CLASSIFICATION

$\mathcal{D} = (x_n, y_n)_{n=1,..,N}$

- DISCRIMINANT FUNCTION : $f(x) \in \{1, .., K\}$
- DISCRIMINATIVE APPROACH : $\boxed{p(C_k|x)} = f(\overbrace{w^T\phi(x)}^{h(x)})$

ACTIVATION FUNCTION

f^{-1} LINK FUNCTION

- GENERATIVE APPROACH :

$$p(C_k|x) = \frac{p(x|C_k)\,p(C_k)}{p(x)}$$

CLASS CONDITIONAL GENERATIVE MODEL

— OUTPUT ENCODING —

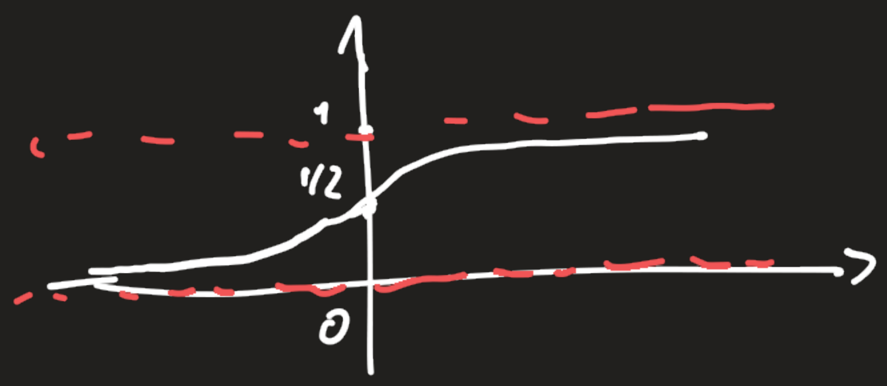
- 2-class problem : $y_n \in \{0, 1\}$
- K-class problem $(K > 2)$: $y_n = (y_{nj})_{j=1..K}$ $\quad y_{nj} \in \{0, 1\}, \sum_j y_{nj} = 1$

ONE-HOT ENCODING

LOGISTIC REGRESSION

$(x_n, y_n)_{n=1,\dots,N}$, $\phi(x) = \phi_0(x), \to \phi_{M-1}(x)$, $p(C_k \mid x) = f(w^T \phi(x))$

f : ACTIVATION FUNCTION
- LOGIT FUNCTION $\quad \sigma(a) = \dfrac{1}{1+e^{-a}}$
- PROBIT FUNCTION $\quad \Psi(a) = \int_{-\infty}^{a} \mathcal{N}(\theta \mid 0,1)\, d\theta$



$$\sigma(w^T \phi(x_i)) = s_i$$

NOISE : BERNOULLI : $p(y_n \mid x_n) = s_n^{y_n}(1-s_n)^{1-y_n}$

LIKELIHOOD $\quad p(y \mid x) = \prod_{n=1}^{N} s_n^{y_n}(1-s_n)^{1-y_n}$

$$E(w) = -\frac{1}{N} \log p(y \mid x) = -\frac{1}{N} \sum_{n=1}^{N} y_n \log s_n + (1-y_n)\log(1-s_n)$$

$$\nabla E(w) = \frac{1}{N} \sum_{i=1}^{N} (s_i - y_i)\phi(x_i)$$, $\nabla E(w) = 0$ NO ANALYTICAL SOLUTION

$w_{ML} = \arg\min_{w} E(w)$ ← CONVEX

GRADIENT DESCENT.

$$w_{n+1} = w_n - \eta_n \nabla E(w_n)$$

$x^0 \quad p(y=1 \mid x^0) = \sigma(w_{ML}^T \cdot \phi(x^0)) \geq \frac{1}{2}$ → CHOOSE 1 . CF? $\quad w_{ML}^T \phi(x) = 0$ SEPARATING HYPERPLANE

