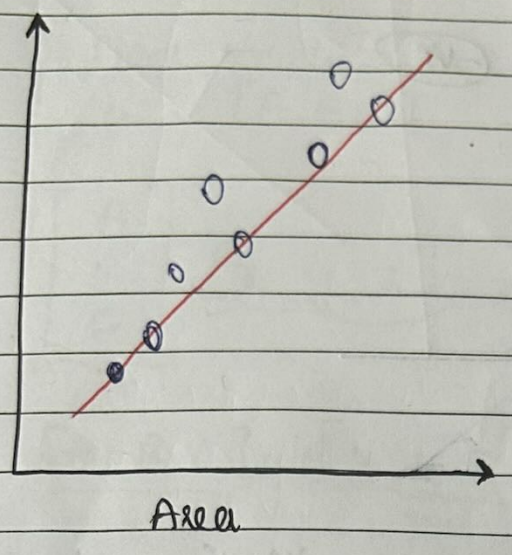
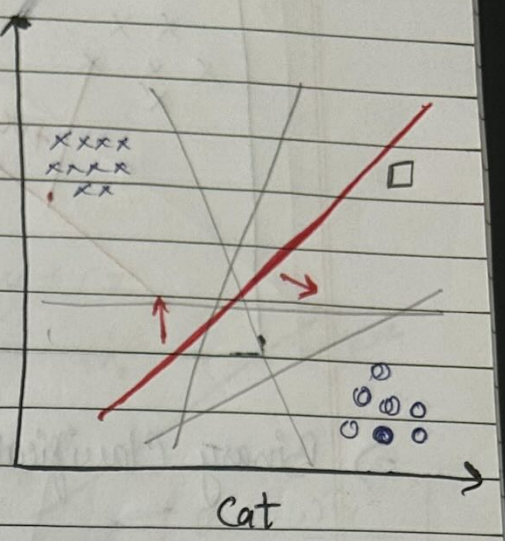
18/1/23

Date / /
Page No.

LOGISTIC REGRESSION





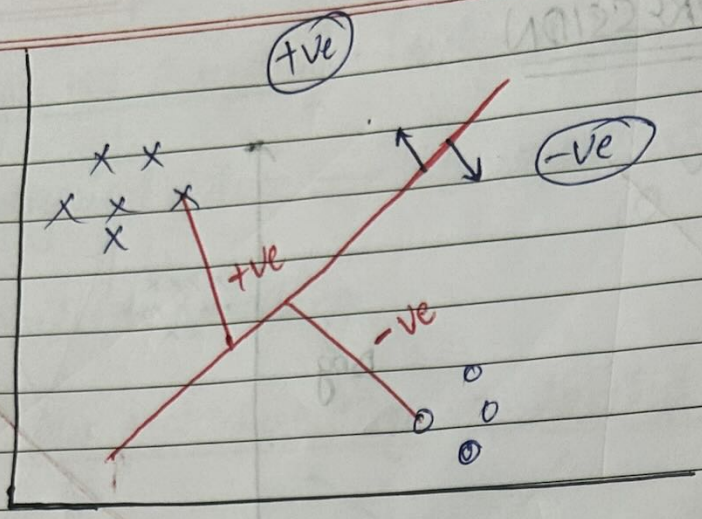
⇒ Supervised Data → Regression
→ Classification

Topics -

→ Logistic Regression
→ Loss function for logistic regression
→ Why logistic regression called regression
→ Precision | Recall | F1 - Score | Confusion Matrix
→ Sigmoid Function | log function | step function | exponential function
→ Ridge | Lasso
→ Bias - variance tarade off
→ Why we can't use linear ~~reg~~ regression instead of logistic regression?
→ Overfitting | underfitting

* In logistic we have to find which line perfectly separates the classification data.

⇒



⇒ Binary Classification

⇒ Logistic regression is a binary classification algorithm which can separate the data (of two diff. classes) with a single line.

⇒ It depends on data, which separation needs to be done.

⇒ Ques How to identify +ve/-ve class with distance?

⇒ 2D → line to divide
3D → (x, y, z) → 3 dimension / Plane
↑
to divide data

more than 3D → Hyperplane
(to divide data)

⇒ $3x - y = 0$ → line
$Ax + By + C = 0$ → line eqⁿ
$Ax + By + Cz + D = 0$ → plane

Date / /
Page No.

* Intitution of plane & distance using vectors.

Plane $\rightarrow w^T x$

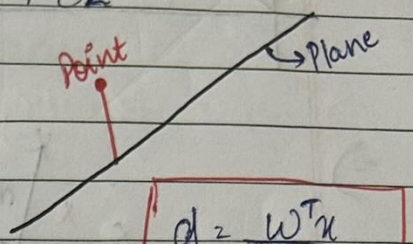
$w = [A, B, C]$

$$\begin{bmatrix} A \\ B \\ C \end{bmatrix} \cdot [x \; y \; z] = Ax + By + Cz$$

(w^T · x)

$\Rightarrow [A \; B \; C] \; [x \; y \; z]$

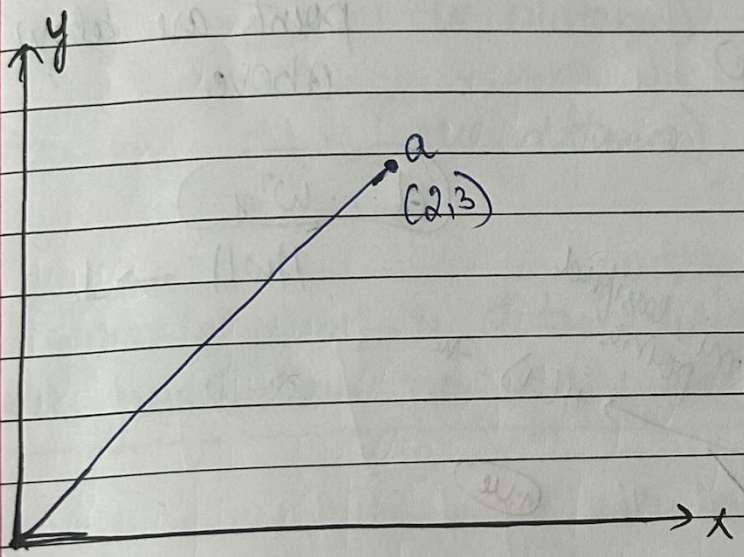
(w · x)



$$d = \frac{w^T x}{||w||}$$

distance of a point from a plane

2)



