Gaussian Process

qhy

2017 年 8 月 18 日

目录

1 Gaussian Process

1.1 数学基础

在GP中需要用到的数学知识,这里给出结论,证明见PRML第二章。

1.1.1 高斯分布

给定一个联合高斯分布$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})$,其中$\boldsymbol{\Lambda}\equiv\boldsymbol{\Sigma}^{-1}$,且

$$\boldsymbol{x}=\begin{pmatrix}\boldsymbol{x_a}\\ \boldsymbol{x_b}\end{pmatrix},\boldsymbol{\mu}=\begin{pmatrix}\boldsymbol{\mu_a}\\ \boldsymbol{\mu_b}\end{pmatrix} \tag{1}$$

$$\boldsymbol{\Sigma}=\begin{pmatrix}\boldsymbol{\Sigma_{aa}} & \boldsymbol{\Sigma_{ab}}\\ \boldsymbol{\Sigma_{ba}} & \boldsymbol{\Sigma_{bb}}\end{pmatrix},\boldsymbol{\Lambda}=\begin{pmatrix}\boldsymbol{\Lambda_{aa}} & \boldsymbol{\Lambda_{ab}}\\ \boldsymbol{\Lambda_{ba}} & \boldsymbol{\Lambda_{bb}}\end{pmatrix} \tag{2}$$

有以下两个结论:

条件概率分布:

$$p(\boldsymbol{x_a}|\boldsymbol{x_b})=\mathcal{N}(\boldsymbol{x_a}|\boldsymbol{\mu_{a|b}},\boldsymbol{\Lambda_{aa}^{-1}}) \tag{3}$$

$$\boldsymbol{\mu_{a|b}}=\boldsymbol{\mu_a}-\boldsymbol{\Lambda_{aa}^{-1}\Lambda_{ab}}(\boldsymbol{x_b}-\boldsymbol{\mu_b}) \tag{4}$$

或

$$\boldsymbol{\mu_{a_b}}=\boldsymbol{\mu_a}+\boldsymbol{\Sigma_{ab}\Sigma_{bb}^{-1}}(\boldsymbol{x_b}-\boldsymbol{\mu_a}) \tag{5}$$

$$\boldsymbol{\Sigma_{a|b}}=\boldsymbol{\Sigma_{aa}}-\boldsymbol{\Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}} \tag{6}$$

边缘概率分布:

$$p(\boldsymbol{x_a})=\mathcal{N}(\boldsymbol{x_a}|\boldsymbol{\mu_a},\boldsymbol{\Sigma_{aa}}) \tag{7}$$

这里证明了边缘概率和联合分布之间的关系,当我们在求边缘概率的时候,就可以通过求联合分布得到(这里在计算的时候省略的积分的过程,但实际的证明就是通过积分证明得到的。)

1.1.2 高斯变量的贝叶斯定理

给定$\boldsymbol{x}$的一个边缘高斯分布,以及在给定$\boldsymbol{x}$的条件下的$\boldsymbol{y}$的条件高斯分布,形式为:

$$p(\boldsymbol{x})=\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Lambda^{-1}}) \tag{8}$$

$$p(\boldsymbol{y}|\boldsymbol{x})=\mathcal{N}(\boldsymbol{y}|\boldsymbol{Ax}+\boldsymbol{b},\boldsymbol{L^{-1}}) \tag{9}$$

$\boldsymbol{y}$的边缘分布以及给定$\boldsymbol{y}$条件下的$\boldsymbol{x}$的条件分布为:

$$p(\boldsymbol{y})=\mathcal{N}(\boldsymbol{y}|\boldsymbol{Ax}+\boldsymbol{b},\boldsymbol{L^{-1}}+\boldsymbol{A\Lambda^{-1}A^T}) \tag{10}$$

$$p(\boldsymbol{x}|\boldsymbol{y})=\mathcal{N}(\boldsymbol{x}|\Sigma\{\boldsymbol{A^TL}(\boldsymbol{y}-\boldsymbol{b})+\boldsymbol{\Lambda\mu}\},\boldsymbol{\Sigma}) \tag{11}$$

