量 $\vec{x}$ 映射到更高维的空间中变成$\boldsymbol{\phi}(\vec{x})$以引入一些非线性元素，再进行回归，即：

$$y = \boldsymbol{\phi}(\vec{x})^T\vec{w} + \varepsilon = f(\vec{x}) + \varepsilon$$

利用与前述相同的推导思路，并将矩阵A替换掉可得：

$$f_*|\vec{y}\sim N\left(\boldsymbol{\phi}_*^T\Sigma_p\Phi(\mathrm{K}+\sigma_{\mathrm{n}}^2 I)^{-1}\vec{y}, \boldsymbol{\phi}_*^T\Sigma_p\boldsymbol{\phi}_* - \boldsymbol{\phi}_*^T\Sigma_p\Phi(\mathrm{K}+\sigma_{\mathrm{n}}^2 I)^{-1}\Phi^{\mathrm{T}}\Sigma_p\boldsymbol{\phi}_*\right)$$

其中$\Phi(X) = \Phi, K = \Phi^T\Sigma_p\Phi$。上述分布的均值就是预测值，方差则代表此预测值的不确定性。

- 从高斯过程出发（Function-space View）

下面将推导，从高斯过程出发，也能推导与上述一致的结论。高斯过程可以视为函数的分布，其每一条样本轨迹对应着一函数。假设高斯过程为$f(\vec{x})$，回顾高斯过程是由均值和协方差定义：

$$m(\vec{x}) = E[f(\vec{x})]$$

$$k(\vec{x},\vec{x}') = E[(f(\vec{x}) - m(\vec{x}))(f(\vec{x}') - m(\vec{x}'))]$$

已经学习过高斯过程的条件分布为：

$$\overrightarrow{f_*}|\vec{f}\sim N(K(X_*,X)K(X,X)^{-1}\vec{f}, K(X_*,X_*) - K(X_*,X)K(X,X)^{-1}K(X,X_*))$$

其中K为协方差矩阵。要注意的是，如果有引入加性白噪声，协方差会变成$K + \sigma_n^2 I$

与先前推导的预测值分布对比，可以发现若令$K(X_*,X) = \Phi_*^T\Sigma_p\Phi$，两者都是高斯过程的条件分布。借此，就能够理解高斯过程与回归之间的关系。

根据高斯过程定义，需给出初始均值和协方差，即给出函数的先验分布才能进行回归。一般情况下会令均值为 0；协方差函数则有多种选择。需要注意的是，协方差函数就是确保回归性能的关键，接下来对协方差函数进行说明。

- 协方差函数（核函数）

协方差函数涵盖着任意两两个特征向量之间的相关性，正因如此才能够从历史数据中预测未来。通过上述推导可知协方差函数与低维到高维的映射函数$\boldsymbol{\phi}$有关。为了计算方便，不会特别关注$\boldsymbol{\phi}$的具体形式，而是直接定义核函数，此举称为“kernel trick”。

按照常理，在特征空间中，距离越相近的点有越大的相关性；距离越大的点相关性就趋向于 0。常用的核函数包括：

1. Squared Exponential 核函数（也称 RBF）

$$k(\vec{x},\vec{x}') = \sigma^2 \exp\left(-\frac{1}{2l^2}|\vec{x}-\vec{x}'|^2\right)$$

其中σ, l为超参数，分别代表幅度和长度尺度

2. Matern 核函数

$$k(\vec{x},\vec{x}') = \frac{\sigma^2}{\Gamma(\nu)2^{\nu-1}}\left(\frac{\sqrt{2\nu}}{l}|\vec{x}-\vec{x}'|^2\right)^{\nu} K_{\nu}\left(\frac{\sqrt{2\nu}}{l}|\vec{x}-\vec{x}'|^2\right)$$

其中σ, l, ν为超参数，分别代表幅度，长度尺度和曲线的平滑度、$K_{\nu}(*)$为修正贝塞尔函数。Matern 核函数是 squared exponential 核函数的一般化，当$\nu \to \infty$时两者相等。通常令$\nu = \frac{3}{2}$，则核函数变成：

$$k(\vec{x},\vec{x}') = \sigma^2\left(1+\frac{\sqrt{3}}{l}|\vec{x}-\vec{x}'|^2\right)\exp\left(-\frac{\sqrt{3}}{l}|\vec{x}-\vec{x}'|^2\right)$$

