Data Science and Artificial Intelligence

Machine Learning

Classification

Lecture No. 4

By- SIDDHARTH SABHARWAL SIR

Recap of Previous Lecture

- Topic — Logistic Reg.
- Topic — Questions
- Topic — Logit / 1D Question
- Topic
- Topic

Topics to be Covered

- Topic — LR Likelihood, Logistic Reg.
- Topic — Cost fxn, loss function
- Topic — Multiclass LR
- Topic — Logistic Reg: How safe from outlier
- Topic


Inspiration comes from within yourself. One has to be positive. When you're positive, good things happen.
DEEP ROY
BRIAN TRACY
INTERNATIONAL

Let $x = 1$ if an email subject includes the student's name and $x = 0$ otherwise.

There are **350** emails with $x = 1$ of which **161** were opened ($y = 1$), and **400** emails with $x = 0$ of which **140** were opened.

y=1 email opened
y=0 email not opened

Fit a logistic regression for the log-odds of opening:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x.$$

H.W

Which model is correct?

H.W

@ $x=0$ $\beta_0 = \log_e \dfrac{P(y=1|x=0)}{P(y=0|x=0)}$ 140/400, 260/400

@ $x=1$

$\beta_1 + \beta_0 = \log_e \dfrac{P(y=1|x=1)}{P(y=0|x=1)}$ 161/350, 189/350

A) $-0.6190 + 0.4587x$

B) $-0.1603 + 0.6190x$

C) $-0.6190 + 0.1603x$

D) $0.4587 - 0.6190x$

Answer: A.

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

H.W

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

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.W

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 $\frac{9}{10}$

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)?

H.W.

Logistic Regression

What is the meaning of likelihood

2D Training data

x^1	x^2	y
a	b	1
c	d	1
e	f	0
g	h	0
i	j	0
⋮	⋮	⋮

we need β values such that

Probab of class '1' for datapoint 1 ⇒ close to 1

" " " '1' " " " 2 ⇒ Close to 1

" " " '0' " " " 3 ⇒ Close to 1

" " " '0' " " " 4 ⇒ close to 1

⋮

→ if this happen it means error on training data is v. small

Logistic Regression

What is the meaning of likelihood

- we know that class '1' Probab for any data point $P_i = \frac{1}{1+e^{-x_i\beta}}$

Probab of class '0' $\Rightarrow 1-P_i$

$\Rightarrow \left(1-\frac{1}{1+e^{-x_i\beta}}\right)$

2D Training data

	x^1	x^2	y
$\boxed{1}$	a	b	1
	c	d	1
	e	f	0
	g	h	0
	i	j	0

Logistic Regression

What is the meaning of likelihood

So we want to maximize the Probability that Y matrix = $\begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ \vdots \end{bmatrix} \begin{matrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ \end{matrix}$

max Prob of Y matrix = $\begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$

2D Training data

	x^1	x^2	y
1	a	b	1
	c	d	1
	e	f	0
	g	h	0
	i	j	0
	⋮	⋮	⋮

What is the meaning of likelihood

So we want to maximize the Probability that Y matrix =

$$\begin{array}{c c} \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ \vdots \end{bmatrix} & \begin{array}{l} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ \end{array} \end{array}$$

- max Prob of Y matrix = $\begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$

2D Training data

	$x1$	$x2$	y
$\boxed{1}$	a	b	1
	c	d	1
	e	f	0
	g	h	0
	i	j	0

all y's are independent

Now $\max P(y_1=1, y_2=1, y_3=0, y_4=0, y_5=0 \ldots)$

$\max P(y_1=1) \cdot P(y_2=1) \cdot P(y_3=0) \cdot P(y_4=0) P(y_5=0) \ldots$

$$\boxed{\max P_{11} \cdot P_{21} \cdot [1-P_{31}] [1-P_{41}] [1-P_{51}] \ldots}$$

What is the meaning of likelihood

2D Training data

	$x1$	$x2$	y
1	a	b	1
	c	d	1
	e	p	0
	g	f	0
	i	h	0
		j	

Now $\max P(y_1=1, y_2=1, y_3=0, y_4=0, y_5=0 \ldots)$

$\max P(y_1=1) \cdot P(y_2=1) \cdot P(y_3=0) \cdot P(y_4=0) P(y_5=0) \ldots$

$$\boxed{\max P_{11} \cdot P_{21} \cdot [1-P_{31}][1-P_{41}][1-P_{51}] \ldots}$$

all y's are independent

for any datapoint i^{th} datapoint $\rightarrow \begin{cases} P_{i1} & \text{if } y_i=1 \\ (1-P_{i1}) & \text{if } y_i=0 \end{cases}$

$\rightarrow$ for any i^{th} datapoint $(P_{i1})^{y_i}(1-P_{i1})^{(1-y_i)}$

for any datapoint
i^{th} datapoint $\Rightarrow \begin{cases} P_{i1} & \text{if } y_i = 1 \\ (1-P_{i1}) & \text{if } y_i = 0 \end{cases}$

for any i^{th} datapoint $(P_{i1})^{y_i}(1-P_{i1})^{(1-y_i)}$

if $y_i = 1$

$(P_{i1})^1 \cancel{(1-P_{i1})}^{0 \to 1}$

P_{i1}

$y_i = 0$

$\cancel{(P_{i1})^0}_{\to 1}(1-P_{i1})^{1-0}$

$(1-P_{i1})$

- likelihood meaning : Think you are an artist
Your task is to draw painting of
of the person exactly similar

In logistic
regression similar
task, for every data
Point we want $\hat{y} = y$
we want Predicted class of
every data - y

- we want Ke pooni ki
Pooni same $\hat{Y}$ matrix generate
Karne ki Probab max
Karo

