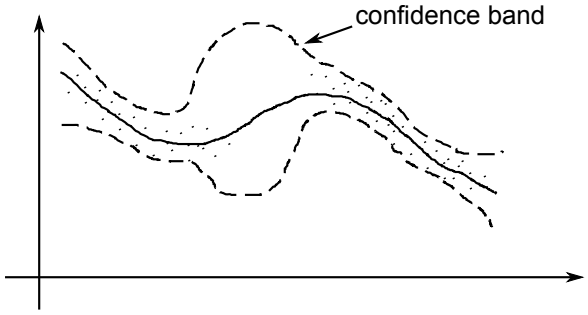
CS/CNS/EE 253: Advanced Topics in Machine Learning
Topic: Nonparametric learning #4: Gaussian Process Regression **Lecturer:** Andreas Krause
Scribe: Tim Black **Date:** March 1, 2010

15.1 Last Lecture

Want to solve a regression problem.



$$f^* = \underset{f \in H_k}{\operatorname{argmin}} \|f\|^2 + \sum_i (y_i - f(x_i))^2 \tag{15.1.1}$$

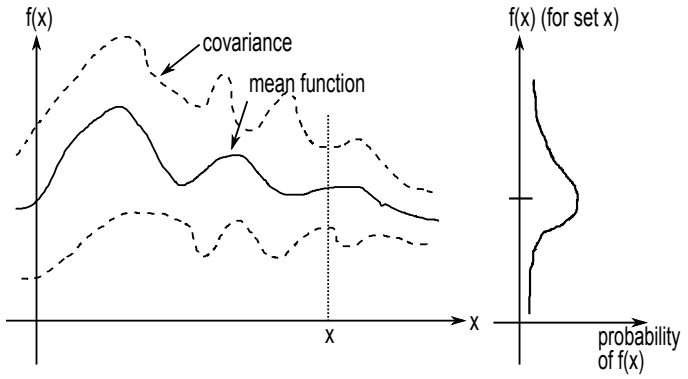
This trades off complexity and goodness of fit. It does not encapsulate the probability of being right.

Idea: Prior $P(f)$. Get data (x_i, y_i), where $y_i = f(x_i) + \epsilon_i$ is function plus some error. Posterior $P(f|D)$.

15.2 Gaussian Processes (GPs)

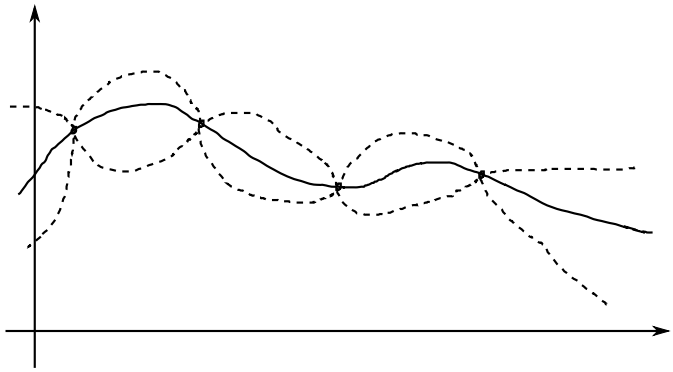
"∞-dimensional Gaussians"

Definition 15.2.1 *Let X be an input set. Let $k : X \times X \to \mathbb{R}$ be a positive definite kernel function. Let $\mu : X \to \mathbb{R}$ be the mean function (with no restrictions; later we will take it to be the 0 function). A random function $f : X \to \mathbb{R}$ is a **GP**, $f \sim \mathrm{GP}(\mu, k)$ if for any $A \subset X$, $A = \{x_1, x_2, \ldots, x_n\}$, $f_A = (f(x_1), \ldots, f(x_n)) \sim \mathcal{N}(\mu_A, \Sigma_{AA})$, where $\mu_A = (\mu(x_1), \ldots, \mu(x_n))$ and the covariance matrix Σ_{AA} is the $n \times n$ matrix whose (i, j) entry is $k(x_i, x_j)$.*



