Today we have a blackboard lecture :)

## Gaussian process regression

Model

$$\underbrace{y}_{\text{output}} = \underbrace{\vec{x}^T}_{\text{input}} \underbrace{\vec{\Theta}}_{\text{parameters}} + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2) \text{ - i.i.d.}$$

Prior

$$\vec{\Theta} \sim N(\vec{0}, \beta I)$$ - prior distribution

Bayes formula

$$p(\vec{\Theta} \mid D) = \frac{p(D \mid \vec{\Theta})\, p(\vec{\Theta})}{p(D)}$$ - Bayes formula

$$p(D \mid \vec{\Theta}) = \prod_{i=1}^{n} p(y_i \mid \vec{x}_i, \vec{\Theta}) \propto p(D \mid \vec{\Theta})\, p(\vec{\Theta}) \to \max_{\vec{\Theta}}$$

$$p(y_i \mid \dots) \sim N(\vec{x}_i^T \vec{\Theta}, \sigma^2)$$

$$p(D \mid \vec{\Theta}) = L_D(\vec{\Theta}) \to \max_{\vec{\Theta}}$$

Maximum aposterior estimate:

$$\vec{\Theta}_{MAP} = \left(X^T X + \frac{\sigma^2}{\beta} I\right)^{-1} X^T \vec{y}$$

Posterior distribution

$$p(\vec{\Theta} \mid D) = N\left(\vec{\Theta}_{MAP},\ \sigma^2 \left(X^T X + \frac{\sigma^2}{\beta} I\right)^{-1}\right)$$

distribution $p(\theta \mid D) \sim N(\theta_{map}, \ \ldots \beta I)$

$P(f_\theta) \longrightarrow P(f_\theta \mid D)$

$f_\theta(\vec{x}) = \vec{\theta}^T \vec{x}$


$GP_y$

$\theta \sim N(0, \sigma^2)$

$\vec{\theta}_1, \vec{\theta}_2, \ldots$

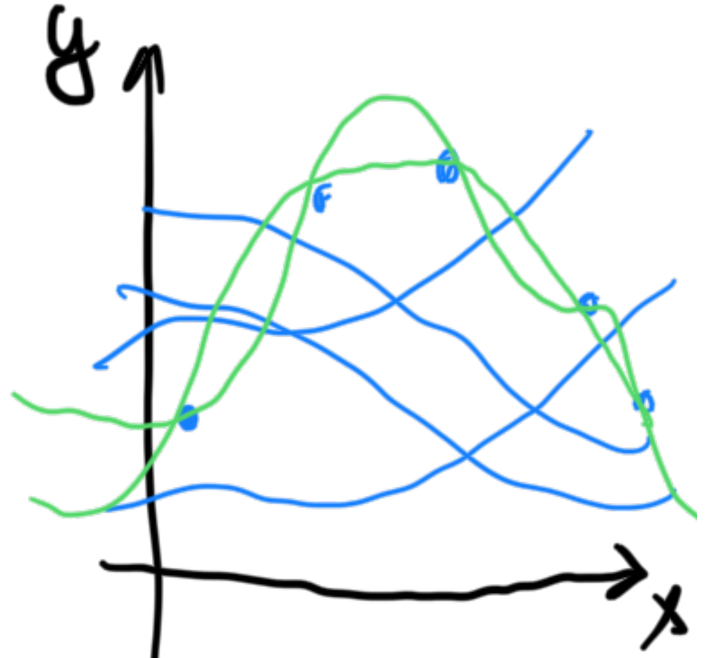
$\vec{f}_{\theta_1}, \quad \vec{f}_{\theta_2} = \vec{\theta}_2^T \vec{x}$

