Data Science and Artificial Intelligence

Machine Learning

Classification

Lecture No. 5

By- SIDDHARTH SABHARWAL SIR

Recap of Previous Lecture

Topic Logistic Regression

Topic Cross Entropy

Topic Log likelihood

Topic

Topic

Topics to be Covered

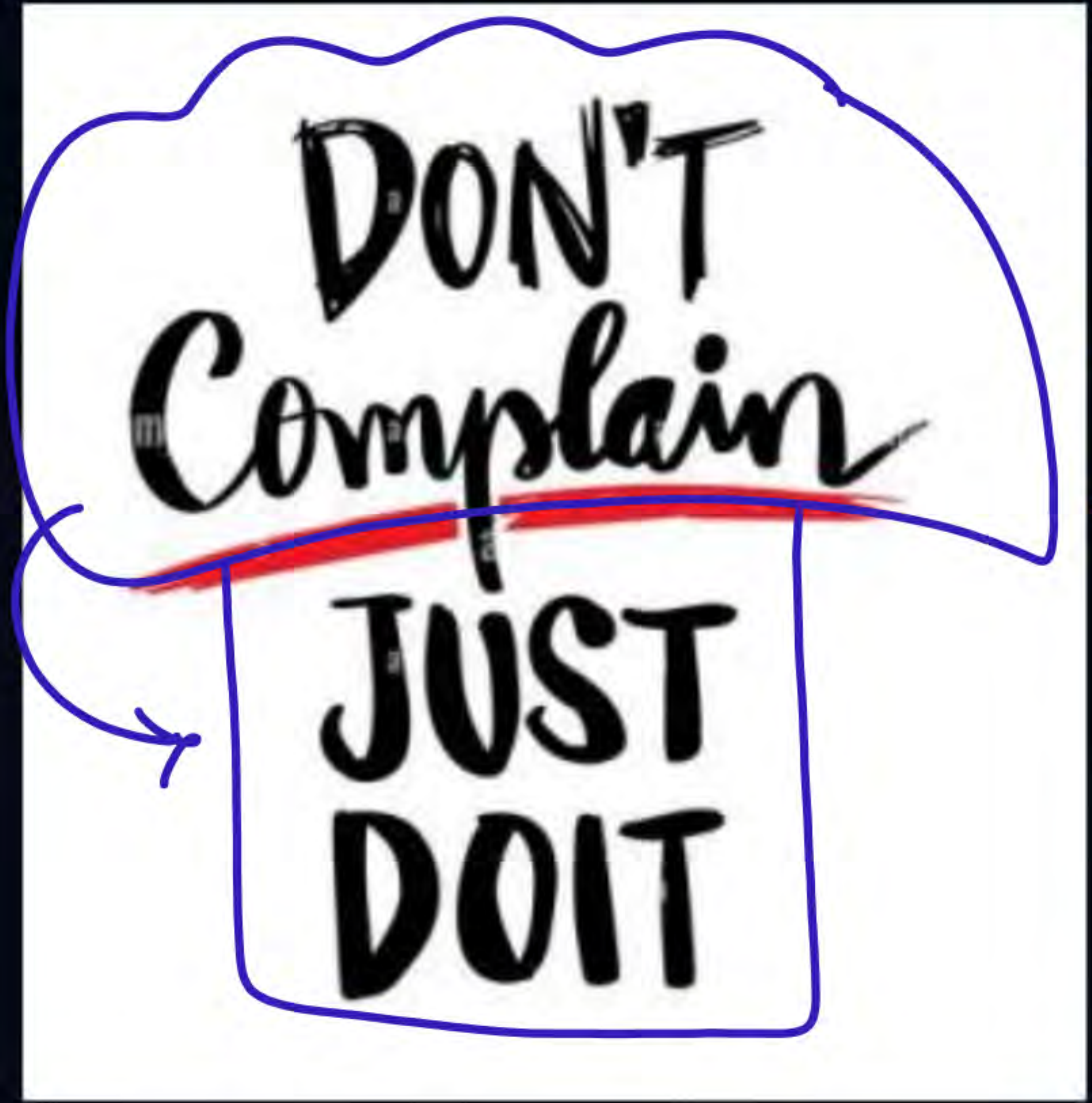
Topic Multiclass → logistic regression

Topic How to make LR → NL

Topic Confusion matrix

Topic

Topic



DON'T Complain

JUST DOIT

Skill

- Naam ← Kaam

- Izzat

About the Faculty

- AIR 1 GATE 2021, 2023 (ECE).
- AIR 3 ESE 2015 ECE.
- M.Tech from IIT Delhi in VLSI.
- Published 2 papers in field of AI-ML.
- Paper 1 : Feature Selection through Minimization of the VC dimension.
- Paper 2 : Learning a hyperplane regressor through a tight bound on the VC dimension.

By- SIDDHARTH SABHARWAL SIR

Join Our Telegram Channel!

Stay updated, get instant alerts & connect with the community

https://t.me/siddharthsirPW

By- SIDDHARTH SABHARWAL SIR

Logistic Regression :

↓ done

- log $\boxed{\text{likelihood}}$ max $\sum_{i=1}^{N} y_i \log_e p_i + (1-y_i)\log_e(1-p_i)$ → concave

$$CE = -\sum_{i=1}^{N} \left(\qquad \right)$$

→ convex

$\boxed{\min CE}$

⇒ mean CE

Logistic Regression : for multiclass case

↓

* logistic regression is for binary

* How to use for multi class

Ex: we have data of 3 class.

we create 3 classifier one for each class.

	y
1.	y_1
2.	y_1
3.	y_2
4.	y_2
.	y_1
.	y_3
.	y_3

for Class 1 : Class 1 points '1'   one versus rest

Class 2, 3 points 0

2 Class data

→ apply logistic regression find β^1

(1 pt) In this question, assume that we are using the logistic regression model $\hat{y} = \sigma(x^T\theta)$.

Suppose we want to modify cross-entropy loss to penalize predictions for observations that are truly positive twice as much as we penalize predictions for observations that are truly negative. Which of the following loss functions could we use? Recall that the average cross-entropy loss is:

Sol.

$$R(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\left(y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)\right)$$

○ $R(\theta) = -\frac{2}{n}\sum_{i=1}^{n}\left(y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)\right)$

○ $R(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\left(2y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)\right)$

○ $R(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\left(y_i \log(\hat{y}_i) + 2(1 - y_i)\log(1 - \hat{y}_i)\right)$

○ $R(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\left((y_i + 2)\log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)\right)$

Suppose after training our model we get $\vec{\beta} = \begin{bmatrix} -1.2 & -0.005 & 2.5 \end{bmatrix}^T$, where -1.2 is an intercept term, -0.005 is the parameter corresponding to passenger's age, and 2.5 is the parameter corresponding to sex.

i. [3 Pts] Consider Sīlānah Iskandar Nāsīf Abī Dāghir Yazbak, a 20 year old female. What chance did she have to survive the sinking of the Titanic according to our model? Give your answer as a probability in terms of σ. If there is not enough information, write "not enough information".

Sol.

$$P(Y = 1 | \text{age} = 20, \text{female} = 1) = \boxed{}$$

ii. [3 Pts] Sīlānah Iskandar Nāsīf Abī Dāghir Yazbak actually survived. What is the cross-entropy loss for our prediction in part i? If there is not enough information, write "not enough information."

Suppose you have a logistic regression model for spam detection, using a dataset with a binary outcome that indicates whether an email is spam (1) or not spam (0). The predictor variables x_1, x_2, and x_3 are boolean values (0 or 1) that indicate whether the email contains the words "free", "order", and "homework", respectively. The model has four parameters: weights w_1, w_2, w_3, and offset b.

You find that emails containing the words "free" and "order" have a higher probability of being spam, while emails containing the word "homework" have a lower probability of being spam.

Sol

Given this information, which of the following signs is most likely for the weights w_1, w_2, and w_3?

(A) All positive

(B) All negative

(C) w_1 and w_2 are positive, w_3 is negative

(D) w_1 and w_2 are negative, w_3 is positive

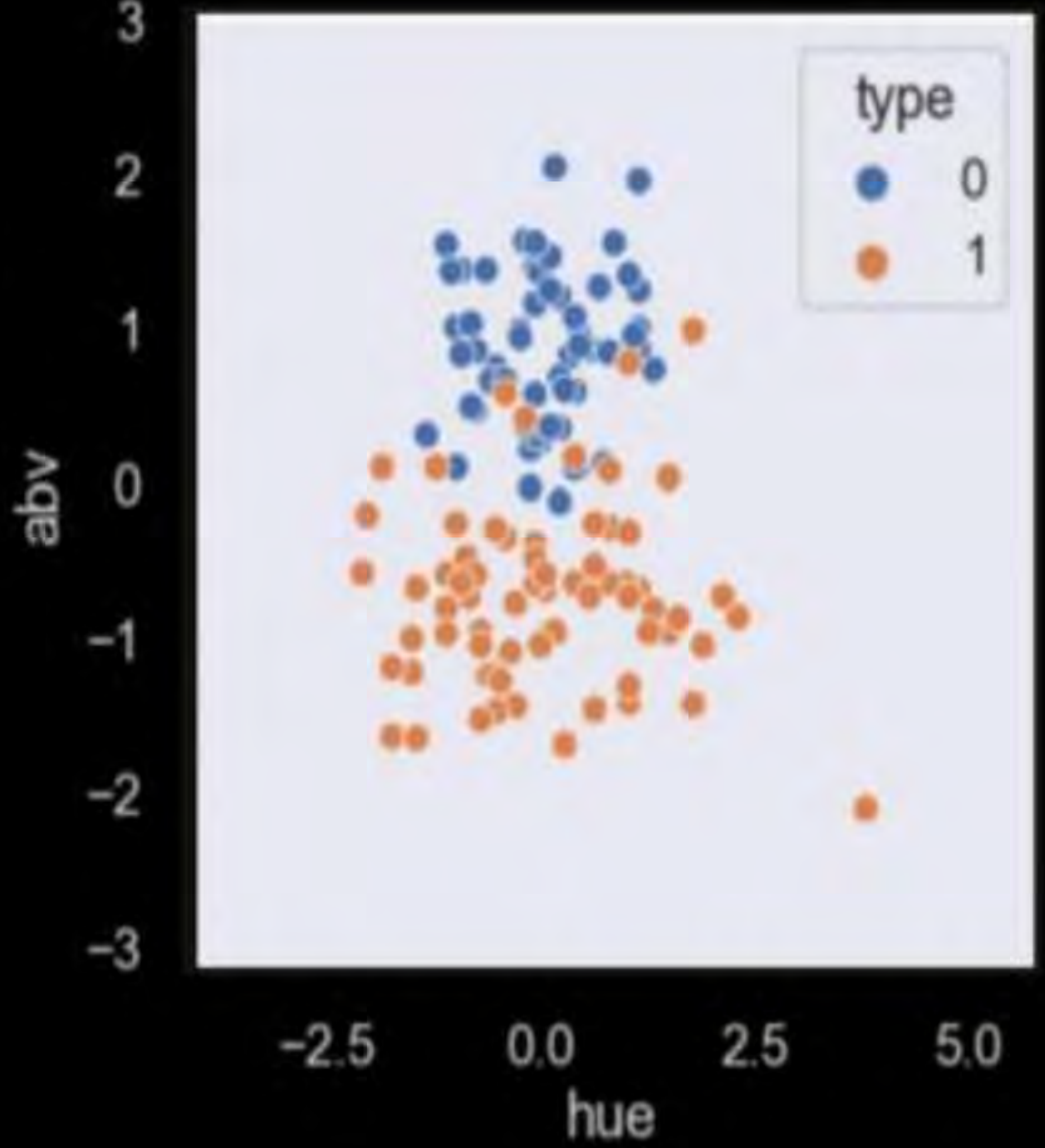
Consider the following scatter plot of our two (standardized) features.

Which of the following statements are true about an unregularized logistic regression model fit on the above data? Select all that apply.

[redacted] Sol

□ After performing logistic regression, the weight for the hue feature will very likely have a negative sign.

□ After performing logistic regression, the weight for the abv feature will very likely have a negative sign.



What is Likelihood.

Example 1: Suppose that X is a discrete random variable with the following probability mass function: where $0 \leq \theta \leq 1$ is a parameter. The following 10 independent observations

X	0	1	2	3
$P(X)$	$2\theta/3$	$\theta/3$	$2(1-\theta)/3$	$(1-\theta)/3$

sol.

were taken from such a distribution: (3,0,2,1,3,2,1,0,2,1). What is the maximum likelihood estimate of θ.

You are walking down Shattuck Ave. when you find a quarter on the ground. You see nothing unusual about this quarter, so you figure it is almost certainly a fair coin, though you realize that manufacturing irregularities in the coin minting process mean that coins are rarely *exactly* fair. You toss the coin 10 times and observe the following outcomes:

H H H H H H H H H T

with H denoting heads and T denoting tails. Assume coin tosses are independent. What is the maximum likelihood estimate of the next toss being heads?

○ $\frac{5}{10}$

○ between $\frac{5}{10}$ and $\frac{9}{10}$

○ $\frac{9}{10}$

○ more than $\frac{9}{10}$

Sol.

There are 5 balls in a bag. Each ball is either red or blue. Let θ (an integer) be the number of blue balls. We want to estimate θ, so we draw 4 balls **with replacement** out of the bag, replacing each one before drawing the next. We get “blue,” “red,” “blue,” and “blue” (in that order).