其中,

$$\boldsymbol{\Sigma}=(\boldsymbol{\Lambda}+\boldsymbol{A^TLA})^{-1} \tag{12}$$

1.2 随机过程与高斯过程

随机过程:许多随机变量的集合

高斯过程:许多高斯随机变量的集合

e.g. 给定n个随机变量:$\boldsymbol{y} = \{y_1, y_2, \cdots, y_n\}$,其中$y_i \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

高斯随机过程就是:

$$\begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_n \end{bmatrix} \tag{13}$$

没错,就是这些随机变量的集合。在求这个随机过程的分布的时候,实际上就是求这些随机变量的联合分布。然后,随机过程到底结束。

1.3 高斯过程回归 Gaussian Process Regression

1.3.1 无噪声GPR

问题描述:给定样本$\mathcal{D} = \{(x_1, y_1), \cdots, (x_n, y_n)\}$,和待预测数据$x_*$,相似性矩阵$\boldsymbol{K}$,求$p(y_*|\mathcal{D})$

高斯过程回归:假设目标函数是一条平滑的曲线,在相近的之间,其输出也相近。也就是相近的随机变量(它们采样的值也相近),而这个距离则是通过x来确定的,这就保证的相邻的x之间的曲线的平滑度(曲线不会突然飙到一个很大或者很小的值)。

下面看高斯过程回归的具体过程:

把每个样本的y_i当作随机变量,并且假设这些随机变量满足高斯分布(即,这些随机变量组成高斯过程)。

i.e. $y_i \sim \mathcal{N}(\mu_i, \Sigma_{ii})$

则这个高斯过程满足以下条件:由前面的数学公式可以得到

$$\boldsymbol{y} \triangleq \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{K}) \tag{14}$$

注:$\triangleq$表示定义的意思

其中,

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \\ \vdots \\ \boldsymbol{\mu}_n \end{bmatrix} \tag{15}$$

其中$\boldsymbol{K}$为相似性矩阵,用来描述随机变量之间的距离。

$$K_{nm} = k(\boldsymbol{x_n}, \boldsymbol{x}_m) = \frac{1}{\alpha}\phi(x_n)^T\phi(x_m) \tag{16}$$

根据前面随机变量的定义,有

$$K_{ii} = \Sigma_{ii} \tag{17}$$

给定x_*,假设

$$\begin{bmatrix} \boldsymbol{y} \\ y_* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu} \\ \mu_* \end{bmatrix}, \begin{bmatrix} \boldsymbol{K} & \boldsymbol{K_*} \\ \boldsymbol{K_*^T} & k_{**} \end{bmatrix} \right) \tag{18}$$

其中,

$$\boldsymbol{K_*^T} = [k(\boldsymbol{x}_1, \boldsymbol{x}_*), \cdots, k(\boldsymbol{x}_n, \boldsymbol{x}_*)] \tag{19}$$

根据第一部分的数学知识,我们有以下结果:

$$p(y_*|\boldsymbol{y}) = \mathcal{N}(\tilde{\mu}, \tilde{\Sigma}) \tag{20}$$

其中,

$$\tilde{\mu} = \mu_* + \boldsymbol{K}_*^T \boldsymbol{K}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}) \tag{21}$$

$$\tilde{\Sigma} = k_{**} - \boldsymbol{K}_*^T \boldsymbol{K}^{-1} \boldsymbol{K}_* \tag{22}$$