What is the meaning of likelihood

2D Training data

$x1$	$x2$	y
a	b	1
c	d	1
e	p	0
g	f	0
i	h	0
	j	

$\boxed{1}$

$\max P_Y$

Now $\max P(y_1=1, y_2=1, y_3=0, y_4=0, y_5=0 \ldots)$

$\max P(y_1=1) \cdot P(y_2=1) \cdot P(y_3=0) \cdot P(y_4=0) P(y_5=0) \ldots$

$\max P(y_1=1 | x_1; \beta) P(y_2=1 | x_2; \beta) P(y_3=0 | x_3; \beta) \ldots$

all y's are independent

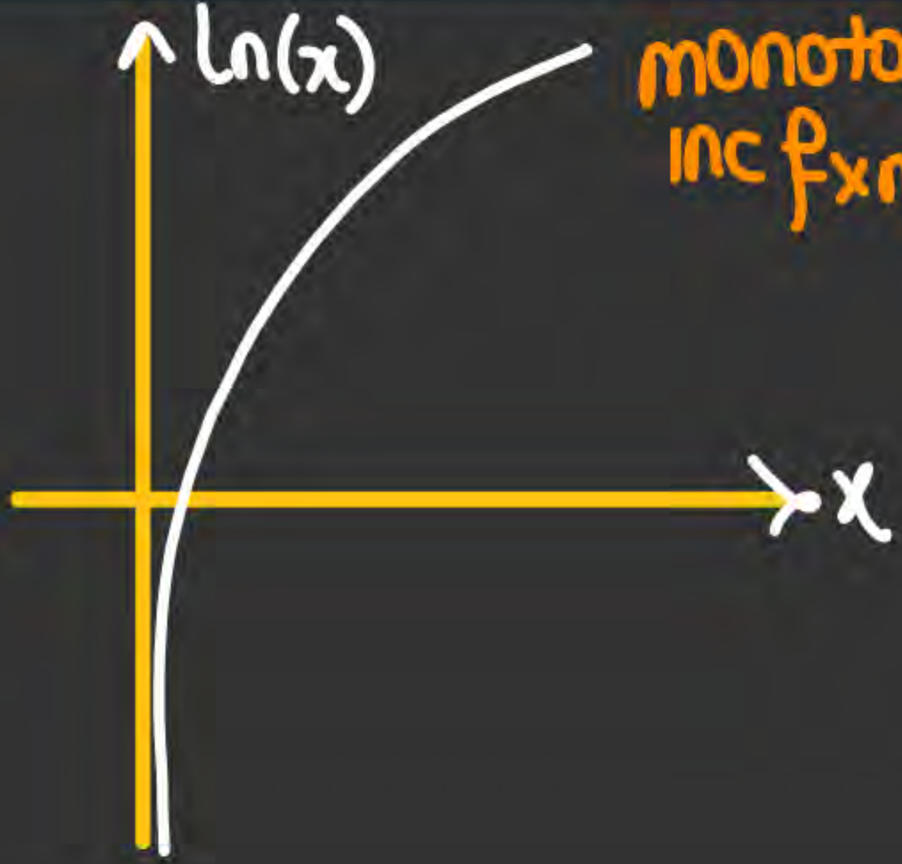
$(P_{11})^{y_1} (1-P_{11})^{(1-y_1)} \; (P_{21})^{y_2} (1-P_{21})^{1-y_2} \ldots$

P_{11} = datapoint 1 Class 1 hone Ki probab

$$\left\{ \max \prod_{i=1}^{N} (P_{i1})^{y_i} (1-P_{i1})^{1-y_i} \right\}$$

max likelihood ⇒

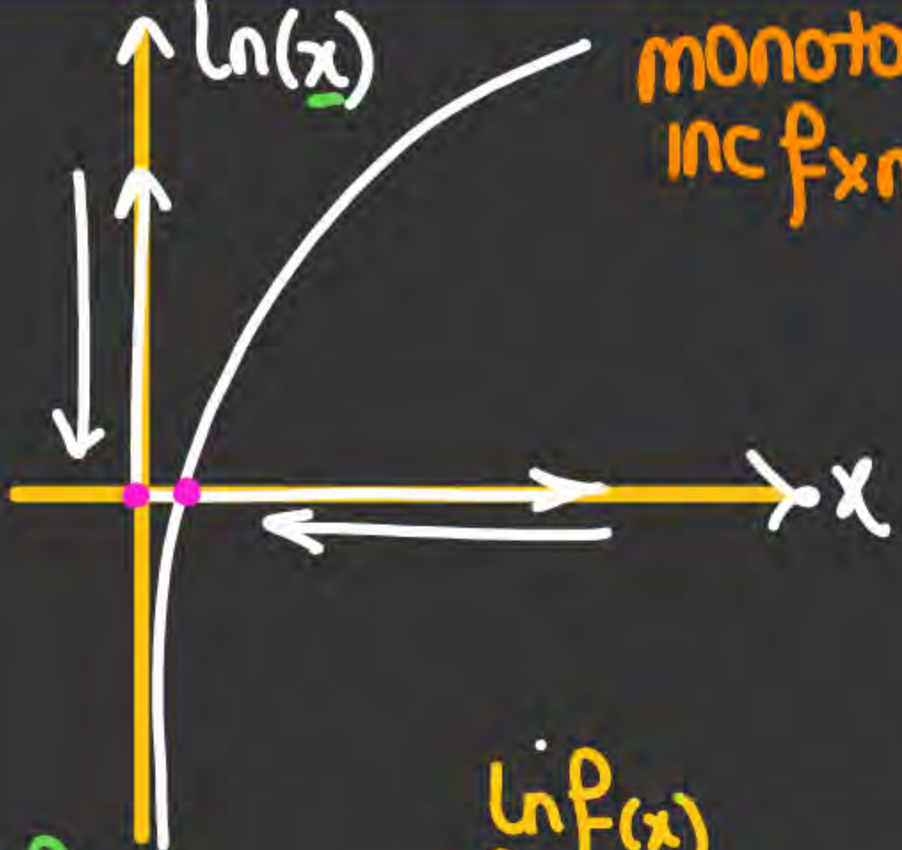
- $$\boxed{\max \prod_{i=1}^{N} (P_{i1})^{y_i} (1-P_{i1})^{(1-y_i)}}$$



monotonically inc fxn.