(x_n, y_n), $y_n = y_{n1}, \dots, y_{nK}$, K = # classes

$$p(C_k|x) = \sigma_k(W^T \cdot \phi(x)) \quad \text{SOFTMAX}$$

$$\Downarrow$$

$$\sigma_k(a) = \frac{\exp(a_k)}{\sum_j \exp(a_j)}$$

W^T = $K \times M$ matrix, $W := (w_{ij})_{\substack{i=1,\dots,M \\ j=1,\dots,K}}$

$$E(W) = -\frac{1}{N}\sum_{n=1}^{N} \sum_{j=1}^{K} y_{nj} \log \underbrace{s_{nk}}$$

$$s_{nk} := \sigma_k(W^T \phi(x_n))$$

$$\nabla_{w_{\cdot j}} E(W) = \frac{1}{N}\sum_{n=1}^{N} (s_{nj} - y_{nj})\,\phi(x_n)$$

GRADIENT wrt

j-th linear combination

LAPLACE APPROXIMATION

1-D case

$$p(z) = \frac{1}{Z} f(z) \quad , \quad Z = \int f(z)\,dz$$

- FIND A MODE OF $f(z)$, say z_0,
- MATCH THE CURVATURE OF f at z_0 WITH A NORMAL DISTRIBUTION.

Let z_0 a point s.t $\frac{d}{dz} f(z_0) = 0$, and z_0 maximum.

We TAYLOR EXPAND $\log f(z)$ around z_0:

$$\log f(z) \approx \log f(z_0) - \frac{1}{2} A (z - z_0)^2$$

$$A = - \frac{d^2}{dz^2} \log f(z_0) \; , \; A > 0 :$$

$$f(z) \approx f(z_0) \cdot \exp\left(-\frac{1}{2} A (z - z_0)^2\right)$$

approx p by a Gaussian $q(z)$

$$q(z) \sim \mathcal{N}(z \mid z_0, A^{-1}) = \left(\frac{A}{2\pi}\right)^{1/2} \exp\left(-\frac{1}{2} A (z - z_0)^2\right)$$

$$p(z) = \frac{1}{Z} f(z) \approx \frac{1}{Z} f(z_0) \exp\left(-\frac{1}{2} A (z-z_0)^2\right)$$

$$\shortparallel$$

$$q(z) = \left(\frac{A}{2\pi}\right)^{1/2} \exp\left(-\frac{1}{2} A (z-z_0)^2\right)$$

$$\Rightarrow Z \approx f(z_0) \left(\frac{A}{2\pi}\right)^{-1/2}$$

n-dim CASE

$p(z) = \frac{1}{Z} f(z)$, z_0 is a mode of f

$$\log f(z) \approx \log f(z_0) - \frac{1}{2} (z-z_0)^T A (z-z_0) \quad , \quad A = -\nabla\nabla \log f(z_0)$$

$$\Rightarrow q(z) = \mathcal{N}(z \mid z_0, A^{-1}) \Rightarrow Z = \frac{(2\pi)^{M/2}}{|A|^{1/2}} f(z_0)$$

$\mathcal{D}$, $\mathcal{M}$ model, depending on ϑ , $p(\vartheta)$ PRIOR

$$p(\vartheta \mid \mathcal{D}) = \frac{\boxed{p(\mathcal{D} \mid \vartheta) \cdot p(\vartheta)}}{p(\mathcal{D})}$$

← computable

← This is hard

$$p(\mathcal{D}) = \int p(\mathcal{D} \mid \vartheta)\, p(\vartheta)\, d\vartheta$$

$f(\vartheta) = p(\mathcal{D} \mid \vartheta) \cdot p(\vartheta)$ $\quad Z = p(\mathcal{D})$ $\quad$ => we can use Laplace.

1) ϑ_{MAP} .

$M = |\vartheta|$

$$\Rightarrow \log p(\mathcal{D}) \simeq \log p(\mathcal{D} \mid \vartheta_{MAP}) + \log p(\vartheta_{MAP}) + \frac{M}{2} \log(2\pi) - \frac{1}{2} \log |A|$$

$$A = -\nabla\nabla\left[\log p(\mathcal{D} \mid \vartheta_{MAP}) + \log p(\vartheta_{MAP})\right]$$

BAYESIAN INFORMATION CONTENT BIC

$$\log p(\mathcal{D}) \simeq \log p(\mathcal{D} \mid \vartheta_{MAP}) - \frac{1}{2} M \log N$$

BAYESIAN LOGISTIC REGRESSION

$(x_n, y_n)_{n=1..N}$ DATA.