15.2.1 Prediction in GPs

Suppose we get to see $f_A = f'$ (that is, we get to see the value of f at a set of points without noise). What is $P(f(x)|f_A = f')$? Conditionals on a Gaussian give a Gaussian: $P(f(x)|f_A = f') = \mathcal{N}(f(x); \mu_{x|A}, \sigma^2_{x|A})$, where $\mu_{x|A} = \mu_x + \Sigma_{xA}\Sigma^{-1}_{AA}(f' - \mu_A)$, $\sigma^2_{x|A} = \sigma^2_x - \Sigma_{xA}\Sigma^{-1}_{AA}\Sigma_{Ax}$, and $\Sigma_x A = (k(x, x_1), k(x, x_2), \ldots, k(x, x_n))$.



What if we have noise?

$y_i = f(x_i) + \epsilon_i$, $\epsilon_i \sim \mathcal{N}(0, \Theta^2)$

$y_A = f_A + \epsilon_A$, $\epsilon_A \sim \mathcal{N}(0, \sigma^2 I)$

$P(y_A) = \mathcal{N}(y_A; \mu_A, \Sigma_{AA} + \sigma^2_{x|A})$

$P(f(x)|y_A) = \mathcal{N}(f(x); \mu_{x|A}, \sigma^2_{x|A})$

$$\boxed{\mu_{x|A} = \mu_x + \Sigma_{xA}(\Sigma_{AA} + \sigma^2 I)^{-1}(y_A - \mu_A)}$$

$$\boxed{\sigma^2_{x|A} = \sigma^2_x + \Sigma_{xA}(\Sigma_{AA} + \sigma^2 I)^{-1}\Sigma_{Ax}}$$

NEVER EVER CALCULATE $(\Sigma_{AA} + \sigma^2 I)^{-1})$ AS `inv(`$\Sigma_{AA} + \sigma^2 I$`)`. Instead, solve the linear

system $\alpha = \tilde{\Sigma}_{AA} \backslash (y_A - \mu_A)$ (Matlab notation), where $\tilde{\Sigma}_{AA} = \Sigma_{AA} + \sigma^2 I$, and $\alpha = \tilde{\Sigma}_{AA}^{-1}(y_A - \mu_A)$.

15.2.2 Connection to RKHS

Assume $\mu = 0$.

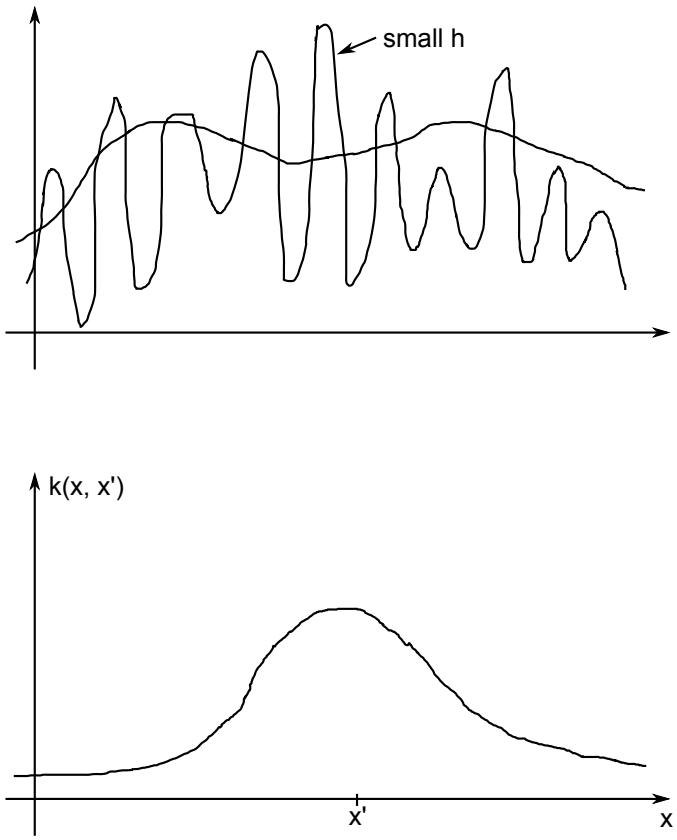
$\mu_{x|A} = \mu_x + \Sigma_{xA}(\Sigma_{AA} + \sigma^2 I)^{-1} y_A = \sum_{i=1}^n \alpha_i k(x_i, x)$.