if we want to solve $\max f(x)$

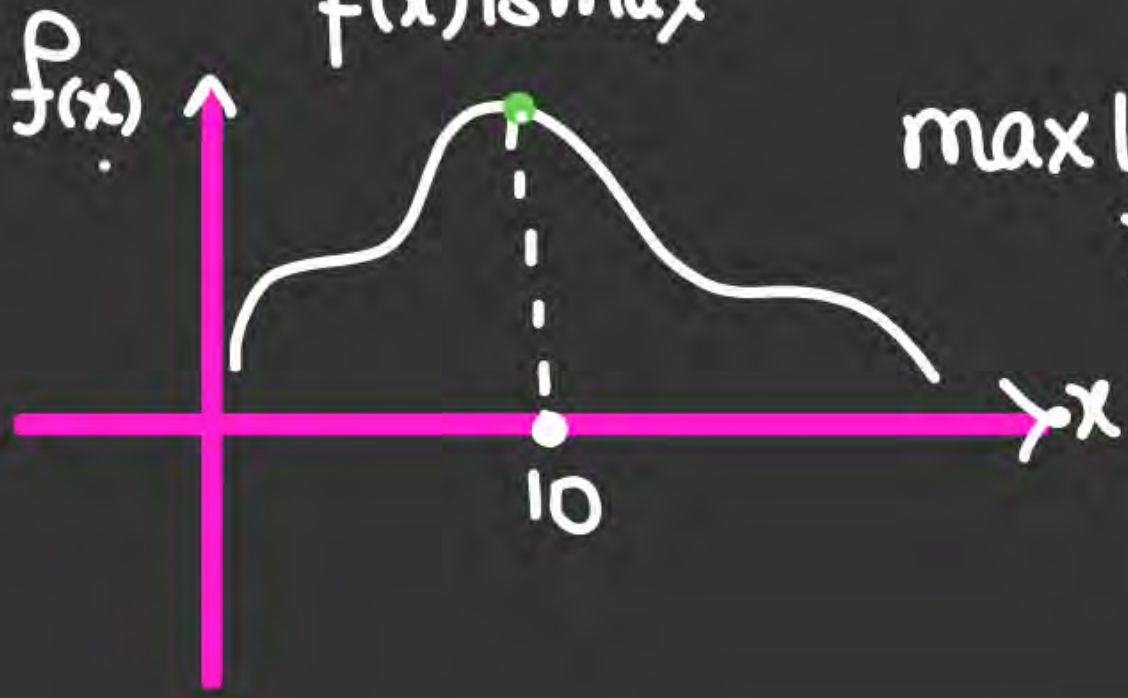
- we want to find value of x where f(x) is max



monotonically inc fxn.

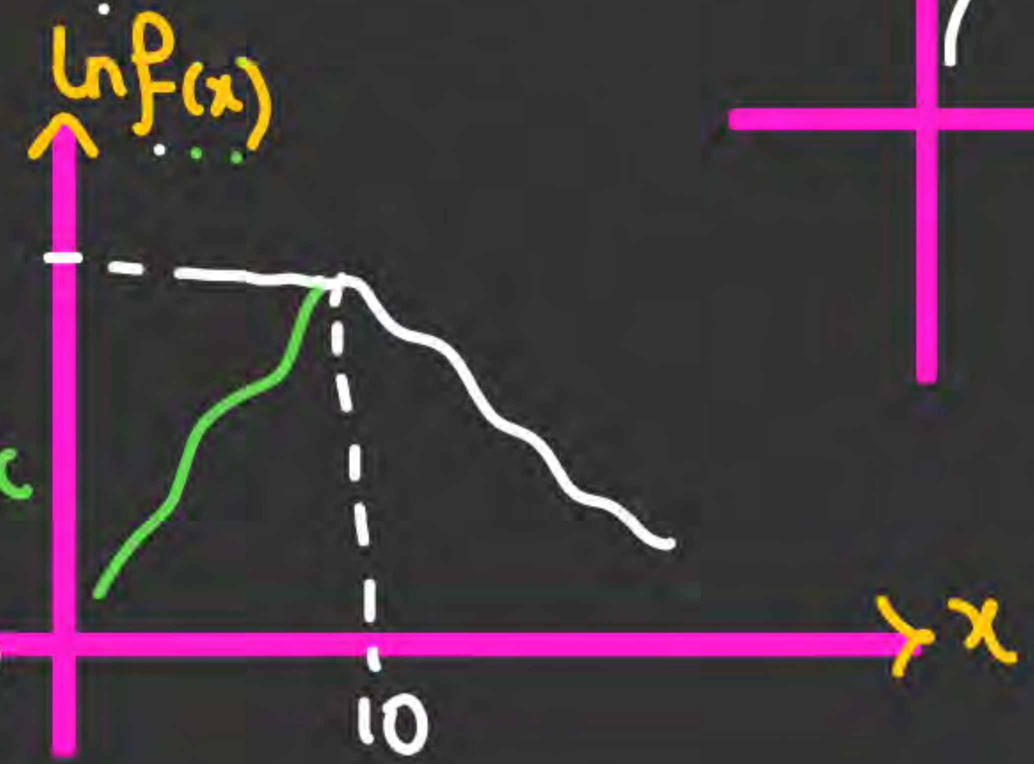
if we want to solve max f(x)

- we want to find value of x where f(x) is max



max location $x=10$

- log is inc fxn
- Jab f(x) inc karega log f(x) will also inc
- when f(x) dec then log f(x) will dec



max likelihood ⇒

$$\max \prod_{i=1}^{N} (P_{i1})^{y_i} (1-P_{i1})^{(1-y_i)}$$

$\ln(abcde)$

$\Downarrow$

$\ln a + \ln b + \ln c + \ln(d)$

Same as $\max \ln \prod_{i=1}^{N} (P_{i1})^{y_i} (1-P_{i1})^{1-y_i}$

max Log likelihood ⇒

$$\max \sum_{i=1}^{N} y_i \ln(P_{i1}) + (1-y_i)\ln(1-P_{i1})$$

aise β's Chahiye that max this equation.

Concave Curve

 $\times(-) =$ 

So loss function $\Rightarrow$ $-$ log likelihood.

Convex Curve $\leftarrow$ Cross Entropy $= -\left[\sum_{i=1}^{N} y_i \ln P_{i1} + (1-y_i)\ln(1-P_{i1})\right]$

- Average Cross Entropy $\Rightarrow$ $-\frac{1}{N}\sum_{i=1}^{N} y_i \ln P_{i1} + (1-y_i)\ln(1-P_{i1})\Big]$

Linear Classification

Logistic Regression

- **The Cost function**

done

log likelihood

How can we use log into this function

Logistic Regression

- The Cost function

Linear Classification

Logistic Regression

- Log likelihood and cross entropy loss function.

done

- Why linear classification is more effected by Outlier
- Max $\sum y_i x_i \beta$

* Outlier has large $x_i \beta$

* $y_i x_i \beta$ will be large -ve hugely effect model

- logistic regression is less effected by the outlier.

$\boxed{CE}$ $y_i = 1, 0$

$$-\left[\sum_{i=1}^{N} y_i \log P_{i1} + (1-y_i) \log (1-P_{i1})\right]$$

• logistic regression is less effected by the outlier.