$\phi(x) = \phi_0(x) \to \phi_{M-1}(x)$ $\quad \sigma(w^T \phi(x))$ LOGIT

$$\underbrace{p(w)}_{\text{PRIOR}} = \mathcal{N}(w \mid m_0, S_0)$$

$$p(\underline{y} \mid w, \underline{x}) = \prod_{i=1}^{N} s_i^{y_i} (1-s_i)^{1-y_i}$$

$$s_i = \sigma(w^T \phi(x_i))$$

$$s_i = s_i(w)$$

$$p(w \mid \underline{y}, \underline{x}) = \frac{p(\underline{y} \mid w, \underline{x}) \cdot p(w)}{p(\underline{y} \mid \underline{x})} \propto p(\underline{y} \mid w, \underline{x}) p(w)$$

$$\log p(w \mid \underline{y}, \underline{x}) = \underbrace{-\frac{1}{2}(w - m_0)^T S_0^{-1} (w - m_0)}_{\propto \log p(w)} + \underbrace{\sum_{i=1}^{N} \left[y_i \cdot s_i(w) + (1-y_i)(1-s_i(w)) \right]}_{\log p(\underline{y} \mid w)} + C$$

NOT ANALYTICALLY TRACTABLE

LAPLACE APPROXIMATION $\Rightarrow$ $p(w) = \frac{1}{Z} f(w)$, $f(w) = p(w)\, p(\underline{t} \mid w)$

1) FIND $w_{MAP} = \arg\max_w \log p(w \mid \underline{y}) = \arg\max_w \log p(w) + \log p(\underline{y} \mid w)$

2) COMPUTE HESSIAN AT w_{MAP}. $\quad S_N^{-1} = S_0^{-1} + \sum_{n=1}^{N} s_n(w_{MAP})(1 - s_n(w_{MAP}))\, \phi(x_n) \phi^T(x_n)$

$$p(\omega \mid \underline{y}) \approx q(\omega) \qquad \underbrace{q(\omega) = \mathcal{N}(\omega \mid \omega_{MAP}, S_N)}$$

PREDICTIVE DISTRIBUTION

$$p(C_1 \mid x^*, \underline{y}, \underline{x}) = \int p(C_1 \mid x^*, \omega, \underline{x}, \underline{y})\, p(\omega \mid \underline{y}, \underline{x})\, d\omega$$

$$\approx \int \sigma(\underbrace{\omega^T \phi(x^*)})\, q(\omega)\, d\omega$$

M-dim integral!

$a = \omega^T \phi(x^*)$ is 1-D GAUSSIAN

$$q(a) = \mathcal{N}(a \mid \mu_a, \sigma_a^2) \qquad \left.\begin{aligned} \mu_a &= \omega_{MAP}^T \phi(x^*) \\ \sigma_a^2 &= \phi^T(x^*) S_N \phi(x^*) \end{aligned}\right]$$

$$\Rightarrow p(C_1 \mid x^*, \underline{y}, \underline{x}) = \int \sigma(a)\, q(a)\, da$$

PROBIT APPROXIMATION $\sigma(a) \approx \Psi(\lambda a)$ $\quad \lambda^2 = \frac{\pi}{8}$ o MATCHING $\sigma'(0) = \Psi'(0)$

$$\int \Psi(\lambda a)\, q(a)\, da = \Psi\left(\frac{\mu_a}{(\lambda^{-2} + \sigma_a^2)^{1/2}}\right) \Rightarrow \boxed{p(C_1 \mid x^*, \underline{y}, \underline{x}) \approx \sigma(K(\sigma_a^2)\mu_a)}$$

$$\boxed{K(\sigma_a^2) = (1 + \pi \sigma_a^2 / 8)^{-1/2}}$$