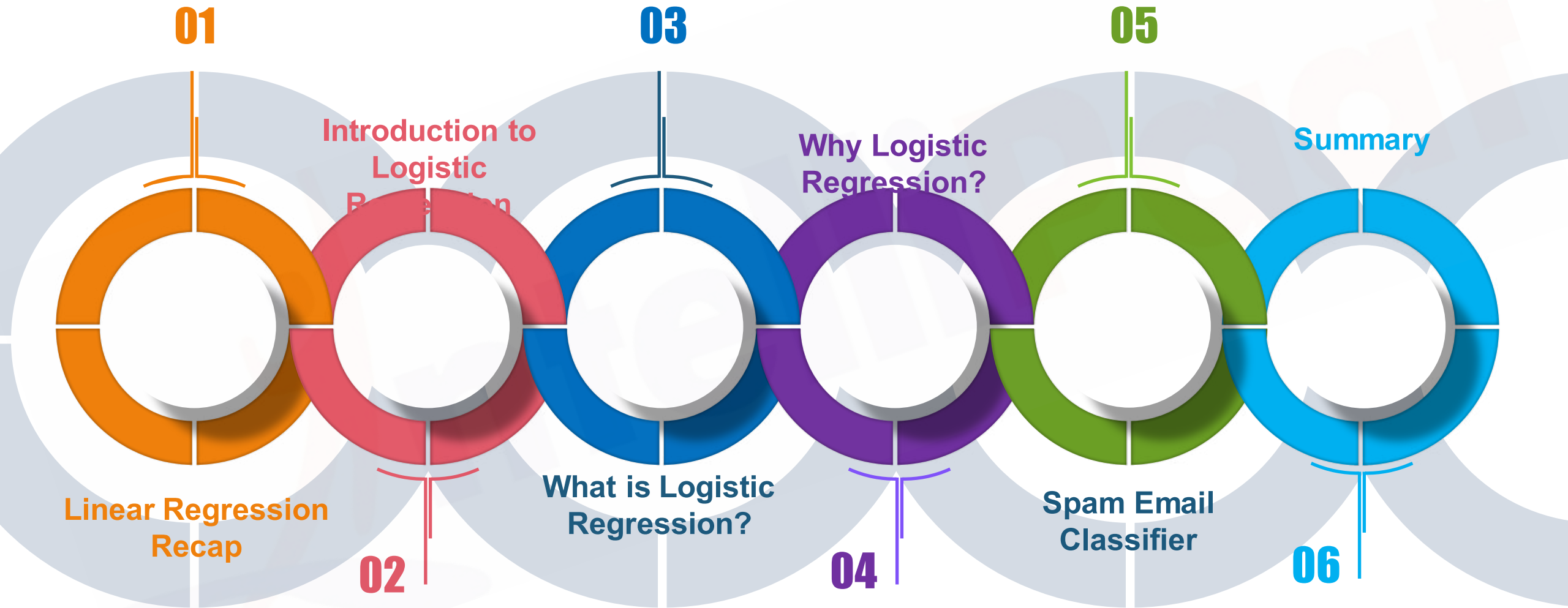
Logistic Regression Algorithm

Agenda for Today's Session



Look back at Linear Regression

IntelliPaat

Regression (Recap)

Online Advertising-Dollars (in 1000s)	Monthly E-commerce Sales-Dollars (in 1000s)
1.7	368
1.5	340
2.8	665
5	954
1.3	331
2.2	556
4	800
3.5	700

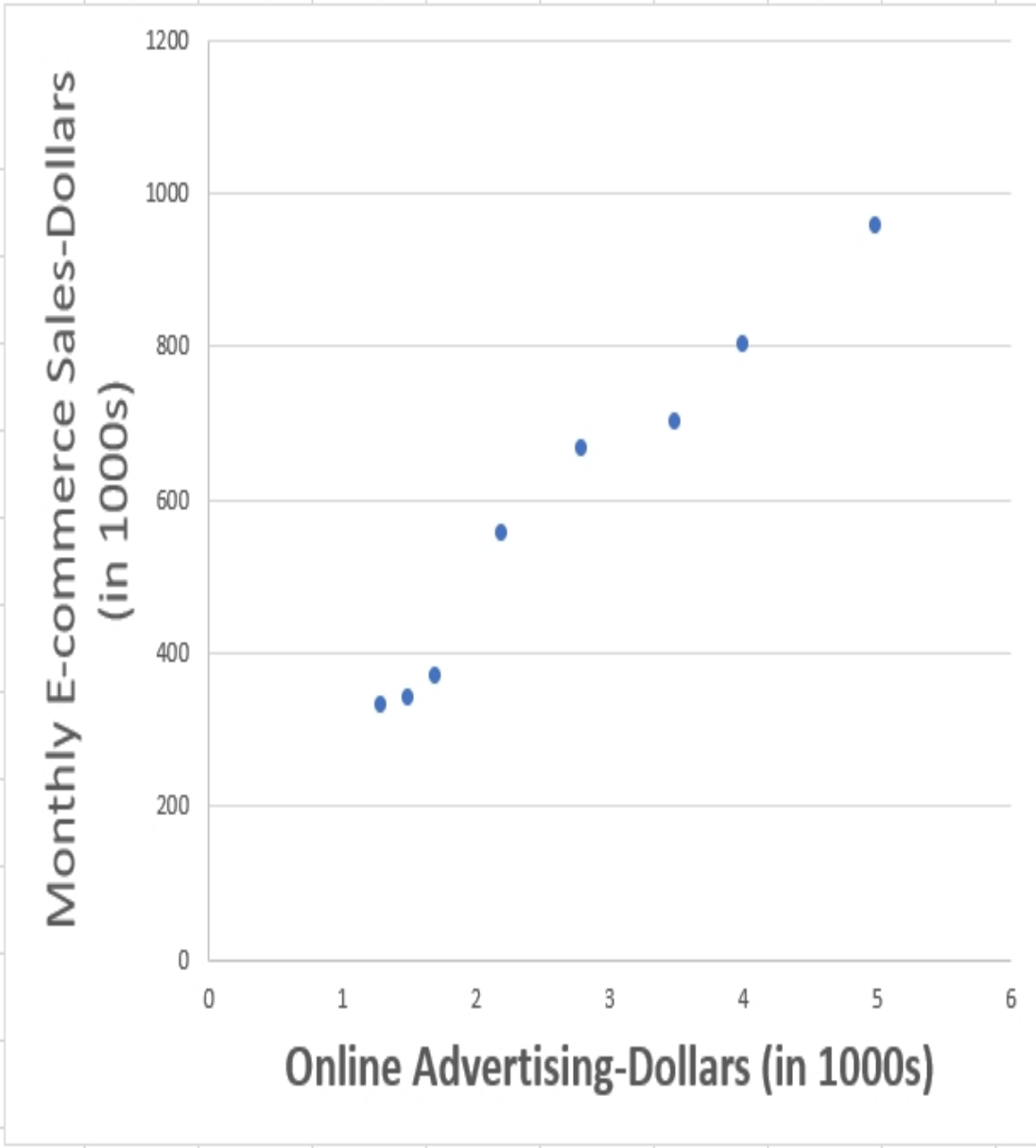
IntelliPaat

Linear Regression (Recap)

Online Advertising-Dollars (in 1000s)	Monthly E-commerce Sales-Dollars (in 1000s)
1.7	368
1.5	340
2.8	665
5	954
1.3	331
2.2	556
4	800
3.5	700
2	
2.5	
3	

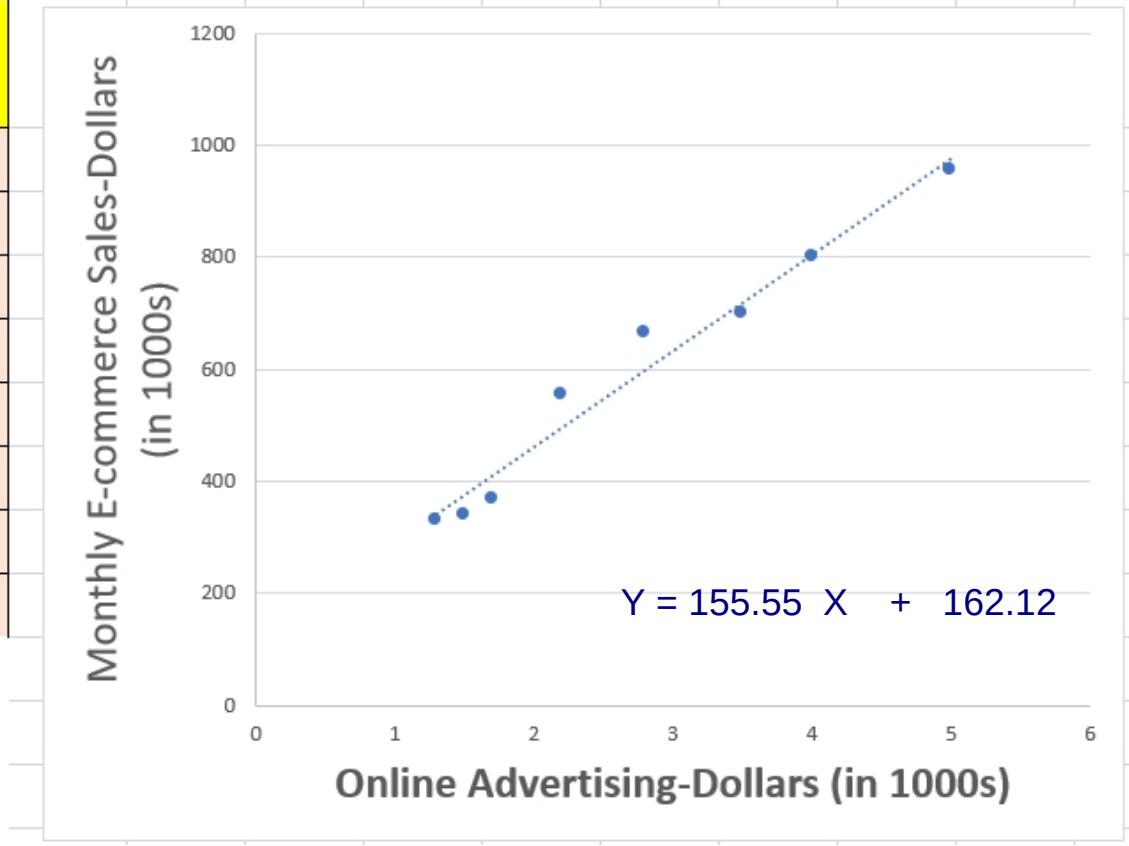
Linear Regression (Recap)

