高斯过程回归

黄政宇
北京大学北京国际数学研究中心
北京大学国际机器学习研究中心

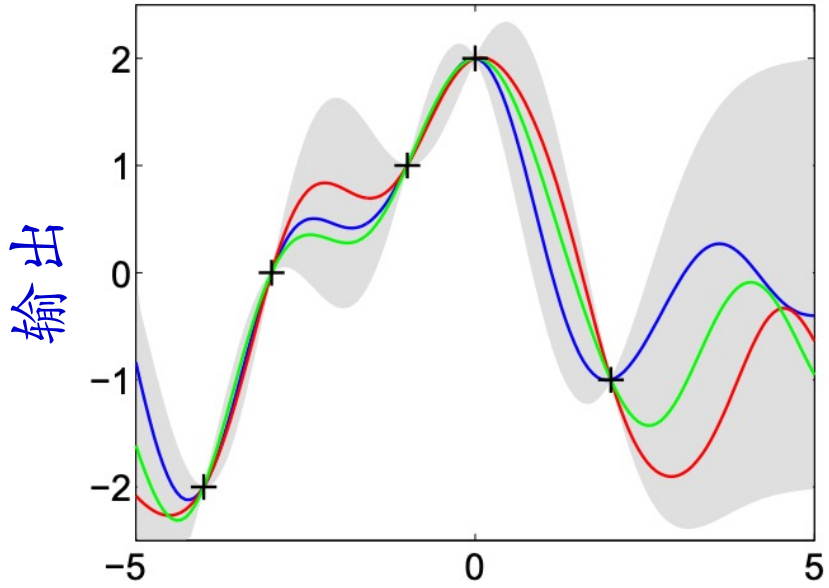
本堂课大纲

- 线性回归
- 贝叶斯线性回归
  - 后验估计
  - 核函数
- 高斯过程
  - 无噪音观测
  - 有噪音观测
  - 期望不为0的高斯过程

高斯过程回归

输入(设计、运行条件等)

输出(受到的力、结构的响应等和相应的不确定性)



线性回归

- 模型

$$f(x) = x^T w \qquad x, w \in R^N$$

- 数据

$$(x_1, y_1), (x_2, y_2), \cdots\cdots, (x_n, y_n)$$
$$y_i = f(x_i) + \epsilon_i \qquad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

- 线性回归

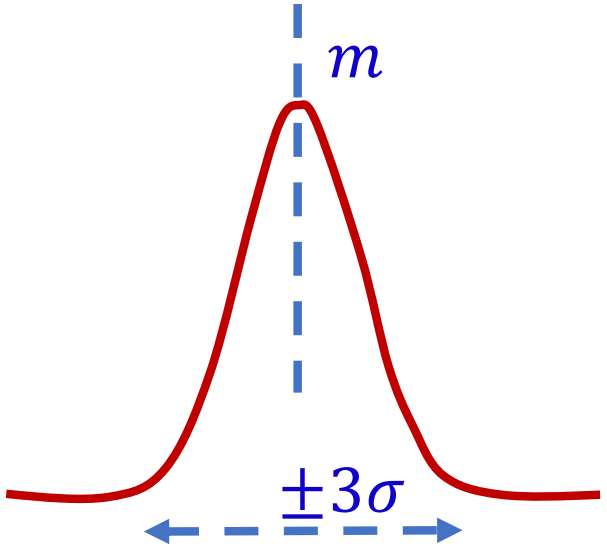
$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \in R^n \quad X^T = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{bmatrix} \in R^{n \times N} \quad y = X^T w$$

$$\hat{w} = (XX^T)^{-1} X y$$

高斯近似

- 高斯分布

$$\mathcal{N}(x; m, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right)$$



$$\mathcal{N}(x; m, C) = \frac{1}{Z} \exp\left(-\frac{1}{2}(x-m)^T C^{-1}(x-m)\right)$$

$$Z = \sqrt{|2\pi C|} = (2\pi)^{N_\theta/2}\sqrt{|C|}$$

$Ax + b$服从什么分布？

本堂课大纲

- 线性回归
- 贝叶斯线性回归
  - 后验估计
  - 核函数
- 高斯过程
  - 无噪音观测
  - 有噪音观测
  - 期望不为0的高斯过程

贝叶斯线性回归

- 模型

$$f(x) = x^T w \qquad x, w \in R^N$$

- 数据

$$(x_1, y_1), (x_2, y_2), \cdots\cdots, (x_n, y_n)$$
$$y_i = f(x_i) + \epsilon_i \qquad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