一般地,我们对均值没有任何的先验知识,根据对称性,我们取$\mu = \mu_* = \boldsymbol{0}$

1.3.2 带噪声高斯过程回归

考虑目标观测值的噪声,其形式为:

$$t_i = y_i + \epsilon_i \tag{23}$$

其中,ϵ_i是一个噪声随机变量,它的值对每个观测n是独立的。满足以下形式:

$$\epsilon_i \sim \mathcal{N}(0, \beta^{-1}) \tag{24}$$

因为ϵ_i和y_i是独立的,因此,对于ϵ_i来说可以把y_i看作常量 我们有以下形式的高斯分布的噪声过程:

$$p(t_i|y_i) = \mathcal{N}(t_i|y_i, \beta^{-1}) \tag{25}$$

其中,β^{-1}是一个超参数,表示噪声的精度。

考虑多个样本,以$\boldsymbol{y} = (y_1, \cdots, y_n)$为条件,以$\boldsymbol{t} = (t_1, \cdots, t_n)$为目标值的联合概率分布。这个分布是一个各向同性的高斯分布,形式为:

$$p(\boldsymbol{t}|\boldsymbol{y}) = \mathcal{N}(\boldsymbol{t}|\boldsymbol{y}, \beta^{-1}\boldsymbol{I}_n) \tag{26}$$

根据高斯过程的定义:上一节的内容 $p(\boldsymbol{y}) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{0}, \boldsymbol{K})$

为了找到以输入值$\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$(即相似性矩阵$\boldsymbol{K}$)为条件的边缘概率分布$p(\boldsymbol{t})$,我们需要对$y$积分:

$$p(\boldsymbol{t}) \triangleq \int p(\boldsymbol{t}|\boldsymbol{y}) p(\boldsymbol{y}) d\boldsymbol{y} = \mathcal{N}(\boldsymbol{t}|\boldsymbol{0}, \boldsymbol{C}) \tag{27}$$

这里计算的时候,并没有进行积分,只是求对应的联合分布,在联合分布的表达中,找到对应的值(这个对应关系的证明是直接积分得到的)

其中协方差矩阵$\boldsymbol{C}$为: $C = K + \beta^{-1}\boldsymbol{I}_n$

给定x_*,则对应输出y_*与样本$\boldsymbol{t}$的联合概率分布为:

$$\begin{bmatrix} \boldsymbol{t} \\ y_* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{0} \\ 0 \end{bmatrix}, \begin{bmatrix} \boldsymbol{C} & \boldsymbol{K_*} \\ \boldsymbol{K_*^T} & k_{**} \end{bmatrix} \right) \tag{28}$$

则$p(y_*|\boldsymbol{t})$为高斯分布,形式为:形式和无噪声的情况是一样的,只不过$\boldsymbol{K}$变成了$\boldsymbol{C}$

$$p(y_*|\boldsymbol{t}) \sim \mathcal{N}(\boldsymbol{K}_*^T \boldsymbol{C}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}), k_{**} - \boldsymbol{K}_*^T \boldsymbol{C}^{-1} \boldsymbol{K}_*) \tag{29}$$

