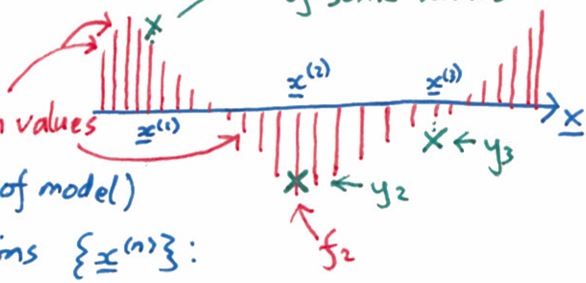
Gaussian Processes

"Function viewpoint"



Prior directly on function values

(These are the parameters of model)

Marginal belief at locations $\{\underline{x}^{(n)}\}$:

$$p(\underline{f}) = N(\underline{f};\ \underline{0},\ K)$$

$$K_{ij} = k(\underline{x}^{(i)}, \underline{x}^{(j)})$$

Posterior at $\{\underline{x}_*\}$

$$p(\underline{f}_* | \underline{y}) = N(\underline{f}_*;\ \bar{\underline{f}}_*,\ V)$$

Expressions in notes

"Weight space view"

$$f(\underline{x}) = \underline{w}^T \underline{\phi}(\underline{x}), \qquad \text{Prior: } \underline{w} \sim N(\underline{0}, \sigma_w^2 \mathbb{I})$$

$$\Rightarrow k(\underline{x}^{(i)}, \underline{x}^{(j)}) = \sigma_w^2\, \underline{\phi}(\underline{x}^{(i)})^T \underline{\phi}(\underline{x}^{(j)})$$

GPs are Bayesian linear regression

+ "kernel trick" to use many/∞ basis functions

Aside Kernel Logistic Regression

Linear case, SGD, "one epoch"

cost $O(DN)$

$\underline{w} \leftarrow \underline{0}$ $\quad D \times 1$

for $n = 1 \ldots N$:

$\quad \underline{w} \leftarrow \underline{w} + u^{(n)} \underline{x}^{(n)} - \lambda \underline{w}$

step size × gradient

$\underline{w}$ is in span of $\{\underline{x}^{(n)}\}$

$\underset{D\times 1}{\underline{w}} = \cancel{\underline{a}^{T}} \underset{N \times D}{X^{T}} \underline{a} = \sum_n a_n \underline{x}^{(n)}$

$\underline{a} \leftarrow \underline{0}$ $\quad N \times 1$

cost $O(N^2)$

for $n = 1 \ldots N$:

$\quad \underline{a} \leftarrow (1-\lambda)\, \underline{a}$

$\quad a_n \leftarrow a_n + u^{(n)}$

At test time, or to find $u^{(n)}$

need $\underline{w}^{T} \underline{x}^{(*)} = \sum_n a_n \underline{x}^{(n)T} \underline{x}^{(*)}$

$= \sum_n a_n k(\underline{x}^{(n)}, \underline{x}^{(*)})$

For 1 prediction:

$$p(f_* \mid \underline{y}) = \mathcal{N}(f_*; m, s^2)$$

Can show:

$$m = \overbrace{\underline{k}^{*T}(K + \sigma_y^2 \mathbb{I})^{-1}}^{\text{linear predictor}} \underline{y}$$

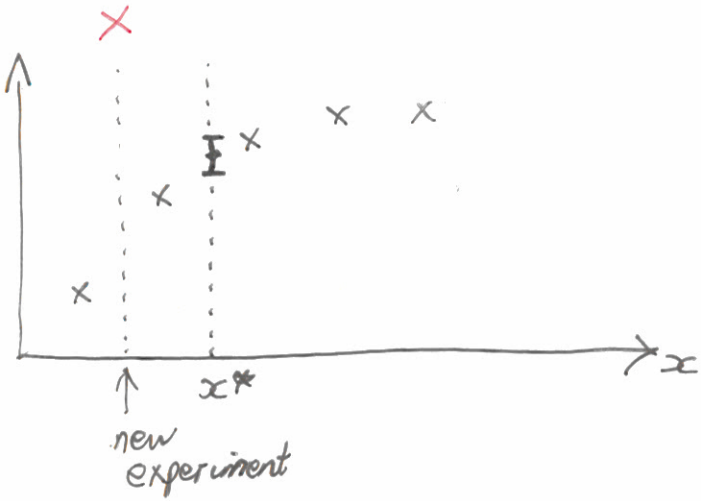
$$(k^*)_i = k(\underline{x}^{(*)}, \underline{x}^{(i)})$$

