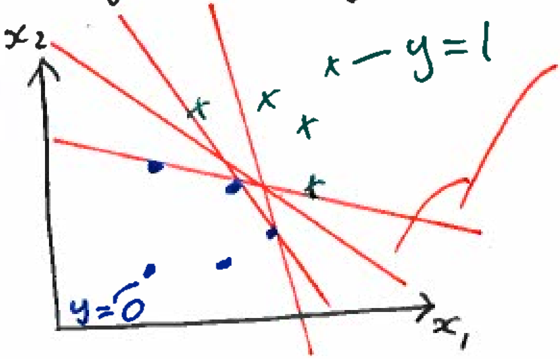
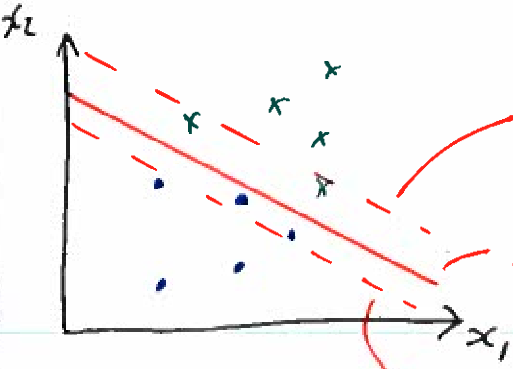
Bayesian Logistic Regression



Decision boundaries:
$p(y=1|\underline{x},\underline{w}) = \sigma(\underline{w}^T\underline{x}) = \frac{1}{2}$
for different plausible $\underline{w}$

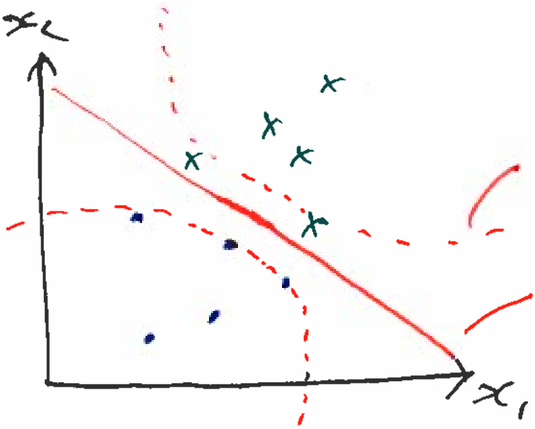


$\sigma(\hat{\underline{w}}^T\underline{x}) = 0.9$

$\sigma(\hat{\underline{w}}^T\underline{x}) = 0.5$

$\sigma(\hat{\underline{w}}^T\underline{x}) = 0.1$

Predictive distribution for a single fitted $\hat{\underline{w}}$, eg L2 regularized fit



$p(y=1|\underline{x},D) = 0.9$

$p(y=1|\underline{x},D) = 1/2$

↑ Training data

Plausible weights described by Posterior

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

D: Train $\{\underline{x}^{(n)}, y^{(n)}\}$

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

M, model choices, hyperparameters, basis $f^{\underline{n}}$'s
(Often miss out)

Likelihood

$$P(D \mid \underline{w}) = \prod_n \cancel{p(x^{(n)} \mid \underline{w})}\; \underbrace{p(y^{(n)} \mid \underline{w}, \underline{x}^{(n)})}_{\sigma(\underline{w}^T \underline{x}^{(n)}(2y^{(n)} - 1))}$$

1 "known" here

Prior

Prior: $p(\underline{w}) = \mathcal{N}(\underline{w}; \underline{0}, \sigma_w^2 \mathbb{I})$

Marginal Likelihood

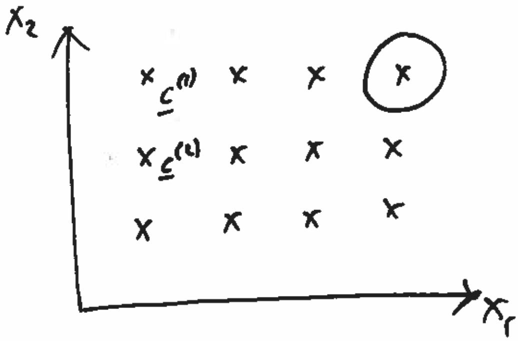
Marginal Likelihood $p(D)$ or $p(D \mid M)$ can compare models

