Page No.
Date :

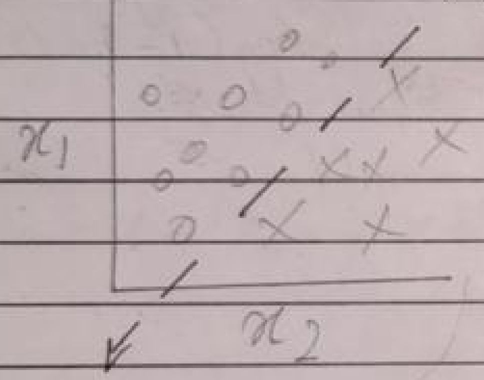
Logistic Regression ⟹ classification Algorithm

- classifies data - binary classification algo
- classify data in 0 or 1
- Supervised learning algorithm need both x & y for training data
- ex - classify email as spam or not, classify image - cat or dog?

Binary class Classification

$x \in R^n$ $\quad y \in \{0, 1, 2, 3 \ldots k\}$ $\quad y \in \{0, 1\}$

$\theta_0 + \theta_1 x_1 + \theta_2 x_2 = 0$



$x \in R^2$

x, y

$\downarrow$

model

$\downarrow$

$x^T \rightarrow g(h_\theta(x)) \longrightarrow y \in (0,1)$

dount discriminate b/w distance of pts

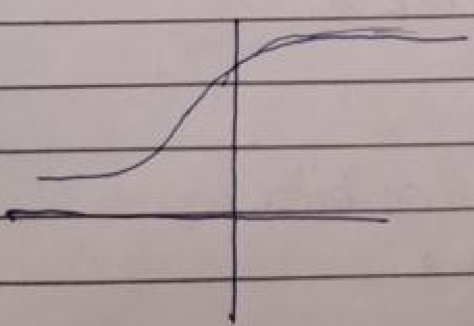
$g(z) = 1$ for $z > 0$

$g(z) = 0$ for $z \leq 0$

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

$z \Rightarrow \infty \quad g(z) = 1$

$z \rightarrow -\infty \quad g(z) = 0$



z = distance of pt from line.

sigmoid(z)

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1+e^{-\theta^T x}}$$

For aways pts - we want correct & confident classification

prediction - 1 confidence ≥ 0.5

- 0 confidence < 0.5

$$h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$$

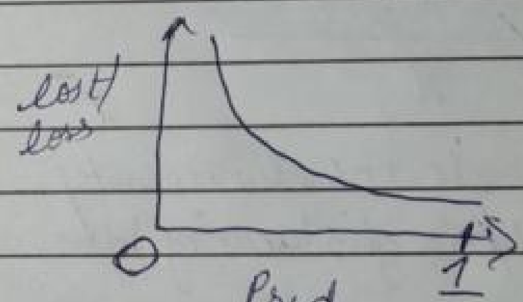
Log Loss / Binary Cross Entropy $\Rightarrow$

$$\text{Avg Loss} = -\frac{1}{m}\sum_{i=1}^{m} y^i \log h_\theta(x^i) + (1-y^i)\log(1-h_\theta(x^i))$$

→ actually minimizes the difference b/w two probability distribution.

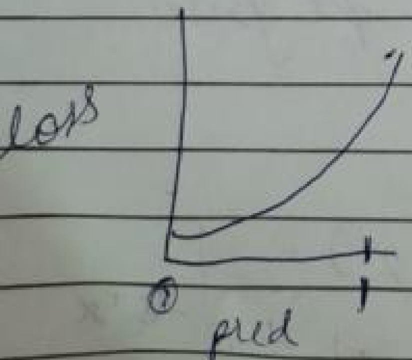
$y^i = 1\ 1\ 1\ 1\ 1$

$$\text{loss} = -\sum y^i \log h_\theta(x^i)$$

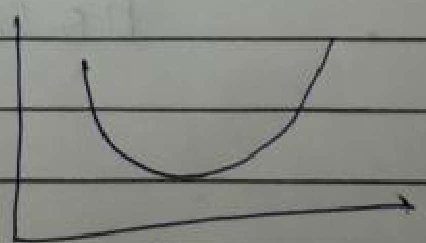


$\text{lossy}^i = 0\ 0\ 0\ 0\ 0$

$$\text{loss} = -\sum (1-y^i)\log(1-h_\theta x^i)$$

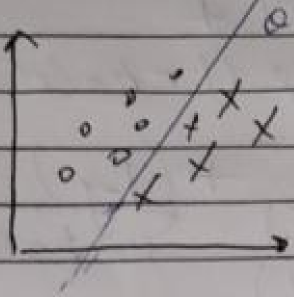


when you combine these two



convex func.