也有求t_*的分布而不是y_*的分布,过程类似,只不过对应的方差k_{**}变成了$c = k_{**} + \beta^{-1}$

1.4 高斯过程的图解

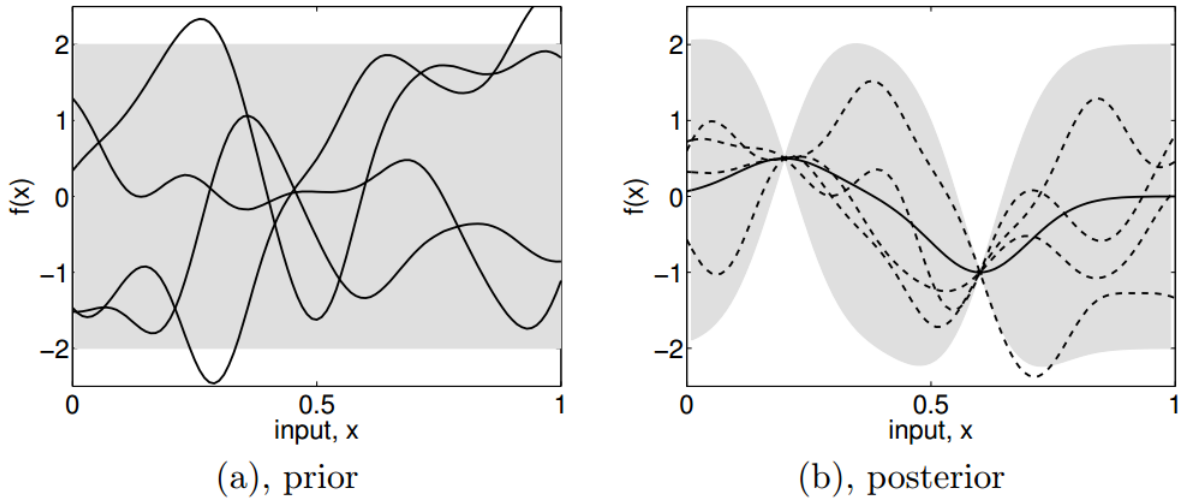

图 1: Panel (a) shows four samples drawn from the prior distribution. Panel (b) shows the situation after two datapoints have been observed. The mean prediction is shown as the solid line and four samples from the posterior are shown as dashed lines. In both plots the shaded region denotes twice the standard deviation at each input value x.

上图是高斯过程的一个图解: 在先验概率的条件下,采样空间有无数个多的函数可以选择,
在后验概率的条件下,则采样的范围就会缩小
而在高斯过程中,我们进一步把采样的范围缩小到高斯分布的范围内。
这里的函数的采样,具体地看,就是对函数的参数$\boldsymbol{w}$的采样,而参数也满足高斯分布,因此,参数是从高斯分布采样得到的。(具体见贝叶斯回归与高斯过程的关系)

1.5 超参数的学习

在高斯过程中,我们抛弃参数模型,直接定义函数上的先验概率分布(具体表现为函数值y_i的概率分布), 而这个分布则取为高斯分布,确定一个高斯分布,只需要找到它的均值和方差即可。

1.5.1 均值

前面已经提到,在大部分应用中,我们对于函数值y_i的均值没有任何的先验知识,因此根据对称性,我们取均值为0

1.5.2 协方差矩阵$\boldsymbol{K}$

$\boldsymbol{K}$矩阵叫相似性矩阵,用来描述随机变量之间的相似性,
i.e. 对于相似的$\boldsymbol{x}_n$和$\boldsymbol{x}_m$,对应的函数值$y(x_n)$和x_m的相关性要大于不相似的点。这里相似性的概念取决于实际应用。
对于高斯过程回归,一个广泛使用的和函数的形式为指数项的二次型加上常数项和线性项,i.e.

$$k(\boldsymbol{x}_n, \boldsymbol{x}_m) = \theta_0 exp\{-\frac{\theta_1}{2}\|\boldsymbol{x}_n - \boldsymbol{x}_m\|\} + \theta_2 + \theta_3 \boldsymbol{x}_n^T \boldsymbol{x}_m \tag{30}$$

1.5.3 学习超参数

学习超参数的方法基于计算似然函数$p(\boldsymbol{t}|\boldsymbol{\theta})$,其中$\boldsymbol{\theta}$表示高斯过程模型的超参数。最简单的方法是通过最大化似然函数的方法进行$\boldsymbol{\theta}$的点估计。

使用**多元高斯分布**的标准形式,高斯过程模型的对数似然函数形式为:

$$\ln p(\boldsymbol{t}|\boldsymbol{\theta}) = -\frac{1}{2}\ln|\boldsymbol{C}| - \frac{1}{2}\boldsymbol{t}_T\boldsymbol{C}^{-1}\boldsymbol{t} - \frac{n}{2}\ln 2\pi \tag{31}$$

1.6 自动相关性确定 automatic relecance determination ARD

上一节用似然函数求参数的方法,这里我们通过为每一个输入变量整合一个单独的参数,在通过最大似然函数方法进行参数的最优化,就能够自动地将不同输入的相对重要性从数据中推断出来。

e.g.考虑二维输入空间$\boldsymbol{x} = (x_1, x_2)$,有以下形式的核函数:

$$k(\boldsymbol{x}, \boldsymbol{x}') = \theta_0 exp\frac{1}{2}\sum_{i=1}^{2}\eta_i(x_i - x_i')^2) \tag{32}$$