(a) [5 pts] Assuming θ is fixed, what is the likelihood of getting exactly that sequence of colors (expressed as a function of θ)?

Sol.

Logistic Regression

- Extending the case for more than 2 class

for class 2 : Class 2 points → '1'

$\underbrace{\text{Class 1, 3 points} \to \text{'0'}}_{\text{2 class data}}$

→ logistic regression find β^2

Linear Classification

Logistic Regression

- Extending the case for more than 2 class

for class 3 : Class 3 points → '1'

$\underbrace{\text{Class 1, 2 points} \rightarrow \text{'0'}}_{\text{2 class data}}$

→ logistic regression find β^3

we create 3 classifier

Now we have a test point x_t

we find $x_t\beta^1, x_t\beta^2, x_t\beta^3$

whichever is max decide class of x_t

Linear Classification

Logistic Regression

- What is softmax

Sigmoid Convert distance ($x\beta$) into poobability, 2 class Case.

we have multiclass then distance Ko poobability conversion

Softmax. x_t is test point, $\beta^1, \beta^2, \beta^3$ show 3 Classifier

$$P_{class_1} = \frac{e^{x_t\beta^1}}{e^{x_t\beta^1} + e^{x_t\beta^2} + e^{x_t\beta^3}}$$

$$P_{class_2} = \frac{e^{x_t\beta^2}}{e^{x_t\beta^1} + e^{x_t\beta^2} + e^{x_t\beta^3}}$$

Sigmoid Convert distance $(x\beta)$ into poobability, 2 Class Case.

→ we have multiclan then distance Ko probability Conversion

Softmax. x_t is test point, $\beta^1, \beta^2, \beta^3$ show 3 Classifier

$$P_{class_1} = \frac{e^{x_t\beta^1}}{e^{x_t\beta^1} + e^{x_t\beta^2} + e^{x_t\beta^3}}$$

$$P_{class_2} = \frac{e^{x_t\beta^2}}{e^{x_t\beta^1} + e^{x_t\beta^2} + e^{x_t\beta^3}}$$

$$P_{class_3} = \frac{e^{x_t\beta^3}}{e^{x_t\beta^1} + e^{x_t\beta^2} + e^{x_t\beta^3}}$$

- The max probab decide the predicted clan

Linear Classification

Logistic Regression

- What is softmax loss – categorical cross entropy loss

CE loss for 2 Class Case

$$-\sum_{i=1}^{N} \underline{y_i \log P_i} + \underline{(1-y_i)\log(1-P_i)}$$

original label

$$\boxed{CE = -\sum_{i=1}^{N} \log P_{y_i}}$$

y :	P_i
1	.8
0	.3
1	.6
1	.7
0	.2

CE = - [1. log class 1 hone ki Pred. Probab +
1. log Class $\underline{0}$ " " " " +
1. log Clan $\underline{1}$ " " " " +
1. log " $\underline{1}$ " " " " +
1. log " $\underline{0}$ " " " "]

So $\boxed{CE = -\sum_{i=1}^{N} \log P_{y_i}}$

This formulae Valid for multiclass Case also

datapoint No	y
1	y_1
2	y_1
3	y_3
4	y_3
5	y_2
6	y_2
7	y_2

$$CE = -\left[\log P_{11} + \log P_{21} + \log P_{33} + \log P_{43} + \operatorname{Log} P_{52} + \log P_{62} + \log P_{72}\right]$$

Consider the following image data point, where the model's predicted probabilities for the classes (Dog, Cat, Mountain) are:

3class

Class	Probability
Dog	0.06 ✓
Cat	0.0 ✓
Mountain	0.94 ✓

Assuming that the true class of the image is 'Dog,' what is the cross-entropy loss for this data point?

Single data point

A) [✓ $-\log(0.06)$] ✓.

B) $-\log(0.94)$

C) $-0.6\log(0.6)$

3clan

$CE = -\sum_{i=1}^{N} \log P_{y_i}$

$CE = -\log \cdot 06$

data belong to Clan DOG

Assume our prediction and ground truth for the three classes for i^{th} point is:

Probab of 3 classes.

$$\hat{y}_i = \begin{bmatrix}0.1\\0.8\\0.1\end{bmatrix} = \begin{bmatrix}\hat{y}_i^1\\\hat{y}_i^2\\\hat{y}_i^3\end{bmatrix}$$

Original label

$$y_i = \begin{bmatrix}0\\1\\0\end{bmatrix} = \begin{bmatrix}y_i^1\\y_i^2\\y_i^3\end{bmatrix}$$

CE =

$-\log(P_{y_i})$

$\boxed{-\log(\cdot 8)}$

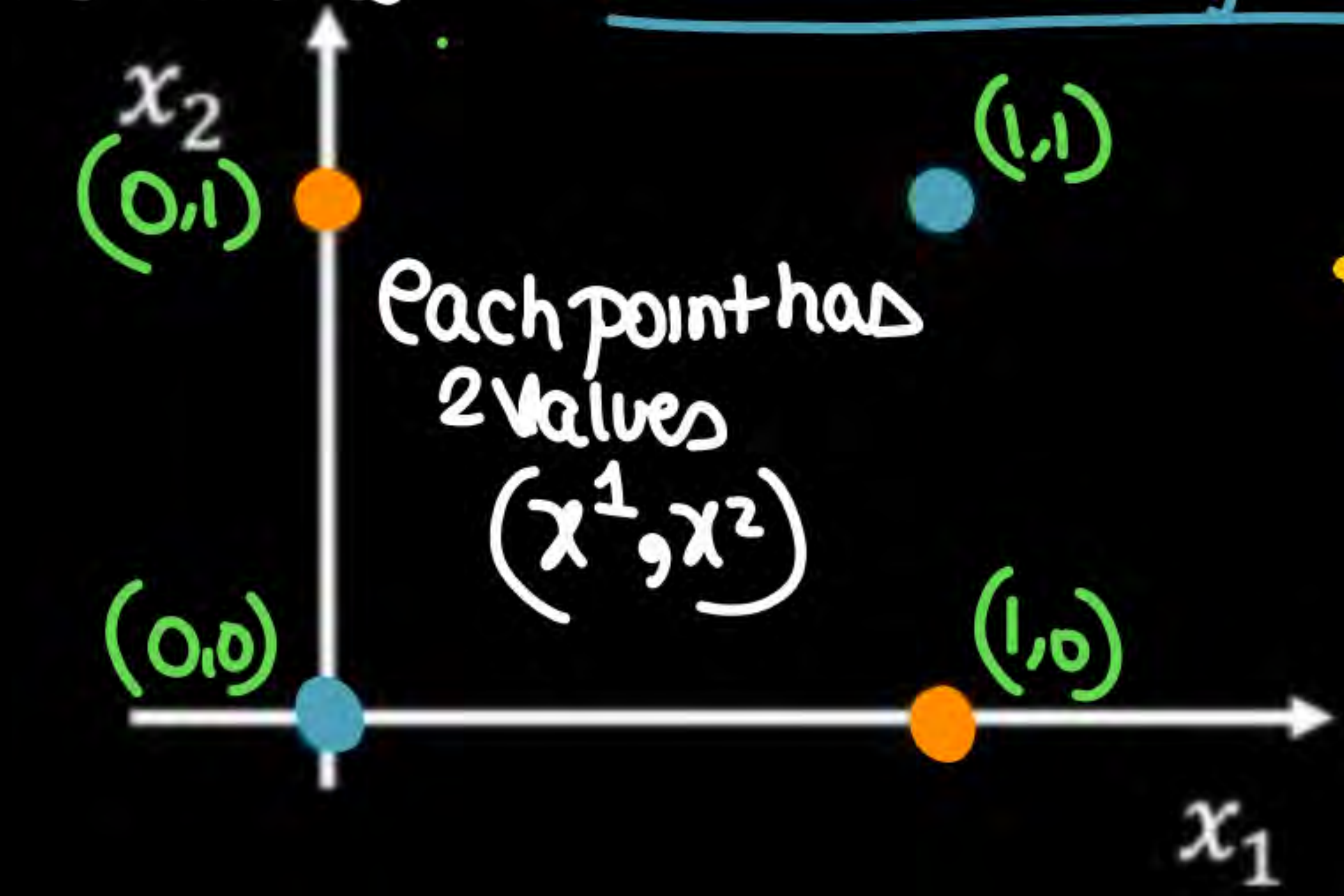
Linear Classification

Logistic Regression

How to classify this ?

2D data

Not linearly classifiable ⇒ logictic Reg and linear Classification cannot work on data



- But we want to use logistic regression only

$(0,1) \rightarrow (1,0,1)$ $(1,1) \rightarrow (2,1,1)$

$(1,0) \rightarrow (1,1,0)$ $(0,0) \rightarrow (0,0,0)$

2D data

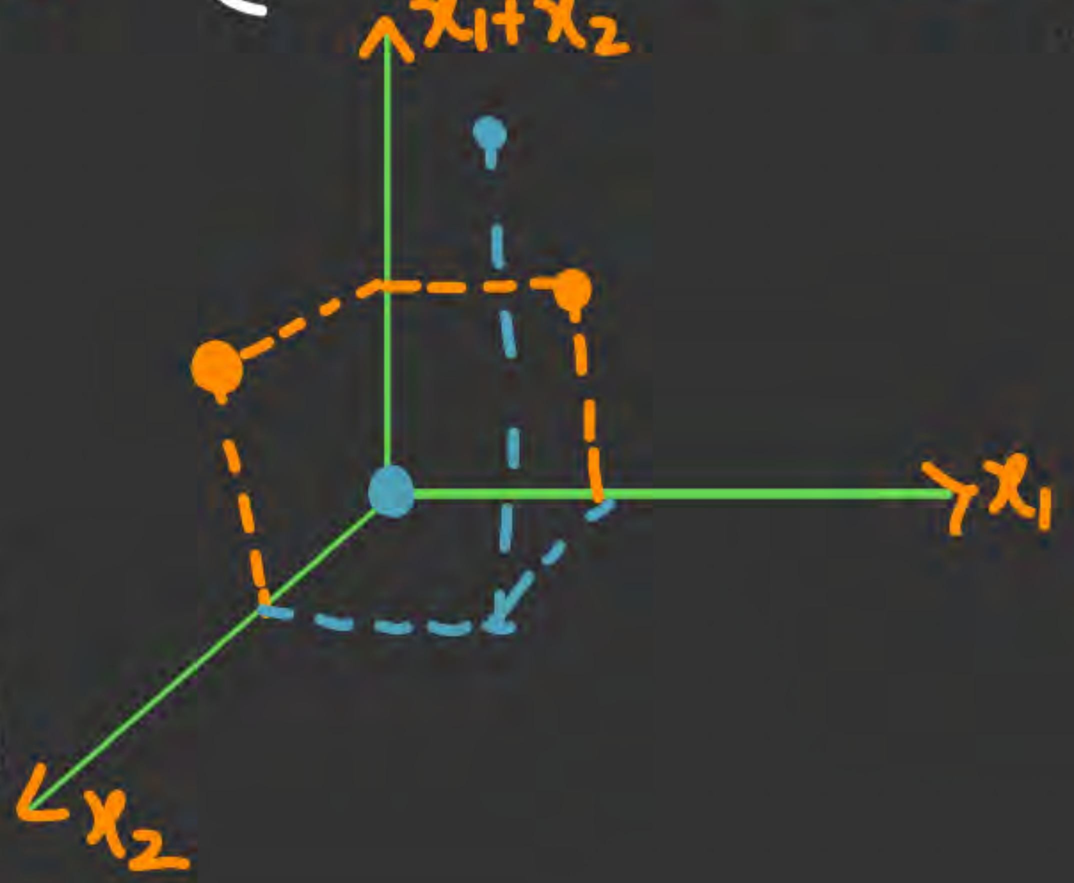
$$\underbrace{(x_1, x_2)}$$

each data has 2 values

$$\xrightarrow[\text{transform}]{\text{data}}$$

3D. space

$$\left[\, \underbrace{x_1 + x_2,\ x^1,\ x^2} \,\right]$$



5D

$$\left[\, x_1^2,\ x_2^2,\ x_1 x_2,\ x_1^3,\ x_2^3 \,\right]$$

2D ⟶ 5D

If after transformation we get linearly separable data ⇒ Now logistic regression is possible on this new (new/old) of 5D (2D/5D)

