高斯过程回归

前言：在2018C中关于1B的解法中，某团队使用了高斯过程回归预测模型（GPR，得到函数f(x)的分布），旨在展现出历史能源的演变。

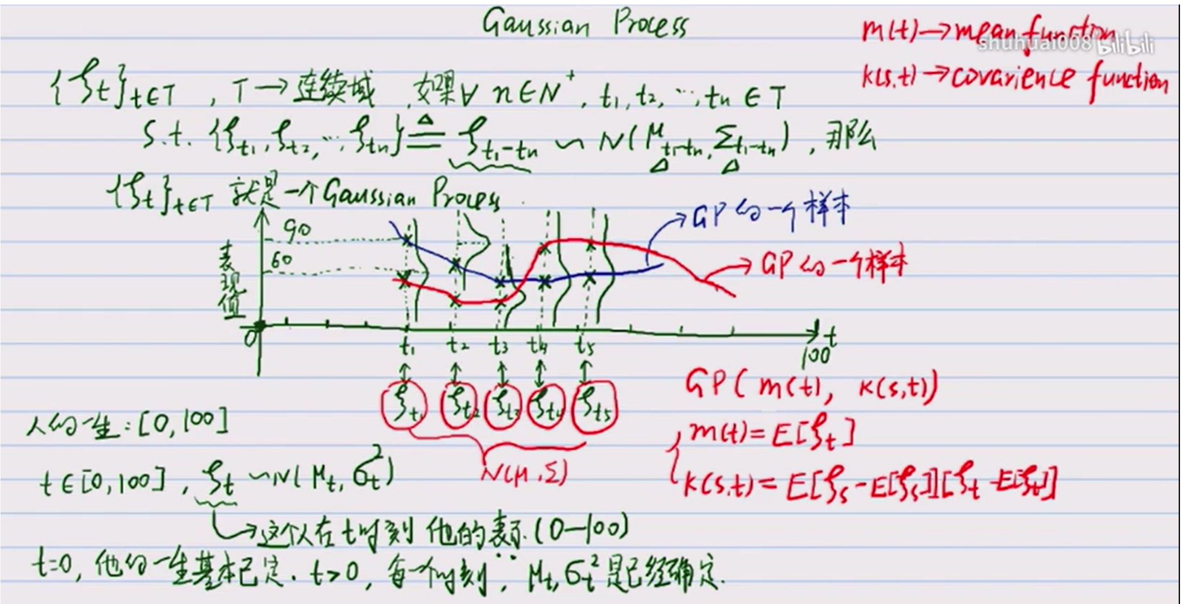
指标---->GPR得到目标值的分布--->演变过程。

数学推导

推导视频

高斯过程是无限维的高斯分布

1、高斯过程解释（一个随机过程）



2、高斯过程回归（基于贝叶斯分布）

前提：变量满足高斯分布

两个视角

- 权重空间
- 函数空间（更容易理解）

GPR:

① weight-space view 关注的是 w

② function-space view：关注的是 f(x)

$x^*, \longrightarrow y^*$

$P(y^*|Data, x^*) = \int_w P(y^*|w, x^*) \cdot P(w) dw$

$P(y^*|Data, x^*) = \int P(y^*|f, x^*) \cdot P(f) df$

回归过程

对于数据集 $D:(X,Y)$，令 $f(x^i)=y_i$，从而得到向量$f=[f(x_1),f(x_2),...,f(x_n)]$，将所需要预测的$x_i$的集合定义为$X*$，对应的预测值为$f*$，根据贝叶斯公式有：

$$p(f*|f)=\frac{p(f|f*)p(f*)}{p(f)}=\frac{p(f,f*)}{p(f)}$$

高斯回归首先要计算数据集中样本之间的联合概率分布，$f\sim N(\mu,K)$，μ为$f(x_1),f(x_2),...,f(x_n)$的均值所组成的向量，K为其协方差矩阵，再根据需要预测的$f*$的先验概率分布 $f*\sim N(\mu*,K*)$与$f\sim N(\mu,K)$，来计算出$f*$的后验概率分布。