Online Advertising-Dollars (in 1000s)	Monthly E-commerce Sales-Dollars (in 1000s)
1.7	368
1.5	340
2.8	665
5	954
1.3	331
2.2	556
4	800
3.5	700



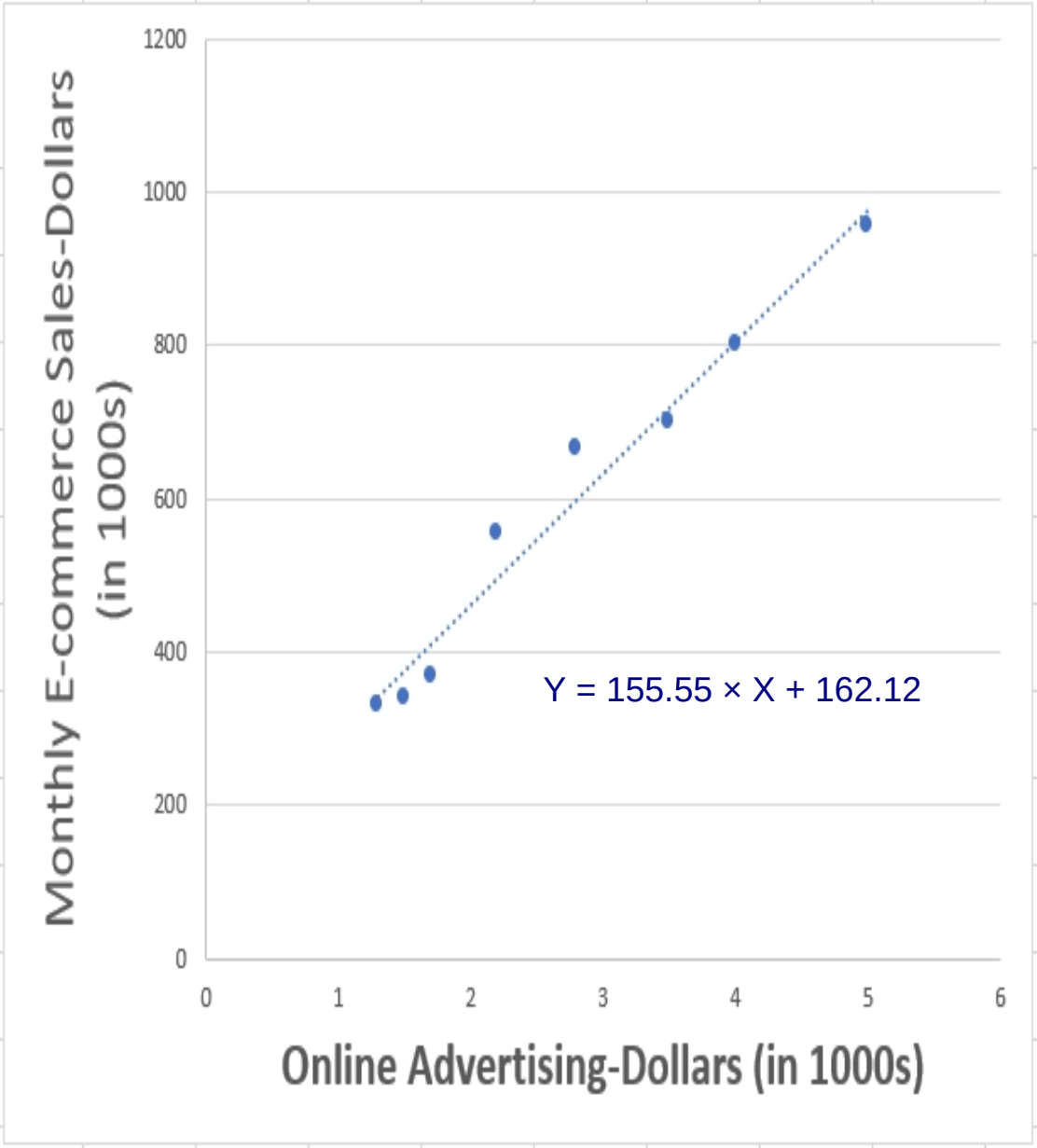
Linear Regression (Recap)

Online Advertising-Dollars (in 1000s)	Monthly E-commerce Sales-Dollars (in 1000s)
1.7	368
1.5	340
2.8	665
5	954
1.3	331
2.2	556
4	800
3.5	700



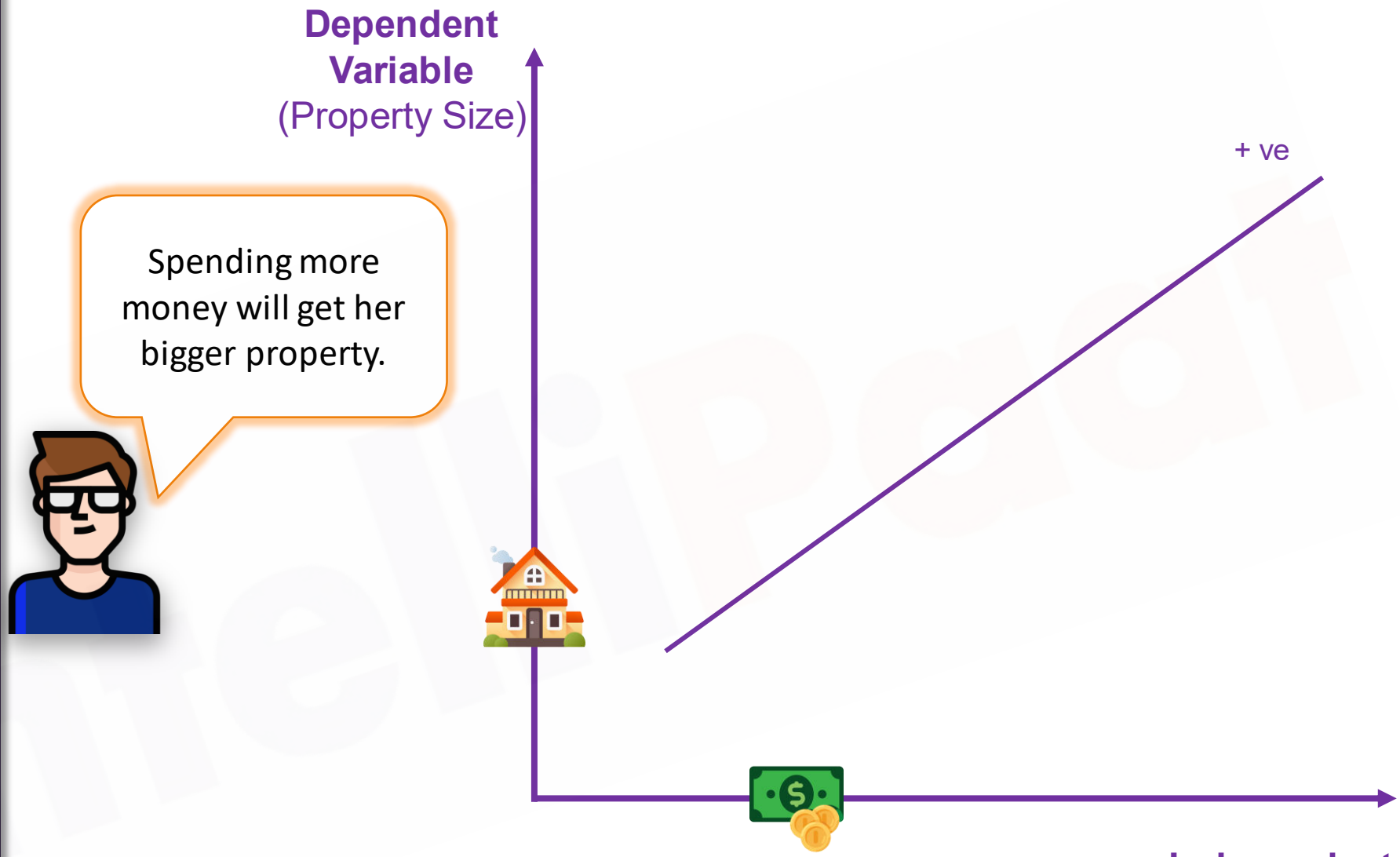
Linear Regression (Recap)

Online Advertising-Dollars (in 1000s)	Monthly E-commerce Sales-Dollars (in 1000s)
1.7	368
1.5	340
2.8	665
5	954
1.3	331
2.2	556
4	800
3.5	700
2	Y = 155.55 × 2 + 162.12 = 473.22
2.5	Y = 155.55 × 2.5 + 162.12 = 550.995
3	Y = 155.55 × 3 + 162.12 = 628.77