- 似然函数

$$\rho(y|X, w) = \prod_{i=1}^{i=n} \rho(y_i|x_i, w)$$

- 先验分布

$$w \sim \rho_0(w)$$

- 后验分布

$$\rho(w|X, y) = \frac{\text{似然函数} \times \text{先验分布}}{\text{归一化常数}} = \frac{\rho(y|X, w)\rho_0(w)}{\rho(y|X)}$$
$$\rho(y|X) = \int \rho(y|X, w)\rho_0(w)dw$$

贝叶斯线性回归

- 似然函数

$$\rho(y|X,w)=\prod_{i=1}^{i=n}\rho(y_i|x_i,w)=\prod_{i=1}^{i=n}\frac{1}{\sqrt{2\pi\sigma^2}}e^{\frac{1}{2\sigma^2}(y_i-x_i^Tw)^2}$$
$$=\mathcal{N}(y;X^Tw,\sigma^2I)$$

$$X=[x_1\ x_2\cdots x_n]\qquad y=[y_1\ y_2\cdots y_n]^T$$

- 先验分布

$$w\sim\rho_0(w)=\mathcal{N}(w;0,\Sigma_0)$$

- 后验分布

$$\rho(w|X,y)\propto\mathcal{N}(y;X^Tw,\sigma^2I)\,\mathcal{N}(w;0,\Sigma_0)$$

贝叶斯线性回归

- 后验分布

$$\rho(w|X,y)$$
$$\propto \mathcal{N}(y;X^Tw,\sigma^2I)\,\mathcal{N}(w;0,\Sigma_0)$$
$$\propto \exp(-\frac{1}{2\sigma^2}(y-X^Tw)^T(y-X^Tw))\exp(-\frac{w^T\Sigma_0^{-1}w}{2})$$
$$\propto \exp\left(-\frac{1}{2}(w-\hat{w})^T\left(\frac{1}{\sigma^2}XX^T+\Sigma_0^{-1}\right)(w-\hat{w})\right)$$

$$\hat{w}=(XX^T+\sigma^2\Sigma_0^{-1})^{-1}Xy$$

$$\rho(w|X,y)=\mathcal{N}(w;\hat{w},\left(\frac{1}{\sigma^2}XX^T+\Sigma_0^{-1}\right)^{-1})$$

贝叶斯线性回归

- 最大后验估计

$$\hat{w} = (XX^T + \sigma^2\Sigma_0^{-1})^{-1}Xy$$

Rigid regression
Tikhonov-Phillips regularization

- 预测 $f_* = f(x_*)$

$$f_* \sim \rho(f_*|x_*, X^T, y)$$
$$= \mathcal{N}(x_*^T(XX^T + \sigma^2\Sigma_0^{-1})^{-1}Xy, x_*^T\left(\frac{1}{\sigma^2}XX^T + \Sigma_0^{-1}\right)^{-1}x_*)$$

贝叶斯线性回归

- 贝叶斯线性回归

表达能力有限

- 特征空间

$$\Psi: R^N \to R^D$$

例子:

$$\Psi: R^1 \to R^D$$
$$\Psi(x) = [1, x, x^2, \cdots\cdots, x^{D-1}]^T$$

- 模型

$$f(x) = \Psi(x)^T w \qquad x \in R^N, w \in R^D$$

$$X = [\Psi(x_1)\ \Psi(x_2)\ \cdots\ \Psi(x_n)] \in R^{D\times n}$$
$$y = [y_1\ y_2 \cdots y_n]^T$$