Predictions

$$p(y = 1 \mid \underline{x}, D) = \int p(y = 1, \underline{w} \mid \underline{x}, D)\, d\underline{w}$$

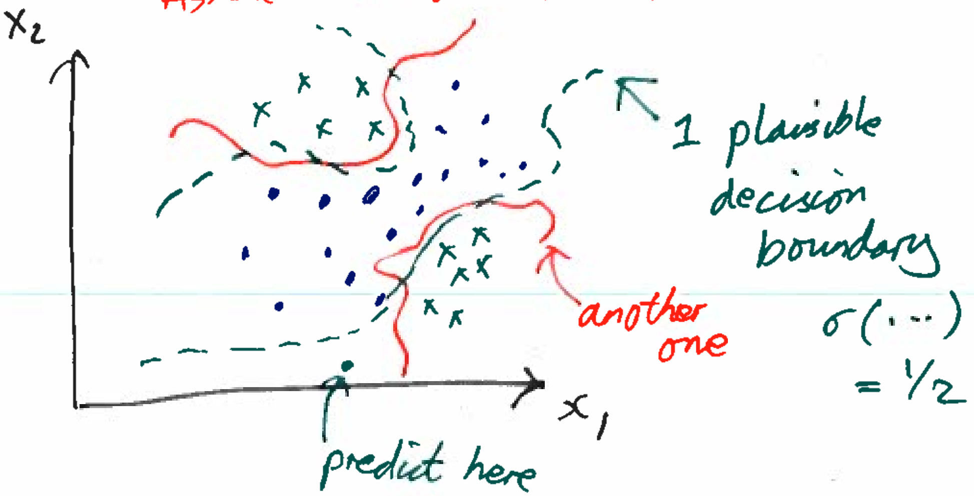
y: test output; $\underline{x}$: test input

$$= \int \underbrace{p(y = 1 \mid \underline{w}, \underline{x}, \cancel{D})}_{\sigma(\underline{w}^T \underline{x})}\; \underbrace{p(\underline{w} \mid \cancel{\underline{x}}, D)}_{\text{posterior}}\, d\underline{w}$$



$$p(y=1 \mid \underline{x}, \underline{w}) = \sigma\left(\underline{w}^T \underline{\phi}(\underline{x})\right)$$

Assume there are 100 RBFs



and different plausible $\underline{w}$'s make different predictions

Marginal Likelihood

$$p(\underset{\substack{\uparrow \\ \text{Training data}}}{D} \mid \text{model choices}) = \int p(D, \underline{w} \mid \text{model})\, d\underline{w}$$