$$||a|| = 2\hat{i} + 3\hat{j}$$
$$= \sqrt{2^2 + 3^2}$$
$$= \sqrt{13}$$

19 | 21 | 23

3) $\frac{3x - y = 0}{A \quad B \quad C}$

$$d = w^T x$$

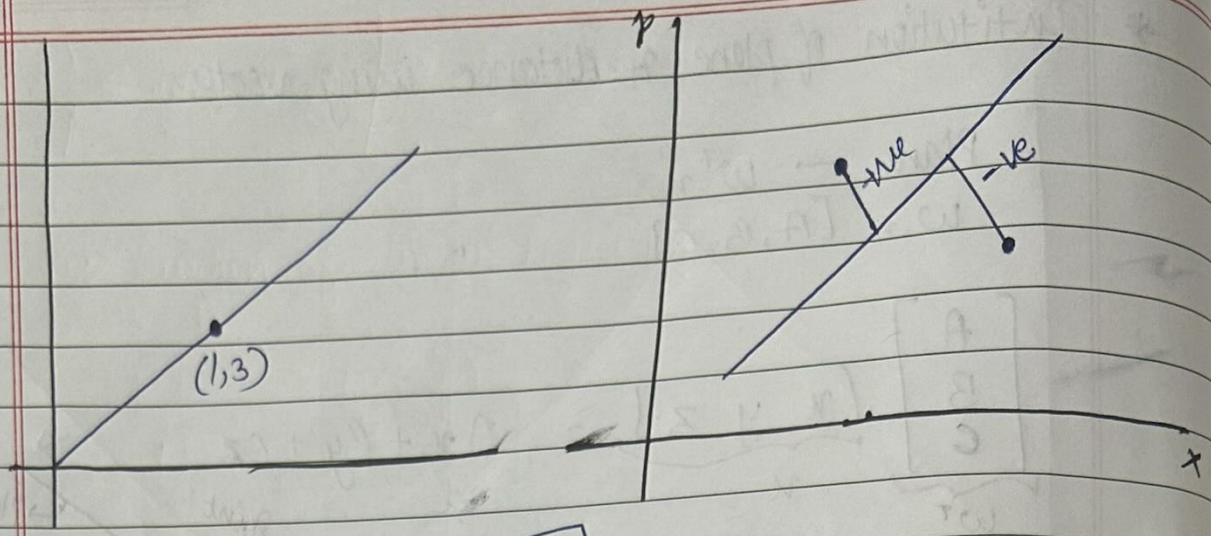
$w = [A, B]$

coefficients | weights | vectors

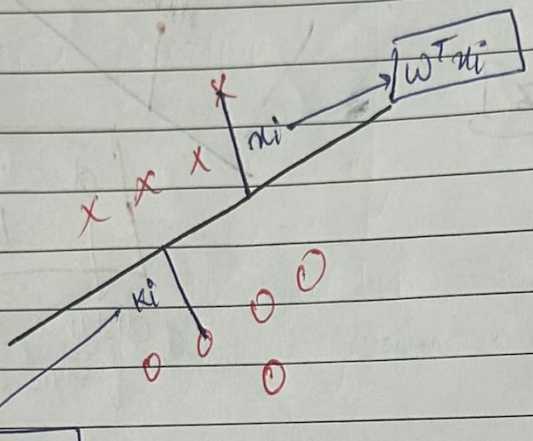
$||w|| = \sqrt{3^2 + 1^2} = \sqrt{10}$

(absolute value of vector)

2)



3)



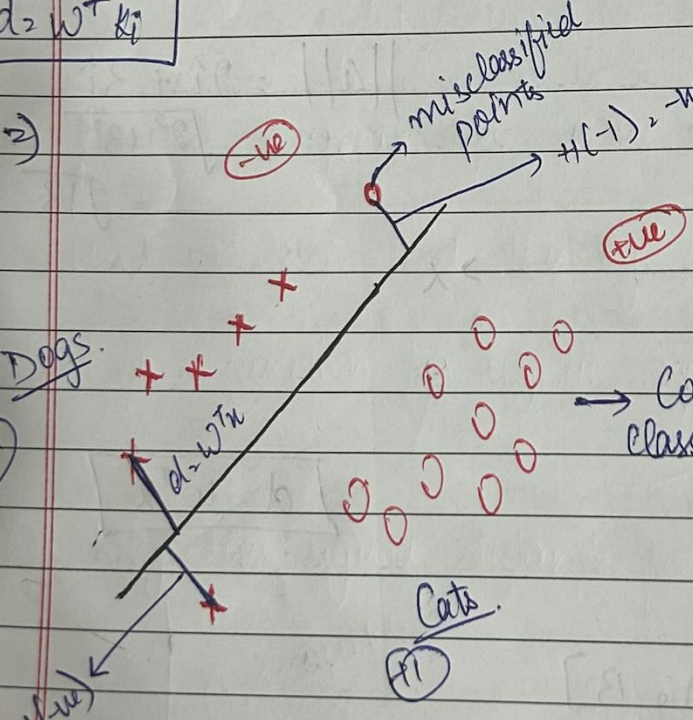
$d_2 = W^T k_i$

Distance will be +ve & -ve based on whether points are below or above.

$d = W^T x$

$\|w\| \rightarrow 1$

3)



Logistic regression is binary classification.

Date / /
Page No.

→ Labels

-1 (Dogs) +1 (Cats)

2 classes

→ $y_i \cdot w^T x$ → Signed distance

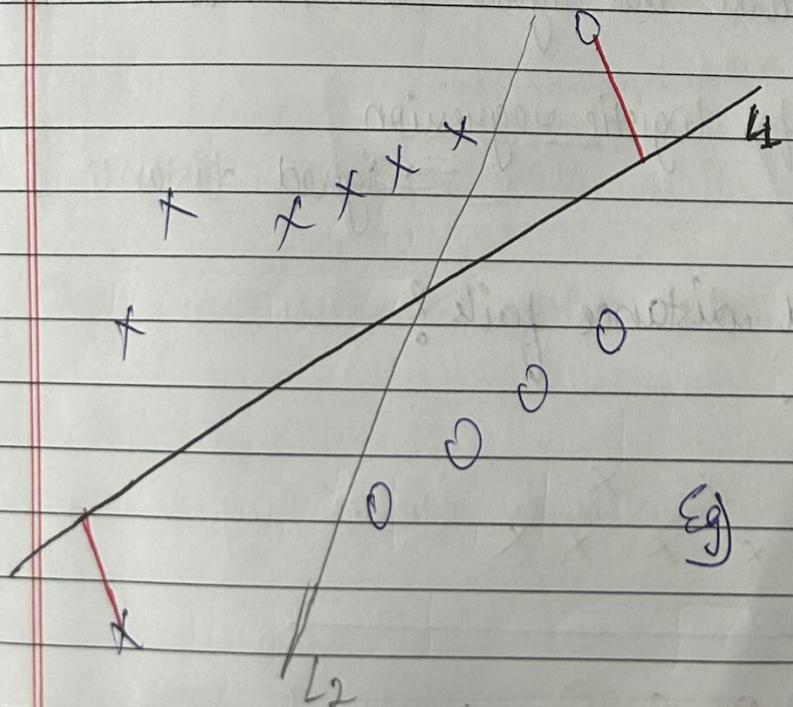
Label Distance

→ Dogs = $\underbrace{-1}_{\text{label}} \cdot$ (-ve distance) = +ve

→ Cats = $\underbrace{+1}_{\text{label}} \cdot$ (+ve distance) = +ve

* Product of label & distance (signed distance) should be greater than zero for correct classification

→



Total points = 12

$\sum_{i=1} y_i \cdot w^T x_i$

misclassification of 2 points

Eg $\alpha_1 = \sum y_i w^T x_i = 100$

$\alpha_2 = \sum y_i w^T x_i > 100$

↓

Correct classification

Date / /
Page No.