$\boxed{CE}$ $y_i = 1, 0$

$$\left[-\sum_{i=1}^{N} y_i \log P_{i1} + (1-y_i)\log(1-P_{i1})\right]$$

$x_i\beta$ for outlier is +ve v.v. large

$$P_{i1} = \frac{1}{1+e^{-x_i\beta}} \quad \boxed{P_{i1} = .8, .9}$$

Logistic Regression

- **Multiclass Logistic Regression – Softmax Regression**

Prediction Rule

$$\text{Predicted class} = \underset{k}{\operatorname{argmax}} \left(\mathbf{w}_k \cdot \mathbf{x} \right)$$

Logistic Regression

- Extending the case for more than 2 classes... (not imp)

We trained a three-way logistic regression and obtained weights

$$w_a = (1, 1, 0), w_b = (-1, 1, 1), w_c = (2, 1, 2).$$

What label would be given to the point $x = (0, 1, 1)$?

(A) A.

(B) B.

(C) C.

Logistic Regression

- How to turn the value into probability

(2 points) What is the definition of softmax$(x_1, \ldots, x_n)$? Recall that this function takes in a list of n real numbers $x_1, \ldots, x_n$ and outputs a list of n real numbers.

A. $\text{softmax}(x_1, \ldots, x_n) = \left[\frac{x_1}{\sum_{i=1}^n x_i}, \ldots, \frac{x_n}{\sum_{i=1}^n x_i}\right]$

B. $\text{softmax}(x_1, \ldots, x_n) = \left[\frac{e^{x_1}}{\sum_{i=1}^n e^{x_i}}, \ldots, \frac{e^{x_n}}{\sum_{i=1}^n e^{x_i}}\right]$

C. $\text{softmax}(x_1, \ldots, x_n) = \left[\frac{e^{x_1}}{e^{\sum_{i=1}^n x_i}}, \ldots, \frac{e^{x_n}}{e^{\sum_{i=1}^n x_i}}\right]$

D. $\text{softmax}(x_1, \ldots, x_n) = [e^{x_1}, \ldots, e^{x_n}]$

Q.

i. [2 Pts] Suppose our true labels are $\vec{y} = [0, 0, 1]$, our predicted probabilities of being in class 1 are $[0.1, 0.6, 0.9]$, and our threshold is $T = 0.5$. Give the total (not average) cross-entropy loss. Do not simplify your answer.

$$CE = -\sum_{i=1}^{3} y_i \log P_i + (1-y_i)\log(1-P_i)$$

Total CE Loss = 0, 0, 1 ; .1, .6, .9

$y_1 = 0$, $y_3 = 1$

$y_2 = 0$

$$-\left[\left(\cancel{y_1 \log P_1} + (1-y_1)\log(1-P_1)\right) + \left(\cancel{y_2 \log P_2} + (1-y_2)\log(1-P_2)\right) + \left(y_3 \log P_3 + \cancel{(1-y_3)\log(1-P_3)}\right)\right]$$

$$-\left[1\log(1-.1) + 1\log(1-.6) + 1\log .9\right] \Rightarrow 1.127$$

Q.

i. [2 Pts] Suppose our true labels are $\vec{y} = [0, 0, 1]$, our predicted probabilities of being in class 1 are $[0.1, 0.6, 0.9]$, and our threshold is $T = 0.5$. Give the total (not average) cross-entropy loss. Do not simplify your answer.

Total CE Loss = 0, 0, 1 $\quad$ .1, .6, .9

$y_1 = 0$, $y_3 = 1$

$y_2 = 0$

$$CE = -\sum_{i=1}^{3} y_i \log P_i + (1-y_i)\log(1-P_i)$$

ii. [2 Pts] For the same values as above, give the total squared loss. Do not simplify your answer.

Squared Loss = $RSS = \sum_{i=1}^{3}(y_i - \hat{y}_i)^2$

$= (0-0)^2 + (0-1)^2 + (1-1)^2 = 1$ ✓

y	P_1	$\hat{y}$
0	.1	0
0	.6	1
1	.9	1

$\hat{y} = 1 \quad P_1 > .5$

$\hat{y} = 0 \quad P_1 < .5$

(b) **(2.0 pt)** Consider the following three rows from our training data, along with their predicted probabilities $\hat{y}$ for some choice of θ:

x^1 hue	x^2 abv	y	P_1 $\hat{y} = \sigma(x^T\theta)$
-0.17	0.24	0	0.45 ✓
-1.18	1.61	0	0.19 ✓
1.25	-0.97	1	0.80 ✓

$\sigma(x^T\theta)$

What is the mean cross-entropy loss on just the above three rows of our training data?

○ $-\frac{1}{3}(\log(0.45) + \log(0.19) + \log(0.20))$

○ $-\frac{1}{3}(\log(0.55) + \log(0.19) + \log(0.80))$

○ $-\frac{1}{3}(\log(0.45) + \log(0.81) + \log(0.80))$

✓ ○ $-\frac{1}{3}(\log(0.55) + \log(0.81) + \log(0.80))$

$$-\frac{1}{3}\left[(1-0)\log_e(1-P_1)+(1-0)\log_e(1-P_2) + 1\log_e P_3\right]$$

(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:

CE

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

• # positive class points ko 2 times penalize.

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

$$CE = -\left[\sum_{i=1}^{N} y_i \log P_{i1} + (1-y_i)\log(1-P_{i1})\right]$$

→ if $y_i = 1$
Term $\log P_{i1}$

→ if $y_i = 0$
Term $\log(1-P_{i1})$

CE is loss we want to minimize it

* CE minimize = 0

if for class 1 point $\log P_{i1} = 0$

if for class 0 point $\log(1-P_{i1}) = 0$

CE = 0

loss = 0

CE min ⇒ for class 1 point $P_{i1} = 1$

" " 0 " $P_{i1} = 0$

$$CE = -\left[\sum_{i=1}^{N} 2 y_i \log P_{i1} + (1-y_i)\log(1-P_{i1})\right]$$

Classifier biased toward class 1 points.

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".

$\frac{1}{1+e^{-x_i\beta}}$ P.W P.W

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

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.

H.W

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.