这里的相关性,某个属性的相关性,而不是某个输入样本的相关性

1.7 怎样从高斯分布中采样

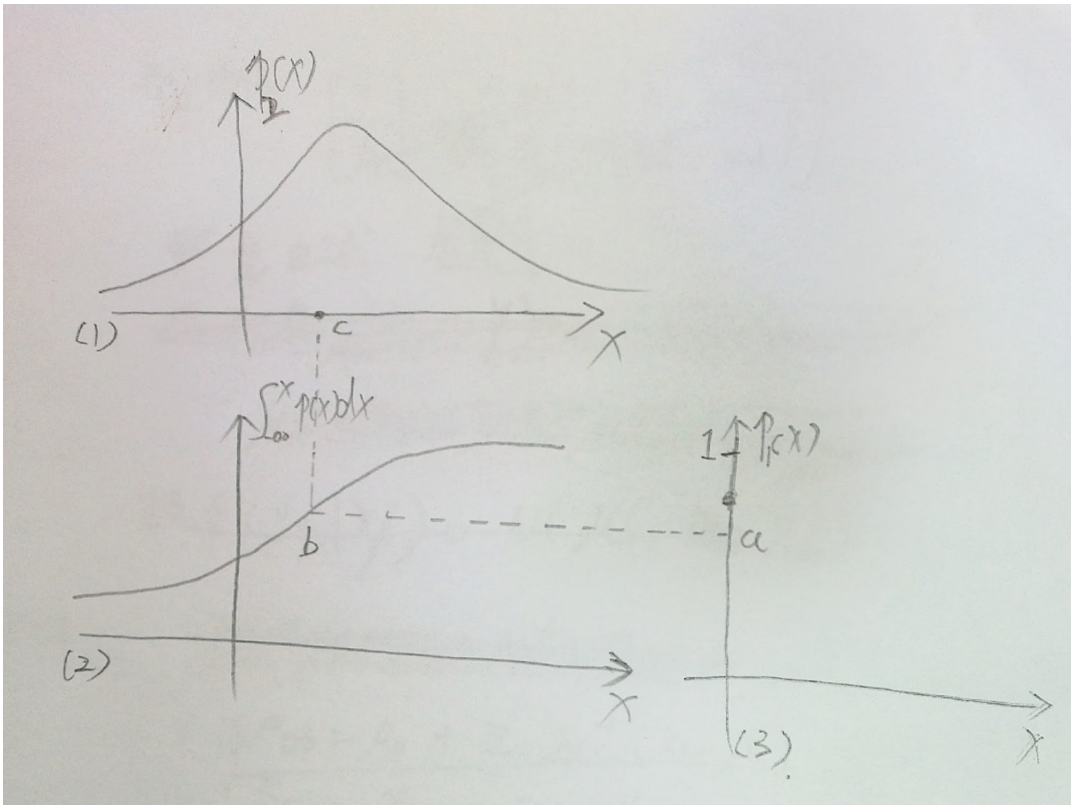

图 2: (1):表示高斯分布,(2)表示从负无穷到x曲线下的面积,(3)表示均匀分布

采样过程:从(3)的均匀分布中采样得到点a,点a对应面积曲线中的点b,点b对应到高斯分布的点c,点c就是高斯分布采样的一个结果。

1.8 与神经网络的关系

在贝叶斯神经网络中,定义参数$\boldsymbol{w}$上的先验分布以及网络函数$f(\boldsymbol{x},\boldsymbol{w})$产生了函数$y(\boldsymbol{x})$上的先验概率分布,其中$\boldsymbol{y}$表示网络的输出向量。

