LOGISTIC REGRESSION

Consider a <u>binary</u> classification problem with labels $y = 0, 1$.

Define

$$\eta(x) =$$

(A)

$$=$$

Then the Bayes classifier may be expressed as

$$f^*(x) =$$

<u>Logistic regression</u> implements the following strategy:

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

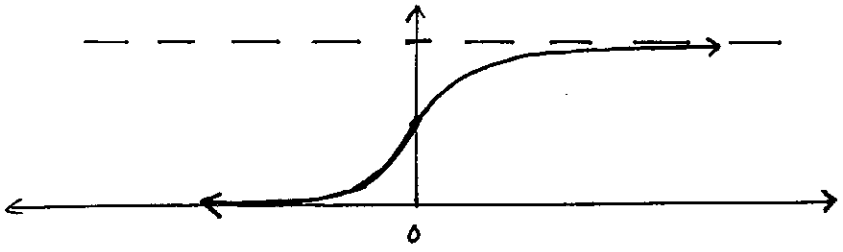
2) Compute the MLE of $\theta = (w, b)$.

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 $\frac{1}{1+e^{-t}}$ is called a _____ function, and also called a _____ function.



Observe that

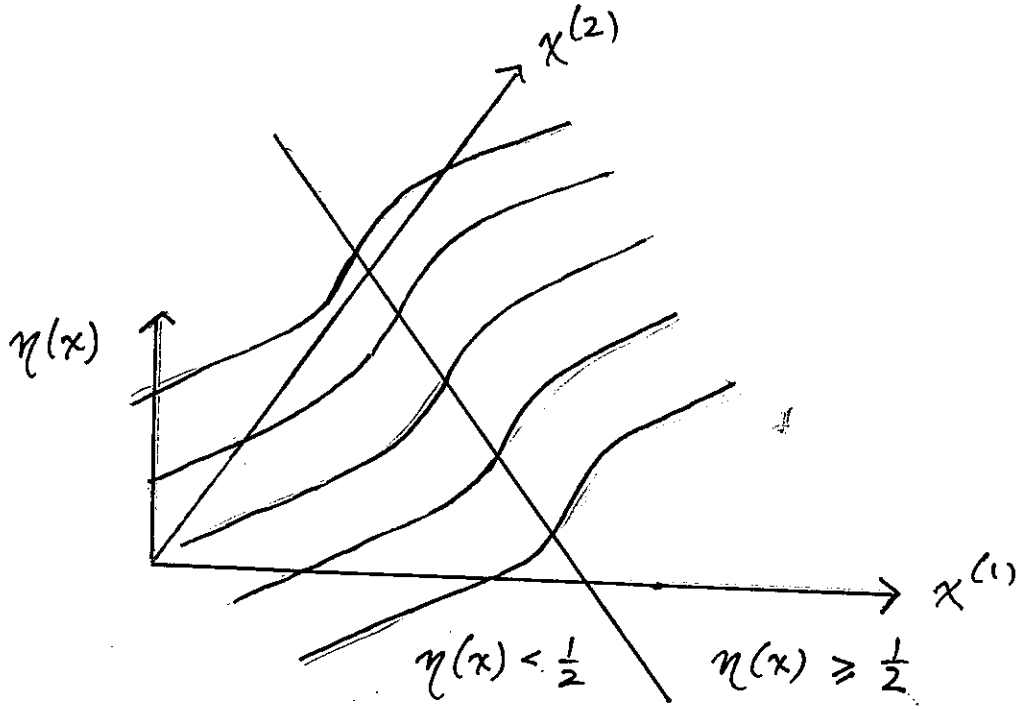
(C) $\hat{f}(x) = 1 \iff$

$\iff$

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

$=$

$=$

Let just write $\ell(\theta)$ for the likelihood.

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

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

is a probability mass function

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

_____ random variable with

(E)

$$p(y \mid x ; \theta) = \begin{cases} \\ \\ \end{cases}$$

$$=$$

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

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

Notation $\quad \tilde{x} = \begin{bmatrix} 1 & x^{(1)} & \cdots & x^{(d)} \end{bmatrix}^T$

$$\theta = \begin{bmatrix} b & w^{(1)} & \cdots & w^{(d)} \end{bmatrix}^T$$

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

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

Note that

(F) $g'(t) =$

$=$

So we have

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

To maximize the likelihood, we can try

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

$$=$$

$$=$$

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>:

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

↑ Hessian

where derivatives are evaluated at θ^{old}

LR vs. LDA

Advantages of LR

- models only the distribution of Y/X, not the joint distribution of (X, Y)

  $\Rightarrow$ discriminative

  $\Rightarrow$ model is valid for a larger class of distributions

- fewer parameters to estimate

Disadvantages

- LDA better if Gaussian assumption is valid

Also note

- Like other "plug-in" approaches to classification, LR yields not just a class label, but also a class probability for every pattern

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

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

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

independent of θ

E. binomial

$$p(y \mid x; \theta) = \begin{cases} \eta(x;\theta) & \text{if } y = 1 \\ 1 - \eta(x;\theta) & \text{if } y = 0 \end{cases}$$

F. $$g'(t) = \frac{e^{-t}}{(1+e^{-t})^2} = g(t)(1-g(t))$$

G. $$\frac{\partial \log \ell(\theta)}{\partial \theta} = \sum_{i=1}^{n} y_i \tilde{x}_i (1 - g(\theta^T \tilde{x}_i)) - (1-y_i)\tilde{x}_i g(\theta^T \tilde{x}_i)$$

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

$$= 0$$