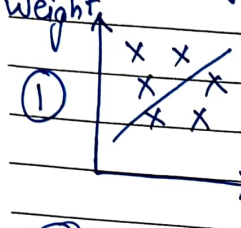
PARSHWANATH CHARITABLE TRUST'S
A.P. SHAH INSTITUTE OF TECHNOLOGY
Department of Computer Science and Engineering
Data Science

Semester : VI Subject :CSC601 Data Analytics and Visualization Academic Year: 2023- 2024

Logistic Regression

Linear Regression Revision

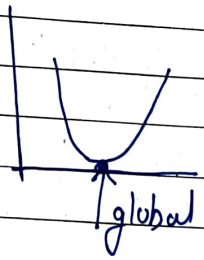


$h_\theta(x) = \theta_0 + \theta_1 x_1$

① Cost function

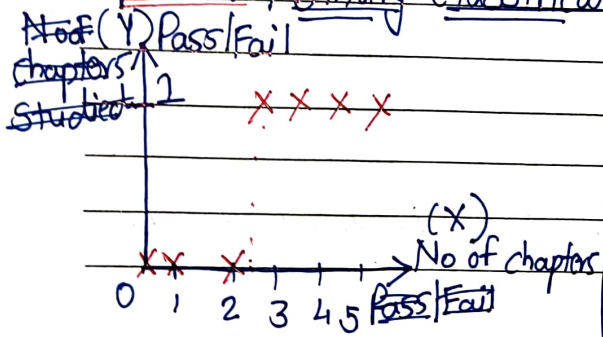
② $J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x)^i - y^i)^2$

③ Gradient decent → mimise cast fun^n to reach to global minima



Linear Regression is used to solve regression problem.

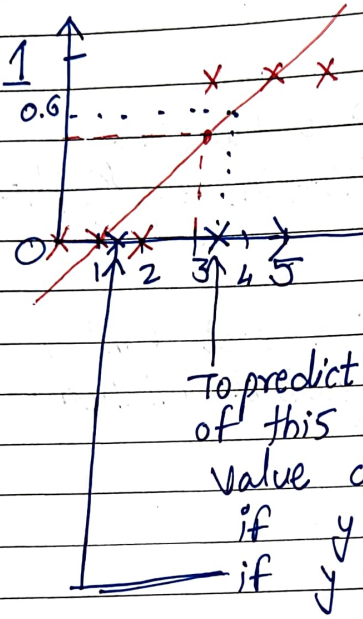
Logistic Regression is used to solve classification problem / Binary classification problem



No of chapters Studied	Pass / Fail
4	Pass
2	Fail
0	Fail
1	Fail
2.5	Pass

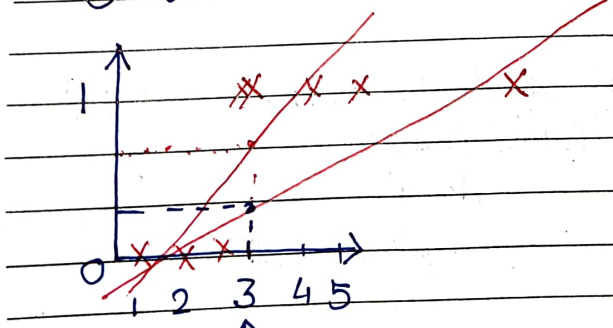
O/P Feature

Can we use linear regression of solving this problem?



To predict output of this value as per regression line value of y axis is 0.6
if y axis value > 0.5 it is 1
if y axis value < 0.5 it is 0