Bayesian linear regression

$\theta \sim p(\theta \mid D)$

$\vec{\theta}_1', \vec{\theta}_2', \ldots$

$\vec{f}_{\theta_2'}(\vec{x}) = \vec{\theta}_2'^T \vec{x}$

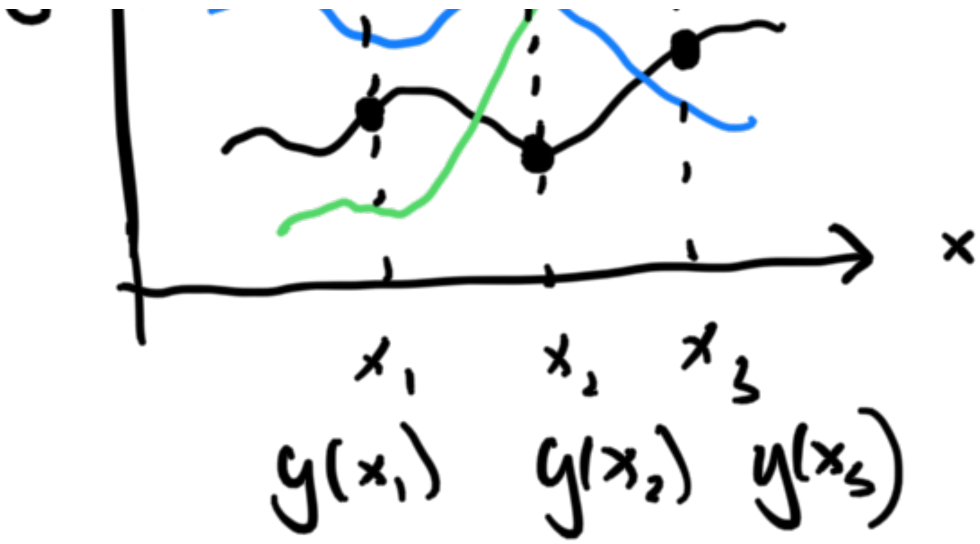



$GP(0, k(x,x'))$

$GP(\mu, k(x,x'))$

Gaussian random process


$\left( y(x_1) \right.$




$$\begin{pmatrix} y(x_1) \\ y(x_2) \\ y(x_3) \end{pmatrix} \sim N(\vec{\mu}, \Sigma)$$

**Def.** Rand. process is Gaussian, if and only if for all possible $\forall k \; \{x_1, \dots, x_k\}$:

$(y(x_1), \dots, y(x_k))$ - Gaussian random vector

1. $N(\underbrace{\mu(\vec{x})}_{\text{mean}}, \underbrace{k(\vec{x}, \vec{x}')}_{\text{covariance}})$

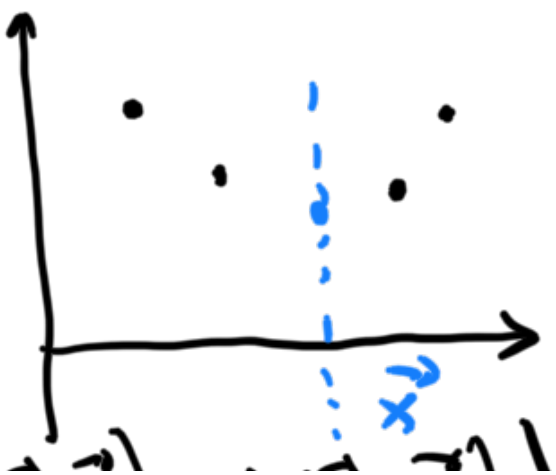
$y(\vec{x}), \; y(\vec{x}')$

$\mathbb{E}(y(x) - \mathbb{E}y(x))(y(x') - \mathbb{E}(y(x')))$

$\vec{x}$

2. $D = \{(\vec{x}_i, y(\vec{x}_i))\}_{i=1}^{n}$

3. $\underline{p(y(\vec{x}) \mid D)}$ V



$$\begin{pmatrix} y(\vec{x}) \\ y(\vec{x}_1) \\ \vdots \\ y(\vec{x}_n) \end{pmatrix} \sim N\left( \begin{pmatrix} \mu(\vec{x}) \\ \vdots \\ \mu(\vec{x}_n) \end{pmatrix}, \begin{pmatrix} k(\vec{x},\vec{x}) & k(\vec{x},\vec{x}_1) & \\ & \ddots & \\ & & \ddots \end{pmatrix} \right)$$

$$P\left( y(\vec{x}) \mid y(\vec{x}_1), \dots, y(\vec{x}_n) \right) = N\left( \hat{y}(\vec{x}), \hat{\sigma}^2(\vec{x}) \right)$$

$O(n^2)$

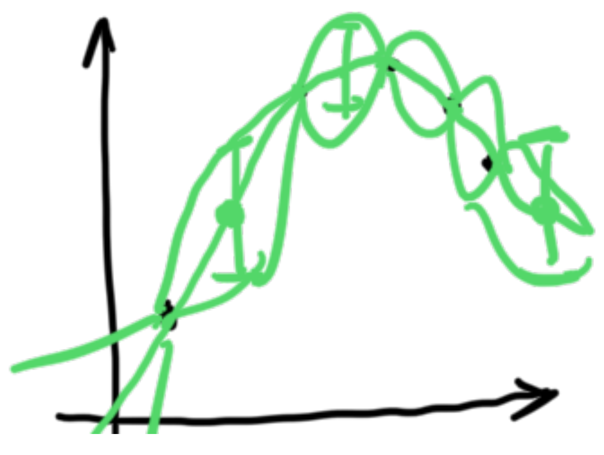
$O(n)$

$$\boxed{\hat{y}(\vec{x}) = \vec{k}^T K^{-1} \vec{y}} = f(\vec{x}) = \underbrace{\sum_{i=1}^{n} \alpha_i k(\vec{x}, \vec{x}_i)}$$