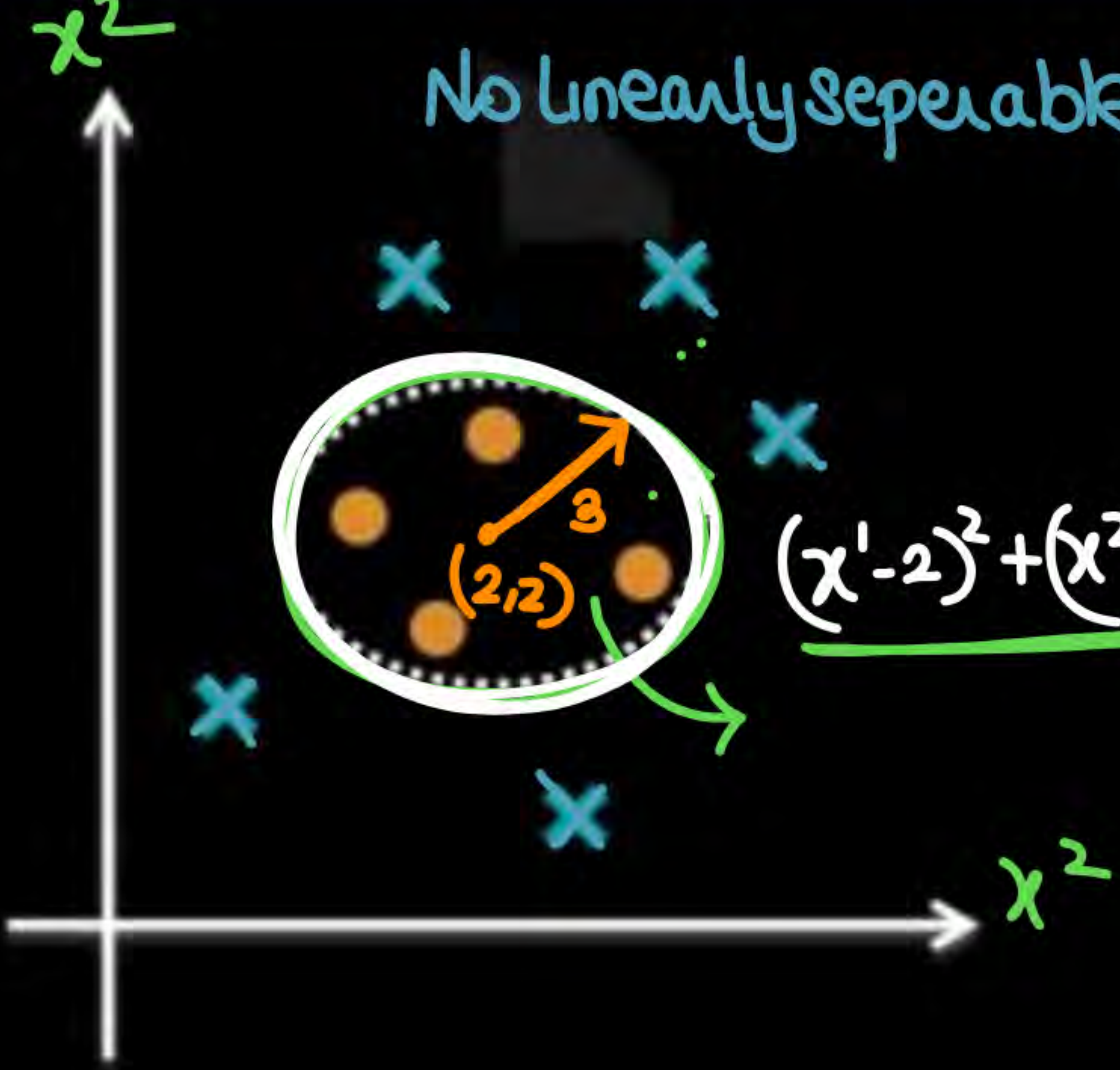
And in

Linear Classification

Logistic Regression

In this case, @ higher dimension 3D, Classifier is linear/hyperlane, But in 2D lower dimension classifier is Nonlinear.

