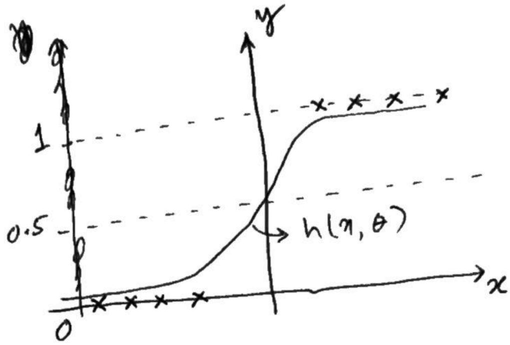
LOGISTIC REGRESSION – I

Classification : Spam vs. Normal Email
Malignant vs. Healthy Tissue
Fake vs. Real News.
Human vs. Animal image.

One of the most popular ML algos.

$y_i \in \{0, 1\}$ → True/Yes, → False/NO



No classifier is perfect.
So, we are looking for a probability estimate for $p(y_i / x_i; \theta)$

Clearly y vs. x is NOT a straight line.
So, makes no sense to ~~fit a str~~ use Linear regression

$\hat{y}_i = h(x_i; \theta)$ → hypothesis

& $0 \le h \le 1 \quad \forall\, x_i, \theta$

Decision boundary (Linear) or HYPERPLANE:

$\theta_0 + \theta^T x_i > 0 \Rightarrow h(x_i; \theta) > 0.5$ (θ → vector)

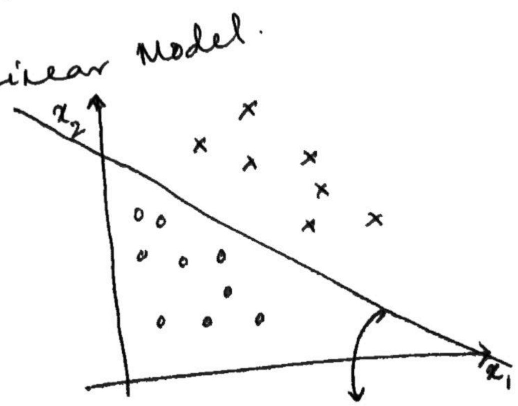
$\theta_0 + \theta^T x_i \le 0 \Rightarrow h(x_i; \theta) \le 0.5$

$\Rightarrow h(x_i; \theta) = h(\theta^T x_i + \theta_0)$

Sigmoid/Logistic function or Logit : $h(x_i; \theta) = \dfrac{1}{1 + e^{-(\theta^T x_i + \theta_0)}}$

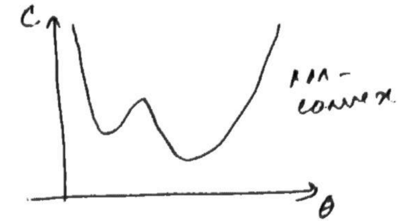
Many other options available but this one is easier to handle & interpret.

$$\log \underbrace{\frac{h(x_i; \theta)}{1 - h(x_i; \theta)}}_{\text{odds}} = \theta_0 + \theta^T x_i \ : \text{Linear Model.}$$



Linear Decision boundary

LSE : $C_{LSE}(\theta) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$



But this not convex.

for a function $f(\theta)$, to be convex,

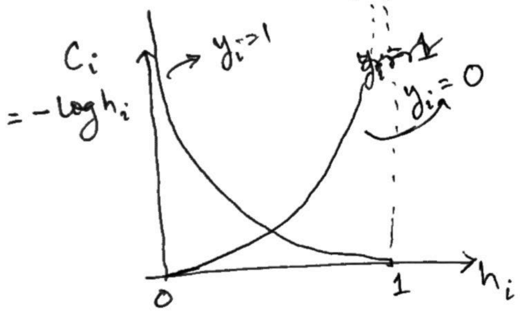
we must have $\frac{\partial^2 f}{\partial \theta^2} > 0 \quad \forall \theta$

for binary classification,

$$\theta_{MLE} = \operatorname{argmin} \frac{1}{N}\sum_{i=1}^{N}\left[-y_i \log h_i - (1-y_i)\log(1-h_i)\right]$$

$$\therefore \quad C(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left[-y_i \log h_i - (1-y_i)\log(1-h_i)\right] \quad ; \; y_i \in \{0,1\}$$

↳ Binary Cross Entropy.



If $y_i = 1$, $C_i(\theta) = -\log h_i$

If $y_i = 0$, $C_i(\theta) = -\log(1-h_i)$

Gradient Descent :

$$\theta_j \longrightarrow \theta_j - \eta \frac{\partial C}{\partial \theta_j}$$

η ↳ Learning Rate

LOGISTIC REGRESSION - II

MULTI-CLASS CLASSIFICATION

$$S = \{\text{Spam, Important, Not Important}\}$$

$$y = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ y^{(3)} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \quad \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \quad \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