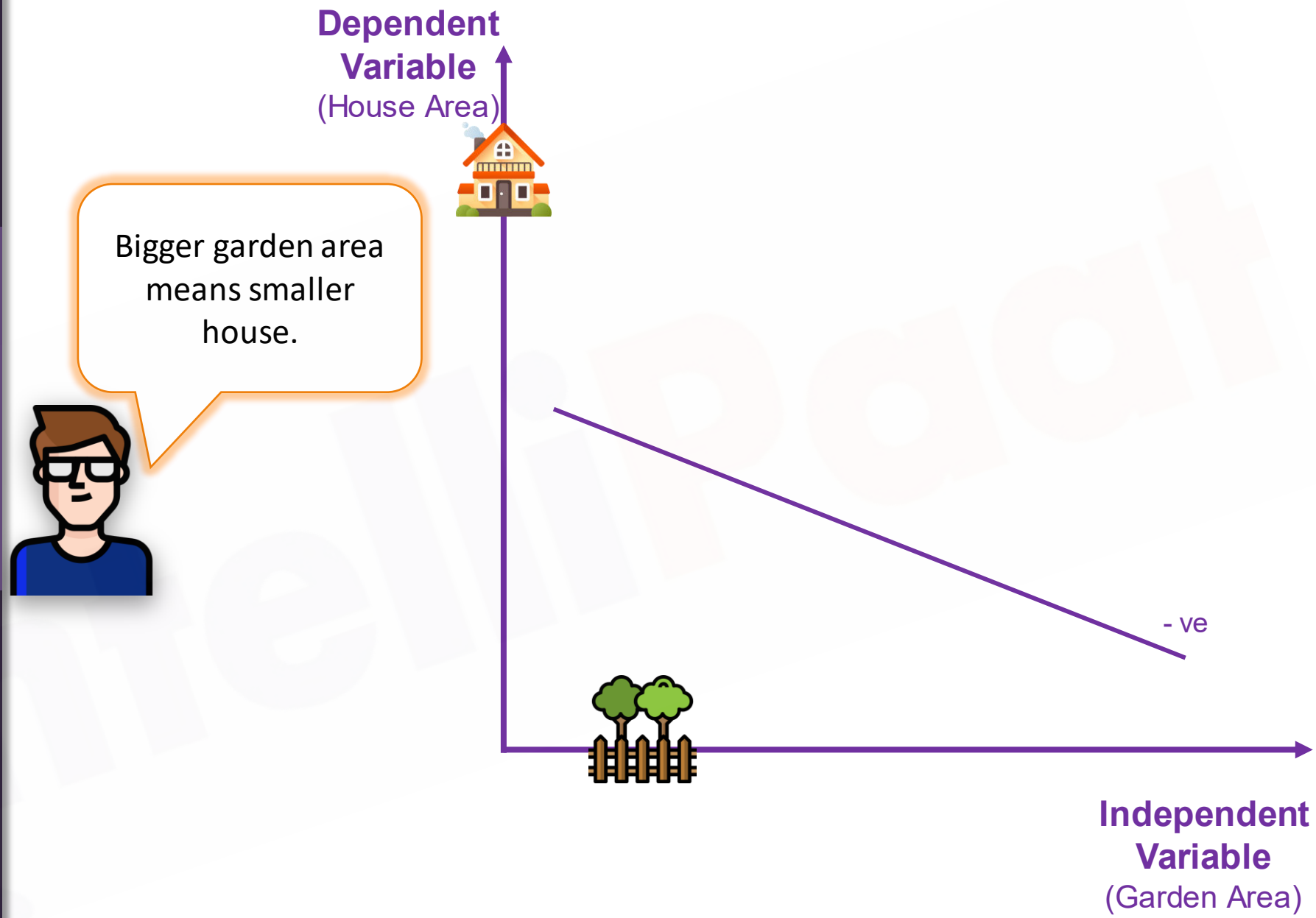


-> Independent Variable (X)
-> Dependent variable (Y)

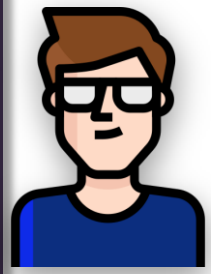
Linear Regression (Recap)

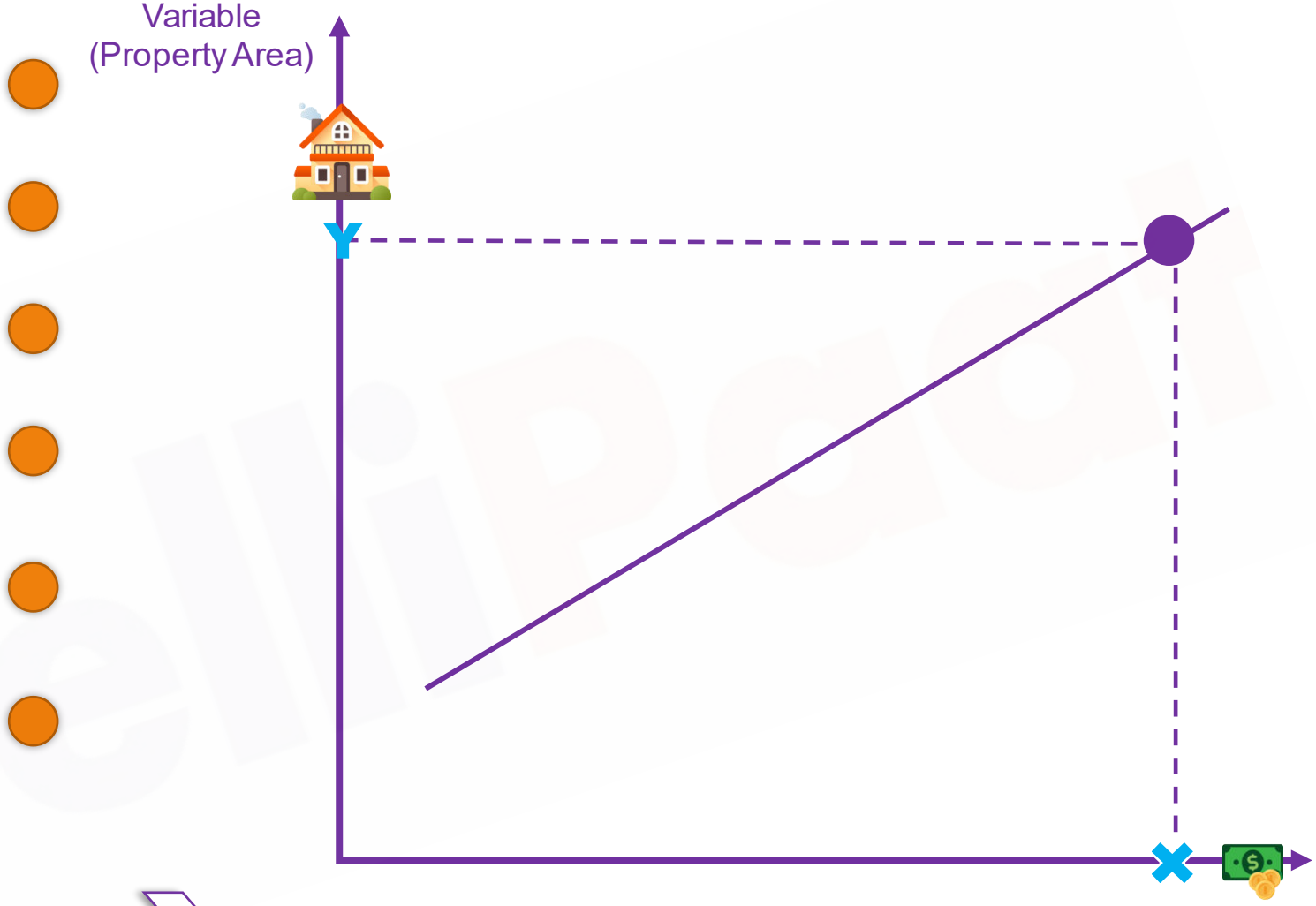


Linear Regression (Recap)



Linear Regression (Recap)





What he can say → If Lauren spends "X" amount of money, she can buy a property of area "Y".

What he can't say → Will the property have a good neighbourhood or not?
Will the location be noiseless suburb or a bustling city?

Introduction to Logistic Regression

Will the property have a good neighbourhood?

Will it rain tomorrow or not?

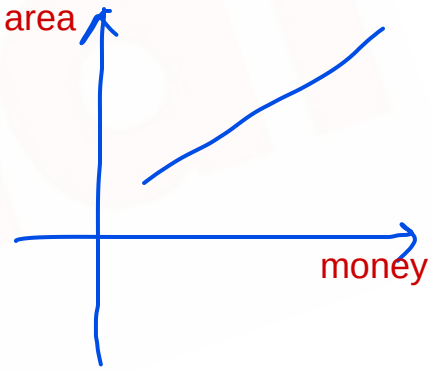
Is the mail spam or not?

Classification Problems

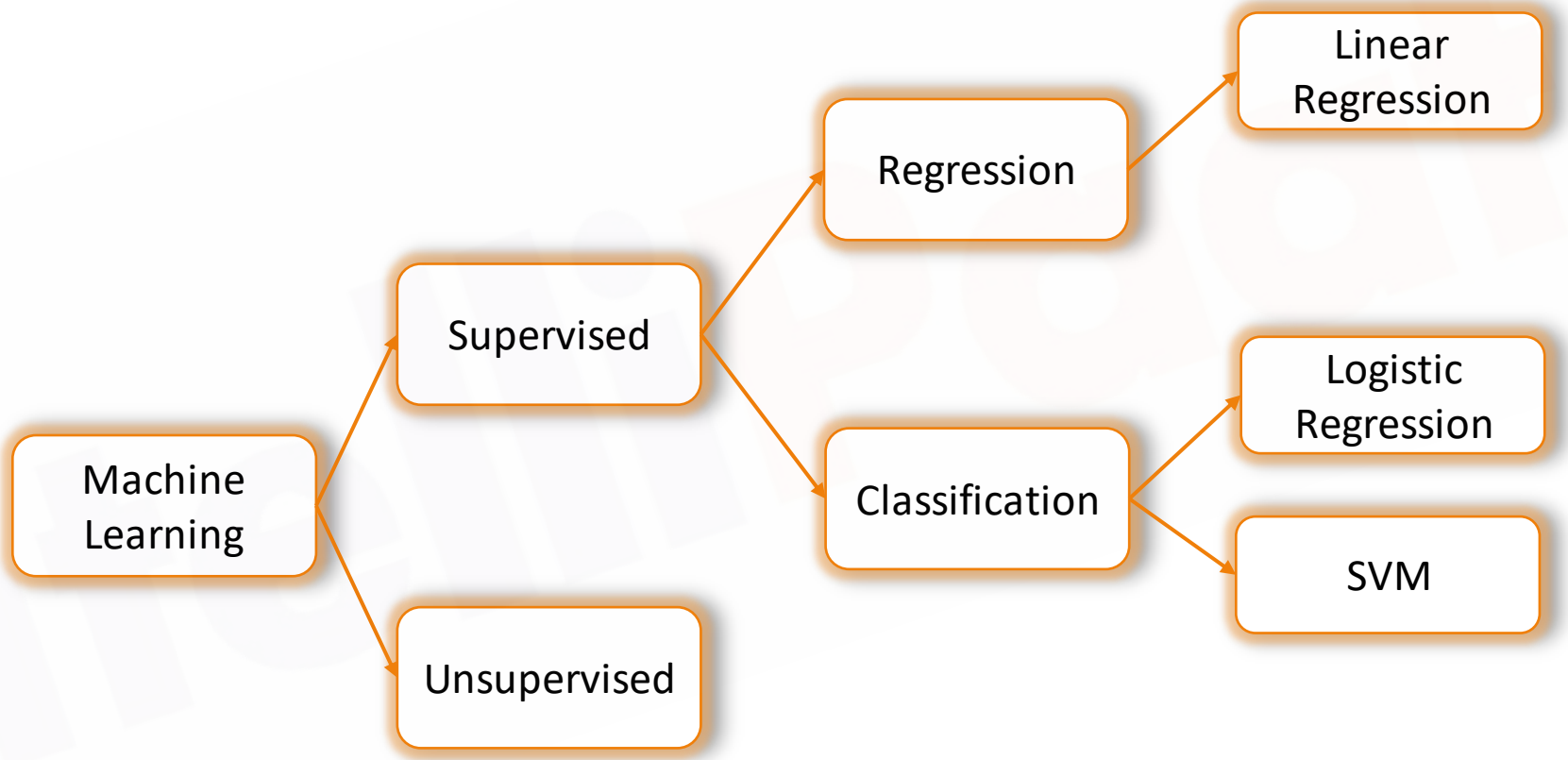
Linear Regression cannot answer

Logistic Regression comes into the picture

Say, in money vs. area problem of regression, we want to find out if the plot area is big or small?