⇒ $\sum y_i \cdot w^T x_i$ → maximise

⇒ **Loss function of Linear Regression-**

$$\frac{1}{N} \sum_{i=1}^{n} (y - \hat{y})^2$$

⇒ **Loss function for Logistic Regression-**

$$\sum_{i=1}^{n} y_i w^T x_i \rightarrow \text{Maximise}$$

Basic loss function for logistic regression

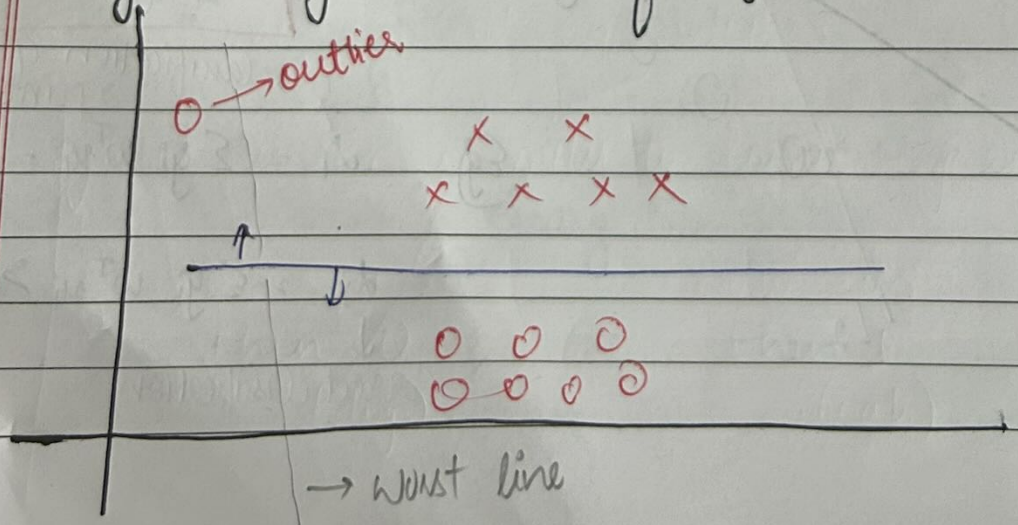
⇒ w^T → weights / vectors

$$w^* = \text{argmax} \sum_{i=1}^{n} y_i w^T x_i$$ → Signed distance

Find w^T such that this signed distance is maximised.

{ Loss function for logistic regression ⇒ Signed distance }

Ques) Why this signed distance fails?



Date / /
Page No.

⇒ Outlier distorts the line

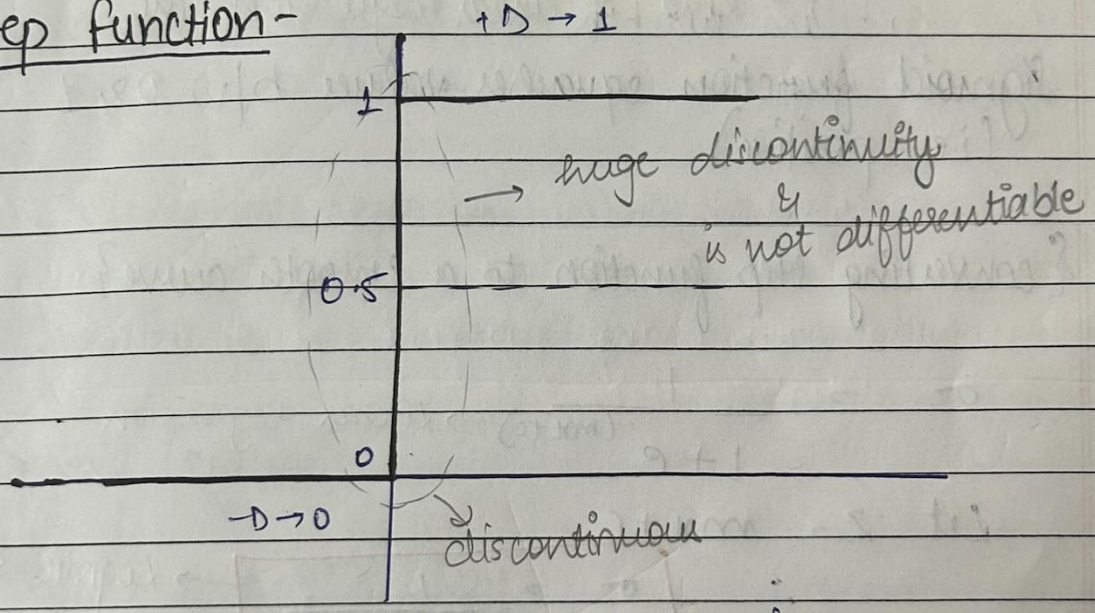
⇒ outlier effects the line, ∴ affects the distance

$y_i w^T x_i$ → outlier can lead the distance towards -ve or can minimise the distance

⇒ outlier impact signed distance, ∴ this is the drawback of signed distance

⇒ solution to signed distance → STEP FUNCTION

Step function -



∴ step function fails as it is discontinuous on y-axis

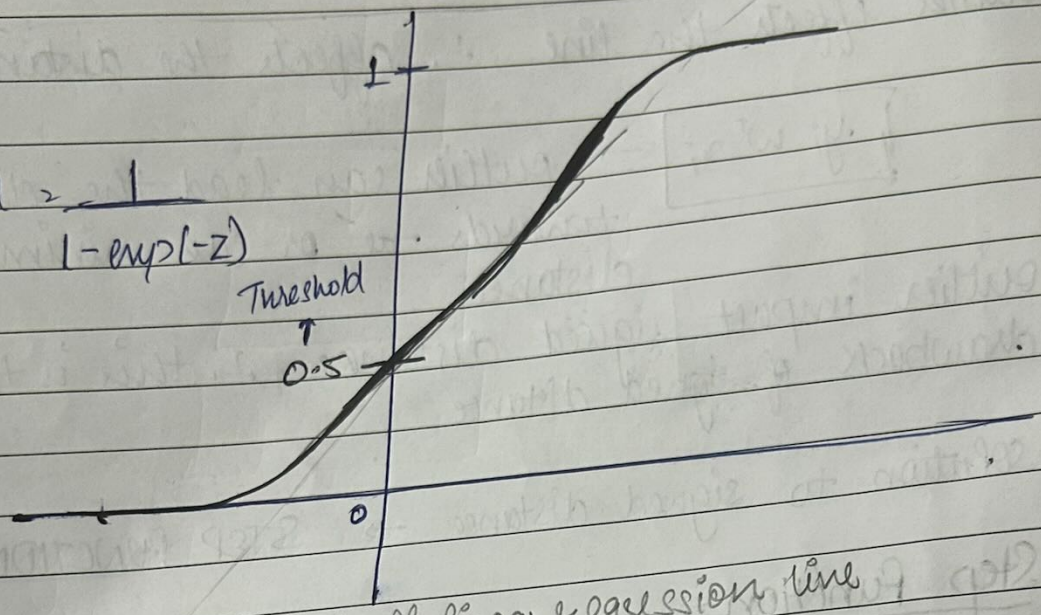
Step function bound values b/w 0 & 1 [0, 1]

[-∞, ∞]

