Probabilistic Models

Often "1" in supervised models. We know $\underline{x}^{(n)}$

Likelihood

$$p(\mathcal{D} \mid \underline{w}) = \prod_{n=1}^{N} p(\underline{x}^{(n)} \mid \underline{w})\, p(y^{(n)} \mid \underline{x}^{(n)}, \underline{w})$$

Data $\{\underline{x}^{(n)}, y^{(n)}\}$

For each can minimize negative log likelihood (+ regularizer)

Examples: $\mathcal{N}(y^{(n)}; f(\underline{x}^{(n)}; \underline{w}), \sigma_y^2)$

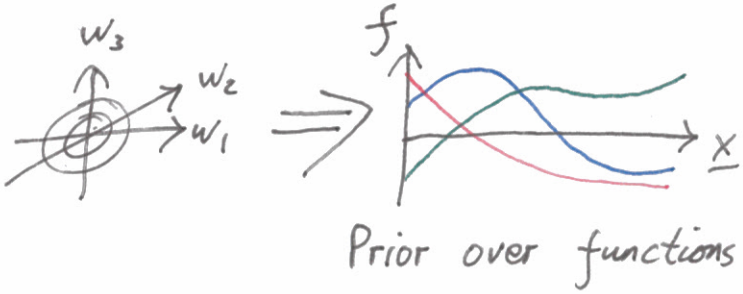
$\mathcal{N}(y^{(n)}; f(\underline{x}^{(n)}; \underline{w}), e^{v}(\underline{x}^{(n)}; \underline{w}))$

Bernoulli$(y^{(n)}; \sigma(f(\underline{x}^{(n)}; \underline{w})))$

Robust versions

...

Prior $p(\underline{w})$



Prior over functions

Bayesian Inference

$\Phi, \sigma_y, \sigma_w, \ldots$

Model assumptions. Can mention everywhere. Often don't.

Posterior

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

Marginal Likelihood (dropping "$\mid M$")

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

Not Bayesian

$$\underset{\underline{w}}{\operatorname{argmax}} \; p(\underline{w} \mid D) = \text{"MAP" params}$$

$$= \underset{\underline{w}}{\operatorname{argmax}} \left[\log p(D \mid \underline{w}) + \underbrace{\log p(\underline{w})}_{\text{L2 regularizer}} \right]$$

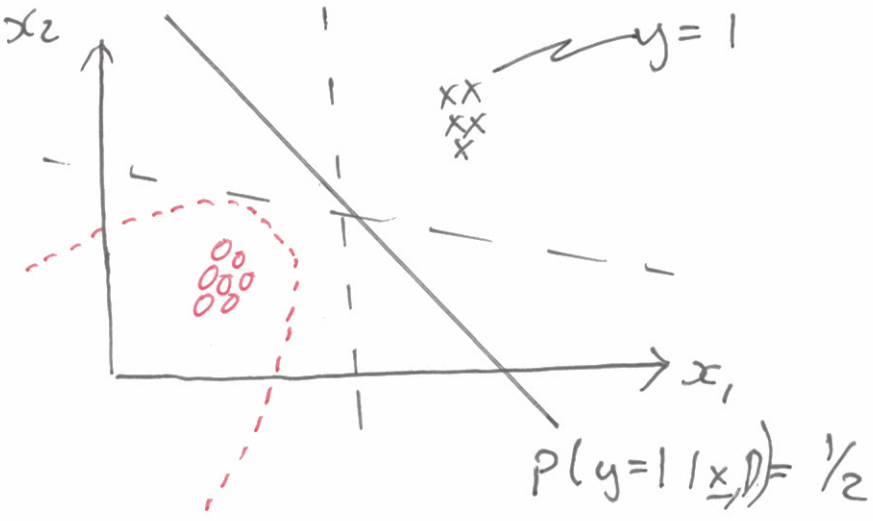
Logistic Regression

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

$$p(\underline{y} \mid X, \underline{w}) = \prod_{n=1}^{N} \sigma(\underline{w}^T \underline{x}^{(n)})^{z^{(n)}}$$

$$= p(D \mid \underline{w}) \quad (\text{ish})$$

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



$p(y=1 \mid \underline{x}, D) = 0.1$

Make predictions

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

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

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

Monte Carlo

$$p(y^* = 1 \mid \underline{x}, D) = \mathbb{E}_{p(\underline{w} \mid D)}\left[p(y^* = 1 \mid \underline{w}, x^*)\right]$$

$$= \mathbb{E}_{p(\underline{w} \mid D)}\left[\sigma(\underline{w}^T \underline{x}^*)\right]$$

$$\approx \frac{1}{S} \sum_{s=1}^{S} \sigma(\underline{w}^{(s)T} \underline{x}^*)]$$

$$\underline{w}^{(s)} \sim p(\underline{w} \mid D).$$

How? $\underline{w}^{(s)} \sim p(\underline{w} \mid D)$

Approximately with "MCMC"

(not this course)

Importance Sampling

$$\int g(x)\, p(x)\, dx = \int \left[g(x) \frac{p(x)}{q(x)} \right] q(x)\, dx$$

$$\mathbb{E}_p[g(x)] = \mathbb{E}_q[\dots] \qquad \left[\text{If } q(x) \neq 0 \text{ if } p(x) \neq 0\right]$$

For logistic regression

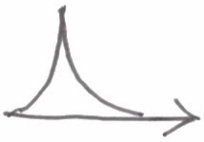
$$\underline{w}^{(s)} \sim q(\underline{w}) = \text{prior } p(\underline{w})$$

$$p(y^* \mid \underline{x}^*, D) \propto \frac{1}{S} \sum_{s=r}^{S} p(y \mid \underline{x}, \underline{w}^{(s)}) \frac{p(D \mid \underline{w}^{(s)})}{\sum_{s'=1}^{S} p(D \mid \underline{w}^{(s')})}$$

Laplace Approximation

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

- Matching mode
- Matching curvature (2nd derivative) of $\log p(\underline{w} \mid D)$



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

MAP Parameters

$$= \underset{\underline{w}}{\operatorname{argmin}} - \log p(\underline{w} \mid D)$$

Numerical methods

"Energy"

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

① $w^* = \underset{\underline{w}}{\operatorname{argmin}}\ E(\underline{w})$

② Find curvature:

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

In general Hessian, $H_{ij} = \dfrac{\partial^2 E}{\partial w_i \partial w_j}$

Laplace Approximation

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