The problem we have here are
① outliers



For 3 chapters the result is shown as fail

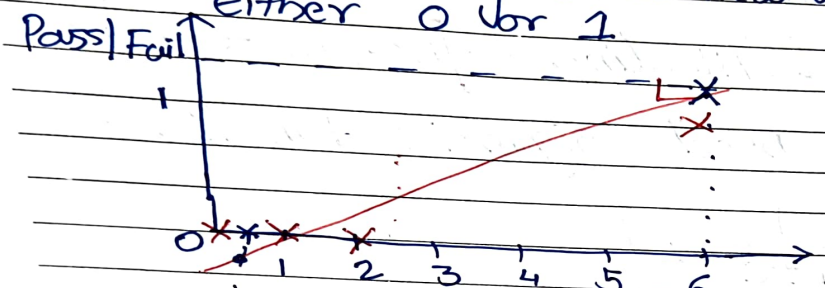
PARSHWANATH CHARITABLE TRUST'S
A.P. SHAH INSTITUTE OF TECHNOLOGY
Department of Computer Science and Engineering
Data Science

CSE DATA SCIENCE

Semester : VI Subject :CSC601 Data Analytics and Visualization Academic Year: 2023- 2024

~~Classif~~

② This binary classification so the result will be either 0 or 1



② The problem > 1 or < 0 (called squashing)

To address these two problems linear regression can not be used.

① outliers
② output values >1 or <0

The classification problems can not be solved using linear regression.

Academic Year: 2023- 2024

Semester : VI Subject :CSC601 Data Analytics and Visualization

For logistic regression, ~~to~~ ~~ad~~ we need to address squashing and outliers

The eqn for best fit line — $h\theta(x) = \theta_0 + \theta_1 x$

Also we will need sigmoid function

$$\text{Sigmoid} = \frac{1}{1+e^{-x}}$$

We need to combine these two ~~lines~~ equations to generate equations of logistic regression that can handle squashing.

$$h\theta(x) = g(\theta_0 + \theta_1 x)$$

Sigmoid activation function

Sigmoid fun^n → $g = \frac{1}{1+e^{-z}}$ here $z = (\theta_0 + \theta_1 x)$

$$h\theta(x) = \frac{1}{1+e^{-(\theta_0+\theta_1 x_1)}}$$

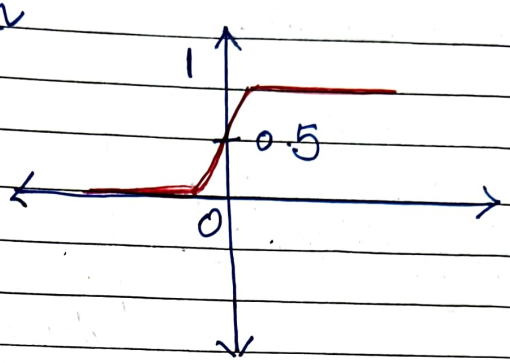
Hypothesis of logistic regression

PARSHWANATH CHARITABLE TRUST'S
A.P. SHAH INSTITUTE OF TECHNOLOGY
Department of Computer Science and Engineering
Data Science

Semester : VI Subject :CSC601 Data Analytics and Visualization Academic Year: 2023- 2024

Sigmoid function

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



If $z \leq 0$
$g(z) \leq 0.5$

If $z \geq 0$
$g(z) \geq 0.5$

The Problem Statement of logistic regression

Training Dataset $\{(x_1, y_1), (x_2, y_2) \ldots (x_m, y_m)\}$

$x \in \begin{Bmatrix} x_0 \\ x_1 \\ \vdots \\ x_m \end{Bmatrix}$ $y \in \begin{Bmatrix} y_0 \\ y_1 \\ \vdots \\ y_m \end{Bmatrix}$

$y \in \{0, 1\}$

$$h_\theta(x) = \frac{1}{1 + e^{-(\theta_0 + \theta_1 x)}}$$

$$\boxed{h_\theta(x) = \frac{1}{1+e^{-(\theta^T x)}}} \quad \ldots \quad z = \theta_0 + \theta_1 x = \theta^T x$$

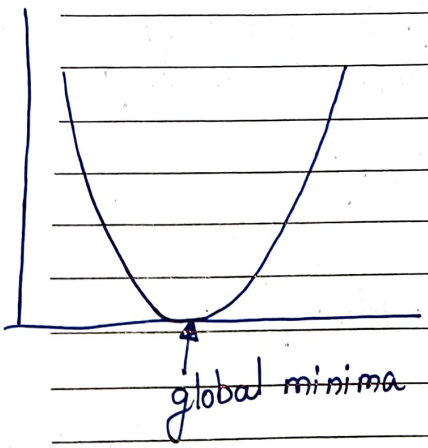
Cost Function

Linear Regression

$$J(\theta_0,\theta_1)=\frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x)^i-y^i)^2$$

