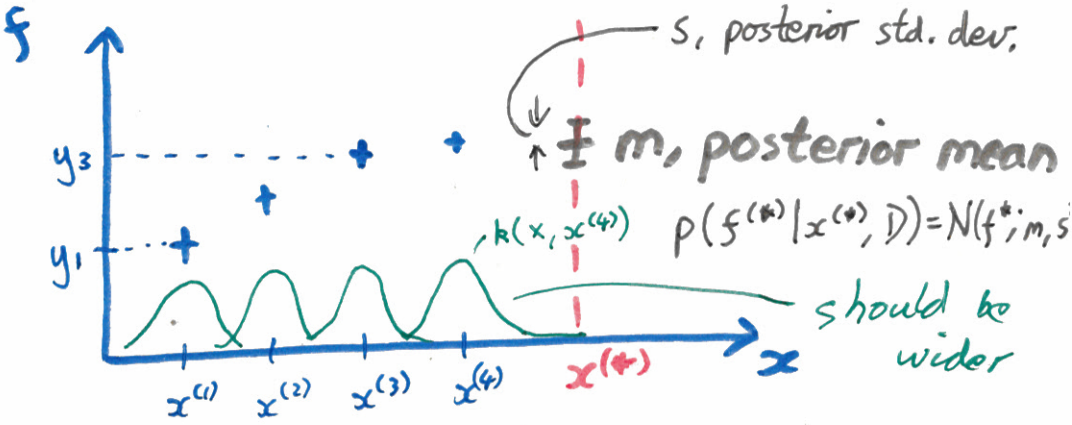
GAUSSIAN PROCESSES



$N=4$ training points

$N \times N$ matrix K, $K_{ij} = k(x^{(i)}, x^{(j)})$

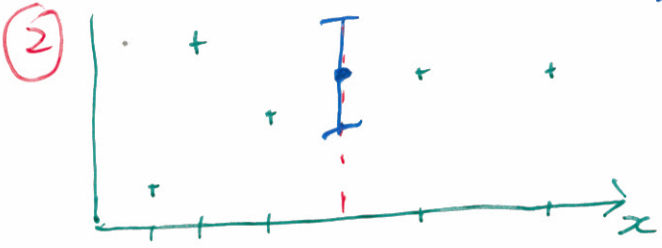
Can show: $(\underline{k}^{(*)})_i = k(\underline{x}^{(*)}, \underline{x}^{(i)})$

$$m = \underline{k}^{(*)T} \left(K + \sigma_n^2 I\right)^{-1} \underline{y}$$

$$s^2 = \underbrace{k(x^{(*)}, x^{(*)})}_{\text{Usually } \sigma_f^2} - \underline{k}^{(*)T} \underbrace{\left(K + \sigma_n^2 I\right)^{-1}}_{M^{-1}} \underline{k}^{(*)}$$

$k(\underline{x}^{(i)}, \underline{x}^{(j)}) = \sigma_f^2 \, e^{-\frac{1}{2}\|\underline{x}^{(i)} - \underline{x}^{(j)}\|}$ Has no dependence on $\underline{y}$

Long lengthscale

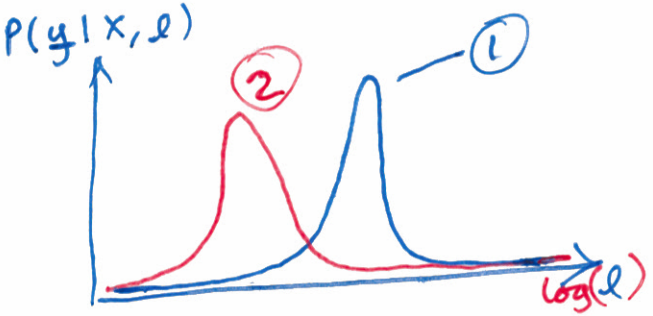


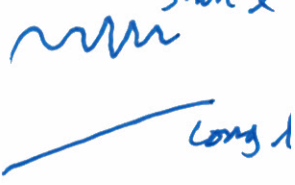
Short lengthscale. Should be more uncertain if y are surprising.

We need learn params of kernel

$$k(x^{(i)}, x^{(j)}) = e^{-\frac{1}{2}(x^{(i)} - x^{(j)})^2 / \ell^2}$$

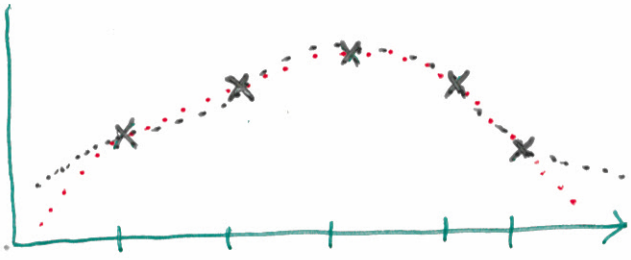
↑ lengthscale





Assuming I know small noise level.