在实验中会利用这两个核函数进行回归。最后，由于核函数有超参数可以进行训练调整，下面总结说明如何进行模型训练和预测。

- 回归模型训练 & 预测

已知核函数的超参数为θ，模型的训练是优化θ以将似然函数$p(\vec{f}|X,\theta)$的自然对数对数最大化，即使用最大似然估计。已知其服从高斯过程，不难写出:

$$\log p\left(\vec{f}\middle|X,\theta\right) = -\frac{1}{2}\vec{f}^{T}K^{-1}\vec{f} - \frac{1}{2}\log|K| - \frac{n}{2}\log 2\pi$$

对上述函数使用常用的优化手段（如梯度上升等）就能得到合适的超参数。另外，对新数据点的预测前面已经推导过，最终预测结果就是高斯过程的条件分布$p(\vec{f_*}|\vec{f})$对应的均值。

3. 实验步骤

- 数据收集和预处理

实验的数据来源于 http://www.nasdaq.com/。选择的股票为 adbe, msft 和 visa。每个股价的原始数据都是从 2012 年 1 月 17 日开始直到 2022 年 1 月 14 日，总数据量为 2518（一年内的交易日约为 252 天）。

本次实验中，对原始数据进行了预处理：将原始的股价按月取平均，将股价的日期转换为一维的特征，并取数据前 80%作为训练数据。取平均的操作有平滑数据的作用，也能减少数据量加快训练和预测时间。

另外，需要注意的是，为了方便，先验分布取了均值为 0，而股价的均值实际上不会为 0，且波动范围很大。因此，为了避免模型训练时误认为数据的方差太大，训练前会将数据标准化（零均值，单位方差）。

预处理方法	取平均后数据量	训练数据量	一维特征的算法
按月取平均	121	$121 \times 0.8 \approx 97$	$x = year + \frac{month}{12}$

- 核函数设计

实验中会采用上述介绍的两种核函数进行回归。实验中使用的核函数基本形式为

$$\text{短期核函数} + \text{长期核函数}$$

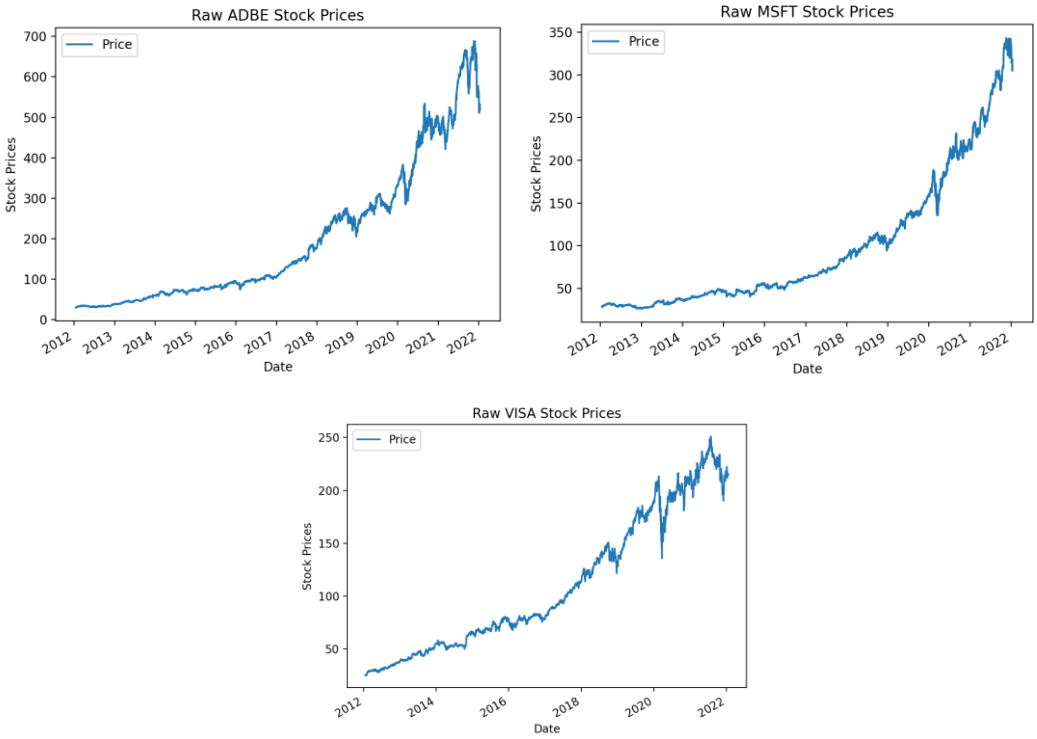
短期与长期核函数的初始长度尺度l分别取 0.1 和 2，前者代表着股价的短期波动，后者代表着股价的长远趋势。至于幅度σ^2则取1.0为初始值。核函数中并没有引入白噪声因为股价并不存在噪声如测量误差等。为了简便，将超参数表示为（长期幅度，长度尺度；短期幅度，长度尺度）：