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

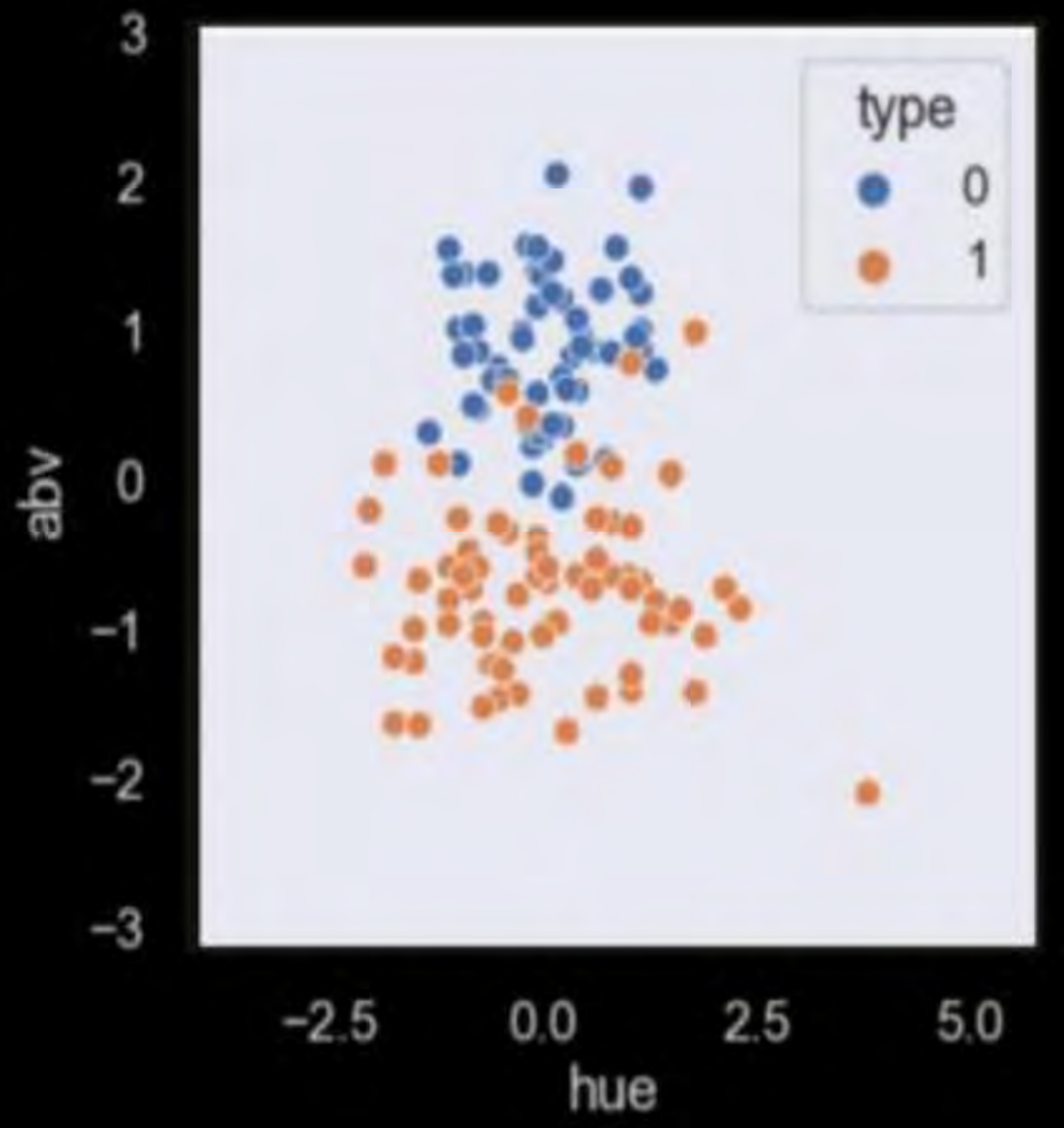
(D) w_1 and w_2 are negative, w_3 is positive

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.

th.w.

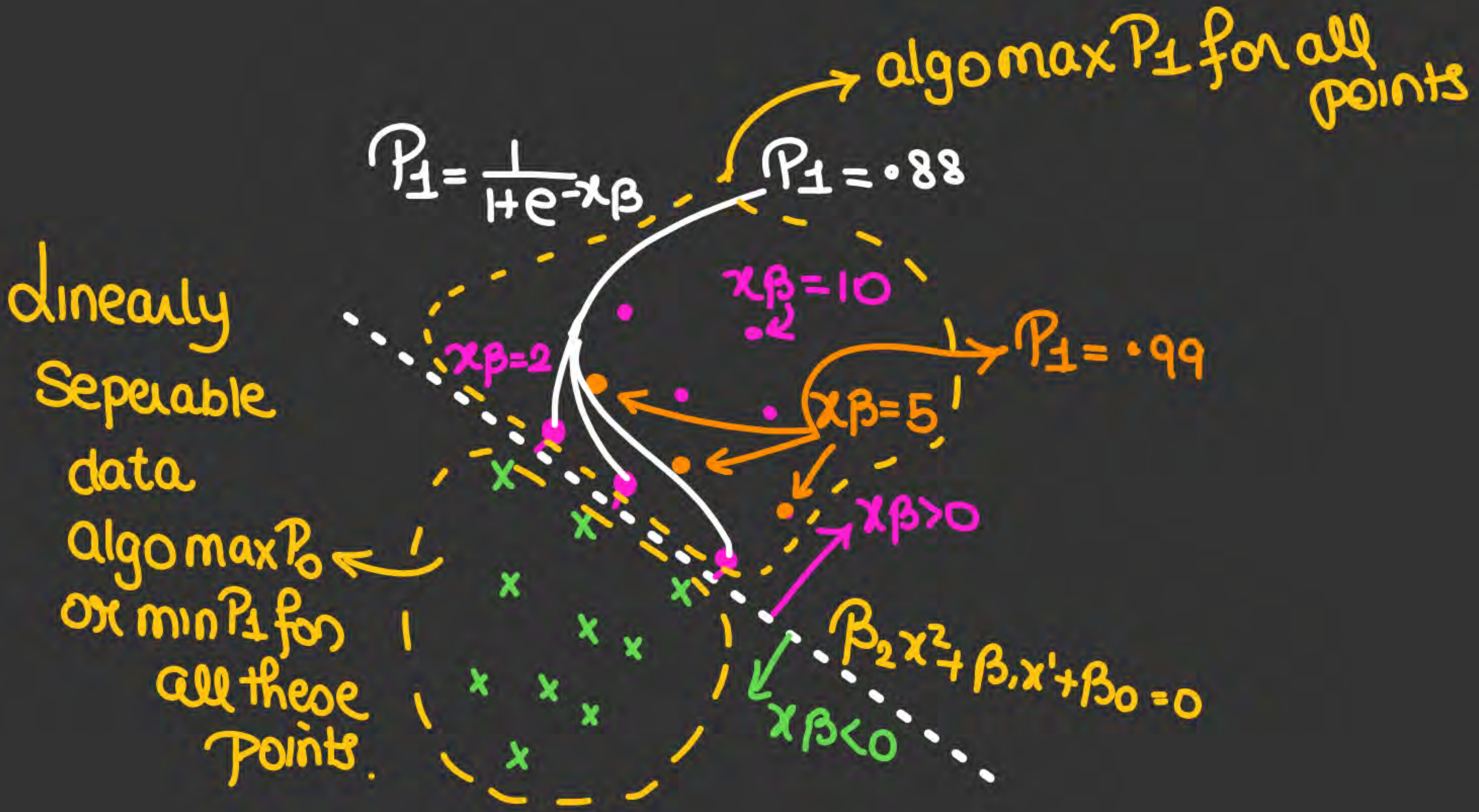
- ☐ After performing logistic regression, the weight for the hue feature will very likely have a negative sign.
- ☐ After performing logistic regression, the weight for the abv feature will very likely have a negative sign.



