LOGISTIC REGRESSION

Consider a binary classification problem with labels $y = 0, 1$.

Define

$$\eta(x) =$$

(A)

$$=$$

Then the Bayes classifier may be expressed as

$$f^*(x) =$$

Logistic regression implements the following strategy:

1) Assume $\eta(x) = \frac{1}{1+e^{-(w^T x + b)}}$, $w \in \mathbb{R}^d$, $b \in \mathbb{R}$

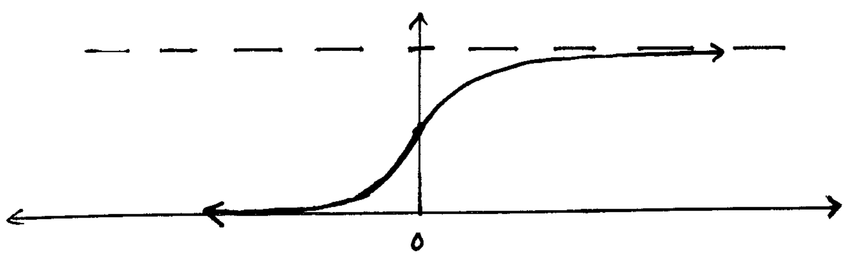
2) Compute the MLE of $\theta = (w, b)$.

3) Plug the estimate

$$\hat{\eta}(x) = \frac{1}{1+e^{-(\hat{w}^T x + \hat{b})}}$$

into the formula for the Bayes classifier

(B) The function $\dfrac{1}{1+e^{-t}}$ is called a ________ function, and also called a ________ function.



Observe that

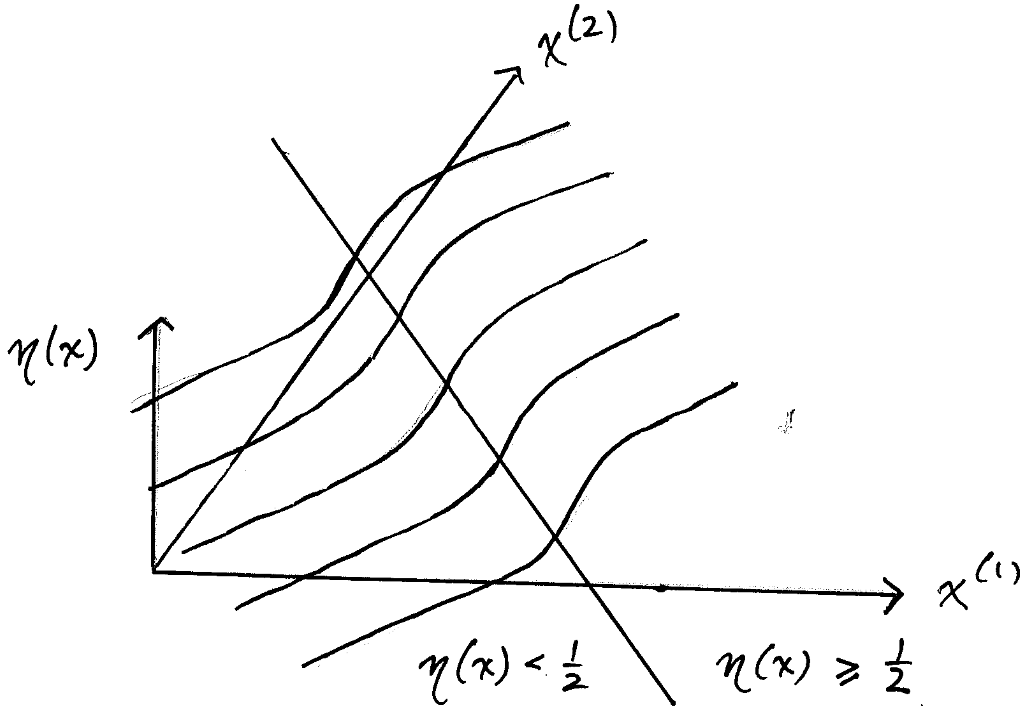
(C) $$\hat{f}(x) = 1 \iff$$

$$\iff$$

Therefore

$$\hat{f}(x) =$$

is ________.



Maximum Likelihood Estimation

Assume the data (x_i, y_i) are independent.