$$\text{初始参数} = (1, 2;\ 1, 0.1)$$

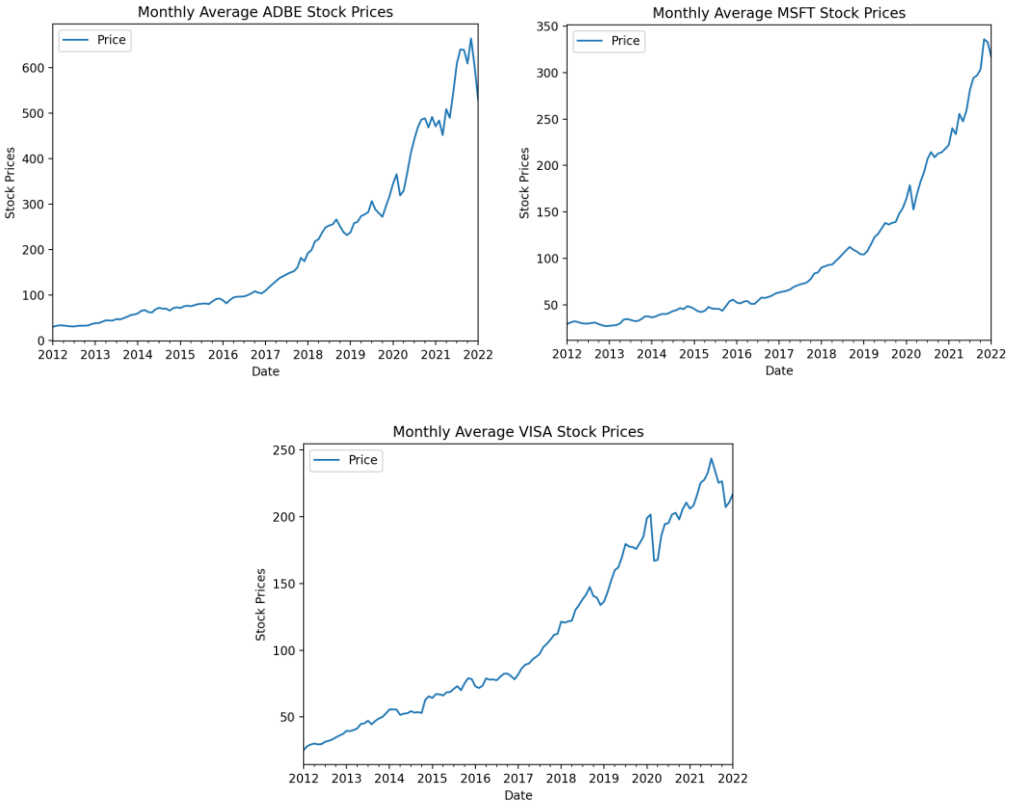
4. 实验结果 & 性能分析与对比

- 实验结果与图片

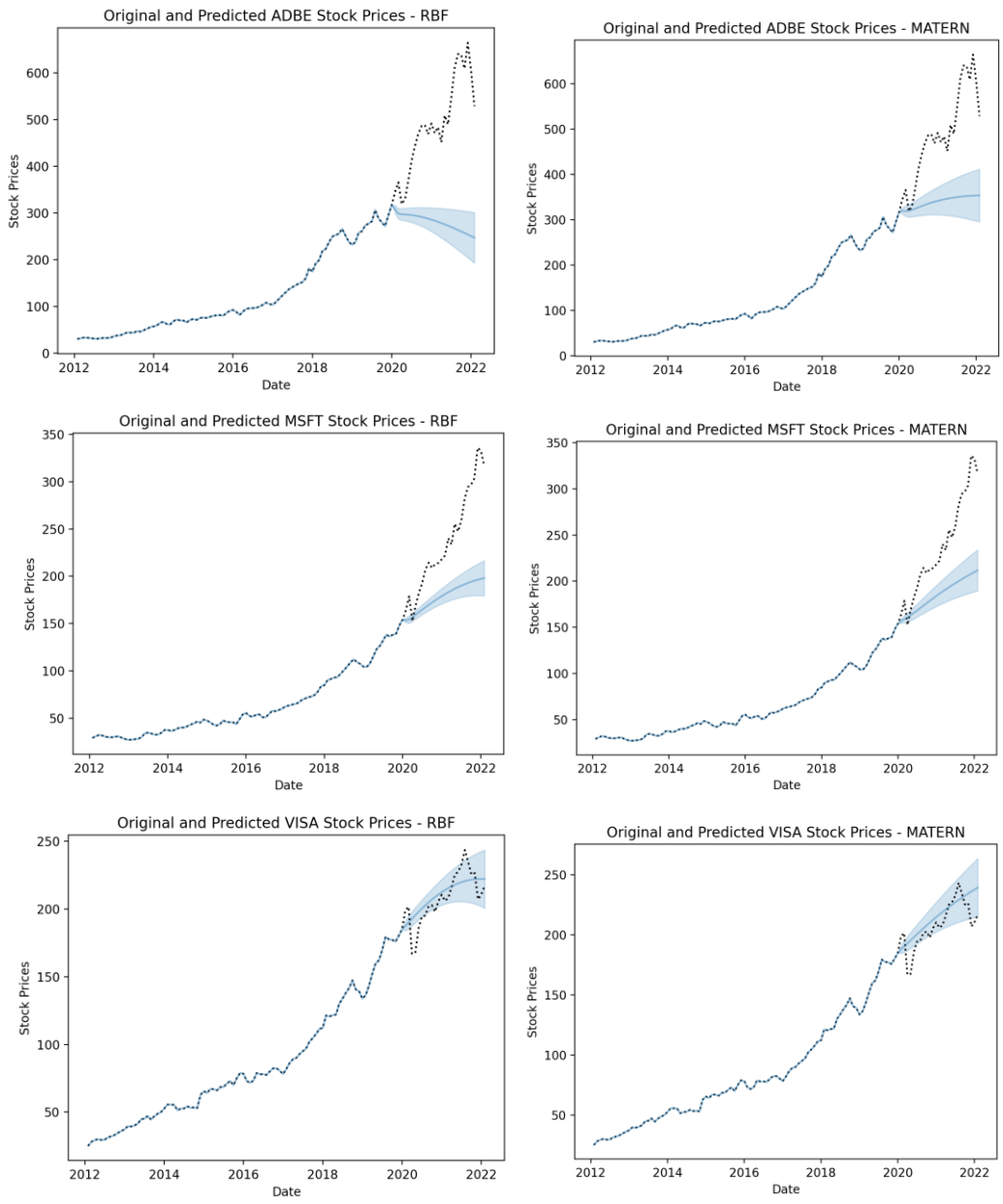
股票	训练后参数； R^2（Squared Exp）	训练后参数； R^2（Matern）
$ADBE$	$(1.21, 2.17;\ 0.01, 0.07); -4.91$	$(3.76, 7.09;\ 0.0104, 0.12); -2.13$
$MSFT$	$(4.71, 3.43;\ 0.01, 0.08); -0.80$	$(20.88, 15.6;\ 0.01, 0.19); -0.46$
$VISA$	$(3.17, 3.33;\ 0.01, 0.08); 0.57$	$(14.21, 15.4;\ 0.01, 0.14); -0.46$



图【1-3】原始股票数据



图【4-6】取月平均后的股票数据



图【7-12】各股票的预测结果

- 性能分析对比

使用 RBF 核函数的结果中，ADBE 股票的预测结果差距最大，股价走势也预测错误。其余两个股票的预测走势都符合实际情况，而 VISA 股价的预测相对较准确。另一方面，使用 Matern 核函数的模型测试结果总体来说较好，对于每一个股票都有判断出正确的趋势。但是，与 RBF 核函数相同，ADBE 和 MSFT 这两个股价的预测值与实际值差距较大。

与文献[3]的模型相比，文献[3]使用的数据预处理是将数据按年区成各个独立的时间序列，然后使用年份与交易天作为特征。这种处理方法似乎是假设股价每一年的走势都

是类似的。无论如何，文献[3]的预测结果与本次实验结果相似，都是走势判断合理但是股票实际值相差较大。

实际上由于初始均值设定为常数，且模型是平稳的，若预测的间隔很长，间隔太远的股价相关性低，股价就很大可能会返回均值，因此实验中的模型比较适合使用在短期预测上。如果想解决此问题，需要补充学习些时间序列处理的知识（即将“趋势”抽出使用非平稳模型描述，波动则使用平稳模型描述）。