Logistic Regression

- Extending the case for more than 2 classes... (not imp)

$$\left.\begin{aligned} &\Rightarrow 4x^2 + 8x^1 + 10 = 0 \\ &\Rightarrow 8x^2 + 16x^1 + 20 = 0 \\ &\Rightarrow 400x^2 + 800x^1 + 1000 = 0 \end{aligned}\right\} \text{Same classifier}$$


algo max P_1 for all points
$P_1 = \frac{1}{1+e^{-x\beta}}$
$P_1 = .88$
$x\beta = 10$
$x\beta = 2$
$P_1 = .99$
$x\beta = 5$
$x\beta > 0$
linearly seperable data
algo max P_0 or min P_1 for all these points.
$\beta_2 x^2 + \beta_1 x^1 + \beta_0 = 0$
$x\beta < 0$

multiply all β's by 1000 ⇒ line same

$P_1 = \frac{1}{1+e^{-x\beta}}$

Linearly Seperable data

$P_1 = \frac{1}{1+e^{+2000}}$

$= 0$

$P_0 = 1$

So all Points have $P_1 = 1$

For all $P_0 = 1$

Logistic Reg $\left\{ \text{overfit} \right\}$ large β's $\left\} \right.$ Unstable model.

Regularisation Solution $\Rightarrow$ loss fxn $\Rightarrow$

$$-\sum_{i=1}^{N} y_i \log_e p_i + (1-y_i)\log_e(1-p_i) + \sum_{i=1}^{2} \beta_i^2$$

β large

overfit, Probab of class 1

1	$x_i\beta > 0$
0	$x_i\beta < 0$

Logistic Regression

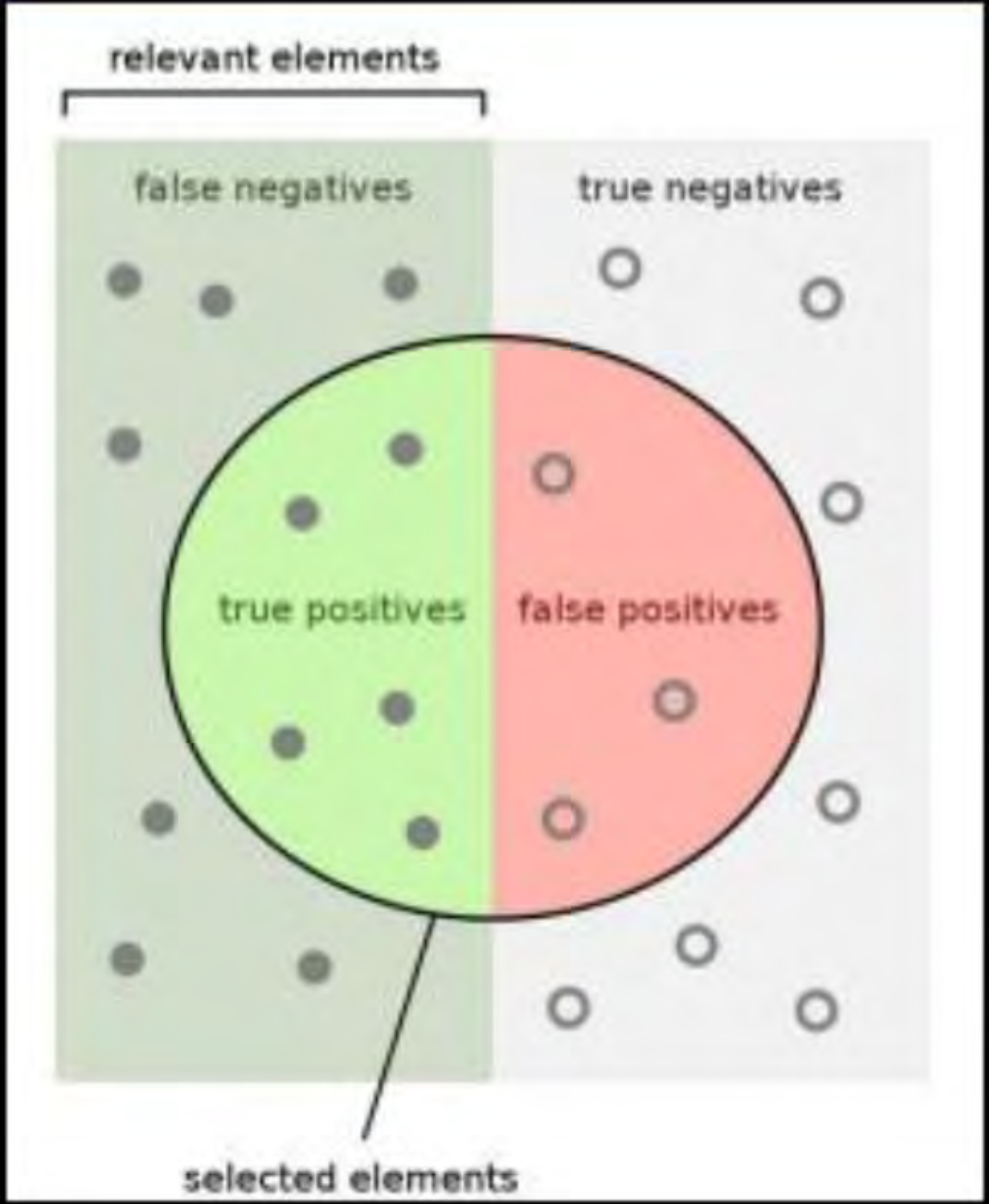
- Why we need regularisation in Logistic regression

