高斯过程公式推导

假设有一个训练集$\mathcal{D}$，其中$\mathcal{D}=\{(x_i,y_i)|\mathrm{i}=1,2,\dots,\mathrm{n}\}$，我们用x来表示输入向量，用y来表示一个标量输出，其中x的尺寸为$\mathrm{D}\times\mathrm{n}$，y的尺寸为$\mathrm{n}\times 1$，那么相应的$\mathcal{D}=(\mathrm{X},\mathrm{y})$。

以标准的线性模型为例，对于：

$$\mathrm{f(x)}=x^T w，\quad \mathrm{y}=\mathrm{f(x)}+\varepsilon$$

x是输入向量，w是权重向量，y是观察到的目标值，一般来说在f(x)中会有偏置向量，但是因为偏置部分可通过$x_{new}^T=(x_{old}^T,1)$来完成，因此此处将偏置隐去。其中ε为在f(x)上的噪声，我们假设它是独立分布的，相应的均值为 0，方差为σ_n^2，那么对应的分布符合：

$$\varepsilon\sim\mathcal{N}(0,\sigma_n^2)$$

相应y的分布如下：

$$\mathrm{y}\sim\mathcal{N}(\mathrm{f(x)},\sigma_n^2)\sim\mathcal{N}(X^T w,\sigma_n^2)$$

即：

$$\mathrm{p(y|X,w)}=\prod_{i=1}^{n}p(y_i|x_i,w)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma_n^2}}\exp(-\frac{(y_i-x_i^T w)^2}{2\sigma_n^2})$$

在贝叶斯分布中，我们需要对于参数设置一个先验分布，对于权重w，我们假设它的分布为：

$$\mathrm{w}\sim\mathcal{N}(0,\Sigma_p)$$

根据贝叶斯定律：

$$\mathrm{posterior}=\frac{likelihood\ \times prior}{marginal\ likelihood}$$

那么：

$$\mathrm{p(w|y,X)}=\frac{p(y|X,w)p(w)}{p(y|X)}，\mathrm{p(y|X)}=\int p(y|X,w)p(w)dw$$

由于边际概率p(y|X)与权重w无关，因此p(y|X)相当于p(w|y,X)的一个常系数，于是：

$$\mathrm{p(w|y,X)}\propto\exp\left(-\frac{1}{2\sigma_n^2}(y-X^Tw)^T(y-X^Tw)\right)\exp(-\frac{1}{2}w^T\Sigma_p^{-1}\mathrm{w})$$

由于：

$$w^T\Sigma_p^{-1}w+\frac{1}{\sigma_n^2}(y-X^Tw)^T(y-X^Tw)$$

$$=w^T\Sigma_p^{-1}w+\frac{1}{\sigma_n^2}w^TXX^Tw+\frac{1}{\sigma_n^2}y^Ty-\frac{1}{\sigma_n^2}y^TX^Tw-\frac{1}{\sigma_n^2}w^TXy$$

$$=w^T\left(\frac{1}{\sigma_n^2}XX^T+\Sigma_p^{-1}\right)w+\frac{1}{\sigma_n^2}y^Ty-\frac{1}{\sigma_n^2}y^TX^Tw-\frac{1}{\sigma_n^2}w^TXy$$

令$\mathrm{A}=XX^T+\sigma_n^2\Sigma_p^{-1}$，有$\mathrm{A}=A^T$，并记上式为：

$$\frac{1}{\sigma_n^2}(w-\overline{w})^T\mathrm{A}(\mathrm{w}-\overline{w})=\frac{1}{\sigma_n^2}(w^TAw+\overline{w}^TA\overline{w}-\overline{w}^TAw-w^TA\overline{w})$$

即：

$$\overline{w}^TA\overline{w}-\overline{w}^TAw-w^TA\overline{w}=y^Ty-y^TX^Tw-w^TXy$$

