

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

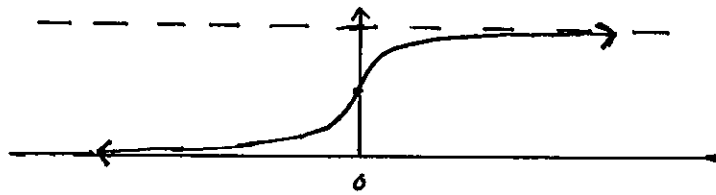
2) Compute the MLE of $\theta = (w, b)$.

3) Plug the estimate

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

into the formula for the Bayes classifier

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



(B)

Denote the logistic regression classifier

$$\hat{f}(x) = \mathbb{1}_{\{\hat{\eta}(x) \geq \frac{1}{2}\}}$$

Observe that

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

$$\iff$$

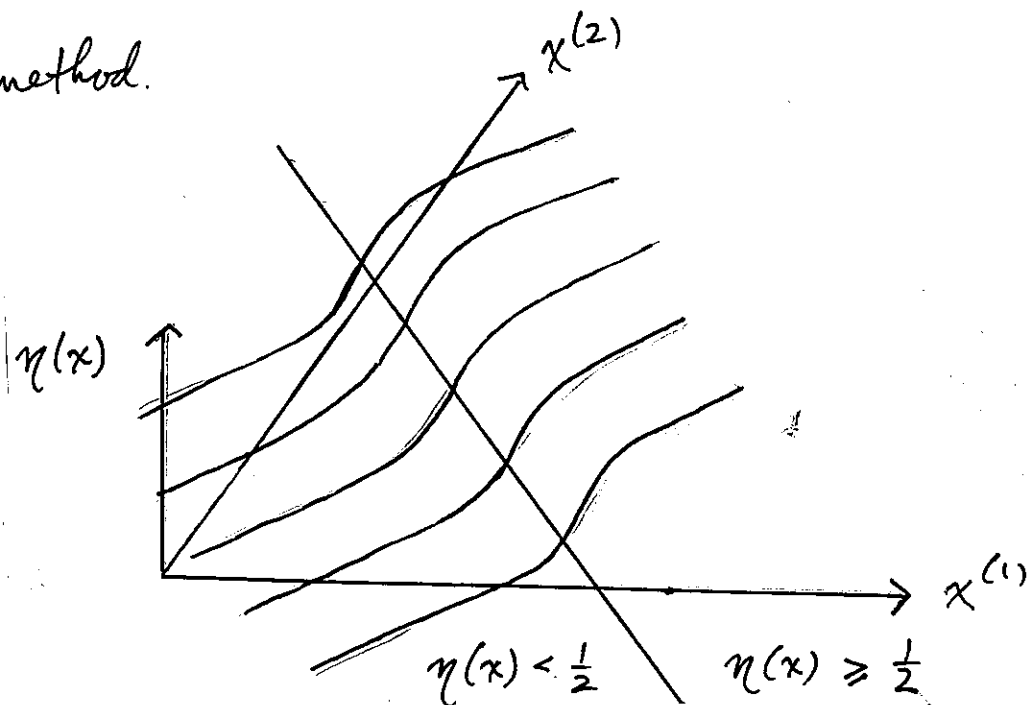
$$\iff$$

(C)

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

a _____ method.

, and LR is



Maximum Likelihood Estimation

The LR model does not describe the feature vector x , so we will view each x_i as fixed and estimate $\theta = (w, b)$ using conditional maximum likelihood estimation.

Denote $\underline{x} = (x_1, \dots, x_n)$, $\underline{y} = (y_1, \dots, y_n)$. Then the conditional log-likelihood is

$$l(\theta; \underline{y} | \underline{x}) =$$

=

where in the second step we assumed that

y_1, \dots, y_n are conditionally independent given \underline{x} .

Let's write $l(\theta)$ for brevity.

Note that y is discrete and therefore

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

is a probability mass function

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

random variable with

⑤

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

$$=$$

$$l(\theta) = \prod_{i=1}^n \eta(x_i; \theta)^{y_i} (1 - \eta(x_i; \theta))^{1-y_i}$$

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

Notation $\tilde{x} = [1 \ x^{(1)} \ \dots \ x^{(d)}]^T$

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

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

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

Note that

(F) $g'(t) =$
 $=$

So we have

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

$$= -\sum_i (1-y_i) (\theta^T \tilde{x}_i + \log(1 + e^{-\theta^T \tilde{x}_i}))$$

To maximize the likelihood, we can try

$$\textcircled{G} \quad \frac{\partial \log l(\theta)}{\partial \theta} = \sum_{i=1}^n$$

=

=

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

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

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

↑
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)
 - ⇒ discriminative
 - ⇒ 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

$$\begin{aligned} \text{A. } \eta(x) &= \Pr\{Y=1 \mid X=x\} \\ &= 1 - \Pr\{Y=0 \mid X=x\} \end{aligned}$$

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

B. logistic, sigmoid

$$\begin{aligned} \text{C. } \hat{f}(x) = 1 &\iff \hat{\eta}(x) \geq \frac{1}{2} \\ &\iff \exp\{-(\hat{w}^T x + \hat{b})\} \leq 1 \\ &\iff \hat{w}^T x + \hat{b} \geq 0 \end{aligned}$$

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

$\Rightarrow \hat{f}$ is linear

$$\begin{aligned} \text{D. } l(\theta; \underline{y} \mid \underline{x}) &= \hat{p}(\underline{y} \mid \underline{x}; \theta) \\ &= \prod_{i=1}^n \hat{p}(y_i \mid x_i; \theta) \end{aligned}$$

E. binomial

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

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

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

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

$$= 0$$