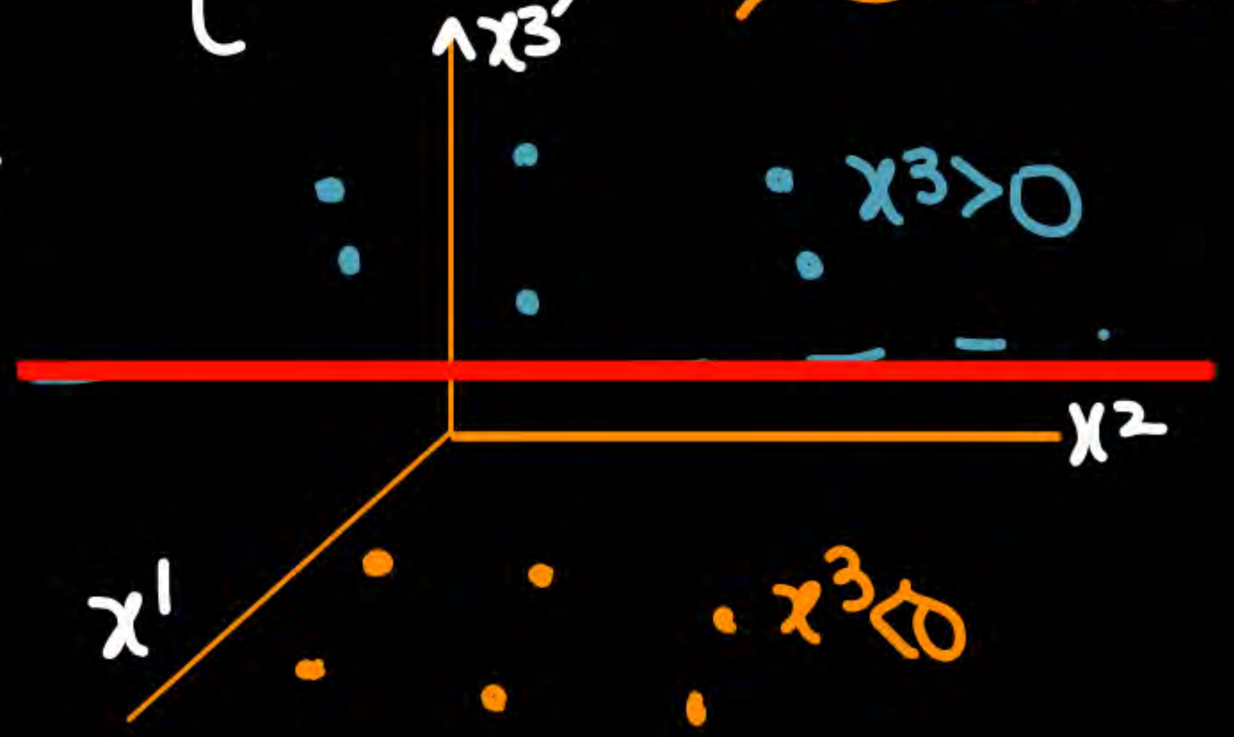
No lineanly seperable



$$(x^1-2)^2+(x^2-2)^2=3^2$$

$$2D \longrightarrow 3D$$

$$\left[x^1, x^2, (x^1-2)^2+(x^2-2)^2-9\right]$$



But $x^3=0$ here is

$$(x_2^2-2)^2+(x_1^2-2)^2-9=0.$$

x^2 Jo ki Circle hai

No linearly seperable

Classification Rule

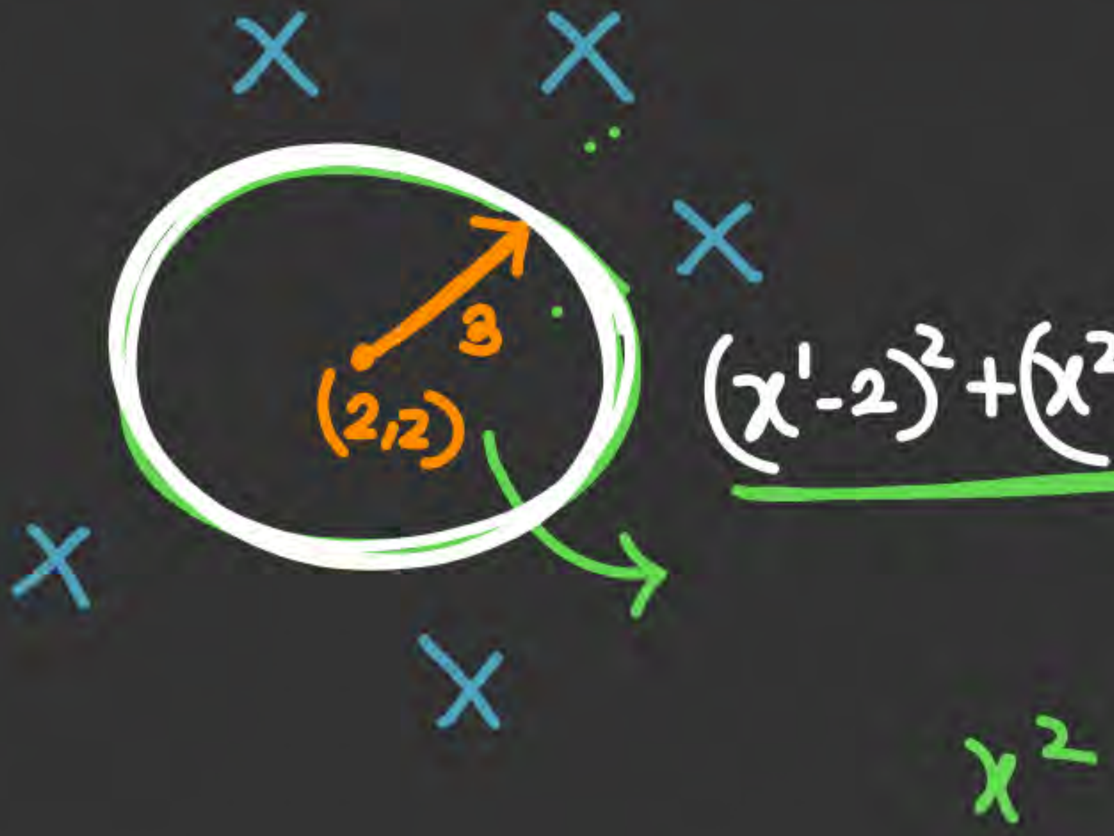
Classification Plane

$\underline{x^3=0 \text{ Plane}}$ $x^3>0 \Rightarrow$ Class 1

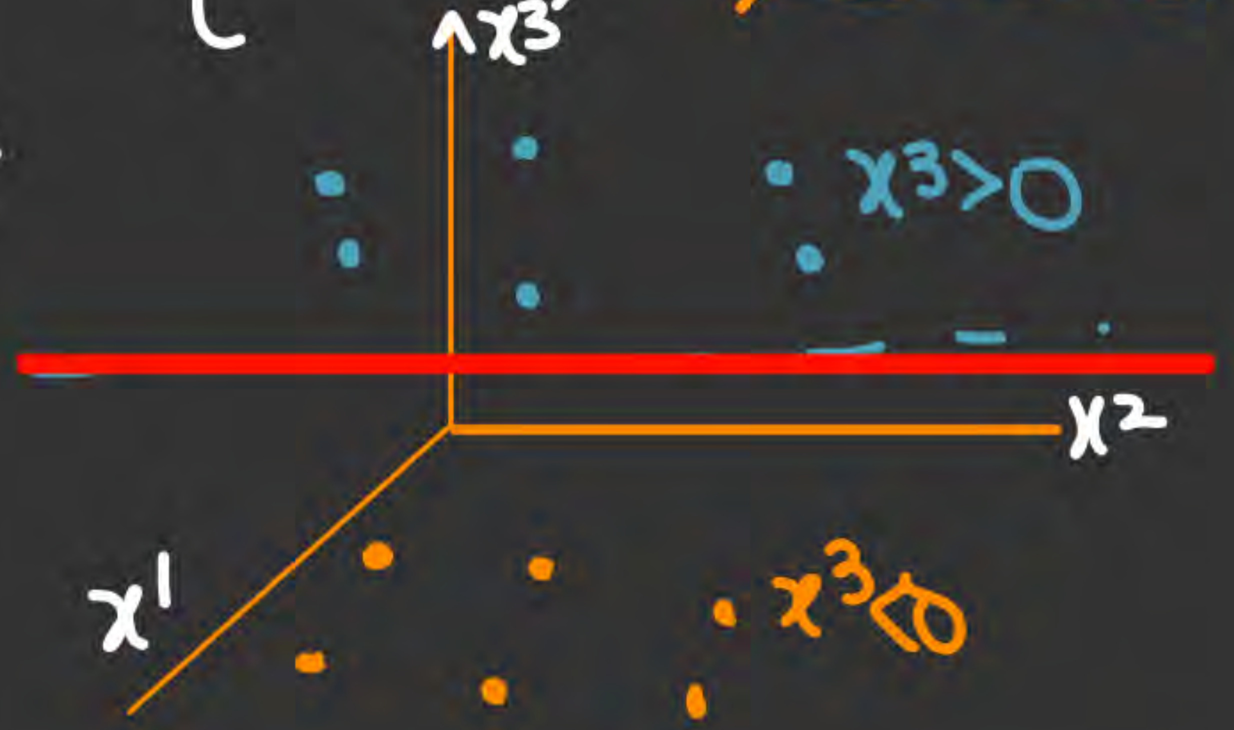
$x^3<0 \Rightarrow$ class 0

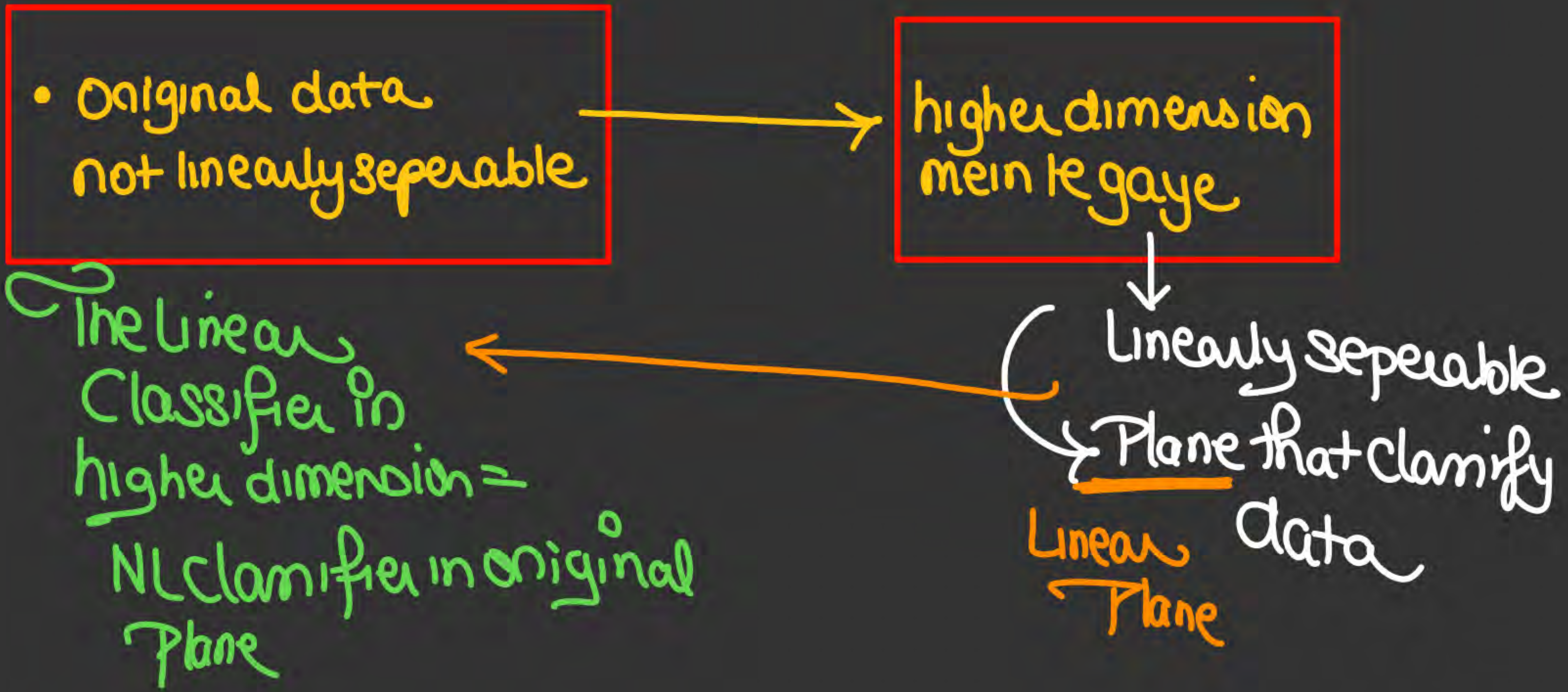
2D → 3D

$$\left[x^1, x^2, \underline{(x^1-2)^2+(x^2-2)^2-9}\right]$$



$$\underline{(x^1-2)^2+(x^2-2)^2}=3^2$$





- Original data not linearly seperable

higher dimension mein le gaye

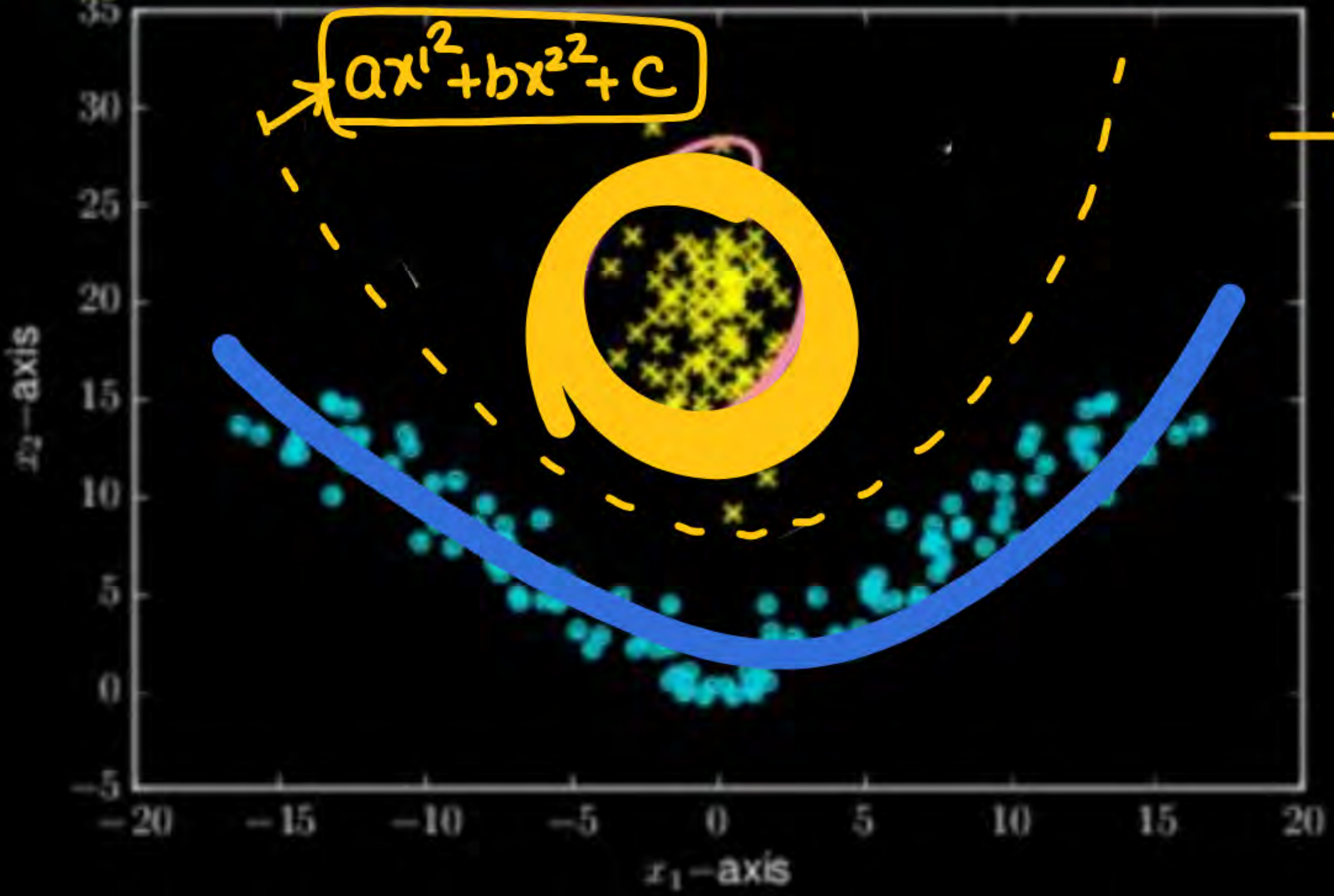
Linearly seperable Plane that classify data

Linear Plane

The Linear Classifier in higher dimension = NL classifier in original Plane

Linear Classification

Logistic Regression



$\boxed{ax_1^2+bx_2^2+c}$

→ 3D में classify linearly

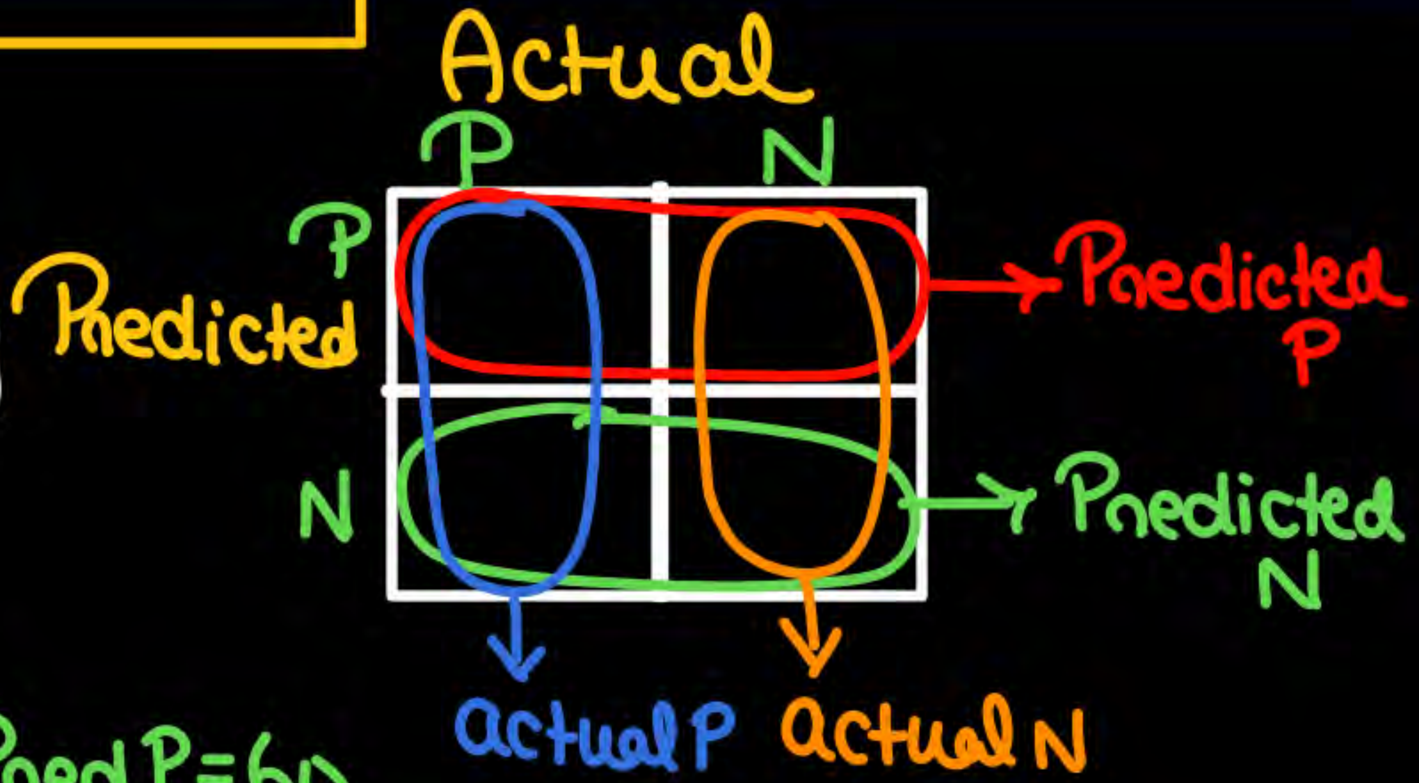
Linear Classification

Confusion Matrix

Confusion Matrix

To measure accuracy of classifier

		Actual P	Actual N	
Predicted	P	50	10	→ Pred P=60
	N	20	80	→ Pred N= 100
		Actual P=70	actual N=90	



Predicted \ Actual	P	N	
P	50	10	→ Pred P=60
N	20	80	→ Pred N= 100
	Actual P=70	actual N=90	

50 → actual P, Predict P, Happy → True Positive

10 → actual N, Predict P, False Positive

20 → actual P, Predict N, False N

80 → actual N, Predict N, True Negative

True Positive Rate

- $TPR :- \frac{TP}{\text{Total actual } P}$

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

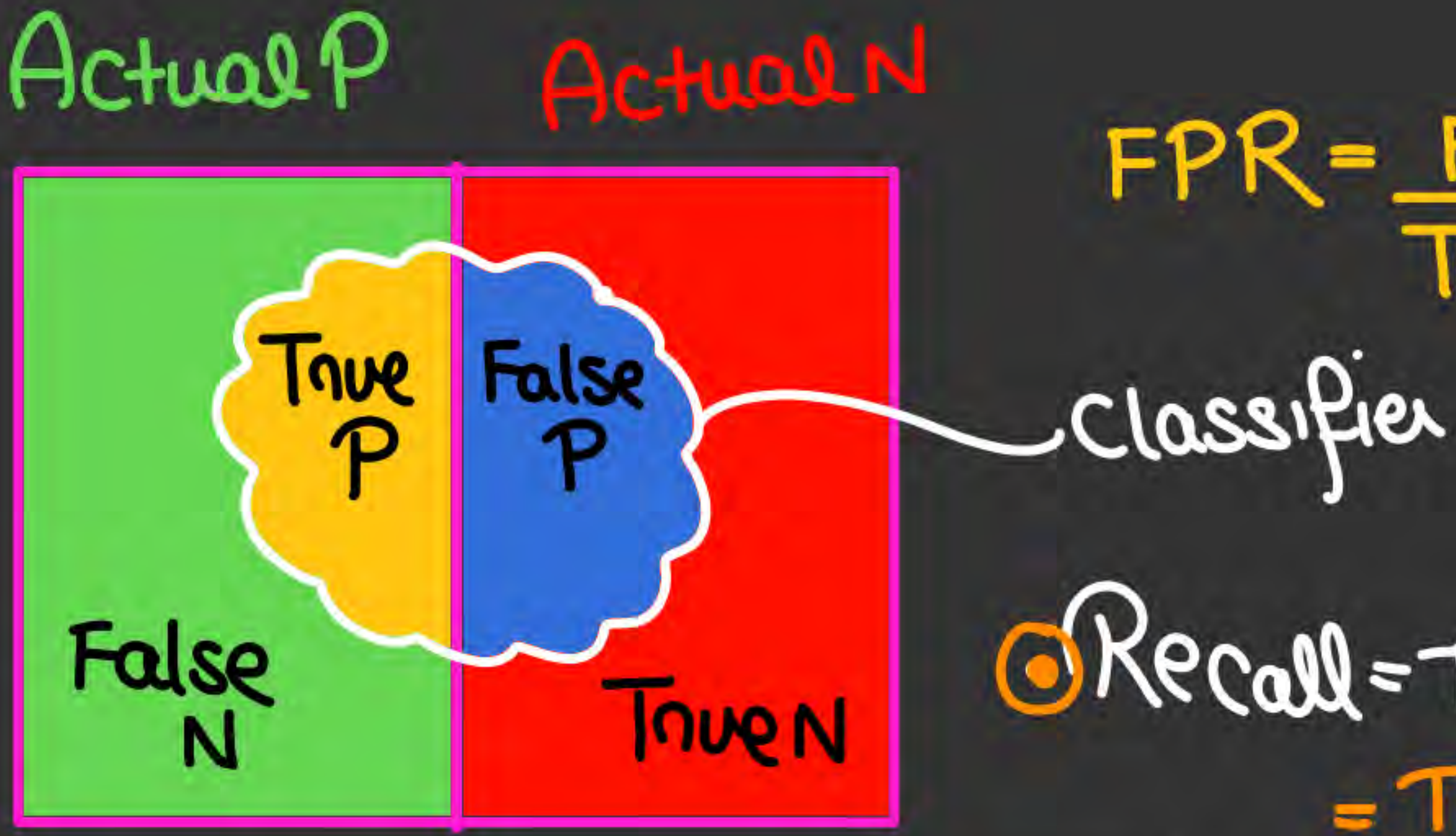
- TNR ⇒ True negative Rate

$$TNR = TN / \text{Total actual } N$$

$$= \frac{TN}{FP+TN}$$

- FPR: False Positive Rate
- $FPR = \frac{FP}{\text{Total actual } N}$
- $FPR = 1-TNR$