Step output → 1 if $f > 0$
↳ 0 if $f < 0$

⇒ solution to step function → SIGMOID FUNCTION

SIGMOID FUNCTION



→ kind of linear regression line

Sigmoid function squashes values b/w 0 & 1 (i.e. positive)

& converting step function to a smooth curve }

⇒ $\sigma = \frac{1}{1+e^{-(mx+c)}}$ → ~~linear regression~~

Let $z = mx+c$

$\therefore \sigma = \frac{1}{1+e^{-z}}$ → Logistic sigmoid function

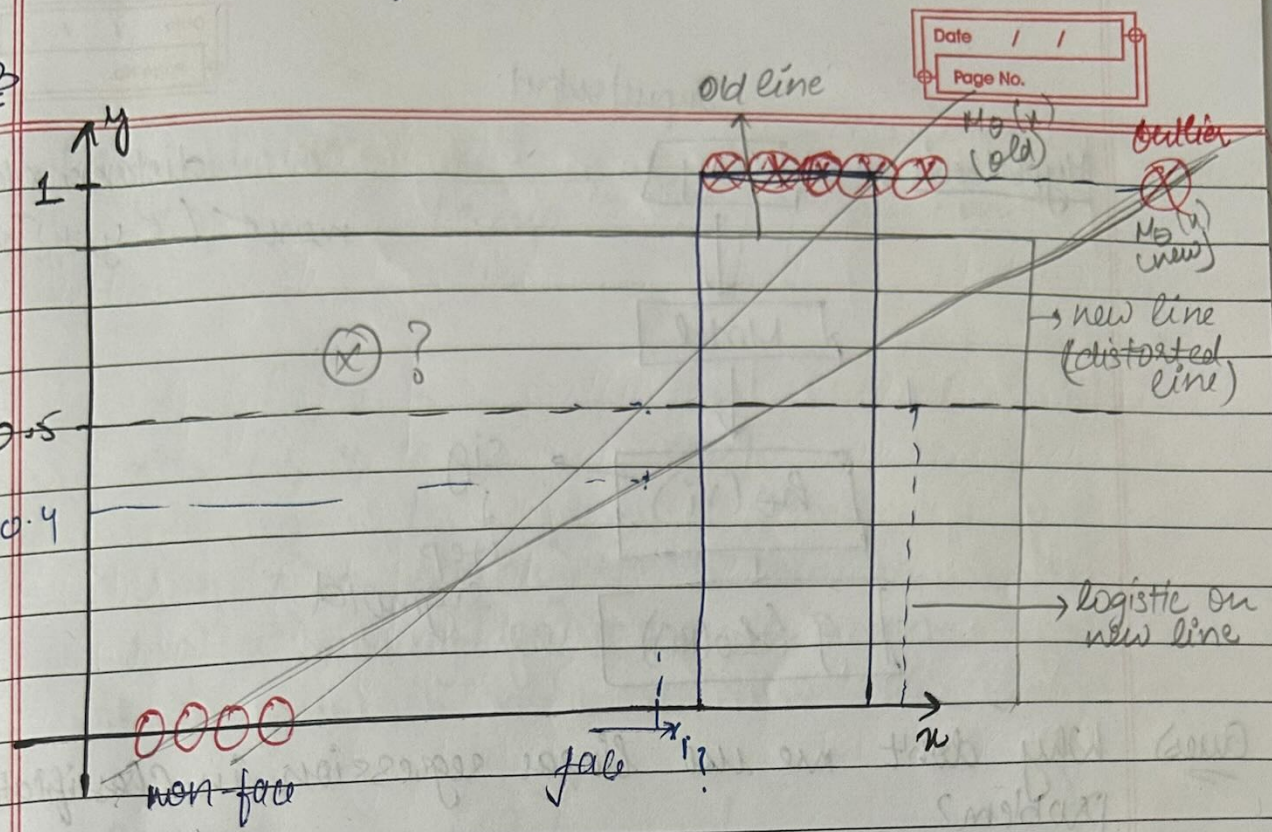
⇒ exponential -ve values ko +ve ki deta hai
- (why we did exponential?)

⇒ sigmoid function is similar to linear regression

⇒ If $w^T < 0.5$ then 0 (dogs)

$w^T > 0.5$ then 1 (cats)

20/1/23



⇒ if ≥ 0.5 → face
< 0.5 → non-face

⇒ Logistic Regression

Geometric | Probabilistic

⇒ Sigmoid (signed)

↓

values : 0.7 | 0.6 | 0.1

⇒ Step function , bound in 0, 1
Sigmoid function , bound b/w 0 & 1

⇒ Sigmoid → $\frac{1}{1+e^{-z}}$ $\{z, w^T x\}$

Date / /
Page No.

Input/output

2) Hypothesis, x, y → Model → $h_\theta(x_i)$ → $g(h_\theta(x_i))$

sig → step sigmoid

distance x label

max $\leftarrow (\sum y_i w^T x_i)$

Ques) Why don't we use linear regression in classification problem?

⇒ line distort hui (beoz of outlier), toh threshold value change hogi

⇒ linear regression issliye use nhi krte, kyuki, outlier ki wajah se line distort ho jaati hai, jis wajah se threshold value change ho jaati hai & woh wrong classification kr deta hai.

23/1/23

(1) $\text{argmax} \sum_{i=1}^{n} y_i w^T x_i$

(w)

find this w

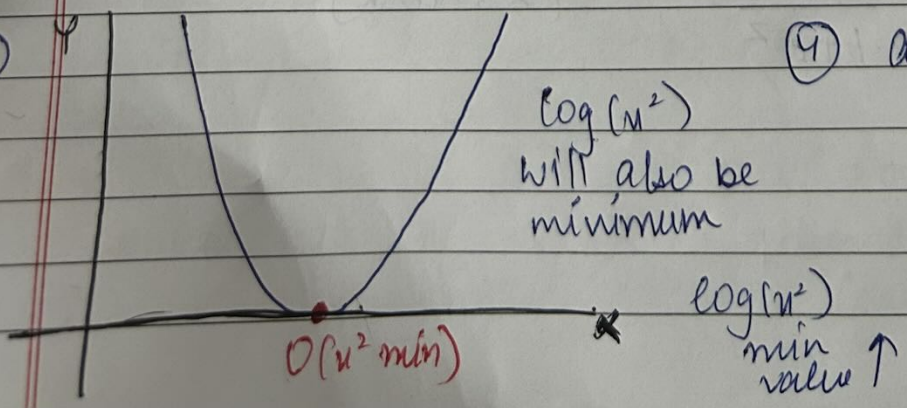
Transforming to →

(2) $f(x) = x^2$

$g(x) = \log(f(x))$

$= \log(x^2)$

(3)



(4) $\text{argmax} \sum \frac{1}{1+e^{(-y_i w x_i)}}$

Date / /
Page No.