Look at the code demo on the website



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:

$$P(\underline{y} \mid X, \underline{w}) = \prod_n \sigma(z^{(n)} \underline{w}^T \underline{x}) \qquad z^{(n)} \in \{\pm 1\}$$

$$z^{(n)} = 2y^{(n)} - 1$$

$$= P(D \mid \underline{w}) \qquad y^{(n)} \in \{0, 1\}$$

↑ Training data.

Minimize a penalized neg. log likelihood

$$\underline{w}^* = \underset{\underline{w}}{\operatorname{argmin}} \left[-\log P(D \mid \underline{w}) + \lambda \underline{w}^T \underline{w} \right]$$

↑ Fit, guess of weights.

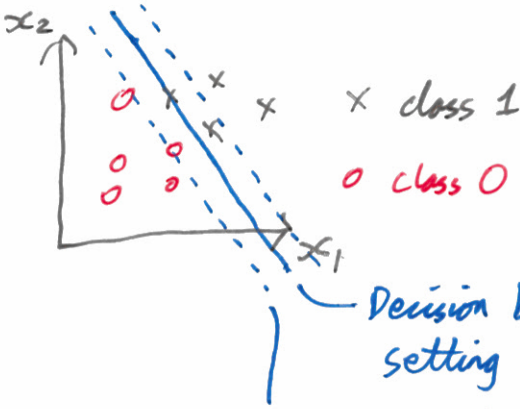
Bayes' Rule

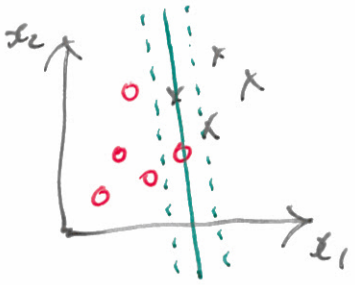
$$p(\underline{w} \mid D) = \frac{p(D \mid \underline{w})\, p(\underline{w})}{p(D)} \propto p(D \mid \underline{w})\, p(\underline{w})$$

$$p(D) = \int p(D \mid \underline{w})\, p(\underline{w})\, d\underline{w} \quad \Big] \text{ we can't do this integral exactly.}$$

$p(D)$ or $p(D \mid M)$

↳ model, choice of basis f^n's
choice of prior





---- contour of $p(y=1 \mid \underline{x}, \underline{w})$

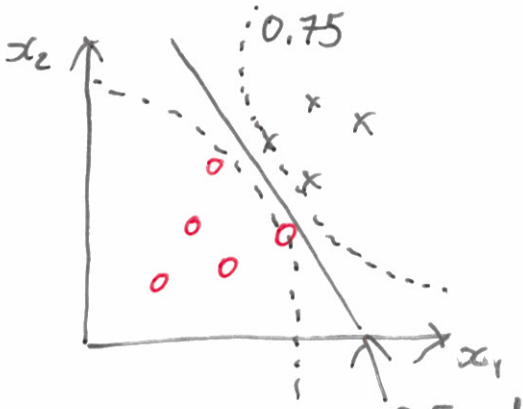
Bayesian Predictions

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

(sum rule)

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

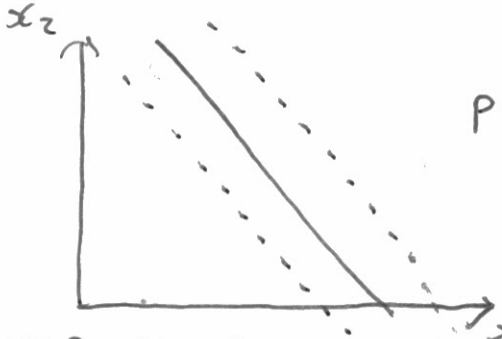
(product rule)



Contours $p(y \mid \underline{x}, D)$

0.5, decision boundary

Compare to MAP predictions



$p(y \mid \underline{x}, \underline{w}^*)$

MAP = Maximum a posteriori.

$$\underline{w}^* = \underset{\underline{w}}{\operatorname{argmax}} \log p(\underline{w} \mid D)$$

$$= \underset{\underline{w}}{\operatorname{argmax}} \left[\log p(D \mid \underline{w}) + \underbrace{\log p(\underline{w})}_{-\frac{1}{2\sigma_w^2} \underline{w}^T \underline{w} + \text{const}} \right]$$

If $p(\underline{w}) = N(\underline{w}; 0, \sigma_w^2 I)$

MAP is not Bayesian,

can be seen as a crude approximation