1、计算p(f)的先验分布。

原始数据、协方差矩阵（选择适合的核函数）

（1）协方差矩阵必须是半正定阵

（2）kernel fountion都是半正定阵。这就意味着我们在学习SVM的时候所学过的核函数形式都可以用

当然应用最广的是RBF kernel，即如下式：

$$k(x,x')=\alpha^2exp(-\frac{1}{2l^2})(x-x')^2)$$

其中α为超参数，l是需要通过学习进行确定的参数，那么我们只需要通过监督学习的方式学习到合适的kernel，即可很方便的计算出f的协方差矩阵。

如何学习kernel的参数？很简单kernel $k(x,x')$ 优劣的评价标准就是在要$f(x)\sim N(m(x),k(x,x^T))$的条件下，让$p(Y|X)$最大，为了方便求导我们将目标函数设为$logp(Y|X)=logN(\mu,K_y)$，接下来利用梯度下降法来求最优值即可：

$$\frac{\partial logp(Y|X)}{\theta}=\frac{1}{2}y^TK_y^{-1}\frac{\partial K_y}{\theta}K_y^{-1}y-\frac{1}{2}tr(K_y^{-1}\frac{\partial K_y}{\theta})$$

2、联合概率分布的先验概率

已知 $f(x)\sim N(\mu,K)$，$f(x*)\sim N(\mu*,K(x*,x*))$，可计算其其联合概率分布的先验：

$$\binom{f}{f_*}\sim(\begin{pmatrix}\mu\\ \mu*\end{pmatrix},\begin{pmatrix}K & K*\\ K*^T & K**\end{pmatrix})$$

其中$K**$为 $f(x*)$的协方差矩阵，$K**=k(X*,X*)$，$K*=k(X,X*)$

3、根据贝叶斯公式，求后验概率并估算

有了$p(f)$的先验分布，以及上面计算的$p(f,f*)$，根据贝叶斯公式可以计算 $p(f*|f)$的后验概率:

$$p(f*|f)=\frac{p(f|f*)p(f*)}{p(f)}=\frac{p(f,f*)}{p(f)}$$

从而得出对于f*的估计，$f*\sim(\mu',K')$

$$\mu'=K^TK^{-1}f$$

$$K'=K*^T\ K^{-1}K*+K**$$

依据原始数据和预测数据之间的协方差，判断y值的差异性，根据概率公式得到预测值

适用范围

一般回归算法给定输入X，希望得到的是对应的Y值；但是该高斯过程回归的目标是求出y的分布，因此可以适用于求解/预测某变量变化/演化情况的题目。

实现

可以调用sklearn库中的gaussian_process库

灰色关联分析

适用范围

用于系统分析：探究对于一个系统，哪些是印象其发展的主要/次要因素。

使用步骤

第一步：确定分析数列
确定反映系统行为特征的参考数列和影响系统行为的比较数列。反映系统行为特征的数据序列，称为参考数列。影响系统行为的因素组成的数据序列，称比较数列。
进行分析的母序列、子序列
第二步，变量的无量纲化
由于系统中各因素列中的数据可能因量纲不同，不便于比较或在比较时难以得到正确的结论。因此在进行灰色关联度分析时，一般都要进行数据的无量纲化处理。主要有一下两种方法：

（1）初值化处理：$x_i(k)=\frac{x_i(k)}{x_i(1)},k=1,2...n;i=0,1,2...m$

（2）均值化处理：$x_i(k)=\frac{x_i(k)}{\bar{x}_i},k=1,2...n;i=0,1,2...m$

其中k 对应时间段，i 对应比较数列中的一行（即一个特征）

第三步，计算关联系数

$$\xi_i(k)=\frac{\min_i\min_k|y(k)-x_i(k)|+\rho\max_i\max_k|y(k)-x_i(k)|}{|y(k)-x_i(k)|+\rho\max_i\max_k|y(k)-x_i(k)|}$$

记$\Delta_i(k)=|y(k)-x_i(k)|$，则

$$\xi_i(k)=\frac{\min_i\min_k\Delta_i(k)+\rho\max_i\max_k\Delta_i(k)}{\Delta_i(k)+\rho\max_i\max_k\Delta_i(k)}$$

ρ 代表 分辨系数，一般取0.5

第四步，计算关联度
因为关联系数是比较数列与参考数列在各个时刻（即曲线中的各点）的关联程度值，所以它的数不止一个，而信息过于分散不便于进行整体性比较。因此有必要将各个时刻（即曲线中的各点）的关联系数集中为一个值，即求其平均值，作为比较数列与参考数列间关联程度的数量表示，

对系数逐一累加再求均值

第五步:关联度排序,找到相似度最大的两序列

ARMA时序模型

[参考链接]

```
from statsmodels.tsa.arima_model import ARMA
```

本质就是用前期的数据预测未来短时间内的数据

重点是对数据的平滑处理和噪音检测，分解出数据的趋势和周期性数据，