(5) $\underset{(w)}{\text{argmax}} \; \Sigma \; \frac{1}{1+\exp^{-(y_i w^T x_i)}}$ → Monotonic increasing function

⇒ $x = [1, 2, 3 \; ---- \; 100]$
$x^2 = [1, 4, 9, \; ---- \; 100]$

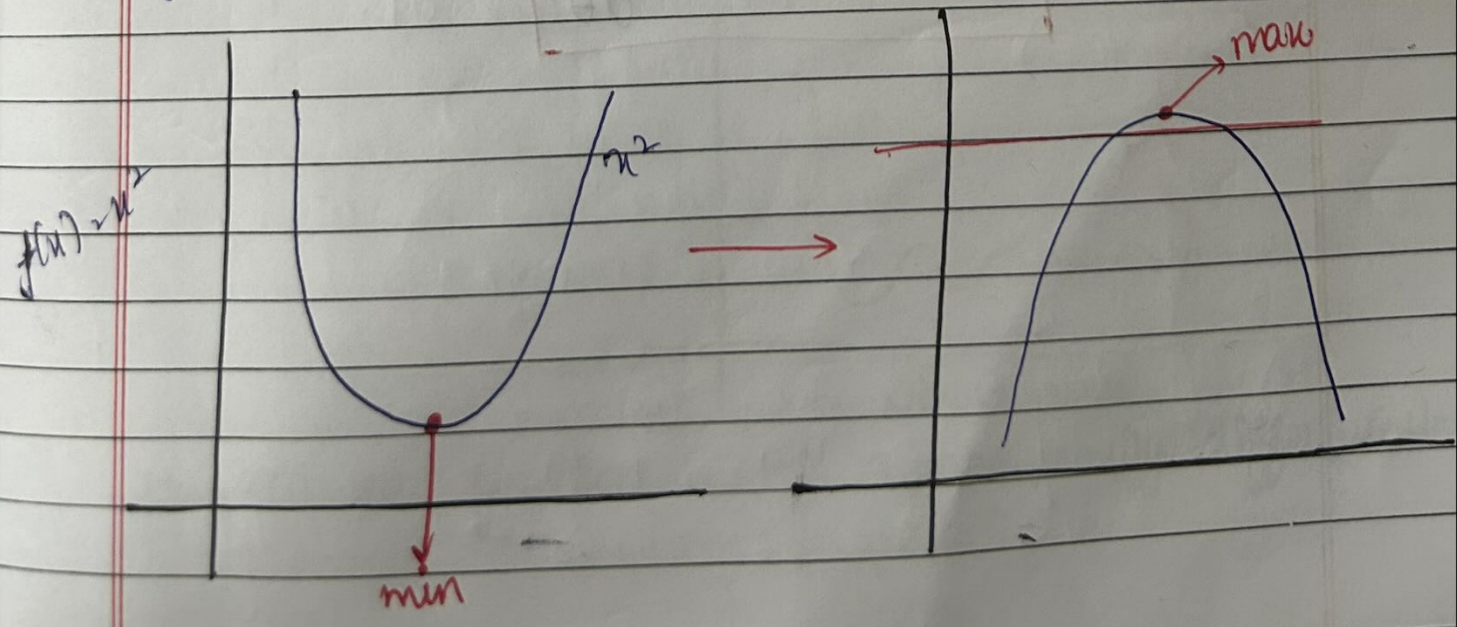
$\log(x) = \log(10) \rightarrow$ maximum
$\log(x^2) = \log(100) \rightarrow$ maximum

⇒ $\underset{w}{\text{argmax}} \; \Sigma \log\left(\frac{1}{1+\exp(-y_i w^T x_i)}\right)$

⇒ $\underset{\text{maximise}}{\underbrace{\text{argmax} - \Sigma \log(1+\exp(-y_i w^T x_i))}}$

$\left\{ \log x^n = n \log x \right.$
$\left. \log\left(\frac{1}{x}\right) = \log x^{-1} = -\log x \right\}$

⇒ $f(x) = \max$



Date / /
Page No.

⇒ Loss function = $\text{argmin} \sum_{i=1}^{N} \log(1+\exp(-y_i w^T x_i))$

* $\{x^2 = \max$
$\log(x^2) \to \max\}$

⇒ log optimisation mein easy rehta hai, issiliye log lete hai
log → monotonic increasing function
⇒ optimisation mein monotonic increasing function use hota hai

⇒ log ki values 0-1 ke beech mein $-\infty$ hoti hai & fir increase krti hai

⇒ Exponential → monotonic increasing function

$(-\infty$ to $2) \to y = 0$
Then starts increasing

⇒ Loss function ko minimise krte hai using gradient descent
then,

$$\theta = \theta - \eta \cdot \frac{dL}{d\theta}$$

Date / /
Page No.

⇒ Probabilistic Approach

⇒ Probabilistic approach uses MLE
(Maximum Likelihood Estimation)

→ Probability & liklihood are proportional.

→ Probabilistic Approach - finding the best parameters for the best fit line using partial derivatives.

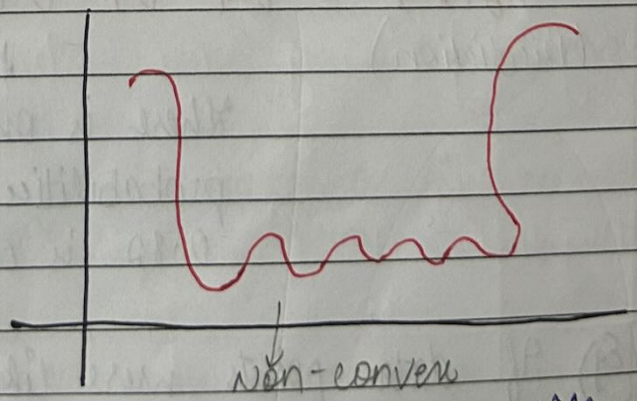
⇒ $h_\theta(x) = \frac{1}{1+e^{-(\theta^T x)}}$ → 100 → 1, 90 → 0.9, 80 → 0.8

$$MSE = \frac{1}{n}\left(y^{(i)} - h_\theta(x)\right)^2$$

MSE ↓ Loss; $(y^{(i)} - h_\theta(x))^2$ ↓ Squared loss

$\left\{\theta = \theta - \eta \frac{dL}{d\theta}\right\}$

⇒ $(1-0.8)^2 + (1-0.2)^2 + (1-0.4)^2 + \dots$ → Curvature



Non-convex function is not always easily differentiable.

→ **Probabilistic approach loss function:**

Log Loss | Binary Cross Entropy

→ **Multiclassification:-**

Categorical cross Entropy

⇒ Avg Loss = $\frac{-1}{m} \sum_{i=1}^{m} \left[y^{(i)} \log h_\theta(x^{(i)}) + (1-y^{(i)}) \log(1-h_\theta x^{(i)}) \right]$

pheli class — doosri class

Actually minimise the diff b/w the probabilities distribution

⇒ Dog ki agar 0.9 toh cat ki 0.1 hogi { pheli class x, then second class 1-x }

Ques Let us see what this function does?

→ Actually minimize diff b/w 2 probabilities distributions

Eg) y = 1 0 1 1 0 {5 data points}

$h_\theta(x^{(i)})$ = 0.9 0.1 0.99 0.82 0.1

(Prediction)

There is only little bit difference in probabilities, like 0.1 is close to 0 & 0.82 is close to 1

Eg) If data points were like

1 0 1 1 0

0.5 0.5 0.6 0.4 0.3 → This difference is large

Date / /
Page No.