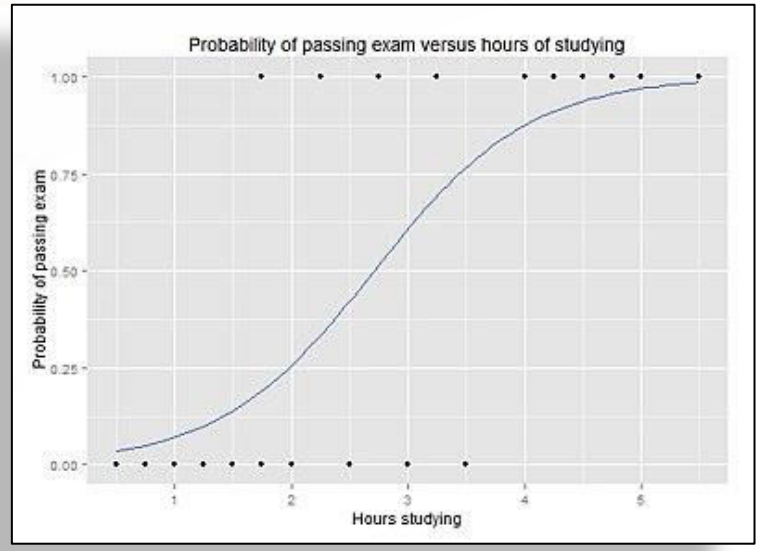


Introduction to Logistic Regression



What is Logistic Regression ?

- A statistical classification model
- Deals with categorical dependent variables
- Could be binary or dichotomous
- Could be multinomial Multinomial Logistic Regression (Softmax Regression) - multiclass classification
- Takes both continuous and discrete input data



Why Logistic Regression ?

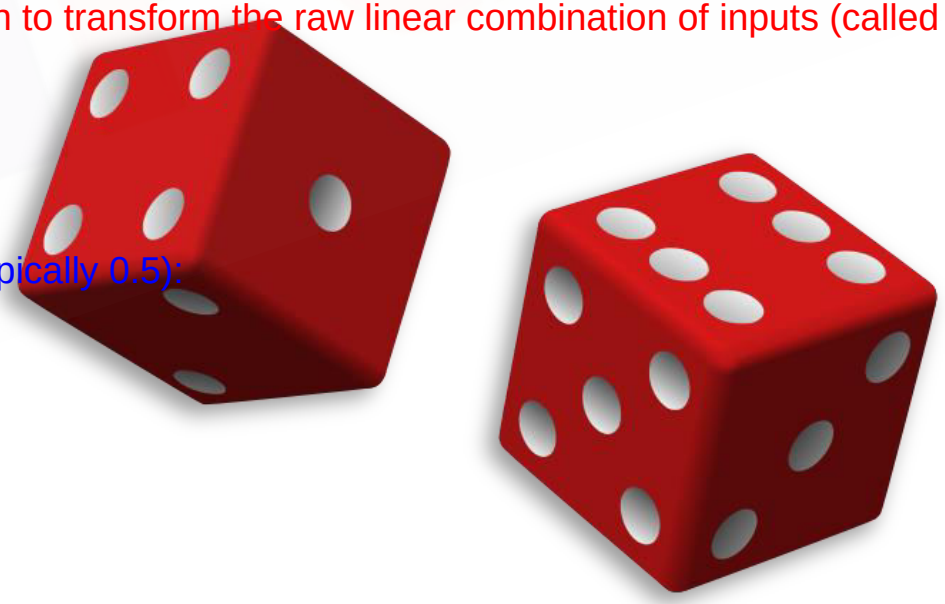
- Tool for applied statistics and discrete data analysis
- Gives outcome in terms of probability
- In-turn helps in classifying the given data

Output as Probability:

Logistic regression uses the sigmoid function to transform the raw linear combination of inputs (called a logit) into a number between 0 and 1:

Thresholding for Classification

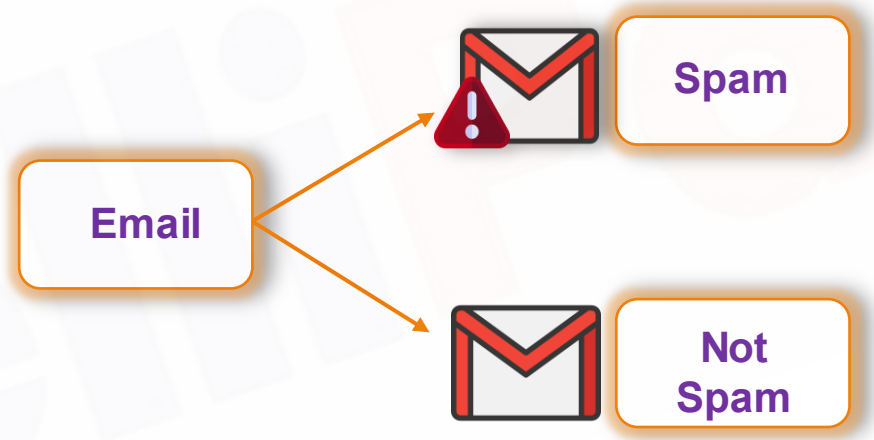
To make a decision, we apply a threshold (typically 0.5):
- If P(y=1) > 0.5, predict Class 1.
- If P(y=1) ≤ 0.5, predict Class 0.



Let’s understand this with an example

Spam Email Classifier

Let us explore with an example of Spam email classifier.



Spam Email Classifier

Steps:

- Understanding the variable
- Plot the labeled data
- Draw regression curve
- Find out the best fitted curve using Maximum Likelihood Estimator(MLE)

Spam Email Classifier

Let's get started.

STEP:1

- Define the variable
- Plot labeled data
- Draw regression line
- Find out the best using MLE

Independent variable:

- Count of spam words (say, x)

Most common spam words:

- Buy
- Get paid
- Guarantee
- Winner
- Unlimited

Bag of spam words

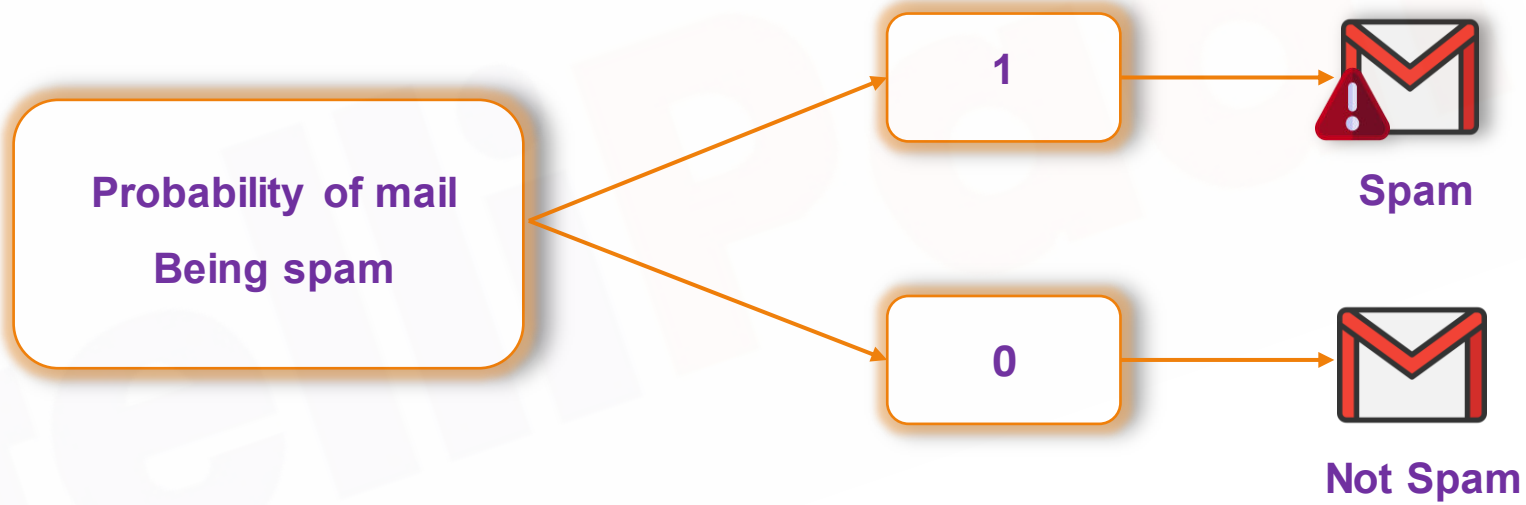
IntelliPaat

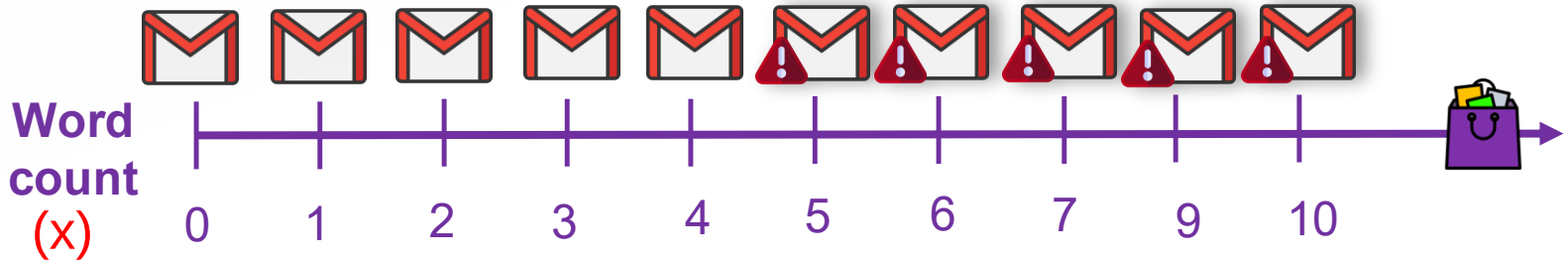
STEP:1

- Define the variable
- Plot labeled data
- Draw regression line
- Find out the best using MLE

Dependent variable:

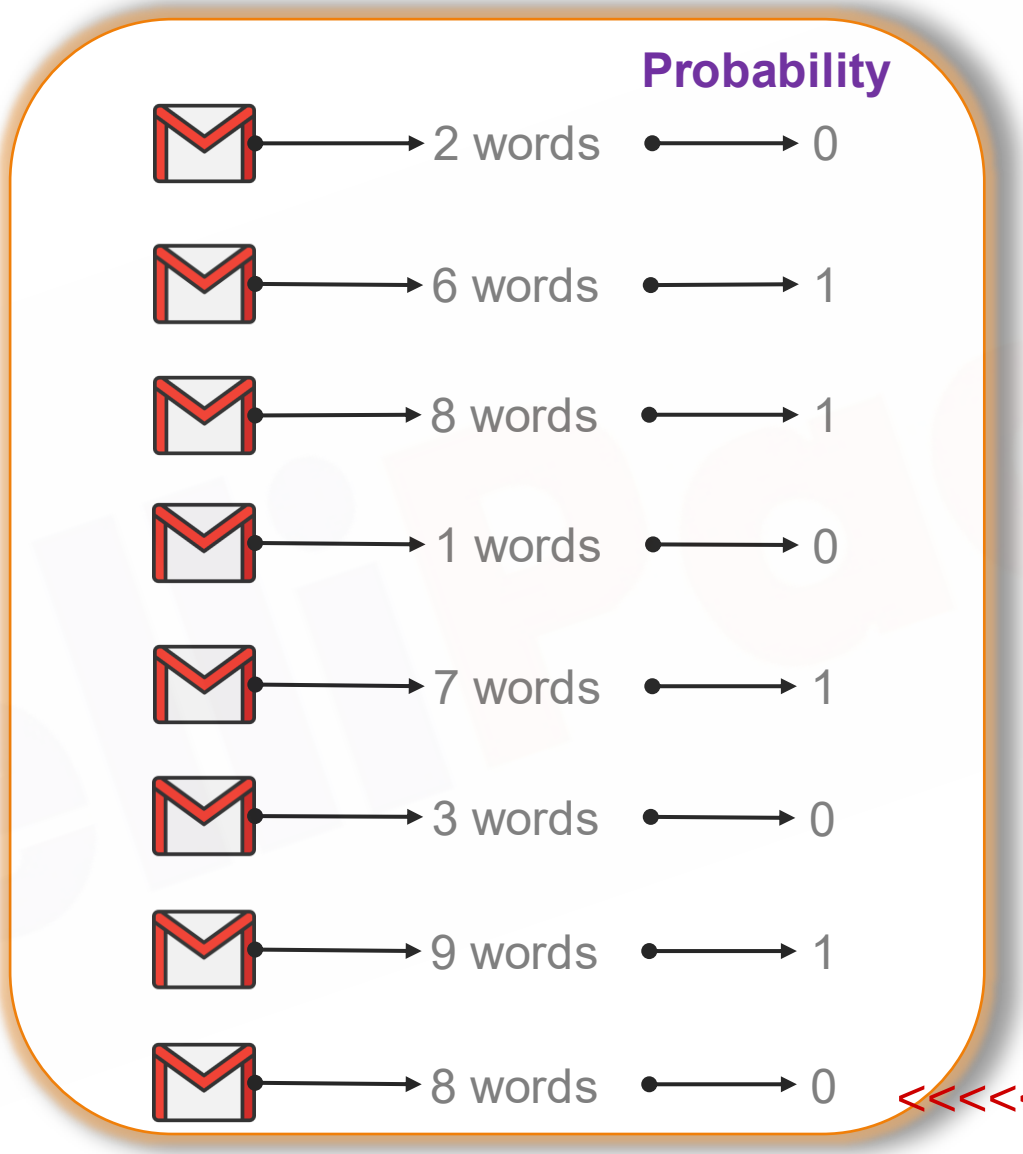
- Probability of mail being spam.
- Binary or dichotomous