贝叶斯线性回归

- 最大后验估计

$$\widehat{w} = (XX^T + \sigma^2\Sigma_0^{-1})^{-1}Xy$$

- 预测$f_* = f(x_*)$

$$f_* \sim \rho(f_*|x_*, X, y)$$
$$= \mathcal{N}\left(\Psi(x_*)^T(XX^T + \sigma^2\Sigma_0^{-1})^{-1}Xy,\ \Psi(x_*)^T\left(\frac{1}{\sigma^2}XX^T + \Sigma_0^{-1}\right)^{-1}\Psi(x_*)\right)$$

贝叶斯线性回归

- Woodbury 矩阵等式

$$(XX^T + \sigma^2\Sigma_0^{-1})^{-1} = \sigma^{-2}(\Sigma_0 - \Sigma_0 X(X^T\Sigma_0 X + \sigma^2 I)^{-1}X^T\Sigma_0)$$

- 预测 $f_* = f(x_*)$

$$f_* \sim \mathcal{N}(\Psi(x_*)^T\Sigma_0 X(K + \sigma^2 I)^{-1}y,$$
$$\Psi(x_*)^T\Sigma_0\Psi(x_*) - \Psi(x_*)^T\Sigma_0 X(K + \sigma^2 I)^{-1}X^T\Sigma_0\Psi(x_*))$$

这里我们定义 $K = X^T\Sigma_0 X$

贝叶斯线性回归

- 预测$f_* = f(x_*)$

$$f_* \sim \mathcal{N}(\Psi(x_*)^T \Sigma_0 X (K + \sigma^2 I)^{-1} y,$$
$$\Psi(x_*)^T \Sigma_0 \Psi(x_*) - \Psi(x_*)^T \Sigma_0 X (K + \sigma^2 I)^{-1} X^T \Sigma_0 \Psi(x_*))$$

- 核函数

$$\Phi(x) = \Sigma_0^{\frac{1}{2}} \Psi(x) \in R^D \qquad \kappa(x, x') = \Phi(x)^T \Phi(x')$$

那么　　所有基底函数：$\Phi(x)$

$$K = X^T \Sigma_0 X \qquad K_{ij} = \kappa(x_i, x_j)$$

$$\Psi(x_*)^T \Sigma_0 X = [\kappa(x_*, x_1)\kappa(x_*, x_2) \cdots \kappa(x_*, x_n)]$$

f_*仅依赖于核函数$\kappa(x, x')$。

贝叶斯线性回归

- 核技巧(Kernel trick)

  如果一个算法完全是用输入空间中的内积来定义的，那么可以通过用替换这些内积来将其提升到特征空间中。

  在计算核比计算特征向量更方便的情况下，这种技术特别有价值。

  这通常会导致将核视为主要关注对象，而相应的特征空间具有次要的实际重要性。

本堂课大纲

- 线性回归
- 贝叶斯线性回归
  - 后验估计
  - 核函数
- 高斯过程
  - 无噪音观测
  - 有噪音观测
  - 期望不为0的高斯过程

高斯过程(Gaussian process)

高斯过程（函数空间观点）

高斯过程是一组随机变量或是随机函数 $f(x)$，其中任何有限个数的变量都具有联合高斯分布。一个高斯过程完全由其均值函数和协方差函数(covariance function)所确定。

给定均值函数$m(x)$，和协方差函数$\kappa(x,x')$，实高斯过程$f(x) \sim GP(m(x),\kappa(x,x'))$满足

$$m(x) = \mathbb{E}[f(x)] \quad \forall x$$
$$\kappa(x,x') = \mathbb{E}[(f(x)-m(x))(f(x')-m(x'))] \quad \forall x,x'$$

高斯过程

- 例子

$f(x) = \Psi(x)^T w$

基底函数 $\Psi(x) = [\psi_1(x); \psi_2(x); \cdots; \psi_D(x)]$： $R^N \to R^D$

高斯随机变量： $w \sim \mathcal{N}(0, \Sigma_0) \in R^N$

$m(x) = \mathbb{E}[f(x)] = \Psi(x)^T \mathbb{E}w = 0$

$\kappa(x, x') = \mathbb{E}[f(x) f(x')^T] = \Psi(x)^T \Sigma_0 \Psi(x)$

高斯过程

- 无噪音观测
  - 训练数据 $\{(x_i, y_i = f(x_i))|i = 1,2 \cdots \cdots n\}$
  - 测试点 $\{x_i^*|i = 1,2 \cdots \cdots n_*\}$
  - 联合分布

$$\begin{bmatrix} f \\ f_* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K(X,X) & K(X,X_*) \\ K(X_*,X) & K(X_*,X_*) \end{bmatrix}\right)$$