⇒ Logistic Regression is a model for binary classification predictive modeling. The parameters of a logistic regression model can be estimated by the probabilistic framework called maximum likelihood estimation.

⇒ Geometric mein → signed distance
Probabilistic mein → Maximum Likelihood Estimation

⇒ The parameters of logistic regression equation are estimated via maximum likelihood, considering the conditional response variable as Bernoulli.

⇒ **Maximum Likelihood Estimation**

i) Sometimes the probability density functions are not known.
ii) So, these probability density functions are to be estimated from available data.
iii) Maximum Likelihood method views the parameters as quantities whose values are fixed but unknown.
iv) The best estimate of their value is defined to be the one that maximises the probability of obtaining the samples actually observed.

24/1/23

$h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$ → it represents line separating two values.

This is our hypothesis function & we know this function lies in range [0, 1]

$$\text{Loss} = \frac{1}{n} \sum \left(y^{(i)} - h_\theta(x^{(i)})^2\right)$$

↓

Squared error loss function

this is non-convex

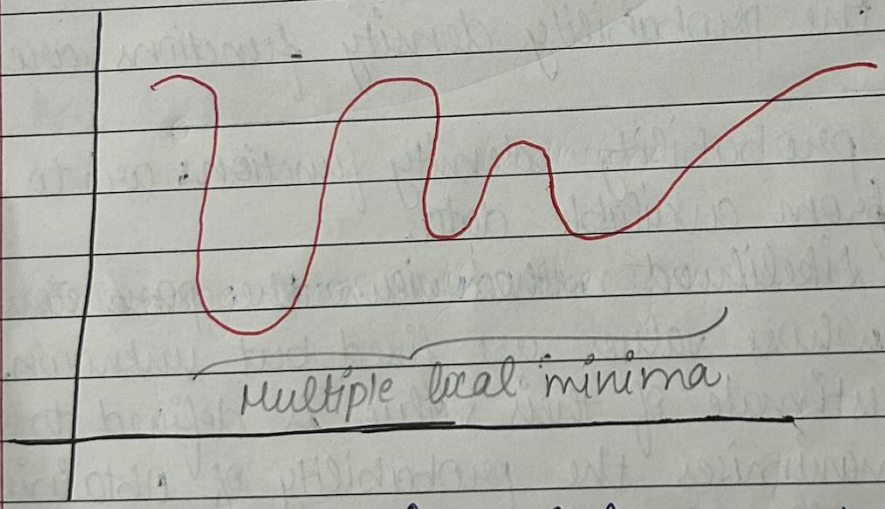
→ if class is +ve , $(1-0.8)^2$

↓

Confidence for pt (a) will be high

→ if class is -ve, confidence will be low

→ $(1-0.2)^2 + (0-0.6)^2$



→ For multiclass → Categorical cross entropy

→ Log loss | Binary Cross Entropy → for binary classification

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

means?

$y^{(i)} \log h_\theta x^{(i)}$ → confidence of point being +ve

$(1-y^{(i)}) \log\left(1-h_\theta(x^{(i)})\right)$ → confidence of point being -ve

$h_\theta x = 0.6$

$1 - 0.6 \rightarrow 0.4$ → it says that there is 60% surity that it belongs to class 1, & 40%, that it belongs to class 0.

* (Jiska confidence jyada hai, woh model jyada acha hai)

⇒ I want my predictions to be as close as the real probabilities.

→ And this loss function will help to do this task.

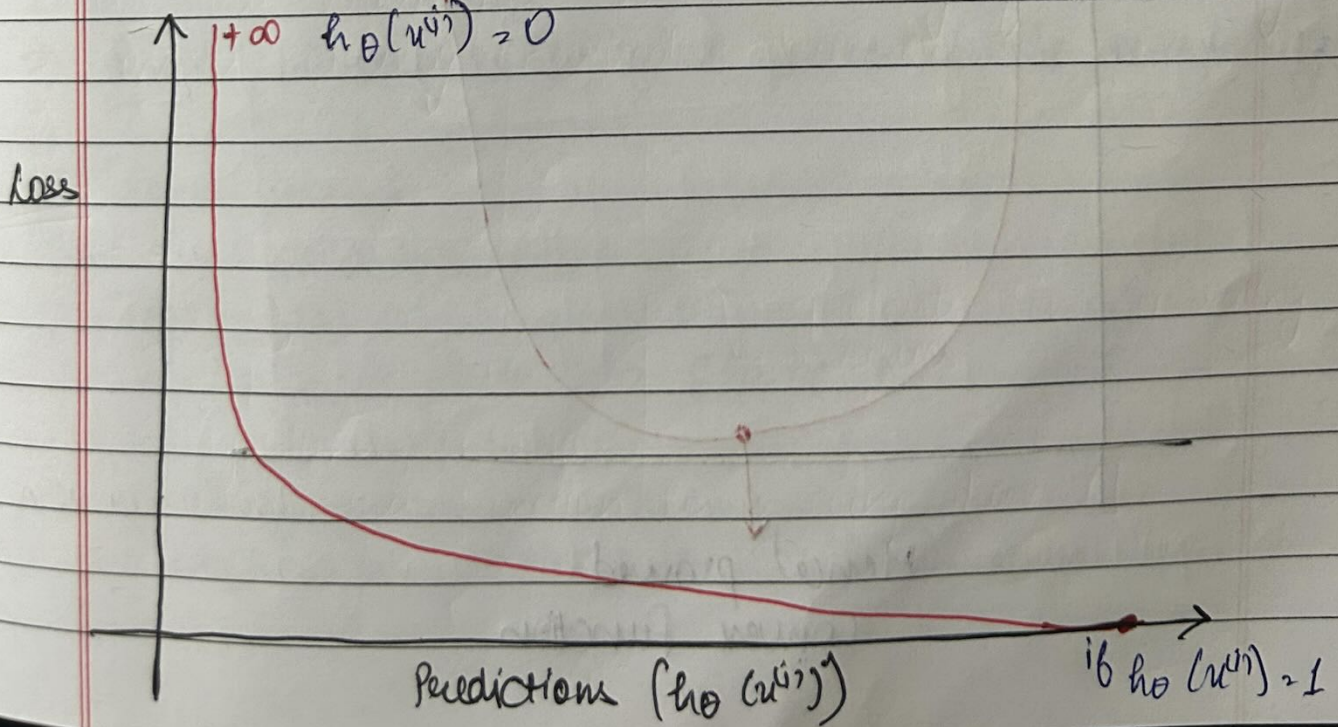
⇒ Let's visualise,

Case 1 $y^{(i)} = 1 \quad 1 \quad 1 \quad 1 \quad 1$

Loss function $= \frac{-1}{m} \sum_{i=1}^{m} y^{(i)} \underbrace{\log h_\theta(x^{(i)})}_{1}$

(second term of loss function will be zero)