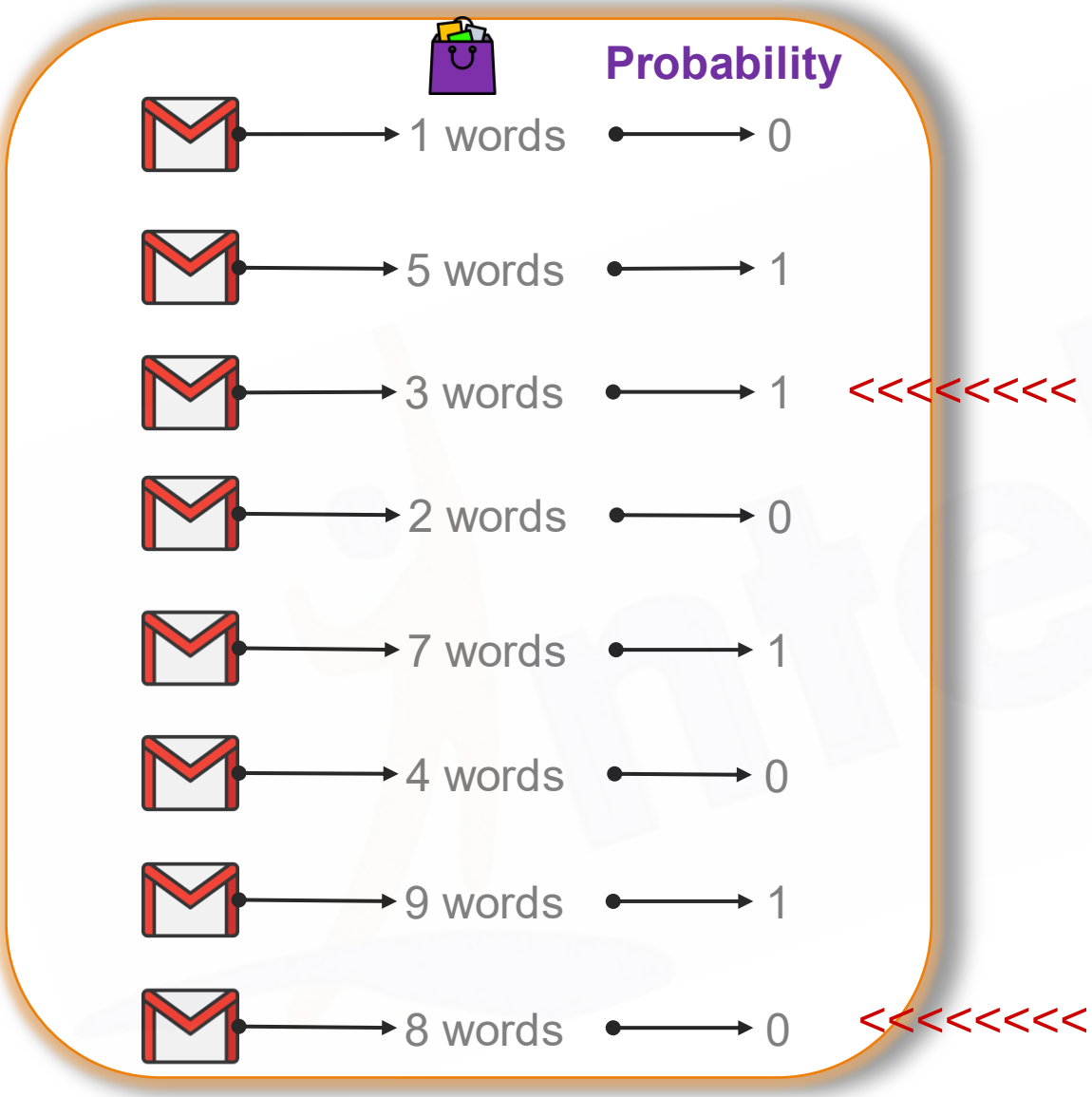


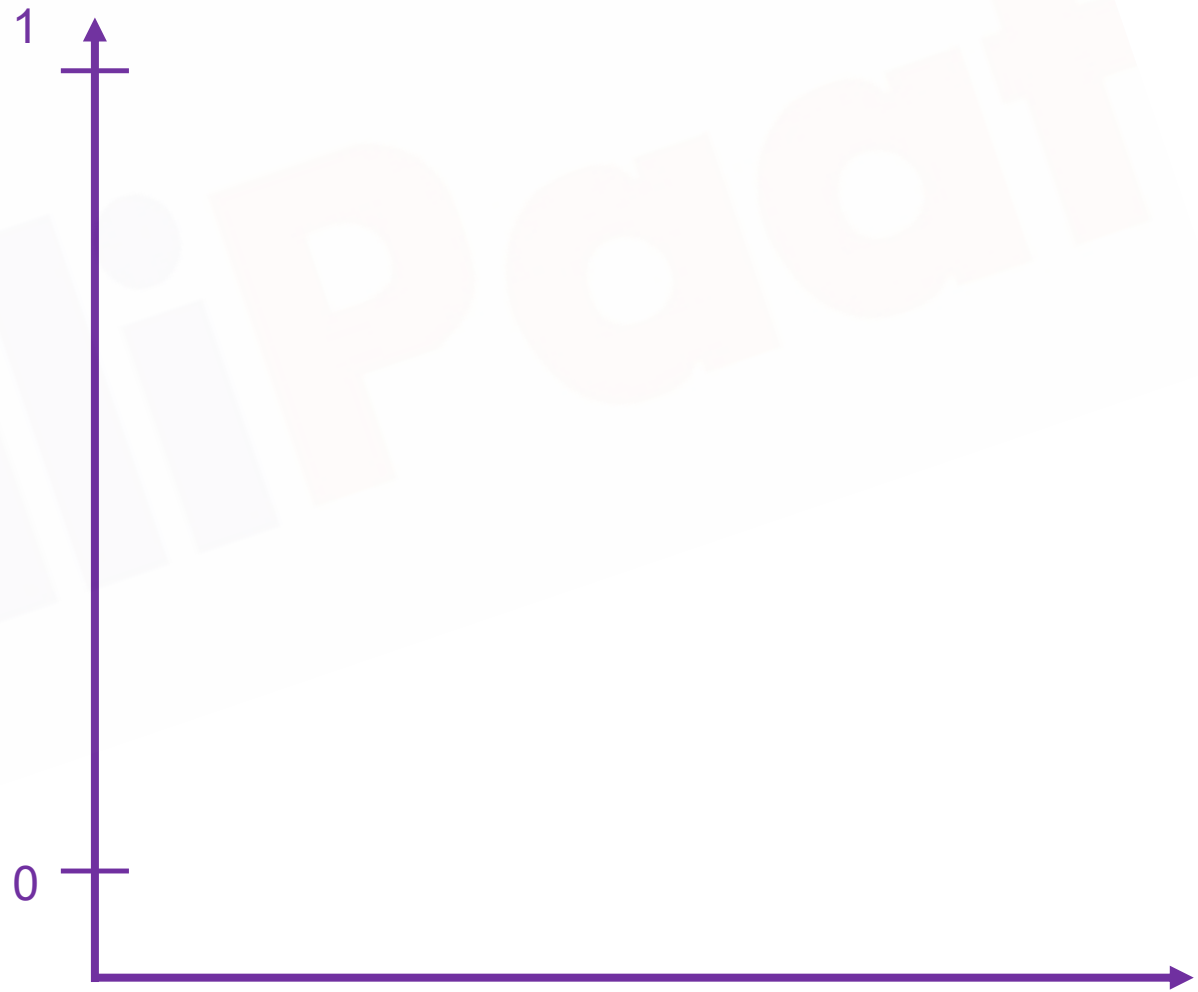
STEP:2

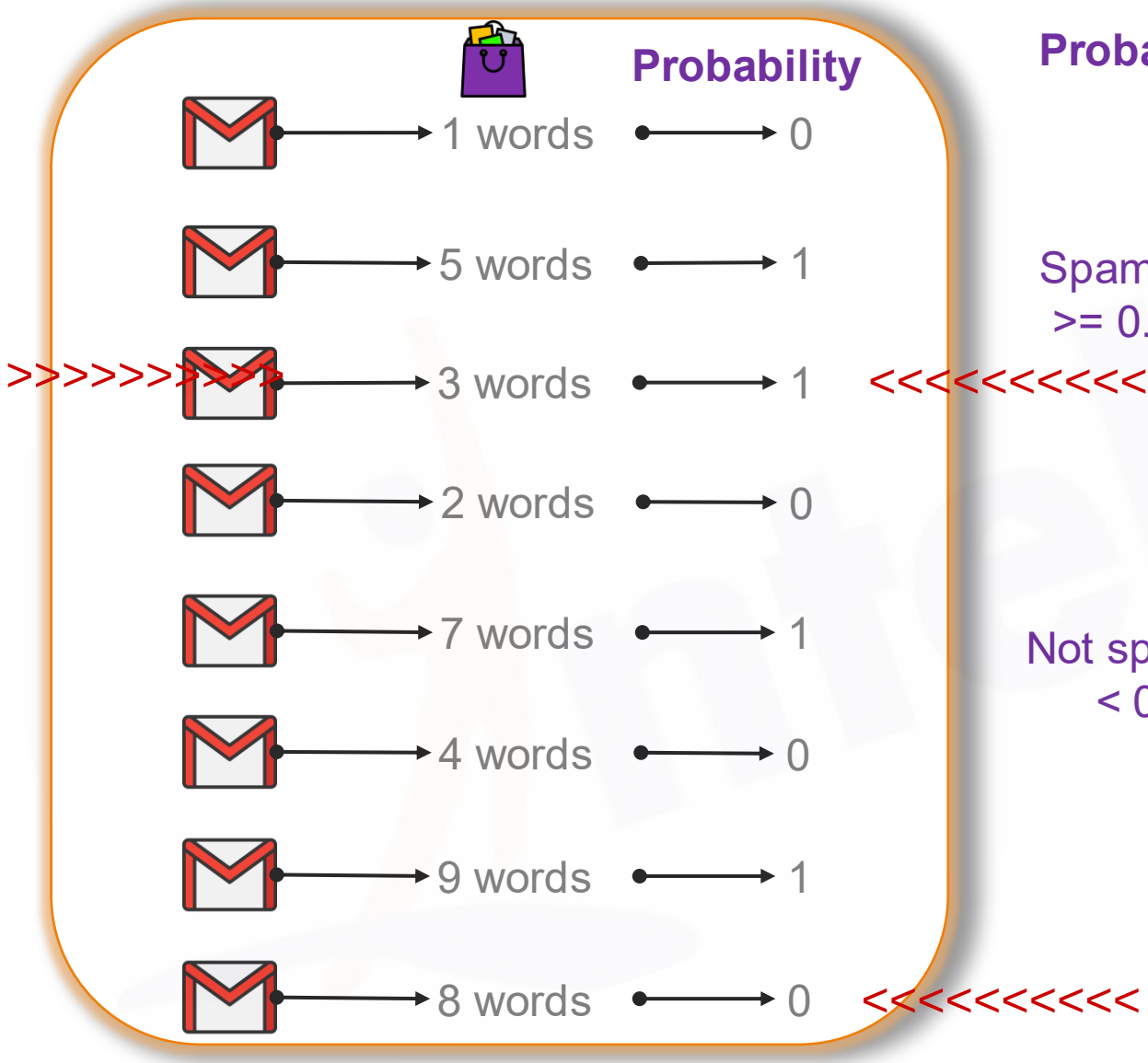
- Define the variable
- **Plot labeled data**
- Draw regression line
- Find out the best using MLE