True P



$$FPR = \frac{FP}{\text{Total Actual N.}}$$

$$\text{Recall} = TPR = \frac{TP}{\text{Total actual P}} = \frac{TP}{TP+FN}$$

$$\text{Precision} = \frac{TP}{\text{Total predicted P}} = \frac{TP}{TP+FP}$$

Confusion Matrix

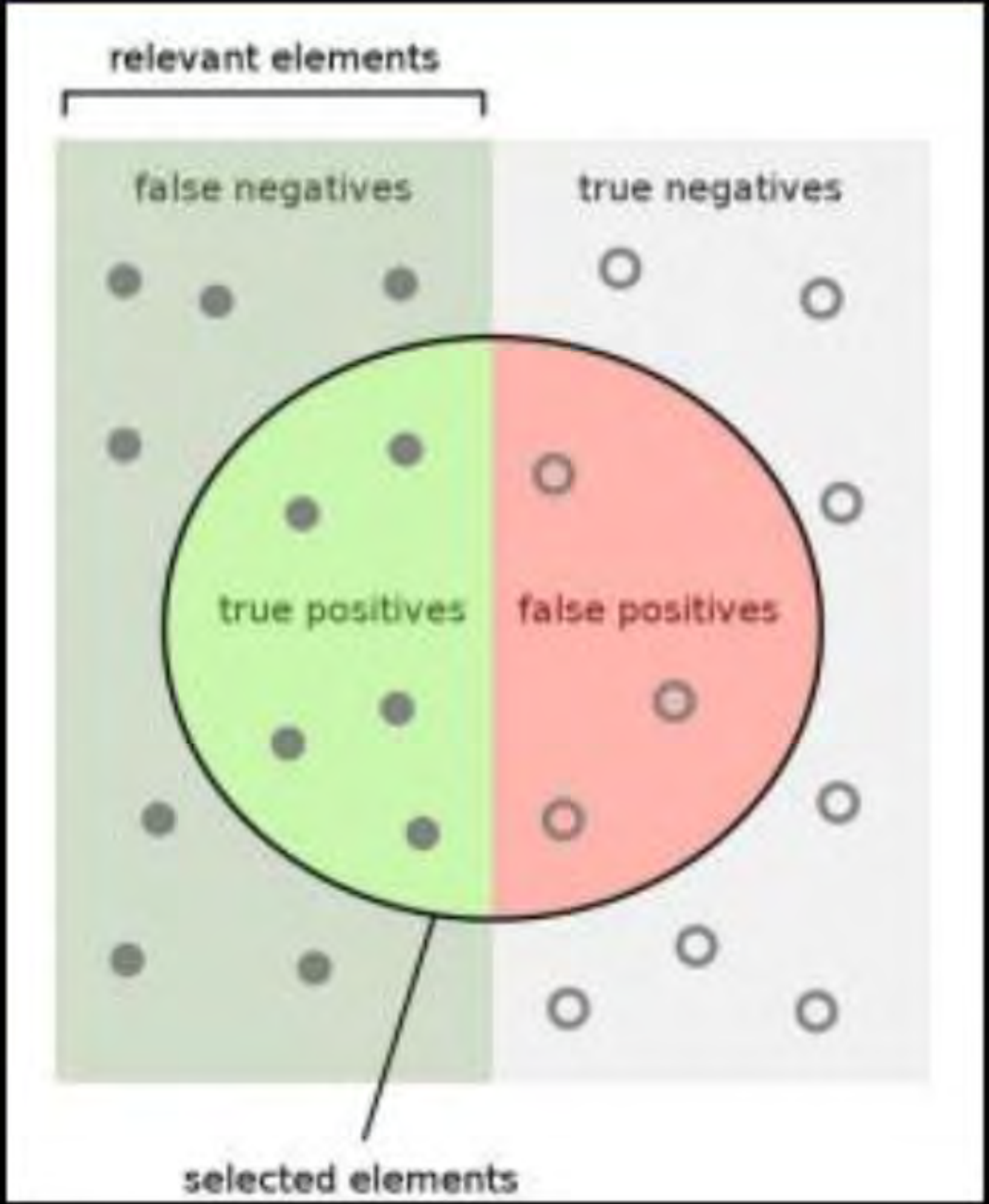
- Recall
- Precision

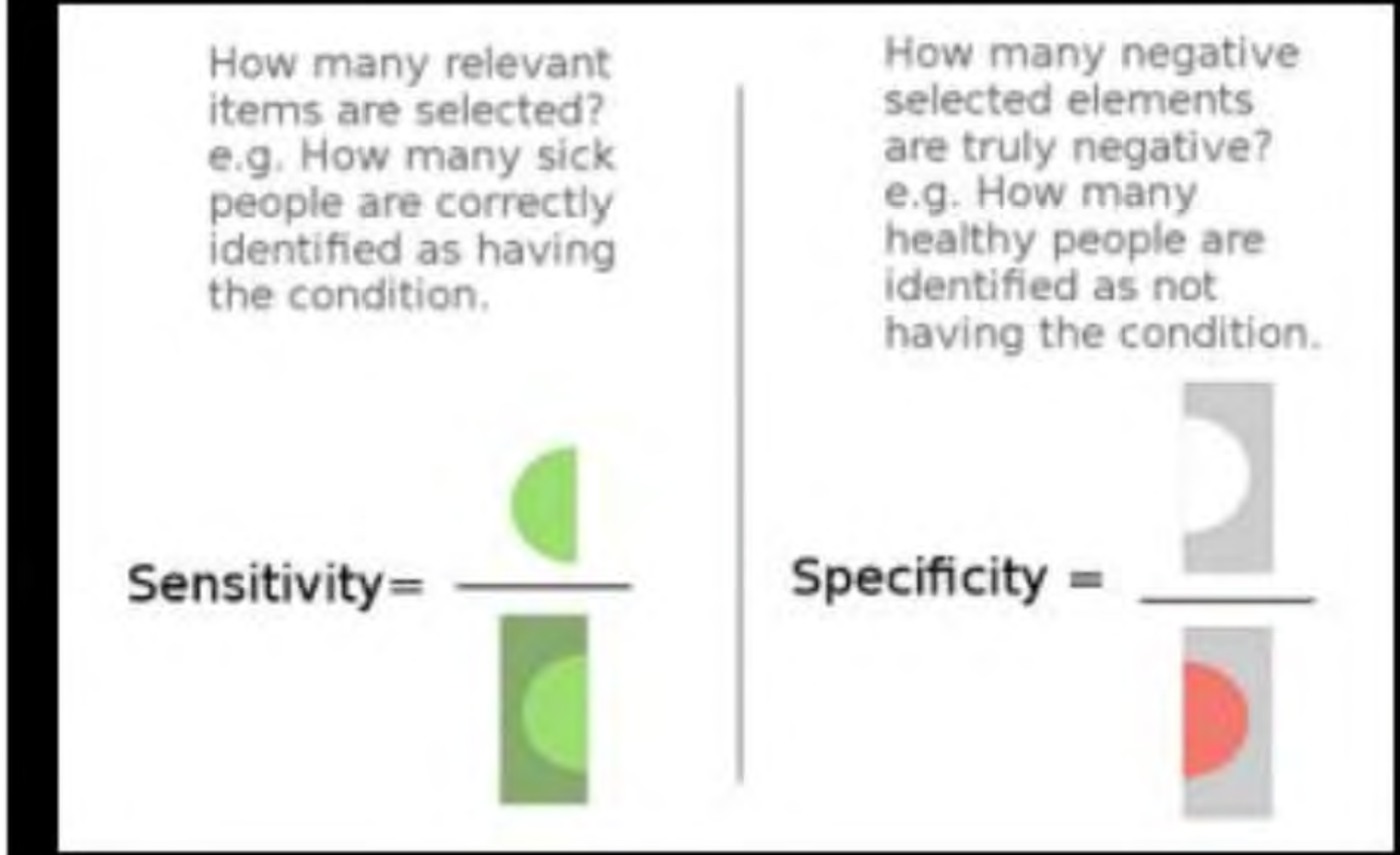
Linear Classification

What is ROC curve (receiver operating characteristic curve)

- A receiver operating characteristic curve, or ROC curve, is a graphical plot that illustrates the performance of a binary classifier model (can be used for multi class classification as well) at varying threshold values.
- The ROC curve is the plot of the true positive rate (TPR) against the false positive rate (FPR) at each threshold setting

What is ROC curve (receiver operating characteristic curve)




How many relevant items are selected? e.g. How many sick people are correctly identified as having the condition.

Sensitivity=

How many negative selected elements are truly negative? e.g. How many healthy people are identified as not having the condition.

Specificity =

Linear Classification

What is ROC curve (receiver operating characteristic curve)

- **Sensitivity** is a measure of how well a test can identify true positives ✓ TPR
- **Specificity** is a measure of how well a test can identify true negatives: TNR

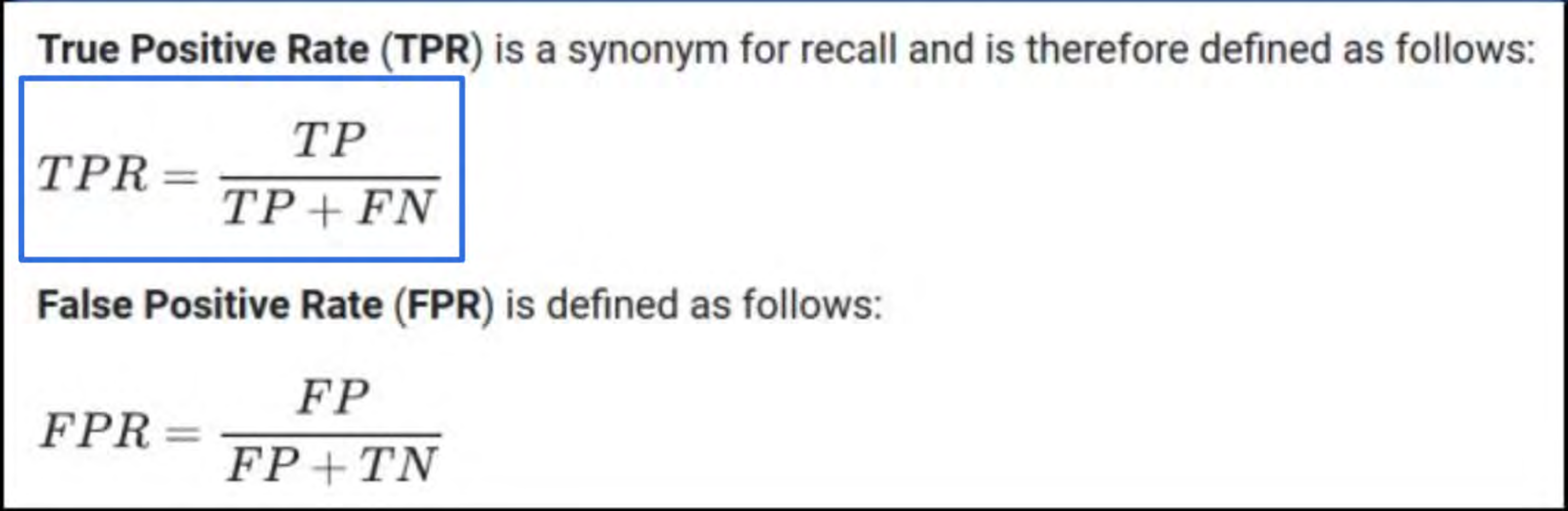
$$\overset{\nearrow \text{TPR}}{\boxed{\text{sensitivity}}} = \frac{\text{number of true positives}}{\text{number of true positives} + \text{number of false negatives}}$$

$$\overset{\nearrow \text{TNR}}{\boxed{\text{specificity}}} = \frac{\text{number of true negatives}}{\text{number of true negatives} + \text{number of false positives}}$$