Training input

Test input

$$s^2 = k(\underline{x}^{(*)}, \underline{x}^{(*)}) - \underbrace{\underline{k}^{*T} \underbrace{(K + \sigma_y^2 \mathbb{I})^{-1}}_{\text{+ve definite}} \underline{k}^{(*)}}_{\text{+ve.}}$$

Has no y dependence



Can be more uncertain at x^*

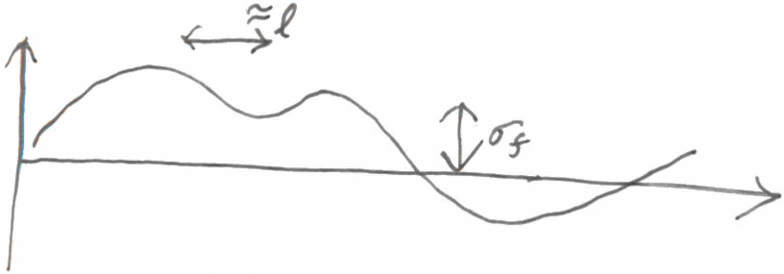
If I change the GP in response to new data. Need to learn the kernel

<u>Example</u>

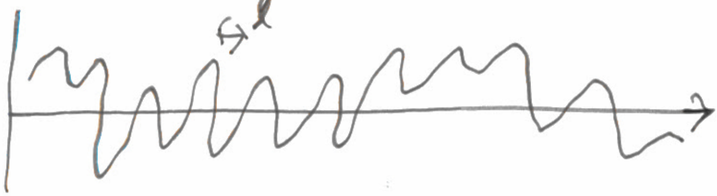
$$k(\underline{x}^{(i)}, \underline{x}^{(j)}) = \sigma_f^2 \exp\left(-\frac{1}{2}\sum_d \left(x_d^{(i)} - x_d^{(j)}\right)^2 / \ell_d^2\right)$$

Also learn σ_y^2, noise level

Function variance



small l



Pick parameters by (marginal) likelihood

$$p(y \mid X, \theta = \{\sigma_y^2, \sigma_f^2, \{l_d\}\})$$

$$= \mathcal{N}(y;\, \underline{0},\, K + \sigma_y^2 \mathbb{I})$$

$\uparrow$ $O(N^2)$ RAM

Factor costs $O(N^3)$ time

$$-\log p(y \mid X \cdots) = -\tfrac{1}{2}\, y^\top (K + \sigma_y^2 \mathbb{I})^{-1} y$$

$$- \log |2\pi (K + \sigma_y^2 \mathbb{I})|$$

Bayesian Logistic Regression

$$p(y=1 \mid \underline{x}, \underline{w}) = \sigma(\underline{w}^T \underline{x}) = \frac{1}{1+e^{-\underline{w}^T \underline{x}}}$$

Maximized Likelihood:

$$p(\underline{y} \mid X, \underline{w}) = \prod_n \sigma(\underline{w}^T \underline{x}^{(n)} z^{(n)})$$

$$= P(D \mid \underline{w})$$

$z^{(n)} \in \{\pm 1\}$

Training data.

Bayesian Logistic Regression

$P(X \mid \underline{w}) = 1$

$$p(\underline{w} \mid D) = \frac{p(D \mid \underline{w})\, p(\underline{w})}{p(D)} \propto p(D \mid \underline{w})\, p(\underline{w})$$

Can't do this exactly

$$\left[\; P(D) = \int P(D \mid \underline{w})\, p(\underline{w})\, d\underline{w}\right.$$

$P(D)$ or $P(D \mid M)$ Marginal Likelihood

Prior σ_w

Choice pre-processing or basis functions

Make predictions

$$p(y^* \mid \underline{x}^*, D) = \int p(y^*, \underline{w} \mid \underline{x}^*, D)\, d\underline{w}$$

test label   test input

$$= \int p(y^* \mid \underline{w}, \underline{x}^*)\, \underbrace{p(\underline{w} \mid D)}_{\text{Posterior over } \underline{w}}\, d\underline{w}$$