ONE-HOT ENCODING

$$\hat{y} = \begin{bmatrix} \hat{y}^{(1)} \\ \hat{y}^{(2)} \\ \hat{y}^{(3)} \end{bmatrix}$$

$$\hat{y}^{(1)} = p[y^{(1)} = 1]$$
$$\hat{y}^{(2)} = p[y^{(2)} = 1]$$
$$\hat{y}^{(3)} = p[y^{(3)} = 1]$$

$$0 \leq \hat{y}^{(k)} \leq 1 \qquad \sum_k \hat{y}^{(k)} = 1$$

$$z^{(k)} = e^{+x^T\theta^{(k)} + \theta_0^{(k)}}$$

SOFTMAX function:

$$\hat{y}^{(1)} = \frac{z^{(1)}}{z^{(1)} + z^{(2)} + z^{(3)}}$$

$$\hat{y}^{(2)} = \frac{z^{(2)}}{z^{(1)} + z^{(2)} + z^{(3)}} \qquad \hat{y}^{(3)} = \frac{z^{(3)}}{z^{(1)} + z^{(2)} + z^{(3)}}$$

COST/ LOSS FUNCTION

for 2-classes : $C = \frac{1}{N}\sum_{i=1}^{N}\left[-y_i \log \hat{y}_i - (1-y_i)\log(1-\hat{y}_i)\right]$

$$= \frac{1}{N}\sum_{i=1}^{N}\left[-y_i^{(1)} \log \hat{y}_i^{(1)} - y_i^{(2)} \log \hat{y}_i^{(2)}\right]$$

for K-classes : $C = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} - y_i^{(k)} \log \hat{y}_i^{(k)}$

CATEGORICAL CROSS ENTROPY

LOGISTIC REGRESSION - II (contd.)

Classification: $y_i \in \{0, 1\}$

$$\hat{y}_i = h(x_i, \theta) = \frac{1}{1 + e^{-(\theta^T x_i + \theta_0)}} \rightarrow \text{SIGMOID, LOGISTIC, LOGIT fn.}$$

$$C(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left[-y_i \log h_i - (1 - y_i)\log(1 - h_i)\right] \rightarrow \text{BINARY CROSS ENTROPY.}$$

Gradient Descent: $\theta_j \longrightarrow \theta_j - \underbrace{\eta}_{\text{Learning Rate}} \frac{\partial C}{\partial \theta_j}$

$$\frac{\partial C}{\partial \theta_j} = \frac{1}{N}\sum_{i=1}^{N}\left[-\frac{y_i}{h_i}\frac{\partial h_i}{\partial \theta_j} + \frac{1 - y_i}{1 - h_i}\frac{\partial h_i}{\partial \theta_j}\right]$$

$$= \frac{1}{N}\sum_{i=1}^{N}\frac{(h_i - y_i)}{h_i(1 - h_i)}\frac{\partial h_i}{\partial \theta_j}$$

$$= \frac{1}{N}\sum_{i=1}^{N}(h_i - y_i)\, x_{ij}$$

Form Exactly same as Linear Regression just with a different expression for h_i

$$\frac{\partial h_i}{\partial \theta_j} = \frac{-1}{(1 + e^{-\theta^T x_i - \theta_0})^2} \cdot \frac{\partial}{\partial \theta_j} e^{-\theta_T x_i}$$

$$= \frac{e^{-\theta_T x_i - \theta_0}}{(1 + e^{-\theta^T x_i - \theta_0})^2} \cdot \frac{\partial}{\partial \theta_j}\theta^T x_i$$

$$= \frac{1}{1 + e^{-\theta^T x_i - \theta_0}} \cdot \frac{e^{-\theta^T x_i - \theta_0}}{1 + e^{-\theta^T x_i - \theta_0}} \cdot x_{ij}$$

$$= h_i(1 - h_i)\, x_{ij}$$

The features x_{ij} used in Logit can actually be nonlinear fn. (eg. polynomial) of other features thereby allowing the decision boundary to be effectively nonlinear.

~~REGULARISATION~~

$$C(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left[-y_i \log h_i - (1 - y_i)\log(1 - h_i)\right] + \underbrace{\frac{\lambda}{N}\|\theta\|_2^2}$$

~~as~~ $N \to \infty$, this term ~~drops~~ out.

$$\therefore \frac{\partial C}{\partial \theta_j} = \frac{1}{N}\sum_{i=1}^{N}(h_i - y_i)\, x_{ij} + \frac{2\lambda}{N}\theta_j$$

MULTI-CLASS LOGISTIC REGRESSION

Email: Spam, Important, Not-Important

Face Identification

ONE vs. ALL [or One vs. Rest]

Let $y_i \in \{S, I, NI\} \equiv K$ - categories

Divide into three problems

$$h^{(1)}(x_i, \theta) = \underline{\underline{S}}^{1} \text{ vs. } \underline{\{I, NI\}}^{0}$$

K - binary classifications

$$h^{(2)}(x_i, \theta) = I \text{ vs. } \{S, NI\}$$

$$h^{(3)}(x_i, \theta) = NI \text{ vs. } \{S, I\}.$$

$$h(x_i, \theta) = \max\{h^{(1)}, h^{(2)}, h^{(3)}\}$$

ONE vs. ONE

S vs. I $\qquad {}^{K}C_2$ binary classifications.

I vs. NI

S vs. NI

Pick category which has most classifications.

These methods do not give a probability estimates

since $h^{(1)} + h^{(2)} + h^{(3)} \neq 1$

One vs. One is also computationally expensive.

Better to have a method which can learn multi-class parameters at one go.