Linear Classification

What is ROC curve (receiver operating characteristic curve)

- **What is TPR and FPR ?**



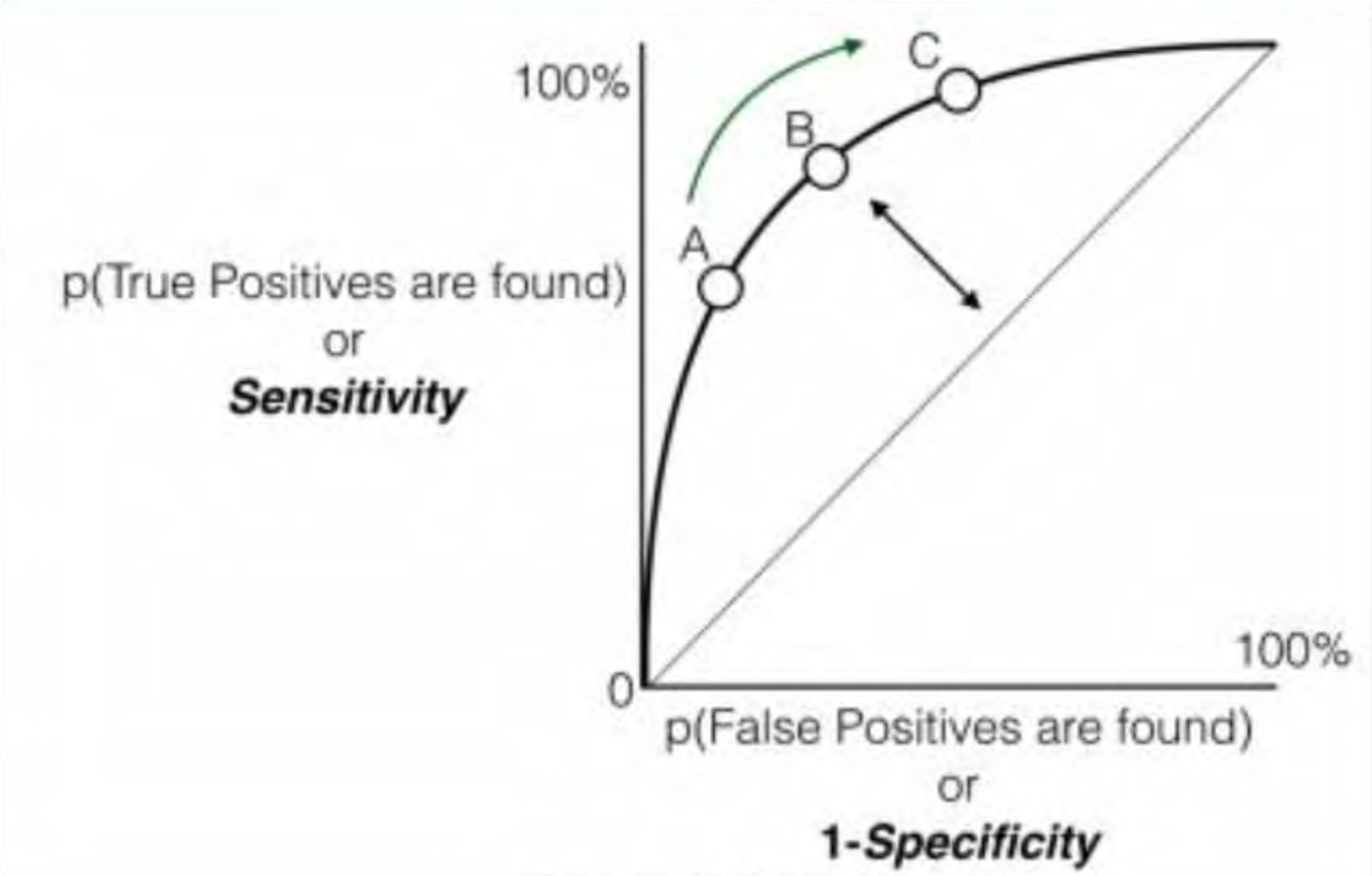
True Positive Rate (**TPR**) is a synonym for recall and is therefore defined as follows:

$$\boxed{TPR = \frac{TP}{TP + FN}}$$

False Positive Rate (**FPR**) is defined as follows:

$$FPR = \frac{FP}{FP + TN}$$

What is ROC curve (receiver operating characteristic curve) an example



Sensitivity versus False Positive Rate plot

What is AUC (Area under the curve)

- AUC stands for the Area Under the Curve, and the AUC curve represents the area under the ROC curve.
- It measures the overall performance of the binary classification model.
- The area will always lie between 0 and 1,
- A greater value of AUC denotes better model performance.
- Our main goal is to maximize this area in order to have the highest TPR and lowest FPR at the given threshold.
- The AUC measures the probability that the model will assign a randomly chosen positive instance a higher predicted probability compared to a randomly chosen negative instance.

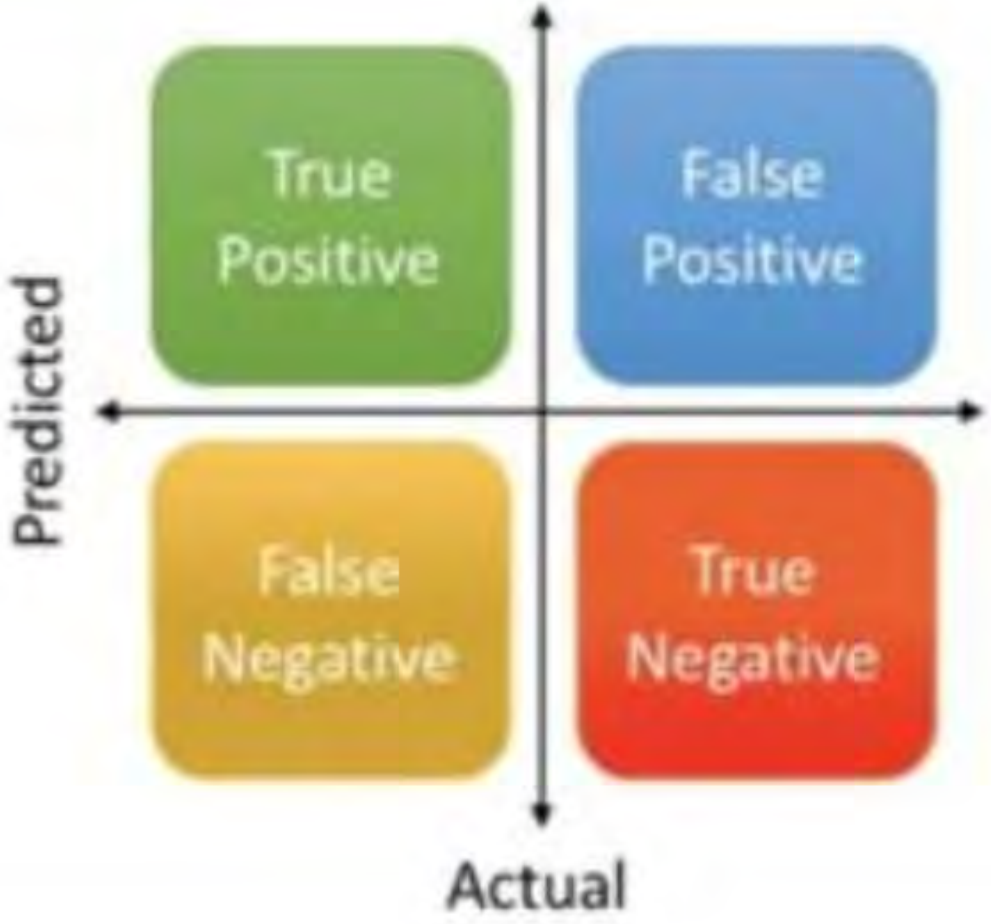
Linear Classification

What is Recall and Precision

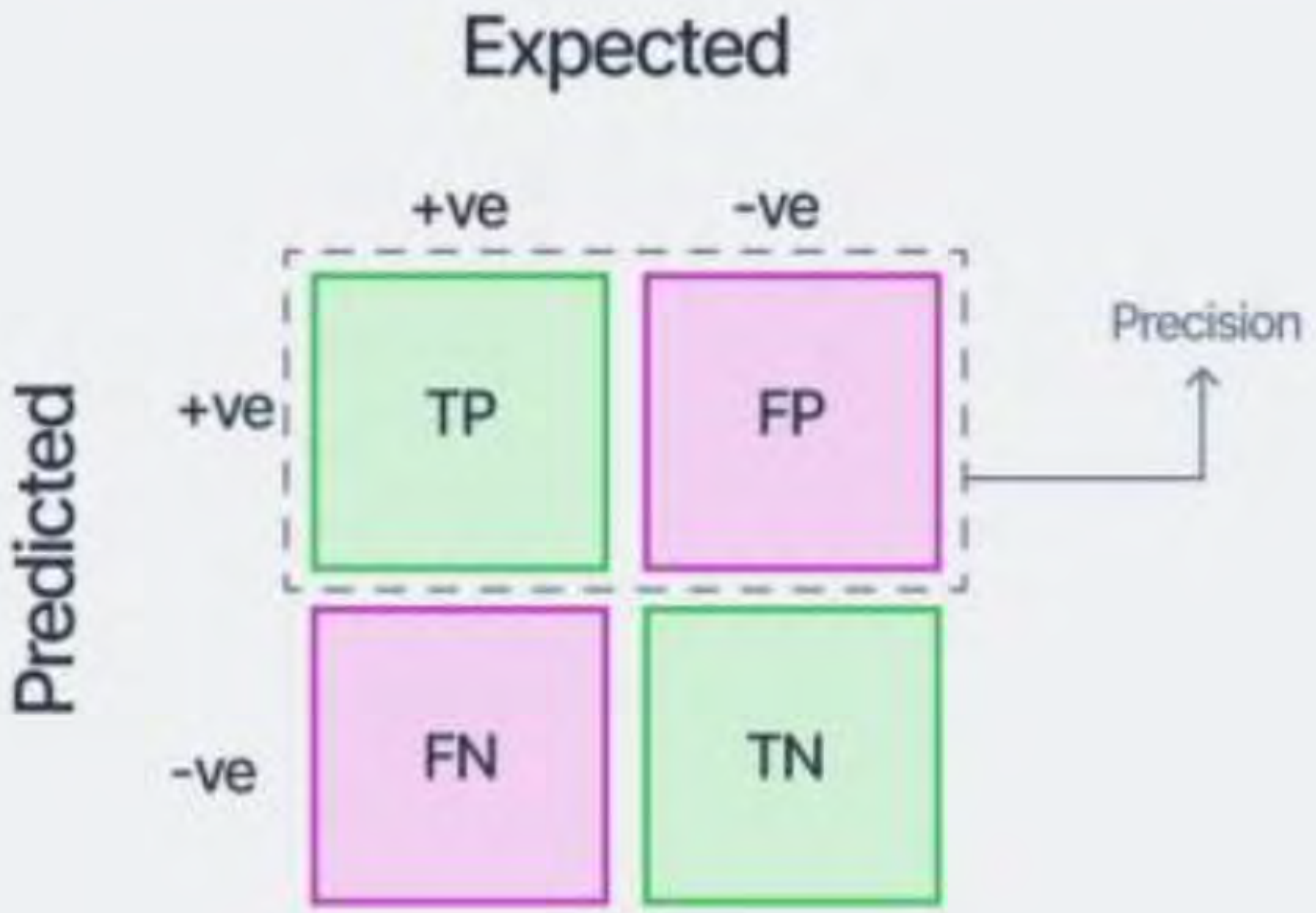
$$\textbf{Precision} = \frac{\text{True Positive}}{\text{Actual Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Positive}}$$

$$\textbf{Recall} = \frac{\text{True Positive}}{\text{Predicted Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Negative}}$$

$$\textbf{Accuracy} = \frac{\text{True Positive + True Negative}}{\text{Total}}$$