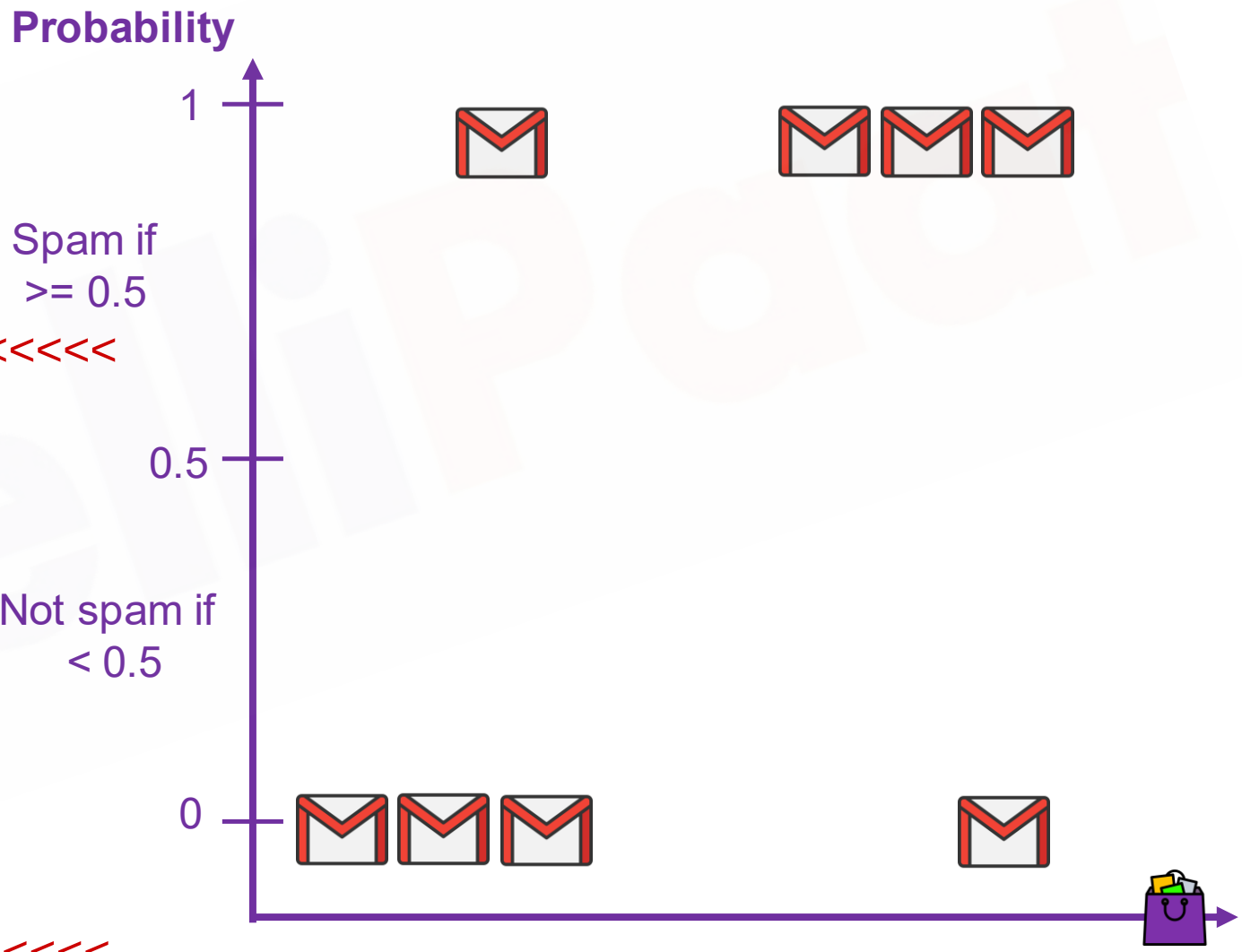


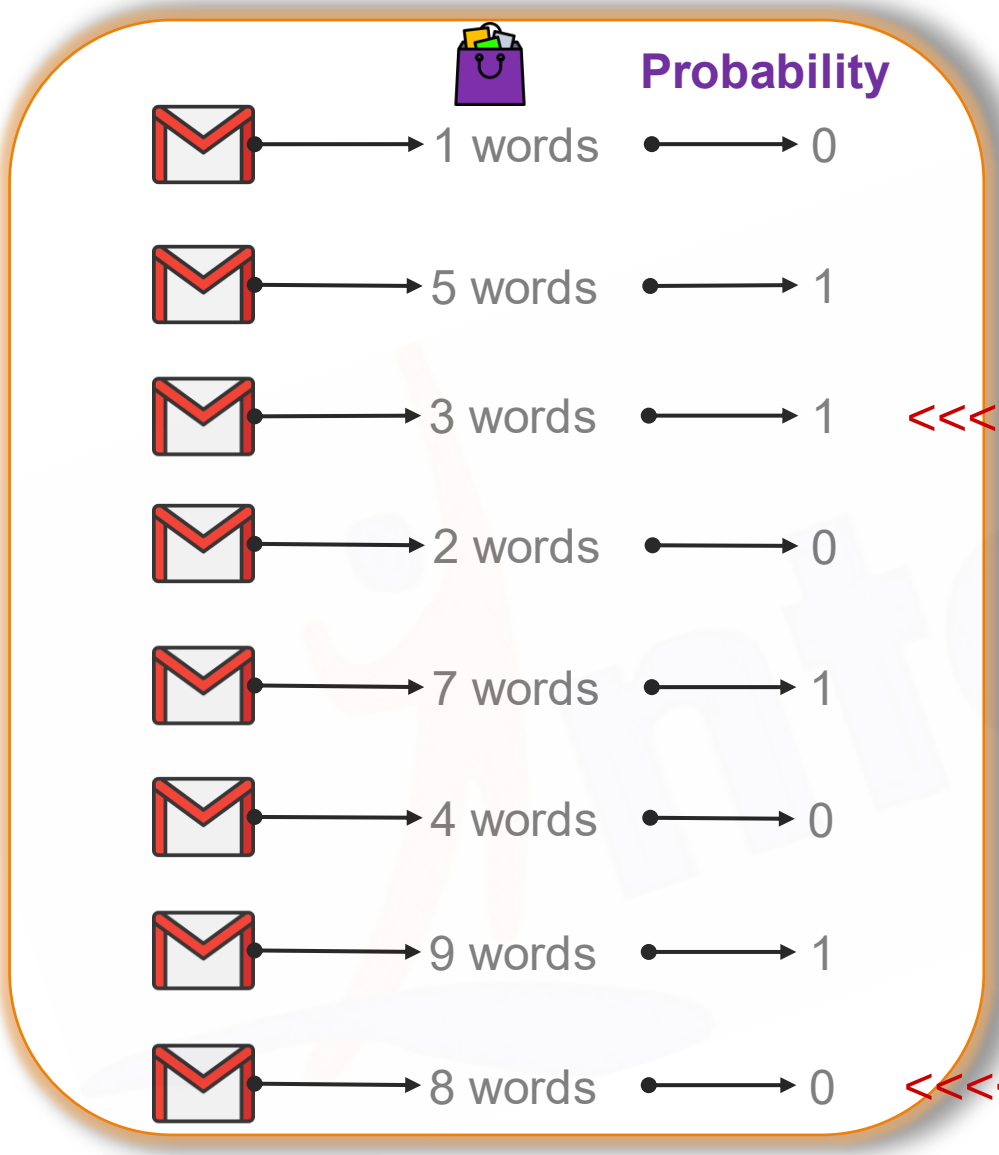
Pre-labeled Dataset

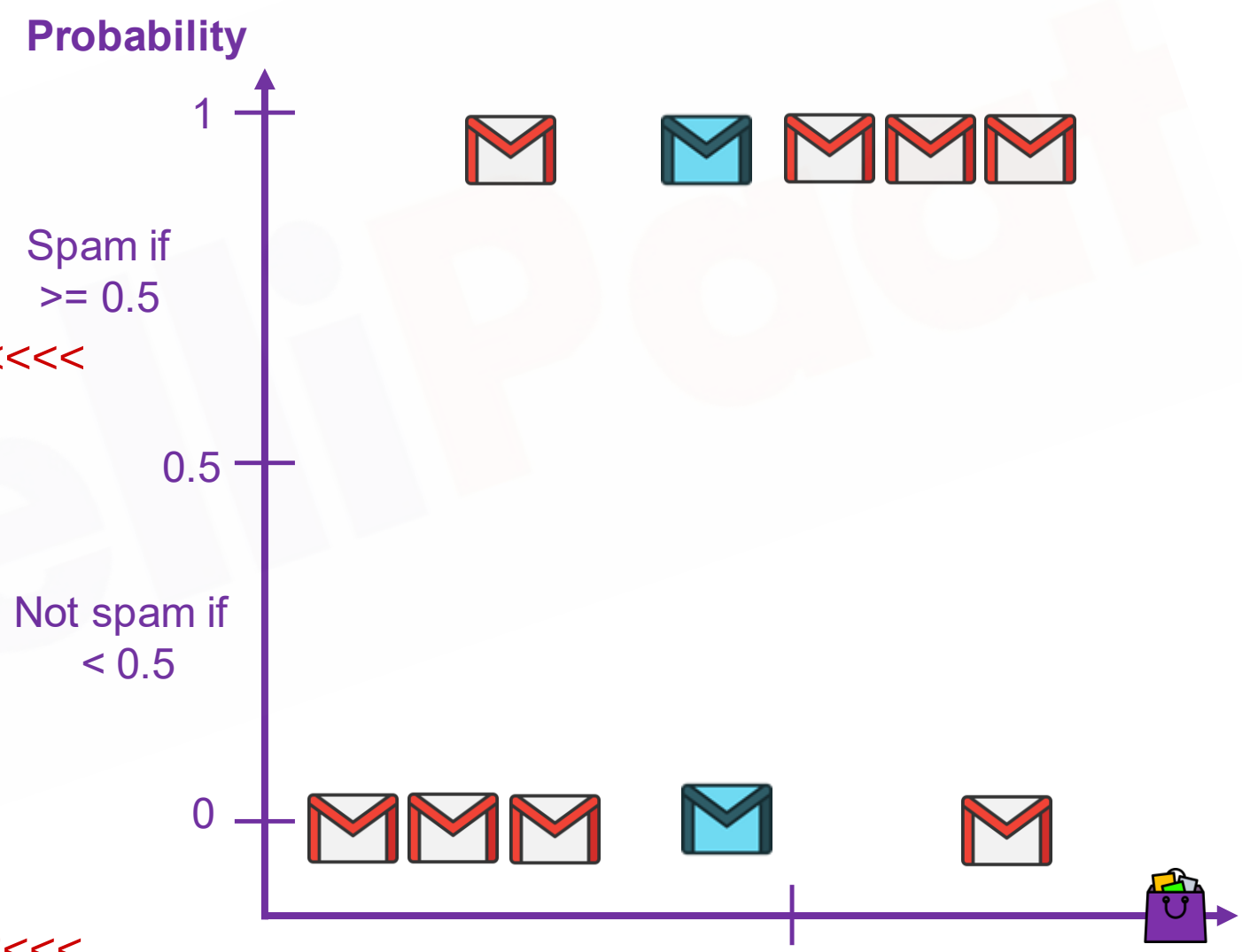

Probability
1 words → 0
5 words → 1
3 words → 1 <<<<<<<<
2 words → 0
7 words → 1
4 words → 0
9 words → 1
8 words → 0 <<<<<<<<


1
0


Probability
1 words → 0
5 words → 1
3 words → 1
2 words → 0
7 words → 1
4 words → 0
9 words → 1
8 words → 0

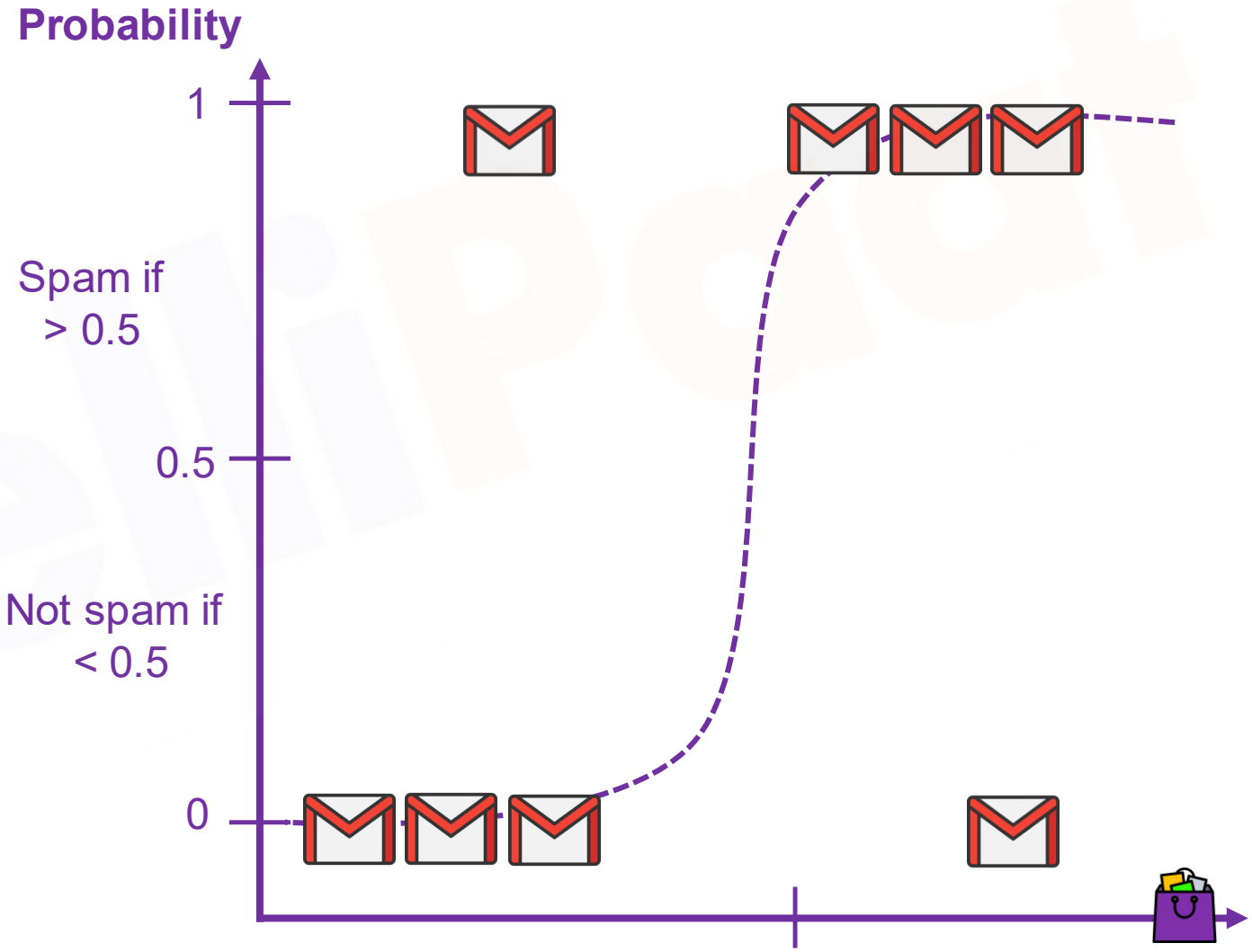

Probability
1
0.5
0
Spam if
>= 0.5
Not spam if
< 0.5


Probability
1 words → 0
5 words → 1
3 words → 1
2 words → 0
7 words → 1
4 words → 0
9 words → 1
8 words → 0

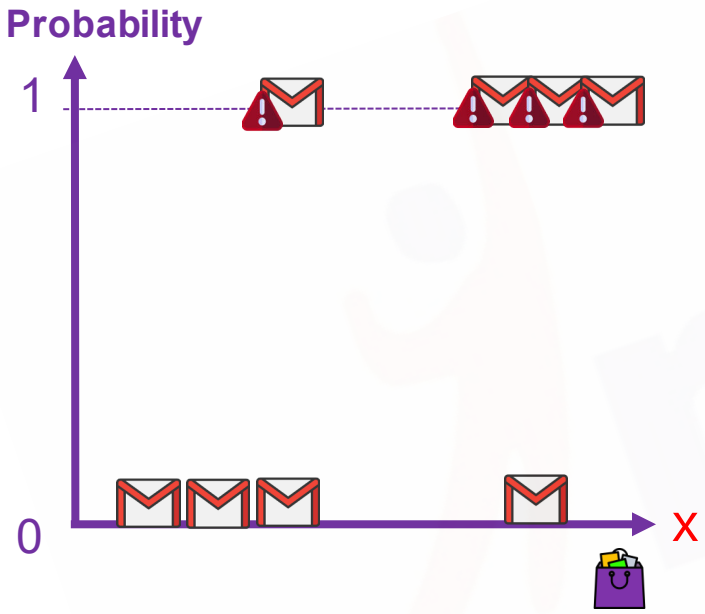

Probability
1
0.5
0
Spam if >= 0.5
Not spam if < 0.5

- Need to find a regression curve which would be the best fit
- That would be our logistic regression curve

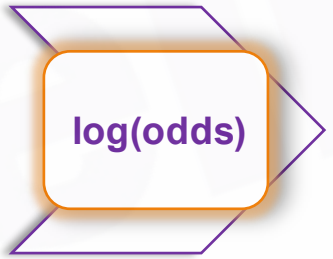
But how do we find the best fit?