那么有$\mathrm{A}\overline{w}=Xy$，即：

$$\mathrm{p(w|y,X)}\propto\exp\left(-\frac{1}{2\sigma_n^2}(w-\overline{w})^T\mathrm{A}(\mathrm{w}-\overline{w})\right)$$

$$\sim\mathcal{N}(\overline{w}=A^{-1}\mathrm{Xy},\sigma_n^2A^{-1})$$

值得注意的是，当我们用传统神经网络 MSE 的方式来表示误差：

$$\mathrm{loss}=\frac{1}{n}\sum_{i=1}^{n}(y_i-x_i^Tw)^2=\frac{1}{n}(y-X^Tw)^T(y-X^Tw)$$

那么对于 loss 求导之后：

$$\frac{\partial loss}{\partial w} = \frac{1}{n} X(y - X^T w) = 0$$

$$\mathrm{w} = (XX^T)^{-1} Xy \approx A^{-1} Xy$$

这样求得的w与高斯分布中w对应的均值相同。

综上分析可以看出：

$$\mathrm{p}(x_{test}{}^T \mathrm{w} | x_{test}, \mathrm{y}, \mathrm{X}) \sim \mathcal{N}(x_{test}{}^T A^{-1} \mathrm{Xy}, \sigma_n^2 x_{test}{}^T A^{-1} x_{test})$$

那么相应的y_{test}分布对应：

$$\mathrm{p}(y_{test} | x_{test}, \mathrm{y}, \mathrm{X}) \sim \mathcal{N}(x_{test}{}^T A^{-1} \mathrm{Xy}, \sigma_n^2 + \sigma_n^2 x_{test}{}^T A^{-1} x_{test})$$

以上所有分析都基于线性模型，那么当我们先把输入向量x通过函数

ϕ(x)映射到M维的特征空间:

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

相应的y_{test}分布为：

$$\begin{cases} \mu(x_{test}) = \phi(x_{test})^T A^{-1} \Phi y \\ \sigma^2(x_{test}) = \sigma_n^2 + \sigma_n^2 \phi(x_{test})^T A^{-1} \phi(x_{test}) \\ \Phi = (\phi(x_1), \phi(x_2), \ldots, \phi(x_n)) \\ A = \Phi\Phi^T + \sigma_n^2 \Sigma_p^{-1} I \end{cases}$$

进一步推导可知，相应的核函数为：

$$\mathrm{E[f(x)]} = \mathrm{E}[\phi(x)^T \mathrm{w}] = \phi(x)^T \mathrm{E(w)} = 0$$

$$\mathrm{k}(x,y)=\mathrm{E}[(f(x)+\varepsilon-E[f(x)+\varepsilon])(f(y)+\varepsilon-E[f(y)+\varepsilon])]$$

$$=\mathrm{E}[(\mathrm{f}(x)+\varepsilon)(\mathrm{f}(\mathrm{y})+\varepsilon)]=\mathrm{E}[\phi(x)^T \mathrm{w}w^T \Phi(\mathrm{y})]+\sigma_n^2$$

$$=\phi(x)^T E[ww^T]\phi(y)+\sigma_n^2=\phi(x)^T cov(w)\phi(y)+\sigma_n^2$$

$$=\phi(x)^T\Sigma_p\phi(y)+\sigma_n^2$$

那么对应的协方差矩阵为：

$$K_\theta=\Phi^T\Sigma_p\Phi+\sigma_n^2 I$$

那么对应的概率分布为：

$$\mathrm{p}(\mathrm{y}|\mathrm{X},\theta)=\frac{1}{(2\pi)^{n/2}|K_\theta|^{1/2}}\exp(-\frac{1}{2}y^T K_\theta^{-1}y)$$

根据《Gaussian Processes for Machine Learning》书中

Appendix.3 可知：

$$K_\theta^{-1}=\left(\sigma_n^2 I+\Phi^T\Sigma_p\Phi\right)^{-1}=\frac{1}{\sigma_n^2}I-\frac{1}{\sigma_n^2}\Phi^T\left(\Sigma_p^{-1}+\Phi\frac{1}{\sigma_n^2}\Phi^T\right)^{-1}\Phi\frac{1}{\sigma_n^2}$$

$$K_\theta^{-1}=\frac{1}{\sigma_n^2}I-\frac{1}{\sigma_n^2}\Phi^T(\sigma_n^2\Sigma_p^{-1}+\Phi\Phi^T)^{-1}\Phi$$