Proof :- Log loss fnx by using Maximum Likelihood Estimates



$\theta^T x \ge 0 \Rightarrow +ve \Rightarrow 1$

confidence $= g(\theta^T x) = \frac{1}{1+e^{-\theta^T x}} \ge 0.5$

$\theta^T x < 0 \Rightarrow -ve \Rightarrow \to 0$

confidence $g(\theta^T x) < 0.5$

$h_\theta x = \frac{1}{1+e^{-\theta^T x}}$

$P(y=1 \mid x;\theta) = h_\theta(x^i)$
$P(y=0 \mid x;\theta) = 1 - h_\theta(x^i)$] Bernoulli distribution

Probablity mass function

$P(y \mid x;\theta) = [h_\theta(x)]^y [1-h_\theta(x)]^{1-y}$

$P(y=1 \mid x;\theta) = h_\theta(x)$ — (1)

$P(y=0 \mid x;\theta) = 1 - h_\theta(x)$ — (2)

$L(\theta) = P(y \mid x;\theta)$

$= \prod_{i=1}^{m} P(y^i \mid x^i;\theta)$ all pts are generated independently

$L(\theta) = \prod_{i=1}^{m} h_\theta(x^i)^{y^i}(1-h_\theta(x^i))^{1-y^i}$

$\log L(\theta) = \log h_\theta(x^i)^{y^i} + \log(1-h_\theta(x^i))^{1-y^i}$

maximise $= y^i \log h_\theta(x^i) + \log(1-h_\theta x^i)(1-y^i)$

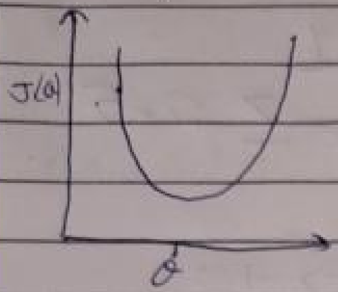
minimize -ve

find a line that maximizes confidence for all data

Page No.

Date :

Gradient Descent Update Rule for Logistic Regression



$$J(\theta) = -\sum_{i=1}^{m} y^i \log h_\theta(x^i) + (1-y^i)\log(1-h_\theta(x^i))$$

$$\frac{\partial J\theta}{\partial\theta_j} = -\sum_{i=1}^{m}\left(\frac{y^i}{h_\theta(x^i)} + \frac{(1-y^i)}{1-h_\theta(x^i)}\right) \times g'(\theta^T x^i)$$

$$\theta_j = \theta_j - \eta\frac{\partial J}{\partial\theta}$$

$$= -\sum_{i=1}^{m}\left(\frac{y^i - y^i h_\theta x^i + (-h_\theta x^i) + y^i h_\theta x^i}{g(\theta^T x^i)(1-g(\theta^T x^i))}\right.$$

$$h_\theta x = g(\theta^T x)$$

Prop of sigmoid

$$g'(z) = g(z)(1-g(z)) \qquad \times\; g(\theta^T x^i)(1-g(\theta^T x^i))x_j^i$$

$$\frac{\partial J\theta}{\partial\theta_j} = -\sum_{i=1}^{m}(y^i - h_\theta(x^i))x_j^i$$

Final update:-

$$\theta_j = \theta_j + \eta\sum_{i=1}^{m}[y^i - \underbrace{h_\theta(x^i)}_{\left(\frac{1}{1+e^{-\theta^T x}}\right)})x_j^i$$

similar update like linear regression

sigmoid function