其中$f \in R^n$代表在训练数据上取值，$f_* \in R^{n_*}$表示在测试点上的取值。

$$K(X,X) = [\kappa(x_i, x_j)] \in R^{n \times n}$$

$$K(X_*,X)^T = K(X,X_*) = [\kappa(x_i, x_{*j})] \in R^{n \times n_*}$$

$$K(X_*,X_*) = [\kappa(x_{*i}, x_{*j})] \in R^{n_* \times n_*}$$

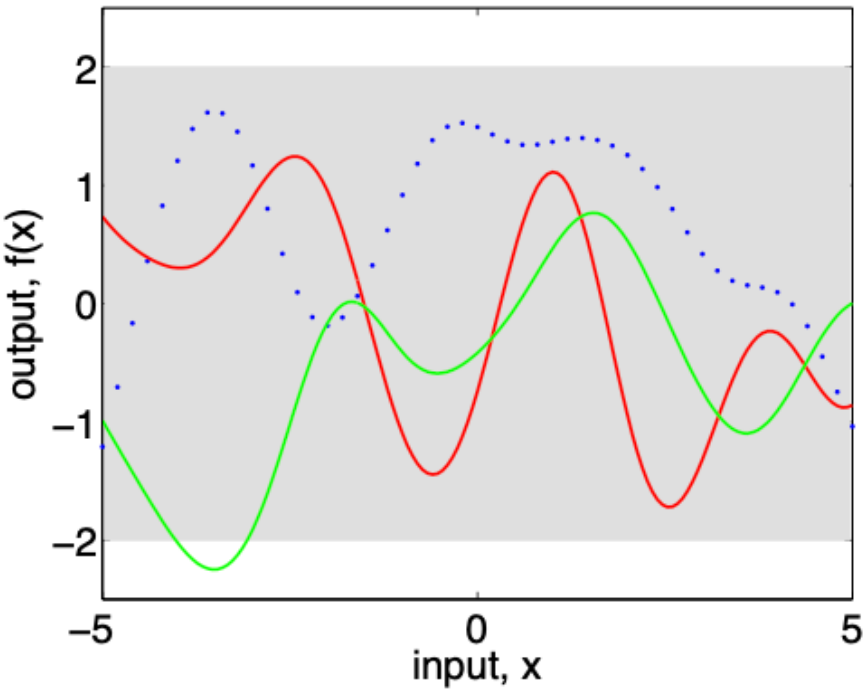
高斯过程

- 预测(无噪音观测)
  - 条件概率

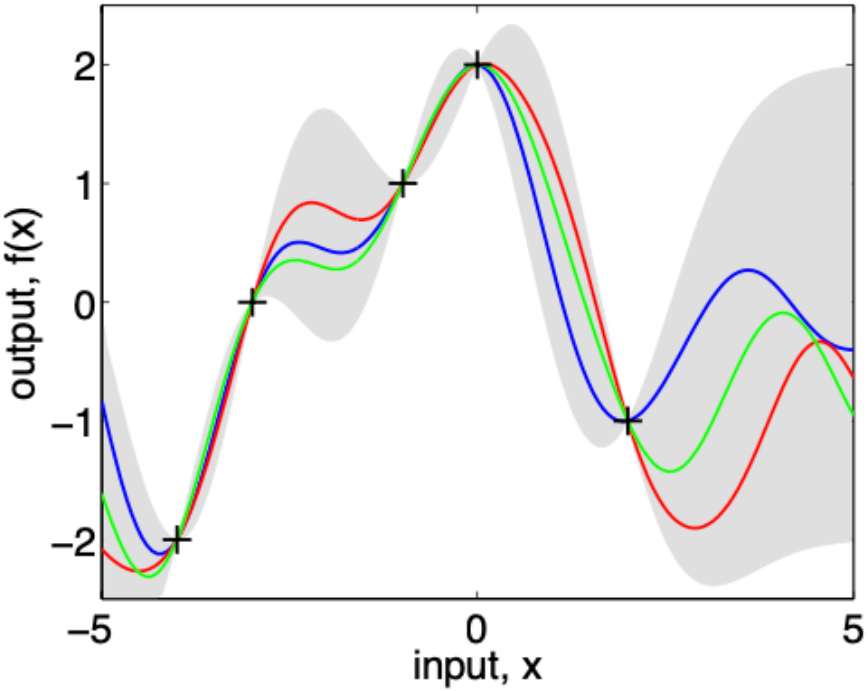
$$\begin{bmatrix} f \\ f_* \end{bmatrix} = \mathcal{N}\left(0, \begin{bmatrix} K(X,X) & K(X,X_*) \\ K(X_*,X) & K(X_*,X_*) \end{bmatrix}\right)$$