Here, $h_\theta(x)=\theta_0+\theta_1 x$

The $h_\theta(x)$ is a convex function



No local minima

Logistic Regression

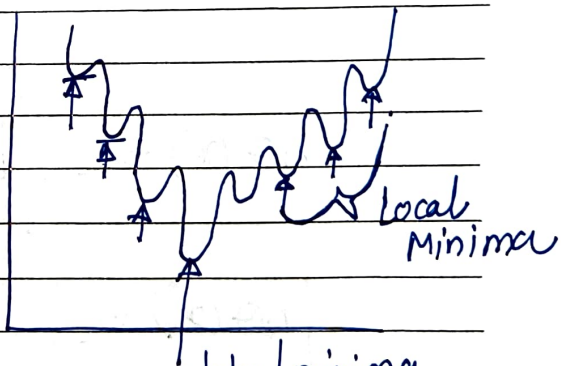
$$J(\theta_0,\theta_1)=\frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x)^i-y^i)^2$$

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

where $\theta^T x=\theta_0+\theta_1 x$

The $h_\theta(x)$ is a non convex function

Due to sigmoid function we are generating non convex function



If it point reaches to local minima it is stucked there only it cannot reach to global minima after that

PARSHWANATH CHARITABLE TRUST'S
A.P. SHAH INSTITUTE OF TECHNOLOGY
Department of Computer Science and Engineering
Data Science

Semester : VI Subject :CSC601 Data Analytics and Visualization Academic Year: 2023- 2024

So, the cost function of logistic regression

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \underbrace{(h\theta(x)^i - y^i)^2}_{cost(h\theta(x^i), y)}$$

$$h\theta(x^i) = \frac{1}{1+e^{-z}}$$

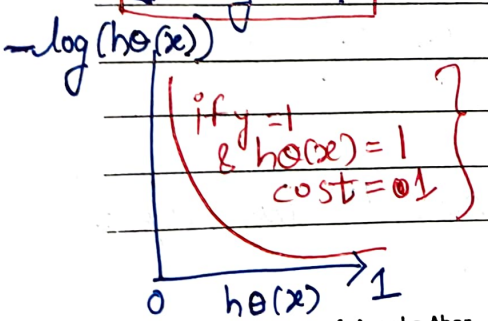
$$z = \theta_0 + \theta_1 x$$

This cost function generates non-convex function so to do generate the convex function we need to rewrite the cost function as

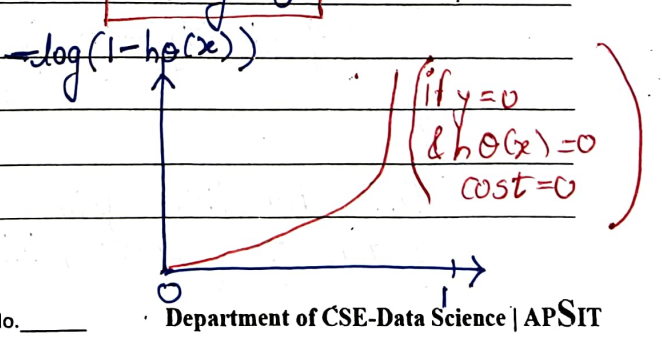
$$cost(h\theta(x)^i - y) = \begin{cases} -\log(h\theta(x)) & \text{if } y=1 \\ -\log(1-h\theta(x)) & \text{if } y=0 \end{cases}$$

This function ~~ge~~ generates a curve like gradient decent.

