CLASSIFICATION

dataset $D = (x_i, y_i), \; i = 1, \dots, N$

- DISCRIMINANT FUNCTION

$$f(x) \in \{1, \dots, K\}$$

- DISCRIMINATIVE APPROACH

$$P(C_k \mid x) = f\big(\overbrace{w^T \phi(x)}^{h(x)}\big)$$

f ACTIVATION FUNCTION

f^{-1} LINK FUNCTION

- GENERATIVE APPROACH

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

---

2 class problem $\quad y_n \in \{0, 1\}$

ONE HOT ENCODING

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

LOGISTIC REGRESSION

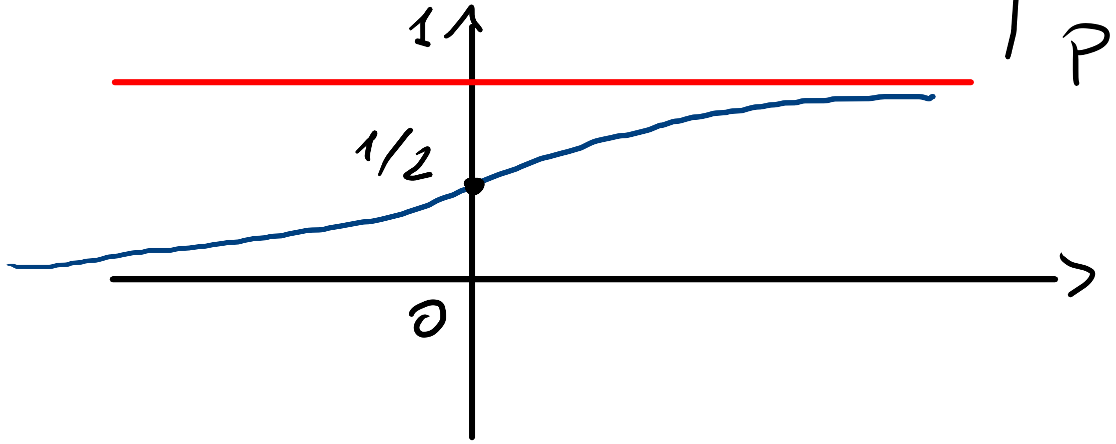
(x_n, y_n), $n=1,\dots,N$; $\phi(x) = (\phi_0(x), \dots, \phi_{M-1}(x))$ basis functions

$$p(C_1|x) = f(w^T \phi(x))$$

f activation function

LOGIT $\quad \sigma(a) = \frac{1}{1+e^{-a}}$

PROBIT $\quad \Psi(a) = \int_{-\infty}^{a} \mathcal{N}(\theta|0,1)\,d\theta$



call $\quad s_n = \sigma(w^T \phi(x_n))$

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

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

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

$$\nabla E(w) = \frac{1}{N}\sum_{i=1}^{N} (s_i - y_i)\phi(x_i) \qquad (\nabla E(w) = 0 \text{ has no analytic solution})$$

$$w_{ML} = \arg\min_{w} E(w) \qquad (\text{convex})$$

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

- SEPARATING HYPERPLANE: $w^T_{ML}\,\phi(x) = 0$

for Bayes decision rule: class 1 iff $\sigma(w^T_{ML}\phi(x)) \geq \frac{1}{2}$

MULTICLASS PROBLEM: (x_n, y_n), with $y_n = (y_{n1}, \dots, y_{nK})$ K classes

$$p(C_k \mid x) = \sigma_k(W^T \phi(x))$$

W^T is a $K \times M$ matrix

where $\sigma_k(\underline{a}) = \dfrac{\exp(a_k)}{\sum_j \exp(a_j)}$ is the SOFTMAX

Now $E(w) = -\dfrac{1}{N} \sum_{n=1}^{N} \sum_{j=1}^{K} y_{nj} \log s_{nj}$, $s_{nj} = \sigma_j(W^T \phi(x_n))$

$$\underbrace{\nabla_{w_j} E(w)}_{\text{GRADIENT FOR CLASS } j \text{ WEIGHTS}} = \frac{1}{N} \sum_{n=1}^{N} (s_{nj} - y_{nj}) \phi(x_n)$$

LAPLACE APPROXIMATION

1-D case

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

- 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

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

$f(\vartheta) = p(\mathcal{D} \mid \vartheta) \cdot p(\vartheta)$ $\quad Z = p(\mathcal{D})$ $\Rightarrow$ 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. $\phi(x) = \phi_0(x) \to \phi_{M-1}(x)$. $\sigma(w^T \phi(x))$ LOGIT

$$p(w) = \mathcal{N}(w \mid m_0, S_0)$$

PRIOR

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

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

$s_i = s_i(w)$

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

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

NOT ANALYTICALLY TRACTABLE

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

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}. $S_N^{-1} = S_0^{-1} + \sum_{n=1}^{N} s_n(w_{MAP})(1 - s_n(w_{MAP})) \phi(x_n) \phi^T(x_n)$

$$p(w|\underline{y}) \approx q(w) \qquad \underbrace{q(w) = \mathcal{N}(w|w_{MAP}, S_N)}$$

PREDICTIVE DISTRIBUTION

$$p(C_1|x^*, \underline{y}, \underline{x}) = \int p(C_1|x^*, w, \underline{x}, \underline{y})\, p(w|\underline{y}, \underline{x})\, dw$$

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

M-dim integral!

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

$q(a) = \mathcal{N}(a|\mu_a, \sigma_a^2)$

$$\left[\begin{aligned} \mu_a &= w_{MAP}^T \phi(x^*) \\ \sigma_a^2 &= \phi^T(x^*) S_N \phi(x^*) \end{aligned}\right]$$

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

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

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

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