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

高斯过程



(a), prior



(b), posterior

高斯过程

- 有噪音观测
  - 训练数据 $\{(x_i, y_i = f(x_i) + \epsilon) | i = 1,2 \cdots \cdots n\}$

$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$

  - 测试点 $\{x_i^* | i = 1,2 \cdots \cdots n_*\}$
  - 联合分布

$$\begin{bmatrix} f \\ f_* \end{bmatrix} = \mathcal{N}\left(0, \begin{bmatrix} K(X,X) + \sigma^2 I & K(X,X_*) \\ K(X_*,X) & K(X_*,X_*) \end{bmatrix}\right)$$

其中$f \in R^n$代表在训练数据上取值，$f_* \in R^{n_*}$表示在测试点上的取值。

$$K(X,X) = [\kappa(x_i, x_j)] \in R^{n \times n}$$

$$K(X_*,X)^T = K(X,X_*) = [\kappa(x_i, x_{*j})] \in R^{n \times n_*}$$

$$K(X_*,X_*) = [\kappa(x_{*i}, x_{*j})] \in R^{n_* \times n_*}$$

高斯过程

- 有噪音观测
  - 条件概率

$$\begin{bmatrix} f \\ f_* \end{bmatrix} = \mathcal{N}\left(0, \begin{bmatrix} K(X,X)+\sigma^2 I & K(X,X_*) \\ K(X_*,X) & K(X_*,X_*) \end{bmatrix}\right)$$

$$f_* \,|X_*, X, y \sim \mathcal{N}(K(X_*,X)(K(X,X)+\sigma^2 I)^{-1} y,$$
$$K(X_*,X_*) - K(X_*,X)(K(X,X)+\sigma^2 I)^{-1} K(X,X_*)$$

  - 方差和观测无关
  - 预处理计算： Cholesky分解 $(K(X,X)+\sigma^2 I)^{-1}$
  - 在线复杂度：$\mathcal{O}(n^2 n_*)$

高斯过程

- 训练集上进行预测

$$\mathbb{E}f(\mathrm{X}) = K(X,X)(K(X,X)+\sigma^2 I)^{-1}y$$

$$K(X,X) = \sum_{i=1}^{n}\lambda_i u_i u_i^T \text{，} y = \sum_{i=1}^{n}\gamma_i u_i$$

$$\mathbb{E}f(X) = \sum_{i=1}^{n}\frac{\gamma_i\lambda_i}{\lambda_i+\sigma^2}u_i \neq y$$

磨光(smoothing)

高斯过程

- 测试集上进行预测

$$f_* \mid X_*, X, y \sim \mathcal{N}(K(X_*, X)(K(X, X) + \sigma^2 I)^{-1} y,$$
$$K(X_*, X_*) - K(X_*, X)(K(X, X) + \sigma^2 I)^{-1} K(X, X_*)$$

$$\mathbb{E} f(x_*) = h(x_*)^T y \qquad \mathrm{Cov} f(x_*) = \kappa(x_*, x_*) - h(x_*)^T \kappa_*$$

$$h(x_*)^T = \kappa_*^T (K(X, X) + \sigma^2 I)^{-1} \qquad \kappa_* = \begin{bmatrix} \kappa(x_1, x_*) \\ \kappa(x_2, x_*) \\ \vdots \\ \kappa(x_n, x_*) \end{bmatrix}$$