在$M\to\infty$的情况下,神经网络产生的函数的分布将会趋于高斯过程。

然而,值得注意的是,在这种期限情况下,神经网络的输出将会变得相互独立。神经网络的优势之一是输出之间共享隐含单元,因此它们可以互相"借统计优势",即,每个隐含节点关联的权值将被所有的输出变量影响,而不只是被它们中的某一个影响。这个性质在极限的高斯过程中丢失了。

1.9 与岭回归的关系

岭回归的优化目标是:

$$\min_{\boldsymbol{\theta}} \|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\theta}\|+\delta^2\|\boldsymbol{\theta}\|^2 \tag{33}$$

有解析解:

$$\boldsymbol{\theta}=\boldsymbol{X}^T\boldsymbol{\alpha} \tag{34}$$

其中,

$$\boldsymbol{\alpha}=(\boldsymbol{X}\boldsymbol{X}^T+\delta^2\boldsymbol{I}_n)^{-1}\boldsymbol{y} \tag{35}$$

那么给定$\boldsymbol{x}_*$,其预测值y_*为:

$$y_*=\boldsymbol{x}_*\boldsymbol{\theta}=\boldsymbol{x}_*\boldsymbol{X}^T\boldsymbol{\alpha}=\boldsymbol{x}_*\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T+\delta^2\boldsymbol{I}_n)^{-1}\boldsymbol{y}=\boldsymbol{K}_*^T\boldsymbol{C}^{-1}\boldsymbol{y} \tag{36}$$

即,最后预测的值和高斯过程中,y_*分布的均值是相等的。

这里体现了高斯过程的一个有点,高斯过程不仅能得到岭回归下的预测结果(均值),还能得到对于每个y_*预测的上下边界。

1.10 与贝叶斯线性回归模型的关系

贝叶斯线性回归模型:具体见PRML3.3章 引入$\boldsymbol{w}$的先验概率分布,形式为:

$$p(\boldsymbol{w})=\mathcal{N}(\boldsymbol{w}|\boldsymbol{m}_0,\boldsymbol{S}_0) \tag{37}$$

在输入$\boldsymbol{t}$下的后验概率分布形式为:

$$p(\boldsymbol{w}|\boldsymbol{t})=\mathcal{N}(\boldsymbol{w}|\boldsymbol{m}_N,\boldsymbol{S}_N) \tag{38}$$

其中,

$$\boldsymbol{m}_N=\boldsymbol{S}_N(\boldsymbol{S}_0^{-1}\boldsymbol{m}_0+\beta\boldsymbol{\Phi}^T\boldsymbol{t}) \tag{39}$$

$$\boldsymbol{S}_N^{-1}=\boldsymbol{S}_0^{-1}+\beta\boldsymbol{\Phi}_T\boldsymbol{\Phi} \tag{40}$$

预测分布:也是求在输入和超参的条件下,预测值的分布

$$p(t|\boldsymbol{t},\alpha,\beta)=\int p(t|\boldsymbol{w},\beta)p(\boldsymbol{w}|\boldsymbol{t},\alpha,\beta)d\boldsymbol{w} \tag{41}$$

预测分布的形式为:

$$p(t|\boldsymbol{t},\alpha,\beta)=\mathcal{N}(t|\boldsymbol{m}_N^T\boldsymbol{\phi}(\boldsymbol{x}),\sigma_N^2(\boldsymbol{x})) \tag{42}$$

其中,

$$\sigma_N^2*(\boldsymbol{x})=\frac{1}{\beta}+\boldsymbol{\phi}(\boldsymbol{X})^T\boldsymbol{S}_N\boldsymbol{\phi}(\boldsymbol{x}) \tag{43}$$

因此,贝叶斯线性回归模型是高斯过程的一个具体例子。

1.11 参考资料

- PRML 2.3,3.35.7,6.4,8.1
- Gaussian Processes for Machine Learning
- 【机器学习技术】高斯过程初探
- 如何通俗易懂地介绍 Gaussian Process？-知乎
- 贝叶斯线性回归（Bayesian Linear Regression）
- Machine learning - Introduction to Gaussian processes-UBC