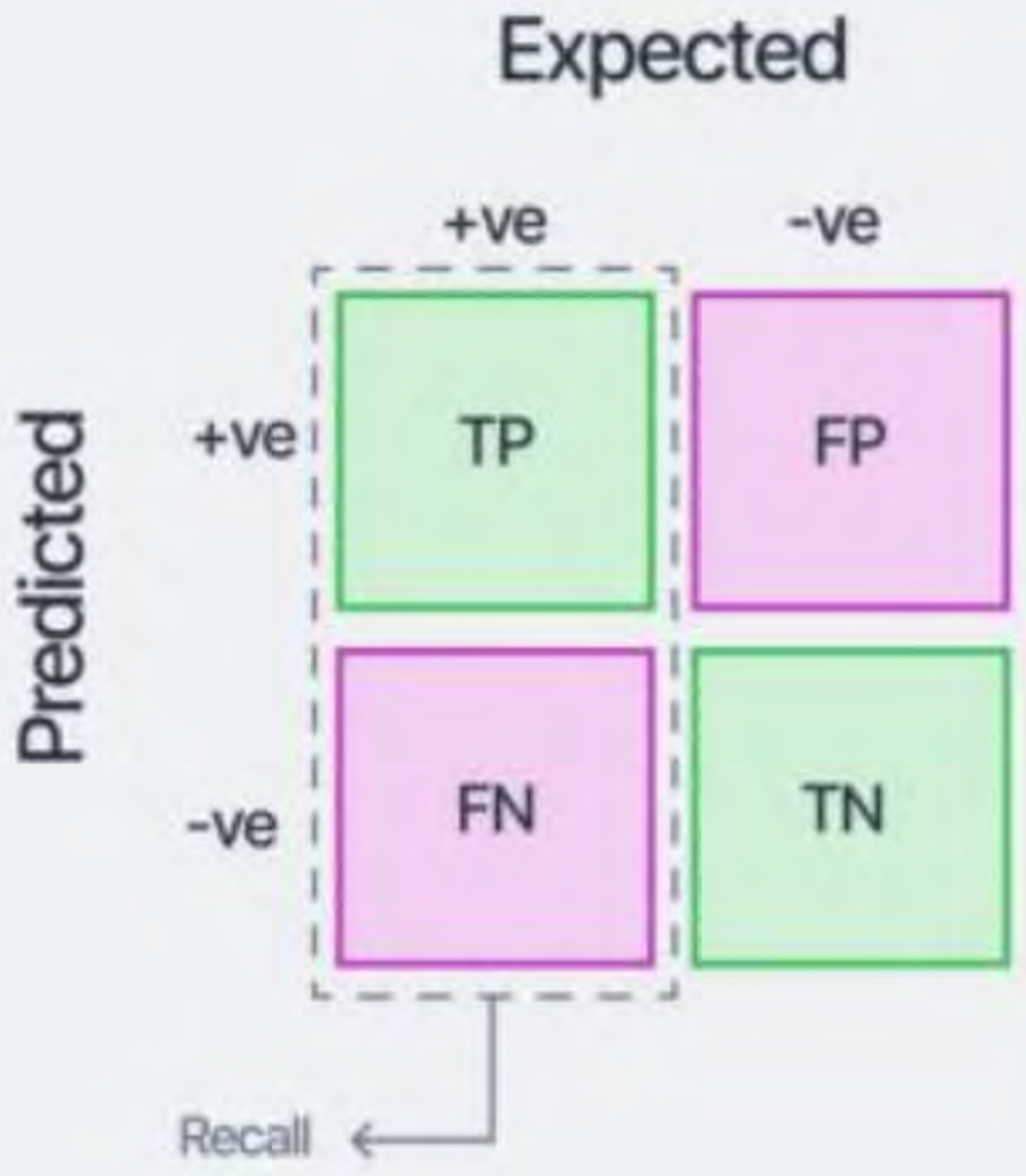


What is Recall and Precision



		Expected	
		+ve	-ve
Predicted	+ve	TP	FP
	-ve	FN	TN

Precision



		Expected	
		+ve	-ve
Predicted	+ve	TP	FP
	-ve	FN	TN

Recall

What is Recall and Precision

Both precision and recall may be useful in cases where there is imbalanced data.

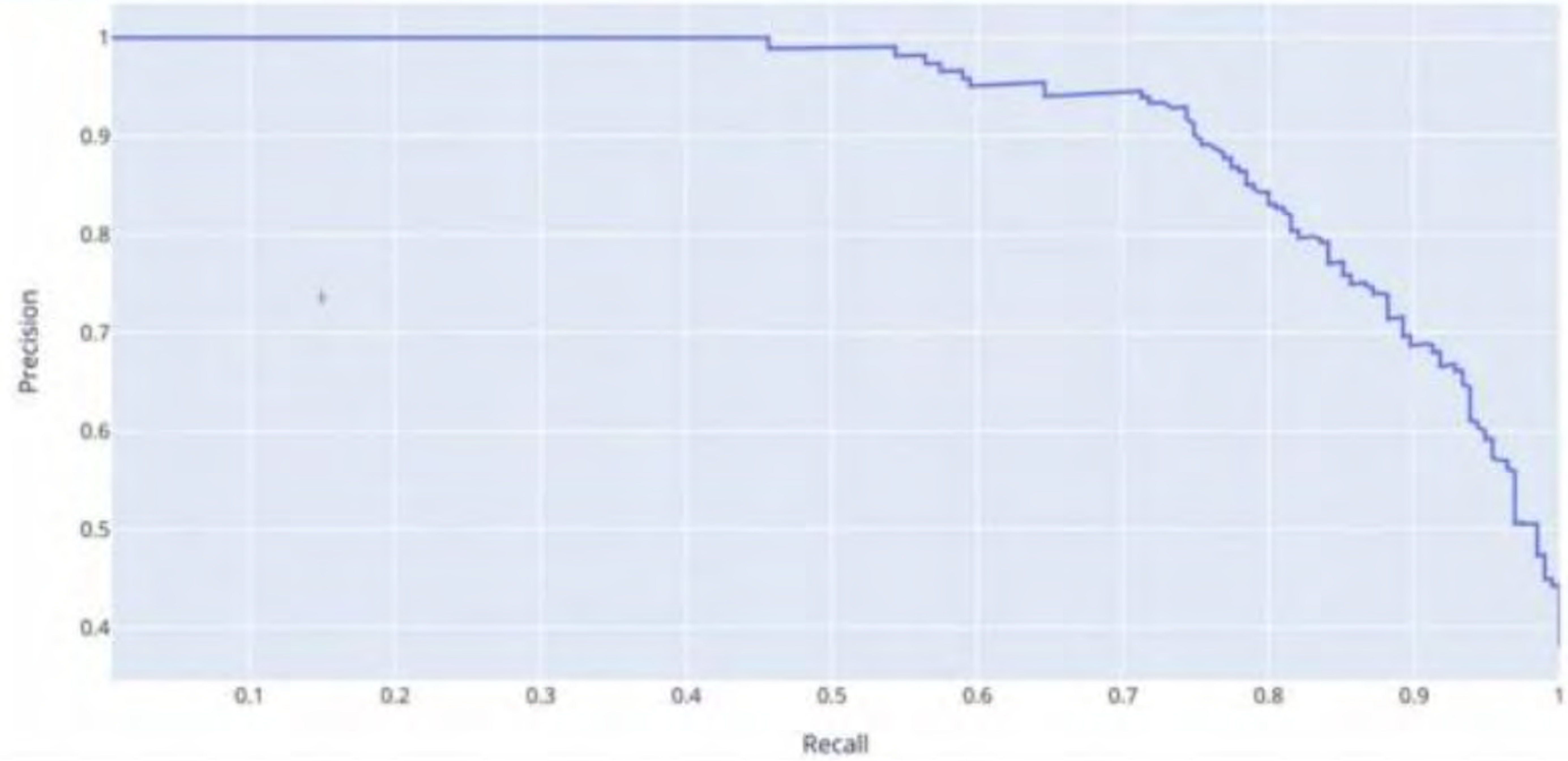
It may be valuable to prioritize one over the other in cases where the outcome of a false positive or false negative is costly.

For example, in medical diagnosis, a false positive test can lead to unnecessary treatment and expenses.

In this situation, it is useful to value precision over recall. In other cases, the cost of a false negative is high.

For instance, the cost of a false negative in fraud detection is high, as failing to detect a fraudulent transaction can result in significant financial loss.

What is Recall and Precision



What is F-1 Score

In most problems, you could either give a higher priority to maximizing precision, or recall, depending upon the problem you are trying to solve. But in general, there is a simpler metric which takes into account both precision and recall, and therefore, you can aim to maximize this number to make your model better. This metric is known as F1-score, which is simply the harmonic mean of precision and recall.

$$\text{F1 Score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

Linear Classification

Practise

The confusion matrix visualizes the ____ of a classifier by comparing the actual and predicted classes.

○ Accuracy

○ Stability

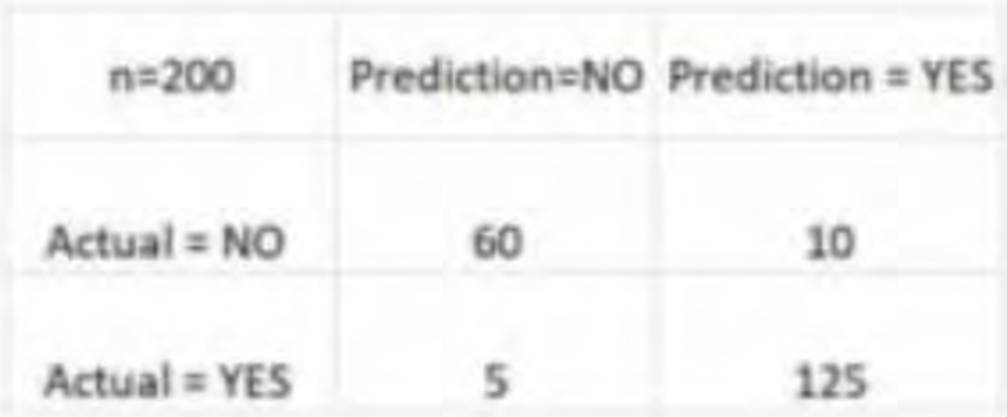
○ Connectivity

○ Comparativity

Linear Classification

Practise

From the above Table



n=200	Prediction=NO	Prediction = YES
Actual = NO	60	10
Actual = YES	5	125

○ In reality, there are totally 135 accounts who have a balance more than $1000 and 70 accounts with balance - less than $1000

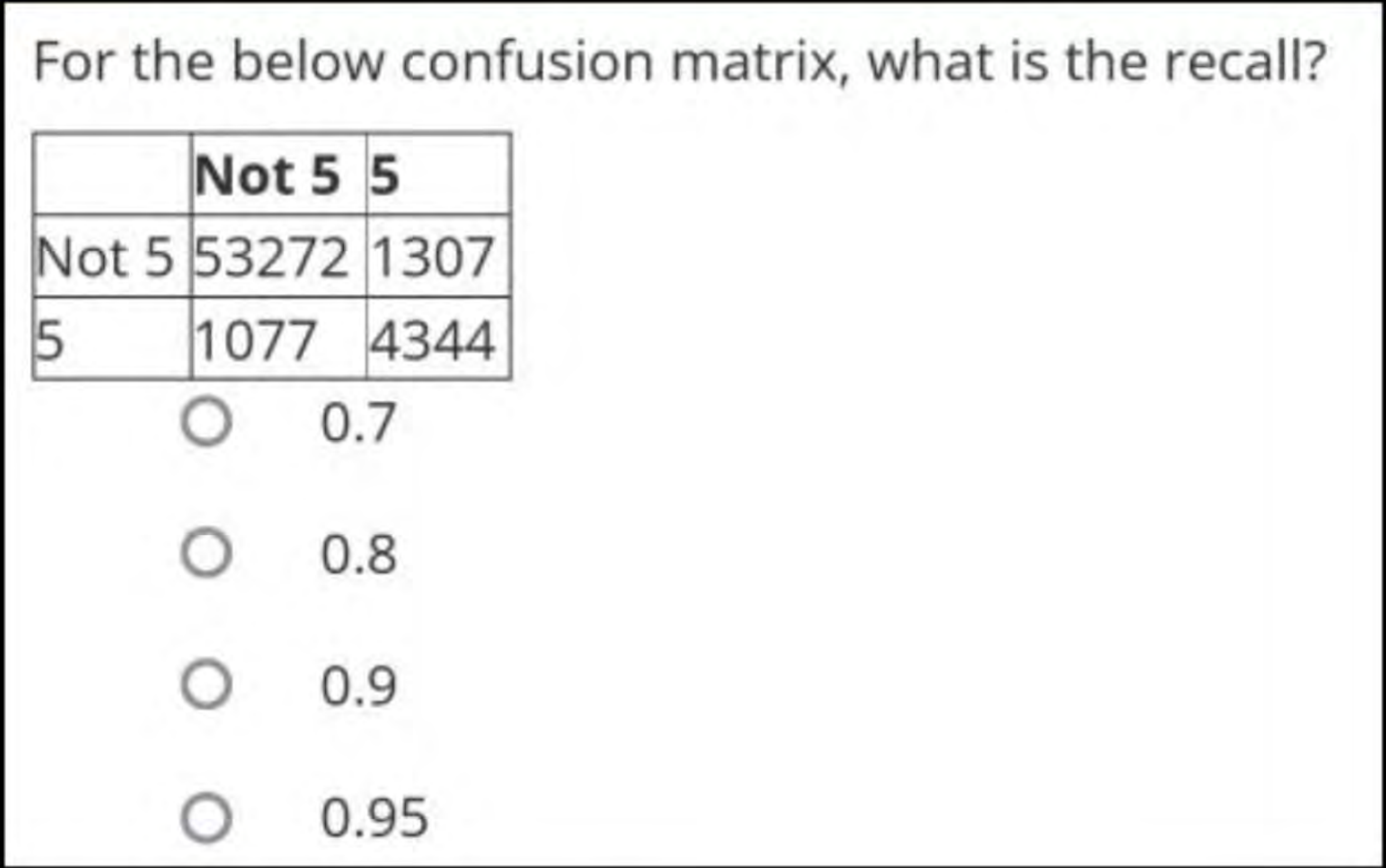
○ In reality, there are totally 60 accounts who have a balance more than $1000 and 70 accounts with balance - less than $1000

○ In reality, there are totally 125 accounts who have a balance more than $1000 and 10 accounts with balance - less than $1000

○ In reality, there are totally 130 accounts who have a balance more than $1000 and 70 accounts with balance - less than $1000

Linear Classification

Practise



For the below confusion matrix, what is the recall?