$$K_\theta^{-1}=\frac{1}{\sigma_n^2}I-\frac{1}{\sigma_n^2}\Phi^T A^{-1}\Phi$$

$$|K_\theta|=|\sigma_n^2 I||\Sigma_p|\left|\Sigma_p^{-1}+\Phi\frac{1}{\sigma_n^2}\Phi^T\right|=|\sigma_n^2 I||\Sigma_p|\left|\frac{A}{\sigma_n^2}\right|$$

$$\log|K_\theta|=log|\sigma_n^2 I|+log\left|\frac{\sigma_p^2}{M}I\right|+\log\left|\frac{A}{\sigma_n^2}\right|$$

$$\log|K_\theta|=N*log(\sigma_n^2)+\log|A|-\mathrm{M}*\log(\frac{M\sigma_p^2}{\sigma_p^2})$$

令$\Sigma_p=\frac{\sigma_p^2}{M}I$，有：

$$\mathrm{logp}(\mathrm{y}|\mathrm{X},\theta)=-\frac{1}{2}y^T K_\theta^{-1}y-\frac{N}{2}\log(2\pi)-\frac{1}{2}log|K_\theta|$$

$$\mathrm{logp}(\mathrm{y}|\mathrm{X},\theta) = -\frac{1}{2}y^T(\frac{1}{\sigma_n^2}I - \frac{1}{\sigma_n^2}\Phi^T A^{-1}\Phi)y - \frac{N}{2}\log(2\pi) - \frac{1}{2}log|K_\theta|$$

$$\mathrm{logp}(\mathrm{y}|\mathrm{X},\theta) = -\frac{1}{2\sigma_n^2}(y^T y - y^T\Phi^T A^{-1}\Phi\mathrm{y}) - \frac{N}{2}\log(2\pi) - \frac{1}{2}log|K_\theta|$$

$$\mathrm{logp}(\mathrm{y}|\mathrm{X},\theta) = -\frac{1}{2\sigma_n^2}(y^T y - y^T\Phi^T A^{-1}\Phi\mathrm{y}) - \frac{N}{2}\log(2\pi\sigma_n^2) - \frac{1}{2}log|A| + \frac{M}{2}\log(\frac{M\sigma_n^2}{\sigma_p^2})$$

以上值得注意的是，在建模的过程中，需要模型尽可能地拟合训练集，目标函数为：

$$\mathrm{maximize.\,logp}(\mathrm{y}|\mathrm{X},\theta)$$

也就是要求：

$$\mathrm{minimize.} -\mathrm{logp}(\mathrm{y}|\mathrm{X},\theta)$$

$$minimize.\ \frac{1}{2\sigma_n^2}(y^T y - y^T\Phi^T A^{-1}\Phi y) + \frac{N}{2}\log(2\pi\sigma_n^2) + \frac{1}{2}log|A| - \frac{M}{2}\log(\frac{M\sigma_n^2}{\sigma_p^2})$$

完成建模后，目标更换为含条件约束的目标优化问题：

$$\mathrm{minimize.\,f(x)}$$

$$s.t.\begin{cases} c_1(x) < 0 \\ \quad\ldots \\ c_{N_c}(x) < 0 \end{cases}$$

暂时先不考虑约束条件，假设目前已经求得的目标函数f(x)最小值为τ，那么求解f(x)更小值的概率可以求得，由于之前我们已经对于y = f(x)进行了高斯过程建模，已知其对应的μ(x)和σ(x)，improvement function 可以表示为：

$$\mathrm{I(y,\tau)} = \begin{cases} \tau - y \quad , y < \tau \\ 0 \quad , otherwise \end{cases}$$

那么相应的 Expected Improvement(EI)为：

$$\mathrm{E[I(y,\tau)]} = \int_{-\infty}^{+\infty} I(y,\tau)p(y|X,\theta)dy$$

$$\mathrm{E[I(y,\tau)]} = \int_{-\infty}^{\tau} (\tau - y)p(y|X,\theta)dy$$

对于x，GP 会预测$\mathrm{y = f(x)} \sim \mathcal{N}(\mu, \sigma^2)$，令$y_* = \frac{y-\mu}{\sigma}$，则$y_* \sim \mathcal{N}(0,1)$为标准正态分布，令$\tau_* = \frac{\tau-\mu}{\sigma}$，进一步令$\Phi(\mathrm{y}) = \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}} \exp(-0.5 * y^2)\, dy$为 CDF(cumulative distribution function)函数，令$\phi(\mathrm{y}) = \frac{1}{\sqrt{2\pi}} \exp(-0.5 * y^2)$为 PDF(probability distribution function)函数，可以进一步推导：

