Gaussian Process 03 Function View

Chen Gong

15 December 2019

在上一小节中，我们从 Weight-Space View 来看 Gaussian Process Regression，好像和 Gaussian Process 并没有什么关系。但是这一小节，我们从函数的角度来看就可以看到了。

1 Recall Gaussian Process

对于一组随机变量 $\{\xi_t\}_{t\in T}$，T : continuous space or time。If: $\forall n \in N^+$ $(n \geq 1)$, Index: $\{t_1, t_2, \cdots, t_n\}$ $\longrightarrow$ random variable: $\{\xi_{t_1}, \xi_{t_2}, \cdots, \xi_{t_n}\}$。令 $\xi_{1:n} = \{\xi_{t_1}, \xi_{t_2}, \cdots, \xi_{t_n}\}^T$。If $\xi_{1:n} \sim \mathcal{N}(\mu_{1:n}, \Sigma_{1:n})$，那么我们称 $\{\xi_t\}_{t\in T}$ is a Gaussian Distribution。并且，$\xi_t \sim GP(m(t), k(t,s))$，$m(t)$ 为 mean function，$k(t,s)$ 为 covariance function。下面我们回到 Weight-Space View 中。

2 Weight-Space view to Function-Space view

在这里 w 是一个先验分布，$f(x)$ 是一个随机变量。$f(x) = \phi(x)^T w$，$y = f(x) + \epsilon$，$\epsilon \sim \mathcal{N}(0, \sigma^2)$。在 Bayesian 的方法中，对于给定的先验信息 (prior)：$w \sim \mathcal{N}(0, \Sigma_p)$。因为，$f(x) = \phi(x)^T w$，所以可以得到：

$$\mathbb{E}_w[f(x)] = \mathbb{E}_w[\phi(x)^T w] = \phi(x)^T \mathbb{E}_w[w] = 0 \tag{1}$$

那么对于 $\forall x, x' \in \mathbb{R}^p$，

$$\begin{aligned} cov(f(x), f(x')) =& \mathbb{E}[(f(x) - \mathbb{E}[f(x)])(f(x') - \mathbb{E}[f(x')])] \\ =& \mathbb{E}[f(x)f(x')] \\ =& \mathbb{E}[\phi(x)^T w \phi(x')^T w] \\ =& \mathbb{E}[\phi(x)^T w w^T \phi(x')] \end{aligned} \tag{2}$$

因为 $\phi(x')^T w$ 的结果是一个实数，所以它的转置就等于它自己。又因为 $w \sim \mathcal{N}(0, \Sigma_p)$，均值为 0，协方差为 Σ_p。并且有 $\mathbb{E}[ww^T] = \mathbb{E}[(w-0)(w^T-0)]$，这个东西不就是协方差矩阵 $cov(w) = \Sigma_p$。

而 $\phi(x)^T \Sigma_p \phi(x')$ 是一个 kernel function，前面我们已经证明过了，$\varphi(x) = \Sigma_p^{\frac{1}{2}}$。而 $\phi(x)\Sigma_p\phi(x') =< \varphi(x), \varphi(x') >= K(x, x')$。

推导进行到了这里，我们就知道了 $f(x)$ 的期望为 0，协方差矩阵由一个核函数 $K(x,x')$ 产生。那么我们是不是惊奇的发现，这个和我们高斯过程的定义：$\xi_t \sim GP(m(t), K(t,s))$，是多么惊人的相似呀。所以，这里可以启发我们：$f(x)$ 的组成是否可以看成一个 GP，而 $\{f(x)\}_{x\in\mathbb{R}^p}$。那么，首先 $f(x)$

是一个 function，而且 $f(x)$ 还是一个服从高斯分布的随机变量，$m(t)$ 是一个 mean function，$K(t,s)$ 是一个 covariance function。为了加深大家的理解，我们做进一步清晰的对比：

$$\begin{cases} t \longrightarrow \xi_t,\ \{\xi_t\}_{t\in T} \sim GP \\ x \longrightarrow f(x),\ \{f(x)\}_{x\in\mathbb{R}^p} \sim GP \end{cases} \tag{3}$$

其实，我这样一对比，就非常的清晰了。在 GPR 的算法中，

1. Weight-Space view 中关注的是 w，即为：

$$x^* \longrightarrow y^* \quad p(y^*|Data, x^*) = \int p(y^*|Data, x^*, w)p(w)dw \tag{4}$$

又因为 w 本身就是从 Data 中，推导得到的，所以 $p(y^*|Data, x^*, w) = p(y^*|x^*, w)$。

2. Function-Space view 中关注的是 $f(x)$，即为：

$$p(y^*|Data, x^*) = \int p(y^*|f(x), x^*)p(f(x))df(x) \tag{5}$$

写到了这里，不知道大家有没有一定感觉了，这里就是把 $f(x)$ 当成了一个随机变量来看的。这里也就是通过 $f(x)$ 来直接推导出 y^*。在 Weight-Space View 中，我们没有明确的提到 GP，但是在 Weight-Space view 中，$f(x)$ 是符合 GP 的，只不过是没有显性的表示出来而已。我们可以用一个不是很恰当的例子来表述一个，Weight-Space view 就是两个情侣之间，什么都有了，孩子都有了，但是就是没有领结婚证，那么他们两个之间的关系就会比较复杂。而 Function-Space view 就是两个情侣之间先领结婚证，在有了孩子，按部就班的来进行，所以他们之间的关系就会比较简单。