$\underset{f \in H_k}{\text{argmax}} P(f|y_A) = \underset{f \in H_k}{\text{argmin}} \|f\|^2 + \frac{1}{\sigma^2} \sum_i (y_i - f(x_i))^2$.

15.2.3 Parameter Estimation

Most kernel functions have parameters.

$k_{SE}(x, x') = c^2 exp(-\frac{(x-x')^2}{h})$. Parameters are $\Theta = (c, h)$. c is the "magnitude" of the functions, and h is the "length scale" of f.



Let $\mathbb{N}_u$ = Number of upcrossings at level in $[0, 1]$ ("How wiggly is this function"). Assume k is **isotropic**, $k(x, x') = k(\|x - x'\|)$, and assume $\mu = 0$.

Theorem 15.2.2 *(Adler).*

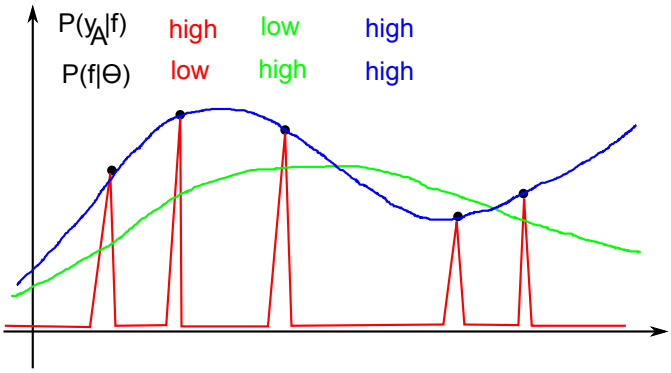
$$\mathbb{E}[N_u] = \frac{1}{2\pi}\sqrt{\frac{k''(0)}{k(0)}}exp\left(-\frac{u^2}{2k(0)}\right) \tag{15.2.2}$$

For SE kernel: $k(0) = c^2$, $k''(0) = -2\frac{c^2}{h^2}$

$\mathbb{E}[N_u] = \frac{1}{2\pi}\sqrt{\frac{2}{h^2}}exp\left(-\frac{u^2}{2c^2}\right) = \frac{1}{\pi\sqrt{2}}\frac{1}{h}exp\left(-\frac{u^2}{2c^2}\right)$.

How do we choose the parameters? Learn from data! Pick parameters to maximize likelihood.

$\Theta^* = \underset{\Theta}{\text{argmax}} P(y_A|\Theta) = \underset{\Theta}{\text{argmax}} \int P(y_A, f|\Theta)df = \underset{\Theta}{\text{argmax}} \int P(y_A|f,\Theta)P(f|\Theta)d\Theta = \underset{\Theta}{\text{argmax}} \int P(y_A|f)P(f|\Theta)d\Theta$



$P(y_A|\Theta) = \mathcal{N}(0, \Sigma_{AA}(\Theta) + \sigma^2 I) = (2\pi(\Sigma_{AA}(\Theta)))^{-n/2}exp(-\frac{1}{2}y_A^T(\Sigma_{AA}(\Theta) + \sigma^2 I)^{-1}y_A)$.

Solve the optimization problem by gradient descent.

$$\boxed{\Theta^* = \underset{\Theta}{\text{argmax}} P(y_A|\Theta)}$$

How do we solve it? Calculate gradient of $log(P(y_A|\Theta))$. Run **conjugate gradient descent**.

15.2.4 Incorporating prior knowlege

$f = g + h \sim \text{GP}(0, k_{lin} + k_{SE})$, where g is parametric and h is nonparametric.

$g(x) = \sum_i w_i\phi_i(x)$. $h \sim \text{GP}(0, k_{SE})$. $w \sim \mathcal{N}(0, I)$. $g \sim \text{GP}(0, k_{lin})$. $k_{lin}(x, x') = \sum_i \phi_i(x)\phi_i(x') = \phi(x)^T\phi(x')$.