If y = 1



If y = 0



$$cost(h_\theta(x), y) = -y \log(h_\theta(x)) - (1-y) \log(1-h_\theta(x))$$

if $y = 1$

$$= -\log h_\theta(x))$$

if $y = 0$

$$= -\log(1-h_\theta(x))$$

The cost function of logistic regression is

$$J(\theta_0, \theta_1) = \frac{-1}{2m} \sum_{i=1}^{m} y^{(i)} \log h_\theta(x)^{i} + (1-y^{(i)}) \log(1-h_\theta(x)^{i})$$

And convergence repeat function will be

Convergence Repeat

{

$$\theta_j \approx \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1)$$

}

This function will continue till we reach to global minima.

PARSHWANATH CHARITABLE TRUST'S
A.P. SHAH INSTITUTE OF TECHNOLOGY
Department of Computer Science and Engineering
Data Science

CSE DATA SCIENCE

Semester : VI Subject :CSC601 Data Analytics and Visualization Academic Year: 2023- 2024

Example 1 (Logistic Regression)

The dataset of pass or fail in an exam of 5 students is given in the table.

Hours Study	Pass(1)/Fail(0)
29	0
15	0
33	1
28	1
39	1

Use logistic regression as classifier to answer the following questions.

1. Calculate the probability of pass for the student who studied 33 hours

2. At least how many hours student should study that makes he will pass the course with the probability of more than 95%

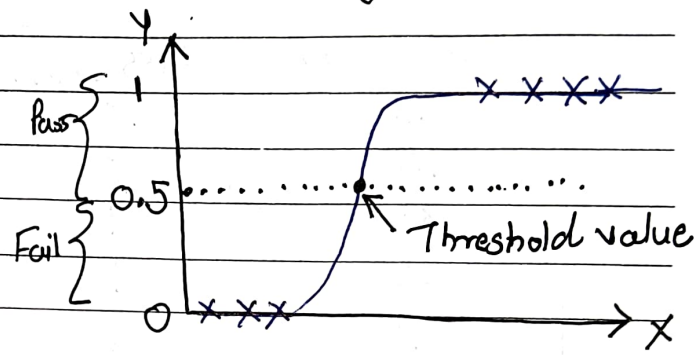
Assume the model suggested by the optimizer for odds of passing the course is,

$$\log(\text{odds}) = -64 + 2 * \text{hours}$$

Solution

We use sigmoid function in logistic regression

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



i) Calculate the probability of pass for the student who studied 33 hours.

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

Given $\log(odds) = z = -64 + 2 * hours$

Let's find value of z,

$z = -64 + 2 * (33)$

$z = -64 + 66$

$\boxed{z = 2}$

Now, let's put value of z in equation given below,

$$h\theta(x) = p = \frac{1}{1+e^{-z}}$$

$$\boxed{p = 0.88}$$

Thus, if student studies 33 hours, then there is 88% chance that the student will pass the exam.

ii) At least how many hours student should study that makes he will pass the course with the probability of more than 95%.

$$h\theta(x) = p = \frac{1}{1+e^{-z}} = 0.95$$

$$0.95\,(1+e^{-z}) = 1$$
$$0.95 \cdot e^{-z} = 1 - 0.95$$
$$e^{-z} = \frac{0.05}{0.95}$$
$$\boxed{e^{-z} = 0.0526}$$

We need to know the value of z

we have $e^{-z} = 0.0526$

Take natural log both sides,

$\ln(e^{-z}) = \ln(0.0526)$ ——— (1)

As per natural logarithmic,

$\ln(e^{x}) = x$

So $\ln(e^{-z}) = -z$

Replace $\ln(e^{-z})$ with $-z$ eq' in eq'n (1)

$-z = \ln(0.0526)$

$-z = -2.94$

$\boxed{z = 2.94}$

The given equation of log(odds) or z

$z = -64 + 2 * hours$

$2.94 = -64 + 2 * hours$

$2 * hours = 2.94 + 64$

$hours = \frac{2.94 + 64}{2}$

$$\text{hours} = \frac{66.94}{2}$$

$$\boxed{\text{hours} = 33.47}$$

The student should study at least 33.47 hours, so that he will pass the exam with more 95% probability.