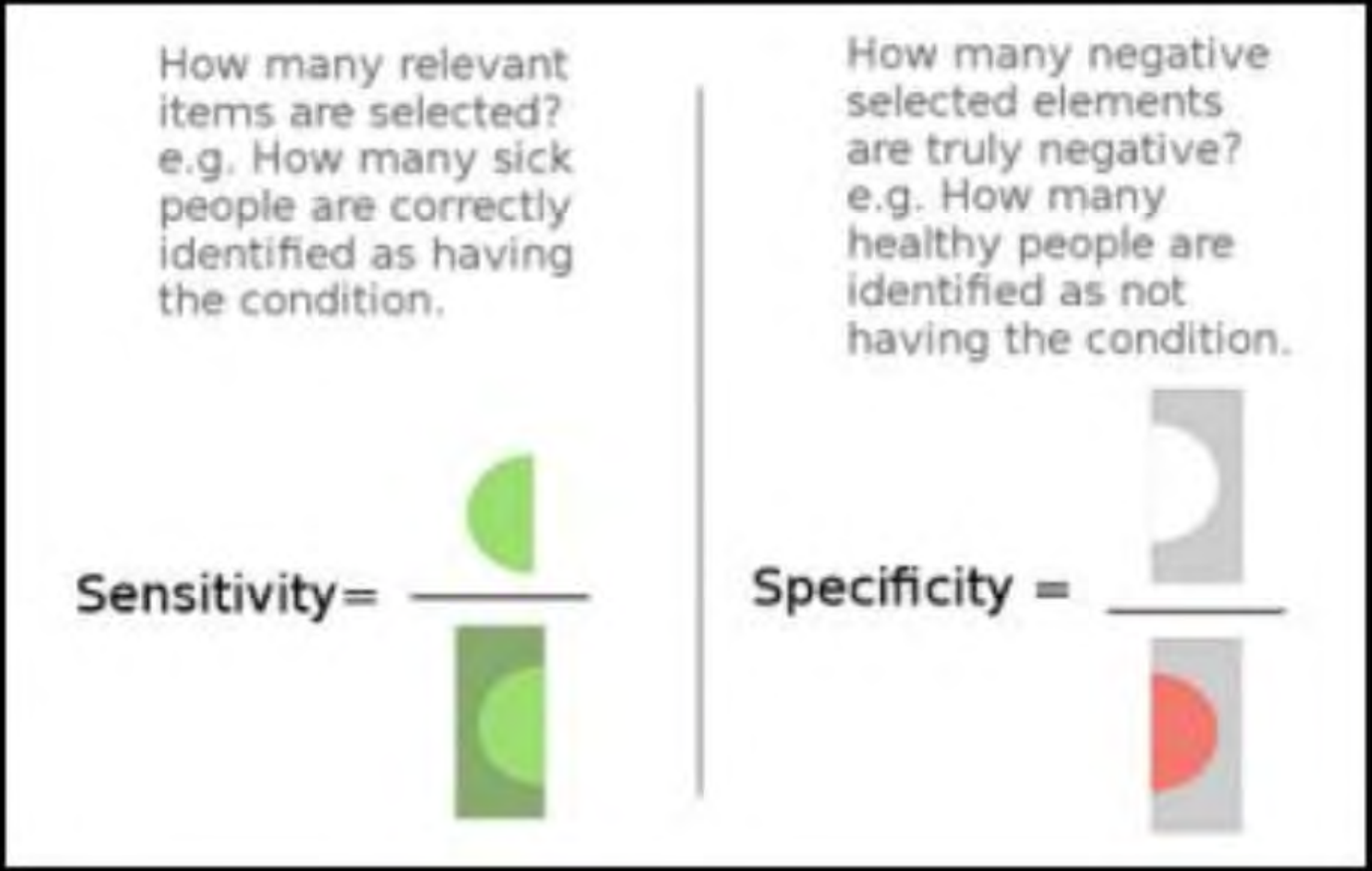
What is ROC curve (receiver operating characteristic curve)

- 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

Linear Classification

What is ROC curve (receiver operating characteristic curve)





What is ROC curve (receiver operating characteristic curve)

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

$$\text{sensitivity} = \frac{\text{number of true positives}}{\text{number of true positives} + \text{number of false negatives}}$$

$$\text{specificity} = \frac{\text{number of true negatives}}{\text{number of true negatives} + \text{number of false positives}}$$

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:

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

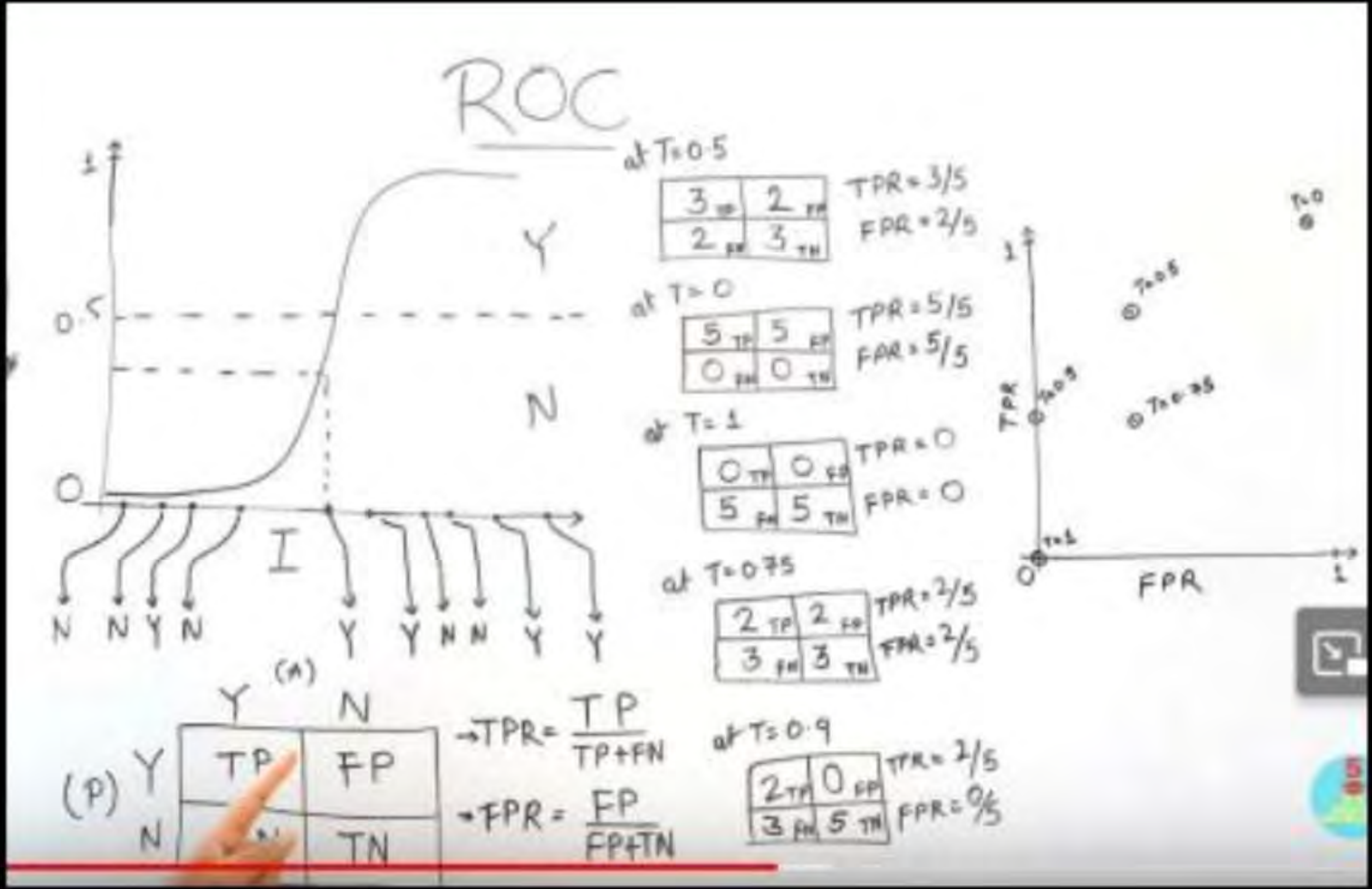
False Positive Rate (**FPR**) is defined as follows:

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

Linear Classification

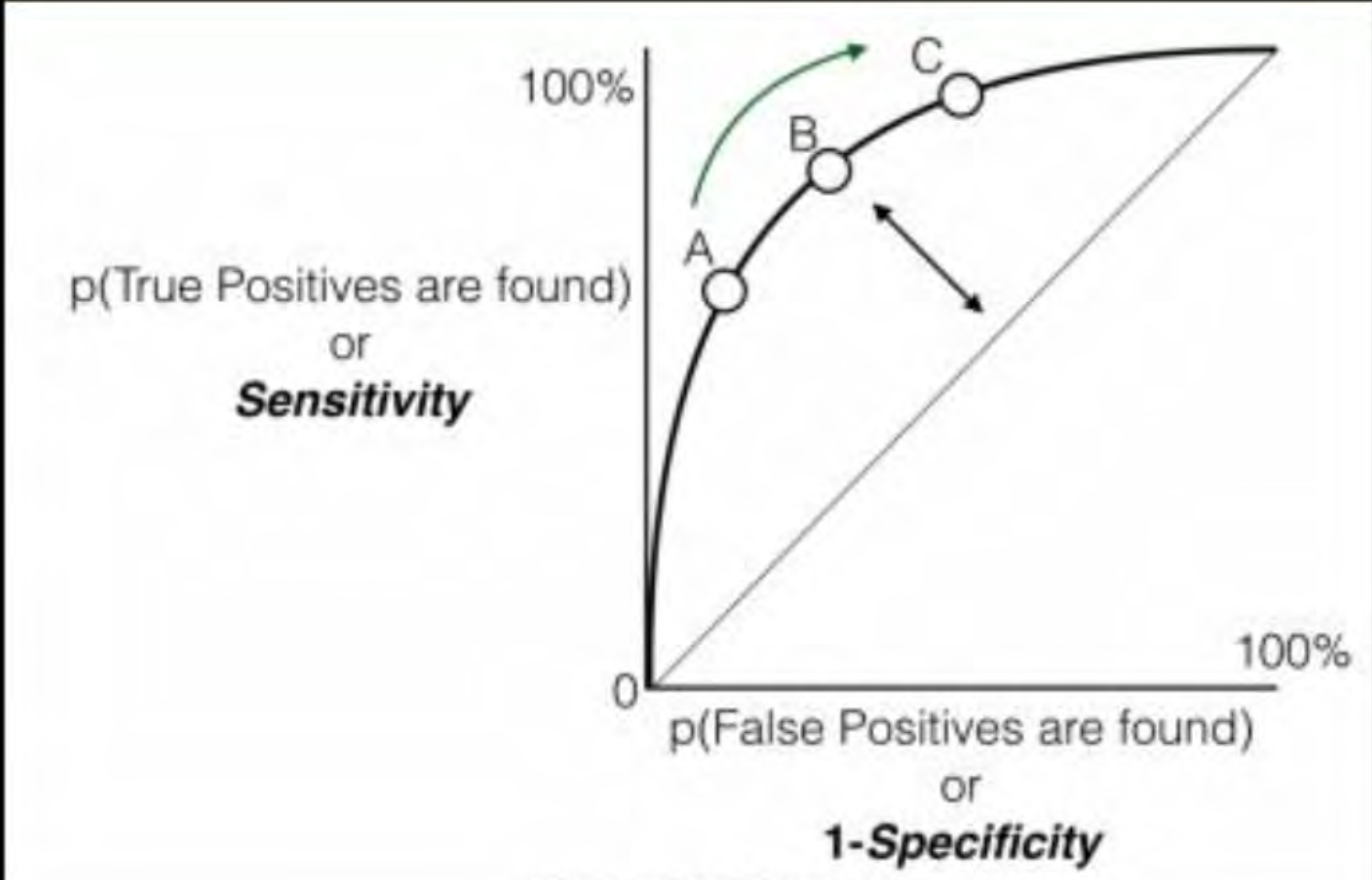
What is ROC curve (receiver operating characteristic curve) an example

The curve between TPR and FPR.



Linear Classification

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.

THANK - YOU