- 算法复杂度

本次实验中使用的是 Python 中的 scikit-learn 库，根据文档 https://scikit-learn.org/stable/modules/gaussian_process.html，其采用的是文献[1]中的算法 2.1。此算法的瓶颈在于协方差矩阵求逆（使用 cholesky 分解），复杂度为$O(n^3)$，其中n为训练数据量。另外，由于需要储存训练数据的协方差矩阵，因此空间复杂度为$O(n^2)$。文献[2]和[3]使用的也是最基本的算法，因此复杂度与本次实验的一致。

补充：Cholesky 分解原理

对于厄米正定矩阵A（协方差矩阵为对称非负定，主要关注正定的情况），其可分解为：

$$A = LL^H$$

其中，L为下三角矩阵，其对角元素皆为正实数（所以逆矩阵存在）。

对于线性方程$A\vec{x} = \vec{b}$，则只需依次求解$L\vec{y} = \vec{b}$，和$L^H\vec{x} = \vec{y}$，即可得到$\vec{x}$。由于这两个方程都是三角矩阵，因此求解非常直接。

Cholesky 分解的复杂度为$O(n^3)$，三角矩阵求解复杂度为$O(n^2)$，因此总复杂度为$O(n^3)$

5. 后续潜在研究方向

高斯过程回归可研究的内容非常丰富，与之紧密相关的还有时间序列分析、机器学习等内容。除了模型预测性能，如何加速高斯过程回归的训练和预测时间也有许多可以学习的内容。受限于时间，本次实验只完成高斯过程回归理论的学习和其在股票价格预测上的应用，还有许多细节部分待研究，接下来列出一些可以完善的细节。

- 特征向量$\vec{x}$的选取

实验中仅使用“时间”作为特征的回归模型与现实情况不符。基本上，股票价格的高低起伏取决于许多因素，最主要的就是企业的运营情况和前景，而“时间”这一特征

并没有涵盖这两个信息。若能将代表当前运营情况和前景好坏的指标纳入特征中，回归模型可能会有显著的性能提升。

- 核函数的选取 & 时间序列分析基础

  选择核函数是回归中关键的一环。要想选取合适核函数并不简单直接，因为除了需要对核函数的特性有充分了解之外，也要了解时序数据中的大致变化规律。很多时候，需要尝试许多不同核函数组合才能够得到满意的结果。参考文献[5]中提出了一个自动搜寻适合核函数的框架，值得进一步深究。另外，也需要补充时间序列方面的知识（其实是一门课，之后有选修的打算），以了解处理时间序列时的常规方法和思路，如此一来可以帮助设计更好的回归模型。

- 高斯过程回归的加速运算

  上面已经分析得知高斯过程回归的瓶颈在于矩阵求逆的复杂度高，不适合训练数据数量较多的情况，因此有许多学者提出了各种近似方法，以加速高斯回归模型的训练与预测。学习设计更快速的算法可以令高斯过程回归更具有实用性（如同 FFT 的情况），总体来说是个值得研究的方向。

6. 总结与感想

总结地说，高斯过程回归相比起一般回归引入了随机性，其有价值的地方在于可以得知预测结果的不确定性。从实验结果可以看到，同一形式的模型在不同股票上的回归效果有较大差别，因此或许需要针对各自股票的特点设计相应模型，并实时为模型输入新的信息才能得到更好的预测结果。当然，也有可能股价实际上是无法预测的（学术界中有这个观点）。反正，这一系列疑惑都只能在继续深入研究后才可能获得解答。

通过本次大作业，我体验了如何将理论知识应用在现实生活上。除了复习随机过程中的高斯过程外，我也加强了贝叶斯预测、回归等方面的知识，因此大作业无疑让我获益良多。有些遗憾的是，实验结果不是特别好，有很大进步空间，因此在接下来的日子中，在补充了更多相关的知识后，我想进行多几次尝试。无论如何，本学期的随机过程课程到此已经正式完结。这门课很有挑战性，但是也特别实用和有趣，在此感谢李刚老师和助教们的一学期付出。

参考文献：

[1] C.E. Rasmussen and K. I. Williams. (2006). Gaussian processes for machine learning.

[2] Farrell, Todd & Correa, Andrew. (2007). Gaussian Process Regression Models for Predicting Stock Trends.

[3] Anonymous. (n.d.). Long-term Stock Market Forecasting using Gaussian Processes.

[4] Ebden, Mark. (2015). Gaussian Processes for Regression and Classification: A Quick Introduction.

[5] Duvenaud, David. (2014). Automatic model construction with Gaussian processes.