$$\mathrm{I(y,\tau) = \sigma I}(y_*, \tau_*)$$

$$\mathrm{E}[I(y,\tau)] = \mathrm{\sigma E[I}(y_*, \tau_*)]$$

$$\mathrm{E[I}(y_*, \tau_*)] = \int_{-\infty}^{\tau_*} (\tau_* - y_*)\phi(y_*)\, dy_*$$

$$\mathrm{E}[\mathrm{I}(y_*,\tau_*)] = \int_{-\infty}^{\tau_*} \tau_* \phi(y_*)\, dy_* - \int_{-\infty}^{\tau_*} y_* \phi(y_*)\, dy_*$$

$$\mathrm{E}[\mathrm{I}(y_*,\tau_*)] = \tau_* \Phi(\tau_*) - \int_{-\infty}^{\tau_*} y_* \phi(y_*)\, dy_*$$

其中$y_* \phi(y_*) = \frac{y_*}{\sqrt{2\pi}} \exp(-0.5 * y_*^2) = \left(\frac{-1}{\sqrt{2\pi}} \exp(-0.5 * y_*^2)\right)'$

$$\int_{-\infty}^{\tau_*} y_* \phi(y_*)\, dy_* = \left(-\phi(\tau_*)\right) - \left(-\phi(-\infty)\right) = -\phi(\tau_*)$$

$$\mathrm{E}[\mathrm{I}(y_*,\tau_*)] = \tau_* \Phi(\tau_*) + \phi(\tau_*)$$

$$\mathrm{E}[\mathrm{I}(\mathrm{y},\tau)] = \sigma \mathrm{E}[\mathrm{I}(y_*,\tau_*)] = (\tau - \mathrm{y}) \Phi\left(\frac{\tau-\mu}{\sigma}\right) + \sigma \phi(\frac{\tau-\mu}{\sigma})$$

以上 Expected Improvement 对应没有考虑约束条件的情况，当将约束条件纳入考虑时，可以采用 weighted Expected Improvement(wEI)计算方式来构造目标函数:

$$I_c(y,c,\tau) = \begin{cases} \tau - y \ \ , y < \tau \ and \ c < 0 \\ \quad 0 \ \ , otherwise \end{cases}$$

对于每一个约束条件$c_i(\mathrm{x})$都构造一个相应的高斯过程模型，那么每一个$c_i(\mathrm{x})$都有$c_i(\mathrm{x}) \sim \mathcal{N}(\mu_i, \sigma_i)$，那么$c_i(\mathrm{x}) < 0$约束满足的概率为：

$$\Pr(c_i(\mathrm{x}) < 0) = \Phi\left(\frac{0-\mu_i}{\sigma_i}\right) = \Phi\left(\frac{-\mu_i}{\sigma_i}\right)$$

相应的期望为：

$$\mathrm{E}[I_c(\mathrm{y},\mathrm{c},\tau)] = \int_{-\infty}^{\tau} (\tau - y) p(y|X,\theta) \prod_{i=1}^{N_c} \Pr(c_i(x) < 0)\, dy$$

$$\mathrm{E}[I_c(\mathrm{y},\mathrm{c},\tau)] = \left\{(\tau - \mathrm{y}) \Phi\left(\frac{\tau-\mu}{\sigma}\right) + \sigma \phi(\frac{\tau-\mu}{\sigma})\right\} \prod_{i=1}^{N_c} \Phi(\frac{-\mu_i}{\sigma_i})$$

为了求得最佳x需要求$\mathrm{maximize.}\ \mathrm{E}[I_c(\mathrm{y},\mathrm{c},\tau)]$。

值得注意的是$\mathrm{E}[I_c(\mathrm{y},\mathrm{c},\tau)]$公式中，我们假设了约束之间相互独立，

在实际问题中往往约束之间存在某种关系，使得其并不相互独立。