线性回归：对输入的线性组合
高斯回归：对数据的线性组合

高斯过程

- 练习

考虑一维高斯过程，期望函数$m(x)=0$，核函数为

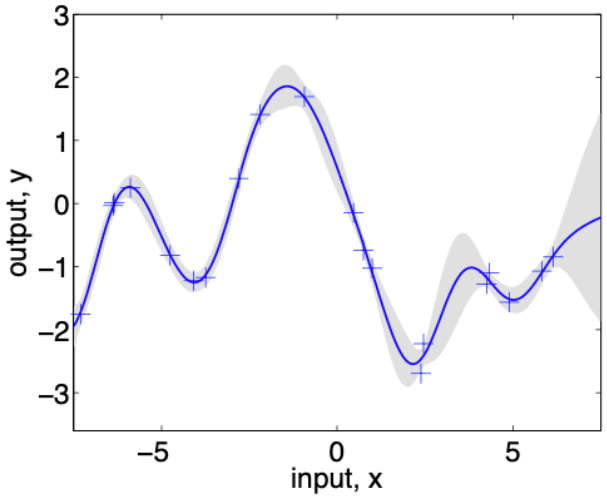
$$\kappa_y(x_p,x_q)=\sigma_f^2\exp(-\frac{1}{2l^2}(x_p-x_q)^2)+\sigma^2\delta_{pq}$$

用$(l,\sigma_f,\sigma)=(1.0,1.0,0.1)$生成20个数据点$x\sim U[-8,8]$，用高斯过程对$x\in[-8,8]$进行预测,并画出%95置信区间，考虑如下高斯过程

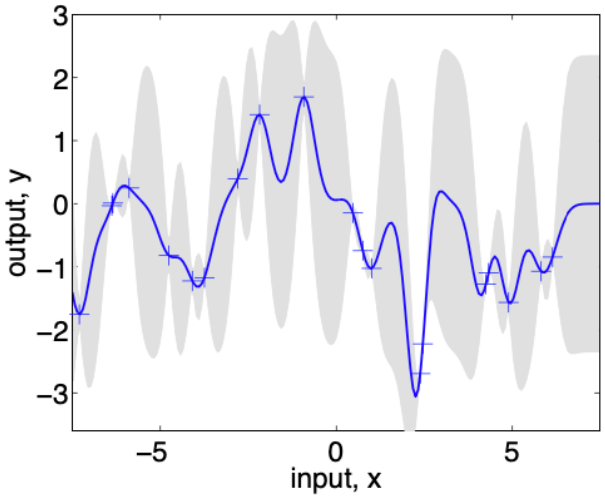
- $(l,\sigma_f,\sigma)=(1.0,1.0,0.1)$
- $(l,\sigma_f,\sigma)=(0.3,1.08,5e-5)$
- $(l,\sigma_f,\sigma)=(3.0,1.16,0.89)$

高斯过程

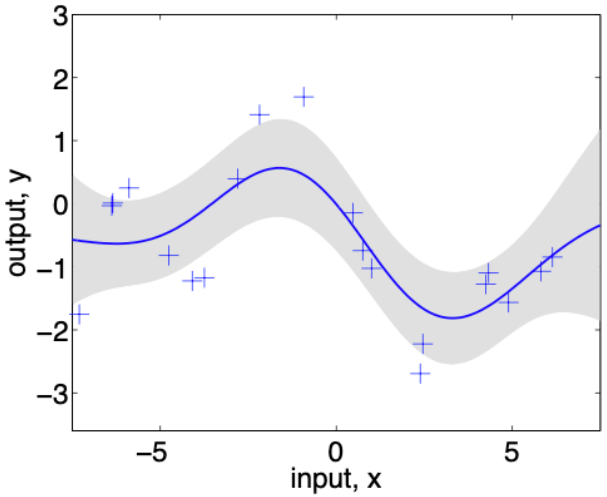
- 练习



(a), $\ell = 1$



(b), $\ell = 0.3$



(c), $\ell = 3$

高斯过程

- 根据回归作决策

  -贝叶斯回归(分布预测)

$$f(x_*)\,|X,y \sim \mathcal{N}(\kappa_*^T(K(X,X)+\sigma^2 I)^{-1}y,$$
$$\kappa(x_*,x_*)-\kappa_*^T(K(X,X)+\sigma^2 I)^{-1}\kappa_*)$$