	Not 5	**5**
Not 5	53272	1307
5	1077	4344

○ 0.7

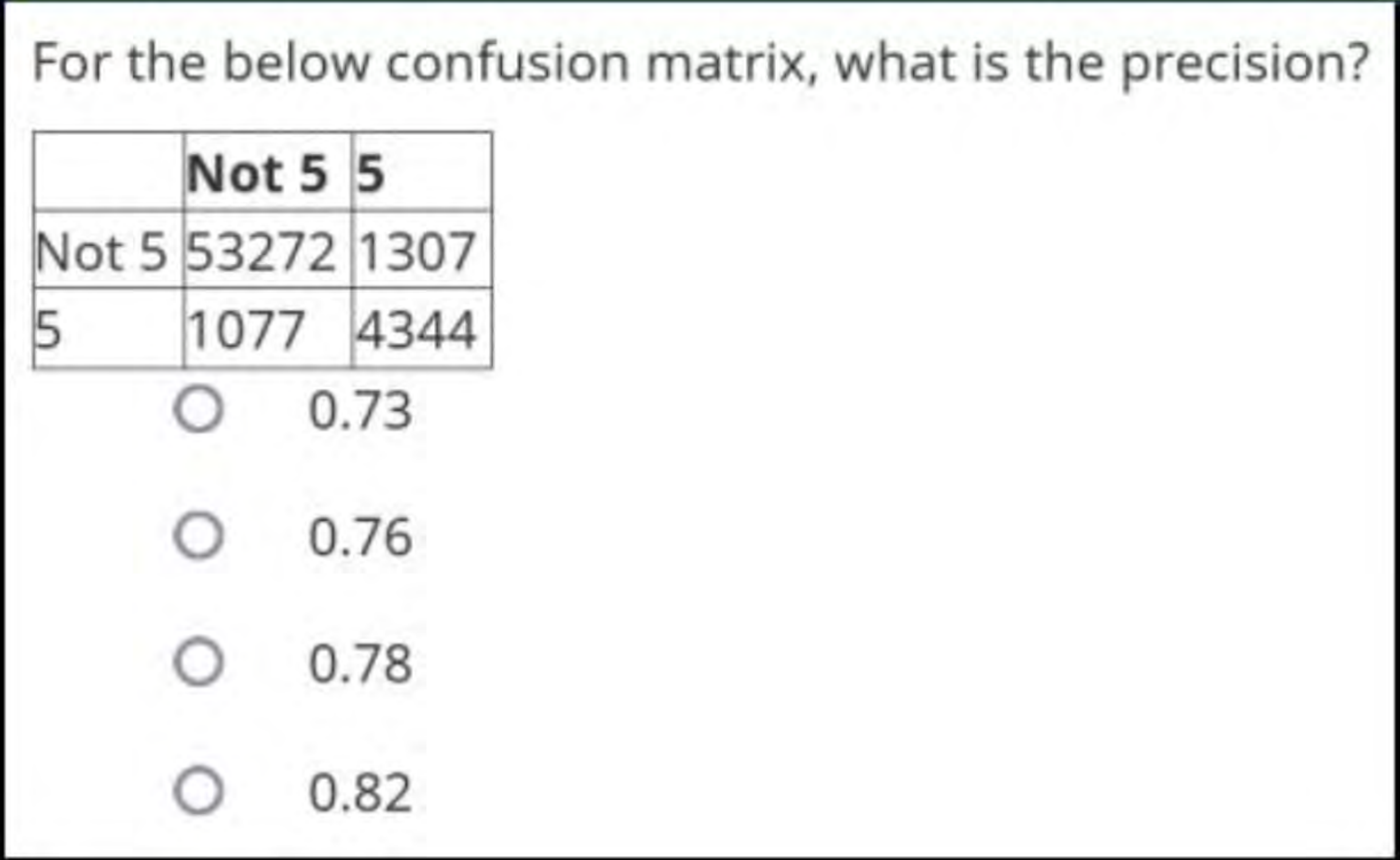
○ 0.8

○ 0.9

○ 0.95

Linear Classification

Practise



For the below confusion matrix, what is the precision?

	Not 5	**5**
Not 5	53272	1307
5	1077	4344

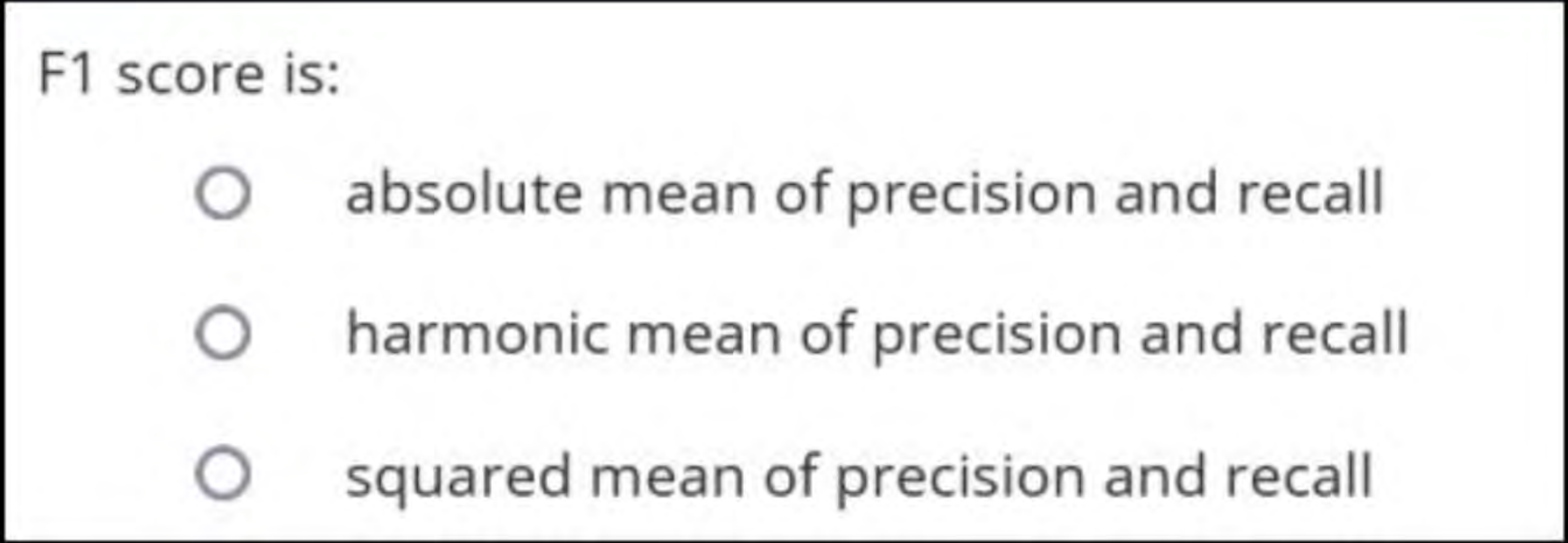
○ 0.73

○ 0.76

○ 0.78

○ 0.82

What is F-1 Score



F1 score is:

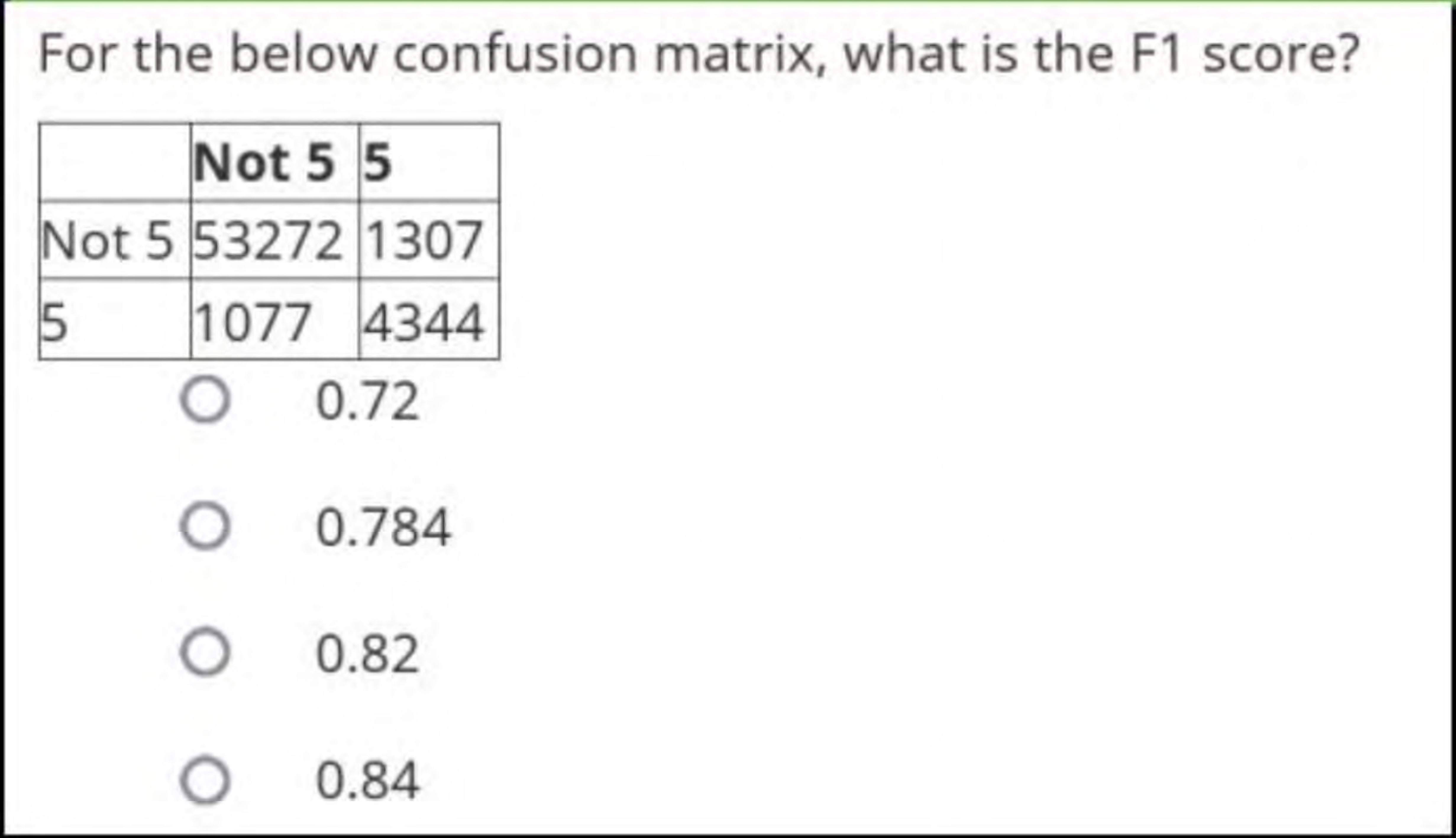
○ absolute mean of precision and recall

○ harmonic mean of precision and recall

○ squared mean of precision and recall

Linear Classification

What is F-1 Score



For the below confusion matrix, what is the F1 score?

	Not 5	**5**
Not 5	53272	1307
5	1077	4344

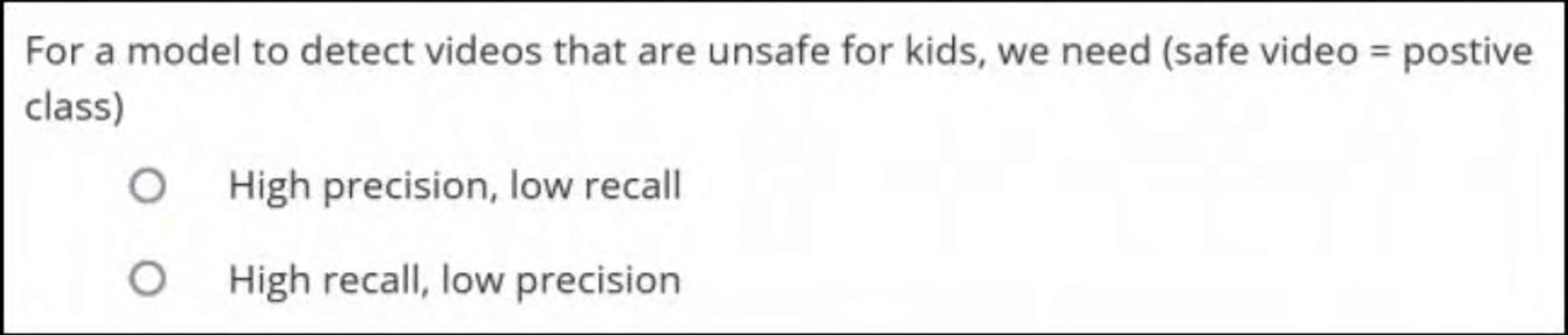
○ 0.72

○ 0.784

○ 0.82

○ 0.84

What is F-1 Score



For a model to detect videos that are unsafe for kids, we need (safe video = postive class)

○ High precision, low recall

○ High recall, low precision

THANK - YOU