$$\underbrace{\vec{\alpha} = K^{-1} \vec{y}}$$

$$\vec{y} = \begin{pmatrix} y(\vec{x}_1) \\ \vdots \\ y(\vec{x}_n) \end{pmatrix}, \quad \vec{k} = \begin{pmatrix} k(\vec{x}, \vec{x}_1) \\ \vdots \\ k(\vec{x}, \vec{x}_n) \end{pmatrix}$$

$$K = \begin{pmatrix} k(\vec{x}_1, \vec{x}_1) & \dots & k(\vec{x}_1, \vec{x}_n) \\ \vdots & & \\ \vdots & & \end{pmatrix}$$



$$\begin{pmatrix} \\ k & \cdots & k(x_n, x_n) \end{pmatrix}$$

$O(n^2)$

$O(1)$

$O(n^3)$

$$\hat{\sigma}^2(\vec{x}) = k(\vec{x}, \vec{x}) - \vec{k}^T K^{-1} \vec{k}$$

$$\mathbb{E}\left(y - \hat{y}(x)\right)^2 \qquad \hat{y}(x) = \sum_{i=1}^{n} w_i y_i \,, \quad y_i = y(x_i) \qquad (*)$$

$$\vec{w} = \vec{k}(\vec{x}) K^{-1}$$

$$\hat{y}_{kern}(x) = \frac{1}{k} \sum_{i=1}^{n} k(x, x_i)\, y_i \quad \text{- Kernel regression}$$

$$k(x, x_i) = \exp\left(- \| \underline{x} - \underline{x}_i \|\right)$$

$$k = \sum_{i=1}^{n} k(x, x_i)$$

A1: $\mu(\vec{x}) = 0$

A2: $k(\vec{x}, \vec{x}') = \exp\left(- \sum_{i=1}^{d} \frac{(x_i - x_i')^2}{\Theta_i^2}\right)$

$\rightarrow$

$\vec{\Theta} = (\Theta_1, \ldots, \Theta_d)$

– how do we estimate $\vec{\Theta}$?

Maximum likelihood estimate!

MLE

$\vec{x}$

$$\vec{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \sim \mathcal{N}(\vec{0}, K_\Theta), \quad K_\Theta = \begin{pmatrix} k_\Theta(x_1, x_1), \ldots & k_\Theta(x_1, x_n) \\ \vdots & \\ & k_\Theta(x_n, x_n) \end{pmatrix}$$

$$L(\Theta) = \log p(\vec{y} \mid \Theta) \propto \underbrace{-\frac{1}{2} \log |K_\Theta|}_{\text{"regularization"}} \underbrace{- \frac{1}{2} \vec{y}^T K_\Theta^{-1} \vec{y}}_{\text{"goodness of fit"}}$$

$$L(\Theta) \to \max_\Theta$$

$$\frac{\partial L(\Theta)}{\partial \Theta} = -\operatorname{tr}\left(K_\Theta^{-1} \frac{\partial K_\Theta}{\partial \Theta}\right) + \frac{1}{2} \vec{y} K_\Theta^{-1} \frac{\partial K_\Theta}{\partial \Theta} K_\Theta^{-1} \vec{y}$$

$$\begin{pmatrix} \partial k_\Theta(x_1, x_1) & \partial k_\Theta(x_1, x_n) \end{pmatrix}$$

$$\begin{pmatrix} \partial \theta & \cdots & \partial \theta \\ & \vdots & \end{pmatrix}$$

$K_\theta$ - size $n \times n$

$\downarrow$

$O(n^3)$

$$\vec{\theta}' = \vec{\theta} - \alpha \frac{\partial L(\vec{\theta})}{\partial \vec{\theta}} \qquad \underbrace{O(n^3)}$$

$n \ll 10000$

$n: 100, 50, d: <50$

## Interpolation

$$\hat{y}_i = \vec{k}_i^T K^{-1} \vec{y}$$

$\vdots$

$$\vec{\hat{y}} = \cancel{K}\cancel{K^{-1}} \vec{y} = \vec{y}$$

$f(x_i) + \varepsilon \qquad \varepsilon \sim N(0, \sigma^2)$ - i.i.d.