Sketch Posterior for one data point

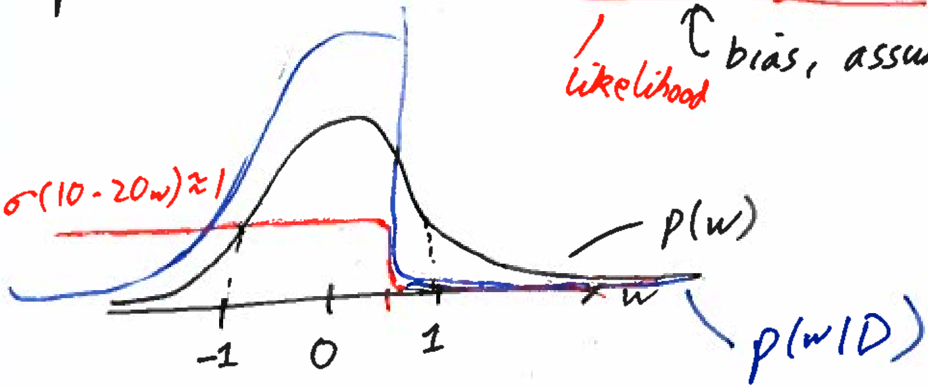
$x = -20$
$y = +1$

$$p(w) = N(w; 0, 1)$$

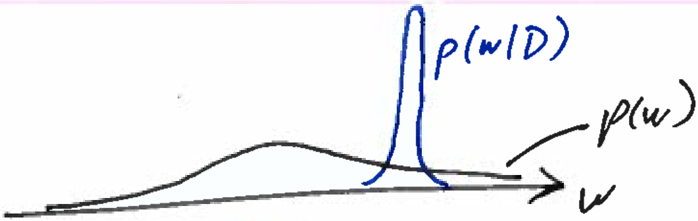
$$p(w \mid D) \propto N(w; 0, 1)\, \sigma(10 - 20w)$$

likelihood

bias, assume known



In the notes $N = 500$ posterior:



Non-exam
"Bayesian central limit theorem"

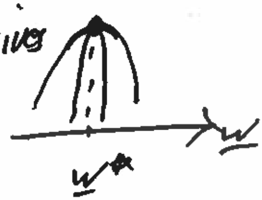
Laplace Approximation

Fits a Gaussian to $p(\underline{w} \mid D)$

$$\log p(\underline{w} \mid D) = \log p(\underline{w}, D) + \text{const wrt } \underline{w}$$

Quadratic in $\underline{w}$ if Gaussian.

Find mode and find 2^{nd} derivatives



"Energy" $E(\underline{w}) = -\log p(\underline{w}, D)$

$$\underline{w}^* = \arg\min_{\underline{w}} E(\underline{w})$$

((L2 regularisation or "MAP" fit)

Hessian $H_{ij} = \left.\dfrac{\partial^2 E}{\partial w_i \partial w_j}\right|_{\underline{w}^*}$

$$\boxed{p(\underline{w} \mid D) \approx \mathcal{N}(\underline{w};\ \underline{w}^*, H^{-1})}$$

Approx. Normaliser $p(D|M)$

$$p(\underline{w}|D) = \frac{p(\underline{w}, D)}{P(D)} \approx \mathcal{N}(\underline{w}; \underline{w}^*, H^{-1})$$

$$= \frac{|H|^{1/2}}{(2\pi)^{D/2}} e^{-1/2(\underline{w}-\underline{w}^*)^T H (\underline{w}\ldots}$$

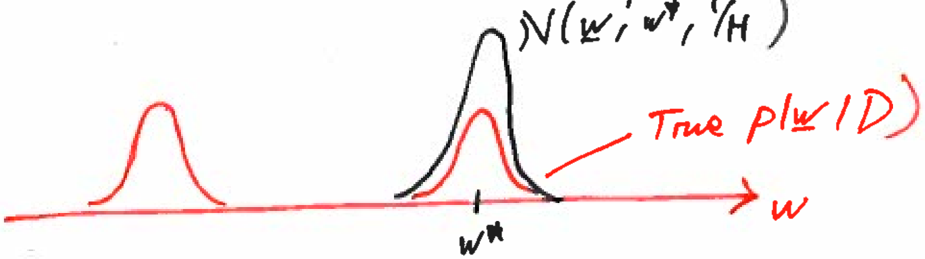
Evaluate approx at $\underline{w} = \underline{w}^*$:

$$\frac{P(\underline{w}^*, D)}{P(D)} \approx \frac{|H|^{1/2}}{(2\pi)^{D/2}}$$

parameters

Training data

$$P(D) = \frac{p(\underline{w}^*, D)}{P(\underline{w}^*|D)} \approx \frac{P(\underline{w}^*, D)}{|H|^{1/2}} (2\pi)^{D}$$



$P(D) \approx \ldots$ from Laplace

A) Too Big; B) Too Small; C) Correct

Z) ???

Modelling game outcomes

Two player game → win / lose (no draws)

Player 1	Player 2	Winner
Fred	Nuria	Nuria
Jo	Tuia	Jo
Tuia	Nuria	Nuria
...	...	...

How might we predict winner given players?

Can we rank players, best → worst?

Think of simple approaches first