Denote $\underline{x} = (x_1, \ldots, x_n)$, $\underline{y} = (y_1, \ldots, y_n)$. Then

(D) $$\ell(\theta; \underline{x}, \underline{y}) =$$

$$=$$

$$=$$

Note that y is <u>discrete</u> and therefore

$$(y \mid x ; \theta)$$

is a probability mass function

In particular, we recognize $y|x$ as a

____________ random variable with

$$p(y \mid x ; \theta) = \left\{ \right.$$

$$=$$

Maximum Likelihood Estimation

$$\ell(\theta) = \prod_{i=1}^{n} \eta(x_i;\theta)^{y_i}\left(1-\eta(x_i;\theta)\right)^{1-y_i} + C$$

$$\Rightarrow \log \ell(\theta) = \sum_{i=1}^{n} y_i \log \eta(x_i;\theta) + (1-y_i)\log\left(1-\eta(x_i;\theta)\right)$$

Notation

$$\tilde{x} = [1 \;\; x^{(1)} \;\; \cdots \;\; x^{(d)}]^T$$

$$\theta = [b \;\; w^{(1)} \;\; \cdots \;\; w^{(d)}]^T$$

$$g(t) = \frac{1}{1+e^{-t}}$$

so that $\eta(x) = g(\theta^T \tilde{x})$

Note that

E

$$g'(t) =$$

$$=$$

So we have

$$\log \ell(\theta) = \sum_i y_i \log g(\theta^T \tilde{x}_i) + (1-y_i)\log\left(1-g(\theta^T \tilde{x}_i)\right)$$

To maximize the likelihood, we can try

(F) $$\frac{\partial \log \ell(\theta)}{\partial \theta} = \sum_{i=1}^{n}$$

$$=$$

$$=$$

Unfortunately, this is a nonlinear system of equations and has no closed-form solution.

However, the log-likelihood is <u>concave</u> and therefore has a global maximum. Typically the log-likelihood is maximized iteratively using the <u>Newton-Raphson algorithm</u>:

$$\theta^{new} = \theta^{old} - \underbrace{\left(\frac{\partial^2 \log \ell(\theta)}{\partial \theta \, \partial \theta^T} \right)}_{\text{Hessian}}^{-1} \frac{\partial \log \ell(\theta)}{\partial \theta}$$

where derivatives are evaluated at θ^{old}

$\boxed{\text{Key}}$

A. $\eta(x) = \Pr\{Y=1 \mid X=x\}$

$$= 1 - \Pr\{Y=0 \mid X=x\}$$

$$f^*(x) = \begin{cases} 1 & \text{if} \quad \eta(x) \geq \frac{1}{2} \\ 0 & \text{if} \quad \eta(x) < \frac{1}{2} \end{cases}$$

B. logistic, sigmoid

C. $\hat{f}(x) = 1 \iff \hat{\eta}(x) \geq \frac{1}{2}$

$$\iff \exp\{-(\hat{w}^T x + \hat{b})\} \leq 1$$

$$\iff \hat{w}^T x + \hat{b} \geq 0$$

$$\hat{f}(x) = \begin{cases} 1 & \text{if} \quad \hat{w}^T x + b \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

$\Rightarrow \hat{f}$ is linear

D. $\ell(\theta; \underline{x}, \underline{y}) = p(\underline{x}, \underline{y}; \theta)$

$$= \prod_{i=1}^{n} p(x_i, y_i; \theta)$$

$$= \prod_{i=1}^{n} p(y_i \mid x_i; \theta) \underbrace{p(x_i; \theta)}_{\text{independent of } \theta}$$

E. $g'(t) = \dfrac{e^{-t}}{(1+e^{-t})^2} = g(t) \cdot (1 - g(t))$

F. $$\frac{\partial \log l(\theta)}{\partial \theta} = \sum_{i=1}^{n} y_i \tilde{x}_i \left(1 - g(\theta^T \tilde{x}_i)\right) - (1-y_i)\, \tilde{x}_i\, g(\theta^T \tilde{x}_i)$$

$$= \sum_{i=1}^{n} \tilde{x}_i \left(y_i - g(\theta^T \tilde{x}_i)\right)$$

$$= 0$$