$\therefore (1 - y^{(i)}) \log (1 - h_\theta(x^{(i)})) \rightarrow 0$

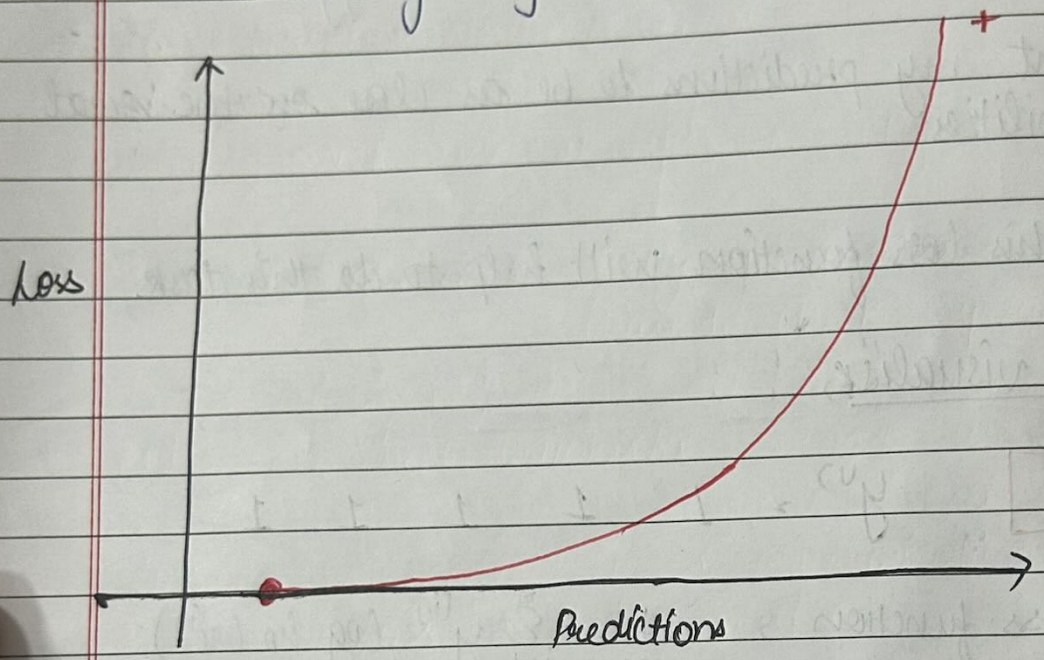


Case 2. $y^{(i)} = 0$ 0 0 0 0 0

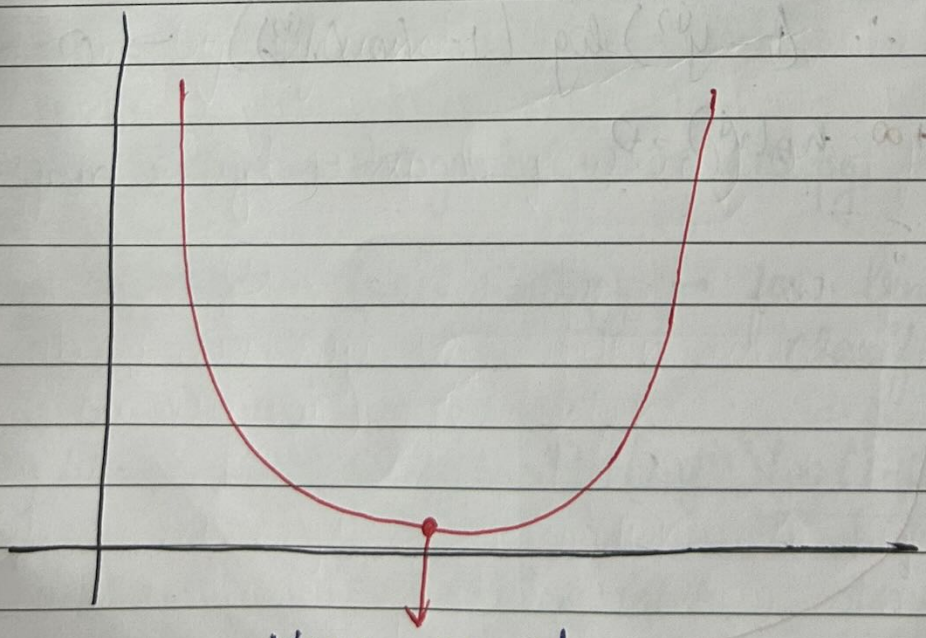
$$\text{Loss} = \frac{-1}{m} \sum_{i=1} (1 - y^{(i)}) \log (1 - h_\theta(x^{(i)}))$$

(first term will be zero)

$\therefore y^{(i)} \log (h_\theta(x^{(i)})) \to 0$



⇒ Combining both the graphs.



Hence proved
Convex function

25/1/23

Date / /
Page No.

⇒ Loss = $\text{argmin} \sum_{i=1}^{n} \log\left(1 + \exp\left(-y_i w^T x_i\right)\right)$

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

$$= -\sum_{i=1}^{m} \left[y^{(i)} - h_\theta(x^{(i)})\right] x_j^{(i)}$$

($x_j^{(i)}$: i → sample, j → feature)

(After doing derivation)

⇒ Final update,

$$\theta_j = \theta_j + \eta \sum_{i=1}^{m} \left[y^{(i)} - h_\theta(x^{(i)})\right]$$

(η: learning rate)

⇒ Learning Rate

Ridge / Lasso Regression
(L2 Loss) (L1 Loss)

$$\text{Loss} = \underset{(w)}{\text{argmin}} \sum_{i=1}^{n} \log\left(1 + \exp\left(-y_i w^T x_i\right)\right)$$

⇒ Ridge / Lasso ⇒ is used to control the overfitting (Regularisation)

Techniques

Ridge + Lasso → Elastic Net

Ridge Loss → L2

Lasso Loss → L1

Date / /
Page No.

⇒ Ridge Regression & Lasso Regression

$$\text{Loss} \rightarrow \text{argmin} \sum_{i=1}^{n} \log\left(1 + e^{-y_i w^T x_i}\right) + \alpha \|w\|^2$$

Penalty term (L2 loss)

Hyperparameter or L2 norm

→ Loss function derive krna hai,
↓
usko optimise krna hai, (GD ke through)

⇒ η, α, c → Hyperparameters
(have to check experimentally)

* Ridge has some drawbacks ∴ we use Lasso

⇒ Loss → minimise (loss + regularisation)
= Loss + $\lambda |w|^2$

→ $\alpha = 0$ = Loss = minimize ⇒ Overfitting

* λ has big influence.

⇒ $\text{argmin} \sum \log\left(1 + \exp\left(-y_i w^T x_i\right)\right) + \alpha |w|_2^2$

find values

① ②

(Jab w max hoga, tab function minimize hoga kyuki (-) lga hua hai) → first section

(Jab w min hoga, tabhi function min hoga) → second section

⇒ Equilibrium to find optimal values of w

Date / /
Page No.

2) $\alpha \uparrow \rightarrow$ underfit
$\alpha \downarrow \rightarrow$ overfit

L-1 regularisation $|w|_1$

$\downarrow$

$< w_1, w_2, w_3 >$

$|w_1| + |w_2| + |w_3|$

If we add this above penalty term,

$\text{argmin} \sum_{i=1}^{N} \log(1 + \exp(-y_i w^T x_i)) + \alpha |w|_1$

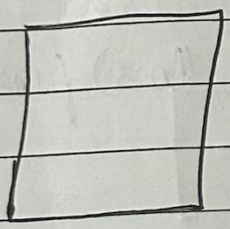
2) $f_1 \quad f_2 \quad f_3$

$\downarrow \quad \downarrow \quad \downarrow$

$w_1 \quad w_2 \quad w_3$

$w_3 > (w_2, w_1)$

2) f_3 is most important feature
(Jiska w jyada, woh jyada imp)



→ 100 features

* L1, not important features ko 0 kardeta hai

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

→ Sparse Matrix

2) Feature selection L1 khud hi kardeta hai

2) Categorical
↳ Chi-square test
↳ Anova

Numerical
↳ z-test
↳ t-test

Date / /
Page No.

