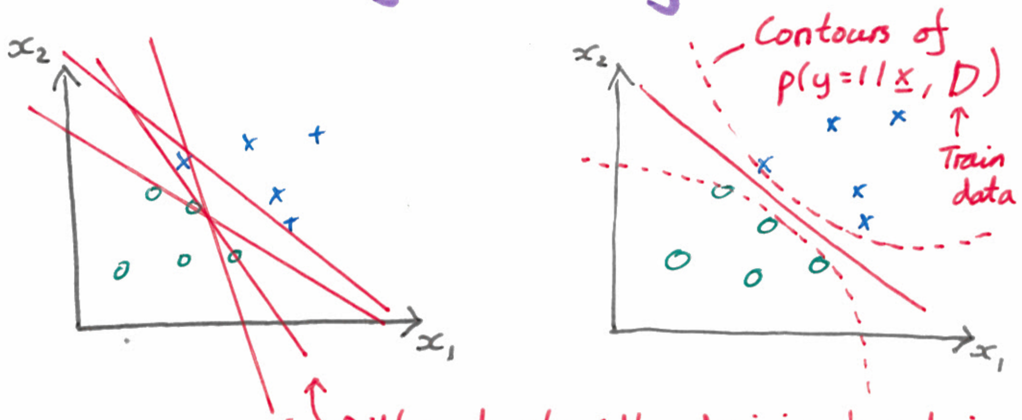
Bayesian Logistic Regression



Different plausible decision boundaries

$p(y=1|\underline{x},\underline{w}) = \sigma(\underline{w}^T\underline{x}) = 1/2$

Posterior

$$p(\underline{w}|D, M) = \frac{\overbrace{p(D|\underline{w}, M)}^{\text{Large product sigmoids}}\, p(\underline{w}|M)}{p(D|M)}$$

Model

$p(D|M)$ } Marginal likelihood $= \int p(D|\underline{w})\, p(\underline{w})\, d\underline{w}$

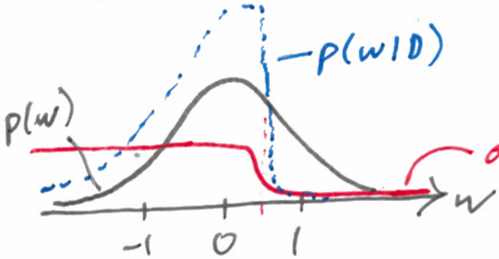
Predictions

$$p(y|\underline{x}, D) = \int p(y|\underline{x}, \underline{w})\, \underbrace{p(\underline{w}|D)}_{\text{Not simple!}}\, d\underline{w}$$

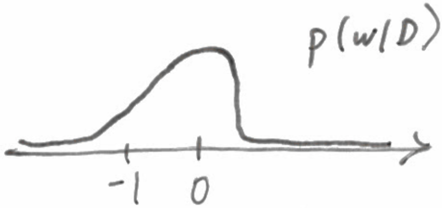
Sketch Posterior for one data point

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

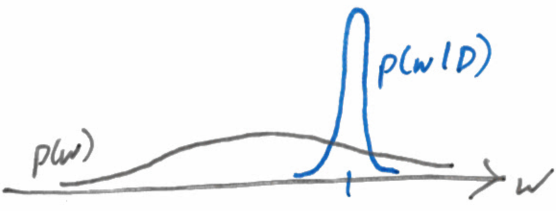
$$p(w|D) \propto N(w;0,1)\,\sigma(\underbrace{10}_{\text{known bias}=10}\ \overbrace{-20w}^{x=-20,\ y=1})$$



⇓



In the notes: sketch N=500 posterior



Laplace Approximation

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

- Matching the mode.
- Matches curvature (2nd derivatives) of $\log p(\underline{w} | D)$

$$\underline{w}^* = \arg\max_{\underline{w}} p(\underline{w} | D) \qquad \text{(MAP parameters)}$$

$$= \arg\max_{\underline{w}} \log p(\underline{w}, D) \qquad \text{Numerical optimization}$$

"Energy"

$$E(\underline{w}) = -\log p(\underline{w}, D) \qquad \text{neg. log prob. up to a constant}$$

Find minimum, then evaluate curvature.

1D: $$H = \left. \frac{\partial^2 E}{\partial w^2} \right|_{w = w^*}$$

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

Compare to Gaussian

"Energy" = $-\log$ prob (throwing away constants)

$$E_{N(w;\mu,\sigma^2)}(w) = \frac{(w-\mu)^2}{2\sigma^2}$$

Minimum of E_N : $w^* = \mu$

$$H = \frac{1}{\sigma^2} \Rightarrow \sigma^2 = \frac{1}{H}$$

$$E_{N(\underline{w};\underline{\mu},\Sigma)} = \frac{1}{2}(\underline{w}-\underline{\mu})^T \Sigma^{-1}(\underline{w}-\underline{\mu})$$

$$\underline{w}^* = \underline{\mu}$$

$$H = \Sigma^{-1} \Rightarrow \Sigma = H^{-1}$$

Laplace Approx

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

Approx. Normalizer $p(D)$

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

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

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

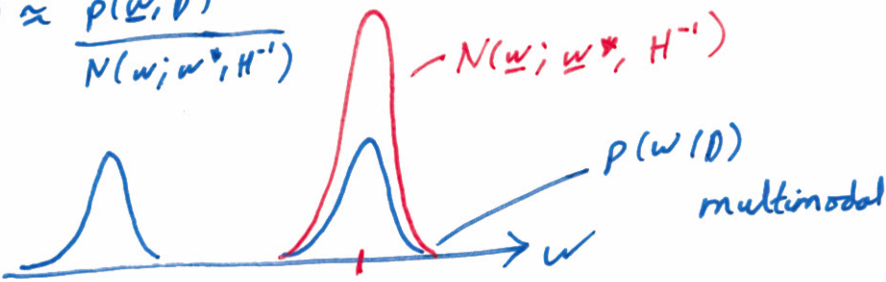
$$\frac{p(\underline{w}^*, D)}{P(D)} \approx \frac{|H|^{1/2}}{(2\pi)^{D/2}}$$

Training data

params (length of $\underline{w}$)

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

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



$P(D)$ approx: A) Too Big; B) Too Small; C) Correct; Z)?