Mathematical Derivation of the Optimization Function for Logistic Regression.

Some Assumptions and notations:-

$$\theta^T x = \sum_{i=1}^{n} \theta_i x_i = \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_n x_n$$

(weighted sum) —— (I)

$$\sigma(z) = \frac{1}{1+e^{-z}}$$

(sigmoid function) (II)

partial derivative of $\sigma(z)$:

$$\frac{\partial}{\partial z}\sigma(z) = \frac{\partial}{\partial z}\left\{(1+e^{-z})^{-1}\right\}$$

$$= (-1)\frac{1}{(1+e^{-z})^2}(-1)e^{-z}$$

$$= \frac{e^{-z}}{(1+e^{-z})^2}$$

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

$$= \sigma(z)[1-\sigma(z)]$$

$$\therefore \frac{\partial}{\partial z}\sigma(z) = \sigma(z)[1-\sigma(z)]$$

- - - - (III)

(5)

(*) Logistic Regression An Overview:-

Logistic Regression falls in the family of discriminative class of algorithms. In a sentence, Logistic Regression is a classification algorithm that works by trying to learn $P(Y|X)$ [probability of Y given X].

Mathematically, for a single training point (x, y) we have:-

$$\left.\begin{aligned} P(Y=1 \mid X=x) &= \sigma(z) \text{ where} \\ \sigma(z) &= \frac{1}{1+e^{-z}} \quad (\text{From } (11)) \\ \text{and } z &= \theta_0 + \sum_{i=1}^{m} \theta_i x_i \end{aligned}\right\} \quad (IV)$$

Now (IV) can be written as:-

$$P(Y=1 \mid X=x) = \sigma(\theta^T x), \text{ where } x_0 = 1 \text{ always.}$$

$$P(Y=0 \mid X=x) = 1 - \sigma(\theta^T x), \text{ by total law of probability.}$$

~~Log Likelihood:~~

(*) In order to choose parameters of LR, we use MLE. For this we use

1. Log Likelihood
2. Find values of θ by Max Likelihood

Bernoulli Distribution:-

A r.v X is said to follow Bernoulli distribution ~~if~~ with parameter p if p.m.f :-

$$P(X=x) = \begin{cases} p^x(1-p)^{1-x}, & x=0,1 \\ 0, & \text{otherwise} \end{cases}$$

Now Logistic Regression predicts labels which are binary and the output follows total law of probability.

Hence $Y \sim B(p)$

$$Y = \sigma(\theta^T x)^y \left[(1-\sigma(\theta^T x))\right]^{(1-y)}$$

~~Summing over all n~~

Log Likelihood of Y

$$L(\theta) = \prod_{i=1}^{n} \sigma(\theta^T x^{(i)})^{y^{(i)}} \left[(1-\sigma(\theta^T x^{(i)})\right]^{(1-y^{(i)})}$$

$$LL(\theta) = \sum_{i=1}^{n} y^{(i)} \log \sigma(\theta^T x^{(i)}) + (1-y^{(i)})\left[(1-\sigma(\theta^T x^{(i)})\right]$$

Now for MLE the only step remaining is to ~~find~~ choose params that maximize θ.

Page

Hence, $\frac{\partial}{\partial \theta_j} LL(\theta) = \frac{\partial}{\partial \theta_j} y \log \sigma(\theta^T x)$

$+ \frac{\partial}{\partial \theta_j} (1-y) \log [1 - \sigma(\theta^T x)]$

$= \left[\frac{y}{\sigma(\theta^T x)} - \frac{1-y}{1-\sigma(\theta^T x)}\right] \frac{\partial}{\partial \theta_j} \sigma(\theta^T x)$

Using equation (111)

$= \left[\frac{y}{\sigma(\theta^T x)} - \frac{(1-y)}{1-\sigma(\theta^T x)}\right] \sigma(\theta^T x)[1 - \sigma(\theta^T x)] x_j$

$= \cancel{y[1 - \sigma(\theta^T x)]}$

$= [y - \sigma(\theta^T x)] x_j$

$$\boxed{J(\theta) = \sum_{i=1}^{n} [y_i - \sigma(\theta^T x^{(i)})] x_j^{(i)}}$$