⇒ To find the values on which the model is working very well → Grid Search CV.

⇒ $\alpha = [10^{-3}, 10^{-2}, 10^{-1}, 1, 10^{2}, 10^{3}]$ } Grid Search CV
$[0.001, 0.01, 1, 100, 1000]$ ↓ (Random Search Industries)

6 Models

→ Jis value pe, good accuracy aati hai, woh model choose kiya jaata hai

⇒ Elastic Net

$$\text{Loss} + \underbrace{\alpha_1 |w|_2^2}_{\text{yeh slow kata hai computation}} + \underbrace{\alpha_2 |w|_1}_{\text{yeh fast kata hai, kyuki yeh unimportant ko zero kr deta hai}}$$

→ hme w ki min value chahiye taaki (loss) function min ho ske.
Isliye hum L2 use nhi karte

6/2/23

⇒ Linear regression → predicted value is continuous (approximate)
Classification → predicted value is discrete

1 (100% sahi predict kr rha hai) 0 (get predict kr rha hai)

Continuous mein accuracy ka formula problem karta hai
∴ hum error nikalte hai

Model Predictions

no. of observations → n = 165	Predicted NO	Predicted YES	
Actual NO	(Model → NO, Real → NO) TN = 50	(Model → YES, Real → NO) FP = 10	60 (Total actually NO obs)
Actual YES	(Model → NO, Real → YES) FN = 5	(Model → YES, Real → YES) TP = 100	105 (Total actual YES obs)
↳ Ground Truth (dataset)	55 (Total predicted NO obs)	110 (Total predicted YES obs)	

~ Confusion Matrix

Ques) Confusion matrix kyu aayi?
→ Kyuki accuracy score aaya.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Accuracy measures how often the classifier correctly predicts

⇒ Eg) 85% accurate

15% Khrab
yeh kahan get hai woh confusion matrix btata hai.

* accuracy (sk learn) → use

⇒ TN & TP values are important (always +ve)
Focus on reducing FN & FP
(100% accuracy if FP = FN = 0)

Date / /
Page No.

Confusion Matrix

Confusion Matrix is a performance measurement for the machine learning classification problems, where the output can be two or more classes. It is a table with combinations of predicted & actual values

* It is done on testing data.

$$\text{Precision} = \frac{TP}{TP + FP}$$

→ Precision - Precision for a label is defined as the no. of true positives divided by the no. of predicted positives.

→ F1 Score - It gives a combined idea about precision & recall metrics. It is a maximum when precision is equal to recall.
F1 score is the harmonic mean of precision & recall.

$$F1 = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

→

Predicted values \ Actual values	Positive (1)	Negative (0)
Positive (1)	TP	FP
Negative (0)	FN	TN

1 → Fraud credit card
0 → Not fraud

Date / /
Page No.

⇒ Actual → NO
Predicted → YES
} Toh fark nhi padega, kyuki in real mein nhi hai

⇒ Actual → YES
Predicted → NO
} Toh problem hogi

⇒ Precision & Recall Trade off → improves the model a little bit.

7/2/23

Confusion Matrix

It is a performance measurement for machine learning classification problem where output can be two or more classes. It is a table with 4 different combinations of predicted & actual values.

Key classifications

→ Accuracy
→ Recall
→ Precision
→ F1 - Score

		Actual Pos (1)	Actual Neg (0)
Predicted	P	TP	FP
	N	FN	TN

⇒ True positive - predicted something yes & it came out to be Yes.

⇒ True Negative - predicted no it comes NO

⇒ False Positive - predicted its Yes, it came out to be NO.

⇒ False Negative - predicted its no, it comes Yes.

→ Accuracy -

$$\frac{\text{True Positive + True Negatives}}{\text{True Pos + True Neg + False Pos + false Neg}}$$

$$\Rightarrow \frac{\text{Correct Prediction}}{\text{Total Prediction}}$$

→ Precision - This will give you how many predictions you made are right

→ Recall - Ability to recall to the number of events

⇒ Precision is the ratio of a no. of events you can correctly recall to the no. of all events you recall (min of ✓ & x events)

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive + False Positive.}}$$

* In general words,
Precision → Accuracy of positive values
Recall → Accuracy of negative values

$$\text{Precision} \uparrow_2 \frac{TP}{TP + FP\downarrow} \qquad \text{Recall} = \uparrow \frac{TP}{TP + FN\downarrow}$$

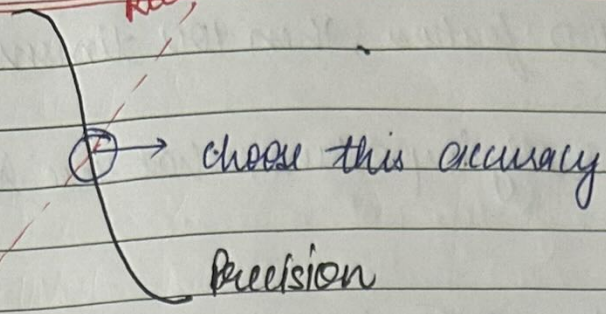
⇒ Applications

Eg) Image Classification
Precision & Recall → fark nhi padta

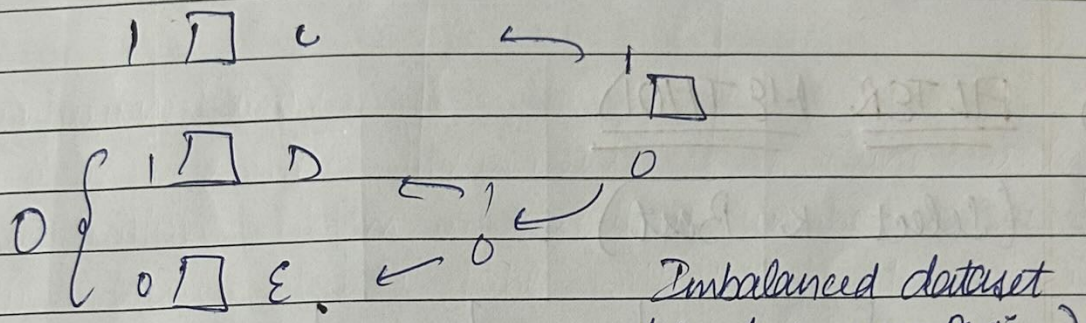
Biometric - Precision & Recall hi kaam karta hai

Date / /
Page No.

Recall

=> 

=> 1 v/s Rest
logistic → for multiple



Code Discussion,
penalty, "none"
uses neither L1 or L2

=> by default uses L2

=> solver ?

9/2/23

FEATURE SELECTION TECHNIQUES In MACHINE LEARNING