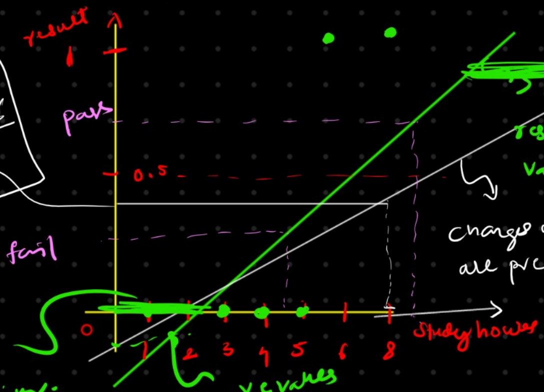
* Logistic Regression (Binary classification)

* can linear regression solve binary classification.



Squashing.

person studying 8 hr is failed

12 hours

we should squash

Best fit line respect to these values

change when outliers are present.

Study hours	result
1 hour	Fail
3 hours	Fail
4 hours	Fail
5 hours	Fail
6 hours	pass
8 hours	pass
12 hours	pass

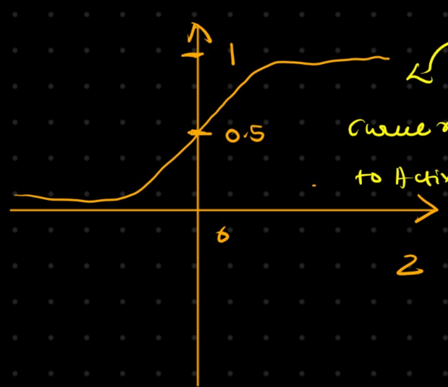
Let's say we have this dataset try to solve this using linear regression.

* But when we introduce any new datapoint or outliers then our best fit line will change

* now if we ... to predict using the new best fit line the output will not be the same.

* why we can't use Lg
* outliers (best fit line changes)

Sigmoid activation



curve respected to Activation function

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

we gonna use this to squash the best fit line

* we create best fit line upon that we gonna apply sigmoid function

$$\boxed{h_\theta(x) = \sigma(\theta_0 + \theta_1 x)}$$

$\boxed{z = \theta_0 + \theta_1 x}$

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

logistic regression hypothesis.

Linear regression cost function

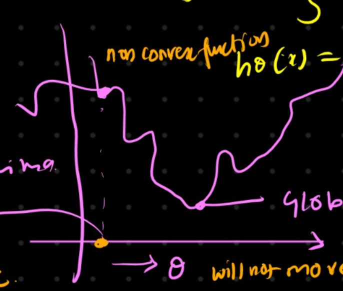
$$\boxed{J(\theta_0,\theta_1) = \frac{1}{2m}\sum_{i=0}^{m}\left(h_\theta(x)^{(i)} - y^{(i)}\right)^2}$$

logistic Regression cf

$$J(\theta_0, \theta_1) = \frac{1}{2m}\sum_{i=0}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)$$

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

non convex function



The θ will be stagnant at here.

→ θ will not move - - >

⇓

so we use log loss

Convergence algorithm

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

Repeat

{

$$\boxed{\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1)}$$

}

$$\boxed{\log loss = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(h_\theta(x)) + (1-y_i)\log(1-h_\theta(x))\right]}$$

convex function //

* Performance metrics

① Confusion matrix

actual values

predicted values	1	0
1	2	2
0	2	1

dataset	y	$\hat{y}$
—	0	1
—	0	1
—	1	1
—	0	0
—	1	0
—	1	1
—	1	0

95

5

Correct predicted no with respect to '1'	TP	FP
	FN	TN

no of missjudged prediction with respect to '1'

no of correct prediction with respect to '0'

no of missjudged prediction with respect to '0'

$$Accuracy = \frac{\underbrace{TP+TN}_{\text{correct predictions}}}{\underbrace{TP+FP+FN+TN}_{\text{total prediction}}}$$

out of the

$$= \frac{2+1}{2+2+2+1}$$

$$\boxed{Accuracy = \frac{4}{7}}$$

② Precision & recall

Dataset

↳ 1000 datpoints — 900 — '1' ; 100 — '0' } imbalanced dataset.