$$y := f(x_i) + \varepsilon, \quad \varepsilon \sim$$

$$\cancel{K_f}\, K_y = \begin{pmatrix} k(x_1,x_1) + \sigma^2 & & k(x_1,x_n) \\ & \ddots & \\ & & k(x_n,x_n) + \sigma^2 \end{pmatrix}$$

$$\text{cond}(K_f) = \frac{\overset{\text{min}}{\lambda_1}}{\underset{\text{max}}{\lambda_n}} \qquad \text{cond}(K_y) = \frac{\lambda_1 + \sigma^2}{\lambda_n + \sigma^2}$$

$$\frac{\lambda_1}{\lambda_n} < \frac{\lambda_1 + \sigma^2}{\lambda_n + \sigma^2} \qquad \cancel{\lambda_1 \lambda_n} + \lambda_1 \sigma^2 \overset{(\le)}{<} \cancel{\lambda_1 \lambda_n} + \lambda_n \sigma^2$$

$$\lambda_1 \le \lambda_n$$

## Cov. functions

$$k(x, x') = \exp\left(- \sum_{i=1}^{d} \frac{(x_i - x_i')^2}{\theta_i^2}\right) \leftarrow \text{squared covariance}$$

$$\leftarrow l \rightarrow \infty$$

d

function

$$= \exp\left(-\sum_{i=1} \frac{|x_i - x_i'|}{\theta_i}\right) \leftarrow \nu = 1$$

Matern covariance functions

$$k(x, x') = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}\, r}{\ell}\right)^{\nu} K_\nu\left(\frac{\sqrt{2\nu}\, r}{\ell}\right),$$

Bessel functional of II kind

$$r = \|x - x'\|$$

$(\nu - 1)$ - differentiable

$\nu = 3/2$, $\nu = 5/2$

1. $k(x, x) = 1$

2. $k(x, x') \geq 0$

3. $k(x, x')$ - positive definite

Kriging, krige ← Geoscience

$\nu = 5/2$

$$(\vec{x}_1, \dots, \vec{x}_n) \to K > 0$$

$$\vec{s}^T K \vec{s} \geq 0 \quad \text{- non-negative definite}$$

Ex. $k(\vec{x},\vec{x}') = \vec{x}^T \vec{x}'$

$\Rightarrow$ Regularized linear regression

Rasmussen, "Gaussian processes in Machine Learning", 2006

Comp. compl.

$O(n^3)$ - prohibitive if $n > 10\,000$

$K \rightsquigarrow \hat{K}$

Nyström approximation

$X_n = (x_1 \dots x_n) \sim X_m = (x_1 \dots x_m)$ $\quad m \ll n$

$K_n \approx \underbrace{K_{nm}}_{O(m, n^2)} \underbrace{K_m^{-1}}_{O(m^3)} \underbrace{K_{mn}}_{O(m^2 n)}$ $\quad m \times n \quad m \times n$

$K_n, K_m$ - c.m. $X_n$ & $X_m$

$$K_{nm} = \begin{pmatrix} k(x_1, x_1) & \dots & k(x_1, x_m) \\ & \vdots & \\ & \vdots & \end{pmatrix}$$ size $n \times m$

$O(m^2 n)$

2) $k(x, x') \approx \sum_{k=1}^{S} \underbrace{\varphi_k(x) \varphi_k(x')}$

- scalar product

$k(\|x - x'\|) = \int s(z) \exp(-iz\|x - x'\|) dz$

$\approx \sum$

$( \ \ )^{-1}$

$$\vec{w} = (X'X)^{-1} X' \vec{y}$$

$$\vec{w}^{*} \ominus \quad d \times d \quad d \times n \quad n \times 1$$

$$O(d^2 n) \quad \underline{O(s^2 n)}$$

$$\boxed{n = 10 \cdot 10^6}$$

Dimension

$$d \leq 50$$

1. Projection : $\tilde{\vec{x}} \to \vec{z} = W_p \vec{x}$

2. Representation : $\vec{x} \to e(\vec{x}) \to \vec{z}$

Bayesian optimization

$$f(\vec{x}) \to \max_{\vec{x}}$$

1. smooth (m.b. with noise)

2. expensive : only n points

2. [illegible]

3. $x \in \mathbb{R}^d$, $d \leq 50$

$$\mathcal{D} = \begin{pmatrix} (\vec{x}_1, \dots, \vec{x}_m) \\ (y_1, \dots, y_m) \end{pmatrix} \qquad \mathcal{D} \to \hat{f}(\vec{x}) \approx f(\vec{x})$$

UCB
upper confidence bound




1. $x^*_k = \arg\max_{x \in X} \hat{f}_k(\vec{x}) + \beta \hat{\sigma}^2_{f_k}(\vec{x})$
2. $f(x^*_k)$
3. $(\hat{x}^*_k, f(x^*_k)) \to \mathcal{D}_k \Rightarrow \mathcal{D}_{k+1}$
4. $\hat{f}_{k+1}$ – from $\mathcal{D}_{k+1}$

Exploration + Exploitation
greedy behavior

Gaussian process
[illegible] to construct $\hat{f}_k(x)$!