  -单点预测y_{opt}

  损失函数$\mathcal{L}(f_*,y_*)$，比如$(f_*-y_*)^2$，目标优化(在不确定性下做决策)：

$$y_{\text{opt}}|x_*,X,y=\underset{y_*}{\text{argmin}}\int\mathcal{L}(f_*,y_*)\rho(f_*|x_*,X,y)df_*$$

$$\mathcal{L}(f_*,y_*)=(f_*-y_*)^2 \quad \rightarrow \quad \text{期望}\mathbb{E}f(x_*)$$

$$\mathcal{L}(f_*,y_*)=|f_*-y_*| \quad \rightarrow \quad \text{中位数}\rho(f_*)$$

高斯过程

- 期望不为0的高斯过程

$$f(x) \sim GP(m(x), \kappa(x, x'))$$

- 联合分布

$$\begin{bmatrix} f \\ f_* \end{bmatrix} = \mathcal{N}\left(\begin{bmatrix} m(X) \\ m(X_*) \end{bmatrix}, \begin{bmatrix} K(X,X) & K(X,X_*) \\ K(X_*,X) & K(X_*,X_*) \end{bmatrix}\right)$$

- 条件概率分布

$$f_* \mid X_*, X, y$$
$$\sim \mathcal{N}(m(X_*) + K(X_*,X)(K(X,X) + \sigma^2 I)^{-1}(y - m(X)),$$
$$K(X_*,X_*) - K(X_*,X)(K(X,X) + \sigma^2 I)^{-1}K(X,X_*))$$

对于实际问题 $m(X)$ 难以确定！

高斯过程

- 期望不为0的高斯过程

  给定基底函数$h(x)$，系数$\beta \sim \mathcal{N}(b, B)$是未知参数

$$g(x) = f(x) + h(x)^T\beta \quad ， \quad 其中 f(x) \sim GP(0, \kappa(x, x'))$$

$$g(x) \sim GP(h(x)^T b, \kappa(x, x') + h(x)^T B h(x'))$$

- 联合分布

$$\begin{bmatrix} g \\ g_* \end{bmatrix} = \mathcal{N}\left(\begin{bmatrix} H^T b \\ H_*^T b \end{bmatrix}, \begin{bmatrix} K + H^T B H & K_*^T + H^T B H_* \\ K_* + H_*^T B H & K(X_*, X_*) + H_*^T B H_* \end{bmatrix}\right)$$

$$K_* = K(X_*, X)$$

$$H_* = [h(x_{*1}) \quad h(x_{*2}) \cdots h(x_{*n_*})]$$

$$H = [h(x_1) \quad h(x_2) \cdots h(x_n)]$$

高斯过程

- 期望不为0的高斯过程

$$g_* \mid X_*, X, y \sim \mathcal{N}\left(H_*^T b + (K_* + H_* B H)(K + H^T B H)^{-1}(y - H^T b),\right.$$
$$\left. K(X_*, X_*) + H_*^T B H_* - (K_* + H_*^T B H)(K + H^T B H)^{-1}(K_*^T + H^T B H_*)\right)$$

- 条件分布

$$\mathbb{E}[g(X_*)] = H_*^T \bar{\beta} + K_* K^{-1}\left(y - H^T \bar{\beta}\right) = \mathbb{E}[f(X_*)] + R^T \bar{\beta},$$
$$\mathrm{Cov}[g(X_*)] = \mathrm{Cov}[f(X_*)] + R^T (B^{-1} + H K^{-1} H^T)^{-1} R$$

修正项

$$\bar{\beta} = (H K^{-1} H^T + B^{-1})^{-1}(H K^{-1} y + B^{-1} b)$$
$$R = H_* - H K^{-1} K_*^T$$

参考文献

参考文献

C. E. Rasmussen & C. K. I. Williams, Gaussian Processes for Machine Learning, Chapter 2