⇓

now if we calculate

Accuracy = 90%

$$Accuracy \rightarrow \frac{TN+TP}{TN+FN+FP+TP}$$

$$\boxed{\text{Precission} = \frac{TP}{TP+FP}}$$

out of all the actual values how many are predicted right

$$90\% = \frac{900 + 0}{1000}$$

but the model not even predicting '0' once.

So the model consider predicting 0 is an error because it is not trained on 0's much

so, we use precision

actual values

		1	0
predicted values	1	TP	FP
	0	FN	TN

$$\frac{TP}{TP+FP}$$ ↓↓↓

ii) Recall

actual values

		1	0
predicted values	1	TP	FP
	0	FN	TN

$$\frac{TP}{TP+FP}$$

Should aim to reduce this always

$$\frac{TP}{TP+FN}$$

should aim to reduce this

$$\boxed{\text{Recall} = \frac{TP}{TP+FN}}$$ ↓↓↓

out of all the predicted values how many were right

use case

tommorow my stock will crash or not

	crash	not crash
crash	TP	FP
not crash	FN	TN

Recall

3. F-Beta score

$$(1+\beta^2)\ \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

→ if both FP & FN is important

$$\boxed{\text{F1score} = 2 \times \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}}$$

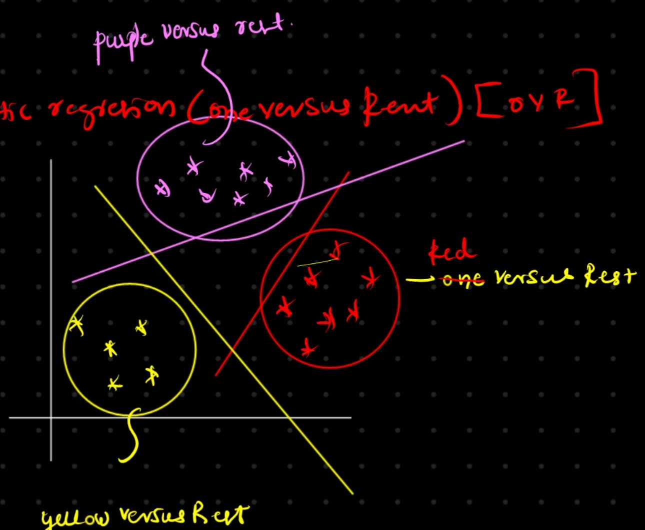
→ if FP is more important

$$\boxed{F_{0.5}\ \text{Score} = (1+0.25)\ \frac{P * R}{P + R}}$$

→ if FN is important.

$$\boxed{F_2\ \text{score} = (1+4)\ \frac{P * R}{P + R}}$$

Logistic regression (one versus Rest) [OVR]



f_1	f_2	f_3	output	O_1	O_2	O_3
—	—	—	O_1	1	0	0
—	—	—	O_2	0	1	0
—	—	—	O_3	0	0	1
—	—	—	O_2	0	1	0
—	—	—	O_1	1	0	0
—	—	—	O_2	0	1	0

One hot Encoding.

on new test data

$M_1 = 0.25$

$M_2 = 0.20$ $\quad P(0.25, 0.20, 0.55) \rightarrow M_3$ output//

$M_3 = 0.55$

Roc and AUC

threshold_values = [0, 0.2, 0.4, 0.6, 0.8]

* what is the threshold ?? (Based on the project)

output y	$\hat{y}$ prob	$\hat{y}$ (0)	$\hat{y}$ (0.2)
1	0.8	1	1
0	0.96	1	1
1	0.4	1	1
1	0.3	1	1
0	0.2	1	0
1	0.7	1	1

$$\text{True-positive-Rate} = \frac{TP}{TP + FN}$$

$$\text{False-positive-Rate} = \frac{FP}{FP + TN}$$

* based on this we will plot the graph.