3 Function-Space View

上一小节中，我们从 Weight-Space View 过渡到了 Function-Space View，而 Weight 指的就是参数。

$$\begin{aligned} &\{f(x)\}_{x\in\mathbb{R}^p} \sim GP(m(x), K(x,x')) \\ &m(x) = \mathbb{E}[f(x)] \quad K(x,x') = \mathbb{E}[(f(x)-m(x))(f(x')-m(x'))] \end{aligned} \tag{6}$$

Regression 问题被我们描述为：

Data：$\{(x_i,y_i)\}_{i=1}^N$，$x = (x_1, x_2, \cdots, x_N)^T_{N\times p}$，$Y = (y_1, y_2, \cdots, y_N)^T_{N\times 1}$。又因为 $f(x)$ 符合一个 GP，所以，$f(x) \sim \mathcal{N}(\mu(x), K(x,x'))$。且 $Y = f(X) + \epsilon$，所以 $Y \sim \mathcal{N}(\mu(x), K(x,x') + \sigma^2 I)$。那么，给定 new input：$X^* = (x_1^*, x_2^*, \cdots, x_N^*)$，我们想要的 Prediction output 为 $Y^* = f(X^*) + \epsilon$。那么，我们可以得到 Y 和 $f(X^*)$ 的联合概率密度分布为：

$$\begin{bmatrix} Y \\ f(X^*) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu(X) \\ \mu(X^*) \end{bmatrix}, \begin{bmatrix} K(X,X)+\sigma^2 I & K(X,X^*) \\ K(X^*,X) & K(X^*,X^*) \end{bmatrix}\right) \tag{7}$$

在这里，我必须要首先列举一下，前面我们曾经提到的更加联合概率密度求边缘概率密度的方法。已知，$x \sim \mathcal{N}(\mu, \Sigma)$，

$$x = \begin{bmatrix} x_a \\ x_b \end{bmatrix} \qquad \mu = \begin{bmatrix} \mu_a \\ \mu_b \end{bmatrix} \qquad \Sigma = \begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{bmatrix} \tag{8}$$

而我们可以得到：

$$
\begin{aligned}
& p(x_b|x_a) \sim \mathcal{N}(\mu_{b|a}, \Sigma_{b|a}) \\
& \mu_{b|a} = \Sigma_{ba}\Sigma_{aa}^{-1}(x_a - \mu_a) + \mu_b \\
& \Sigma_{b|a} = \Sigma_{bb} - \Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab}
\end{aligned}
\tag{9}
$$

我们要求的概率为 $p(f(X^*)|Y, X, X^*)$，就是一个我们要求的条件概率，等价于 $p(f(x^*)|Y)$，为什么这里可以把 X, X^* 给忽略掉了？因为 X 和 Y 相关，因为 $Y = \phi(X)^T w + \epsilon$。而 X^* 涵盖在了 $f(X^*)$ 中，可以把 X^* 当做已知的条件，因为 $f(X^*) = \phi(X^*)^T w$。

所以，我们的目标也就是求 $p(f(X^*)|Y)$，也就是已知联合概率分布的情况下求条件概率分布。

我们对比公式 (8) 和公式 (9) 就可以发现，$Y \rightarrow x_a, f(x^*) \rightarrow x_b, K(X,X) + \sigma^2 I \rightarrow \Sigma_{aa}, K(X,X^*) \rightarrow \Sigma_{ba}, K(X^*,X^*) \rightarrow \Sigma_{bb}$。那么，我们可以令 $p(f(X^*)|Y,X,X^*) \sim \mathcal{N}(\mu^*, \Sigma^*)$，代入之前获得的公式的结果我们就可以得到：

$$
\begin{aligned}
\mu^* = & K(X^*, X)(K(X,X) + \sigma^2 I)^{-1}(Y - \mu(X)) + \mu(X^*) \\
\Sigma^* = & K(X^*, X^*) - K(X^*, X)(K(X,X) + \sigma^2 I)^{-1}
\end{aligned}
\tag{10}
$$

并且，$Y^* = f(X^*) + \varepsilon$。那么 noise-free 的形式可以被我们写完：$p(f(x^*)|Y, X, X^*) = \mathcal{N}(\mu^*, \Sigma^*)$。而 $p(Y^*|Y, X, x^*) = \mathcal{N}(\mu_y^*, \Sigma_y^*)$，$\mu_y^* = \mu^*$，$\Sigma_y^* = \Sigma^* + \sigma^2 I$。

在 Function-Space View 中，$f(x)$ 本身是符合 GP 的，那么我们可以直接写出 *Prediction* 矩阵，并将其转化为已知联合概率密度分布求条件概率密度的问题。Function-Space View 和 Weight-Space View 得到的结果是一样的，但是更加的简单。