How?



log (odds) are linear in the input x.



We don't start with p, we start with a **linear model**:

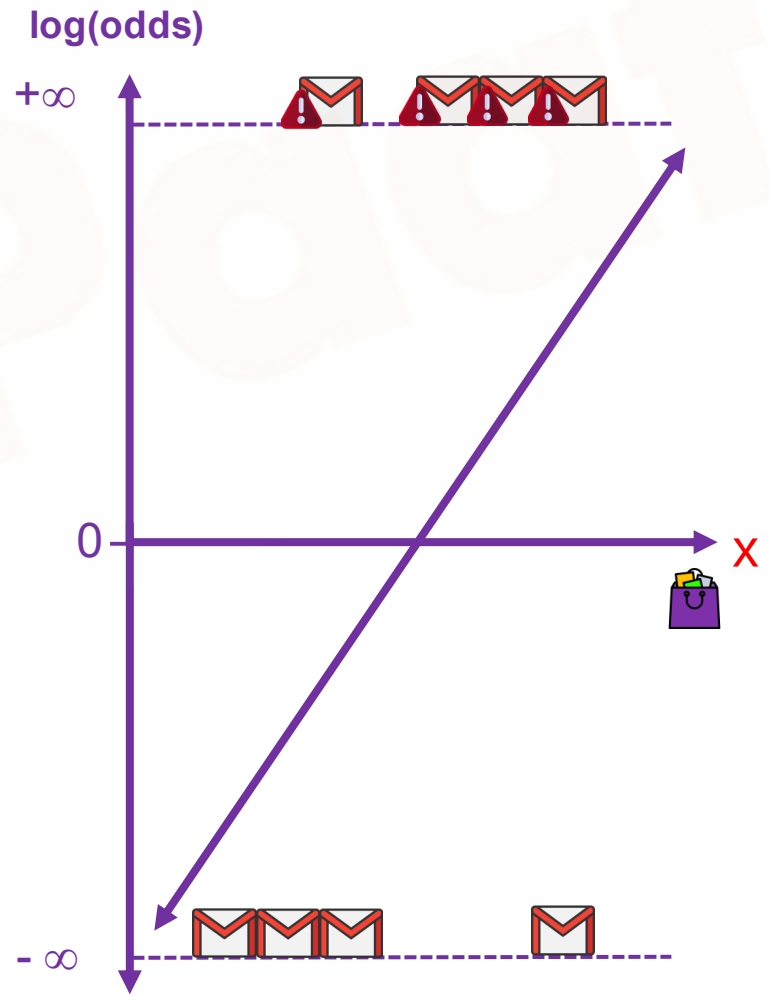
$$z = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

Then define:

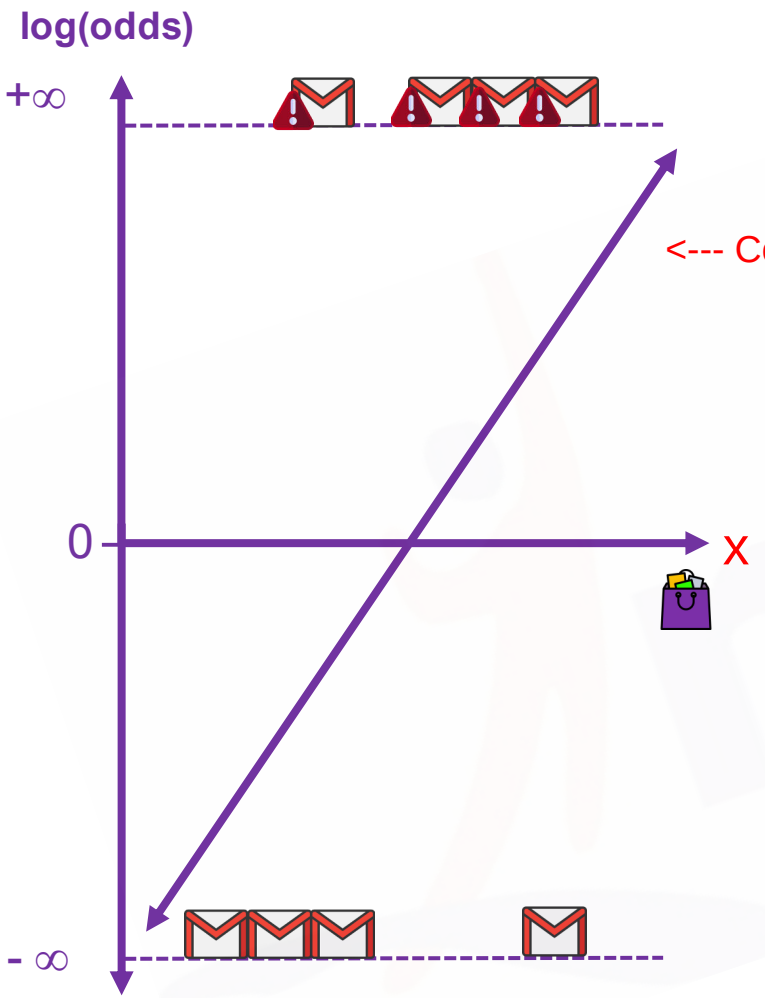
$$p = \frac{1}{1 + e^{-z}}$$

This means:

$$\log\left(\frac{p}{1-p}\right) = z = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$



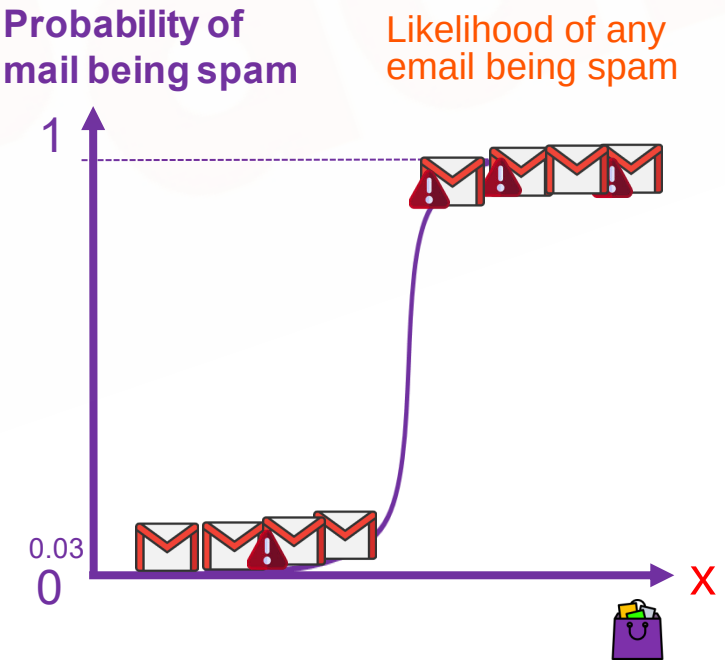
How?



<--- Compute log (odds) for input x.

Sigmoid Function

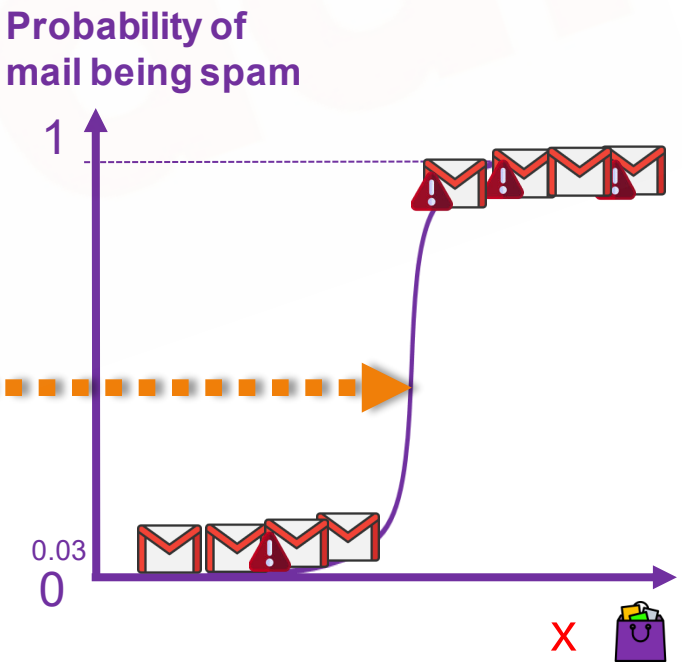
Feed log (odds) to sigmoid to get probability -->



How?

Individual likelihood

log(likelihood)



STEP:3

- Define the variable
- Plot labeled data
- Draw regression line
- Find out the best using MLE

What does log(odds) mean?

IntelliPaat

STEP:3

- Define the variable
- Plot labeled data
- Draw regression line
- Find out the best using MLE

Probability vs odds

- This guy went fishing 5 times a week
- Caught a fish 2 times
- Failed to catch 3 times.

What is the probability and odds of getting a Fish for dinner?

$$\text{Probability} = \frac{\text{Chances for}}{\text{Total chances}} = \frac{2}{5}$$

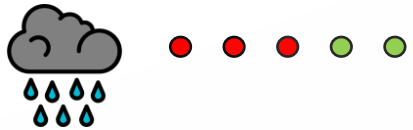
$$\text{Odds} = \frac{\text{Chances for}}{\text{Chances Against}} = \frac{2}{3}$$

IntelliPaat

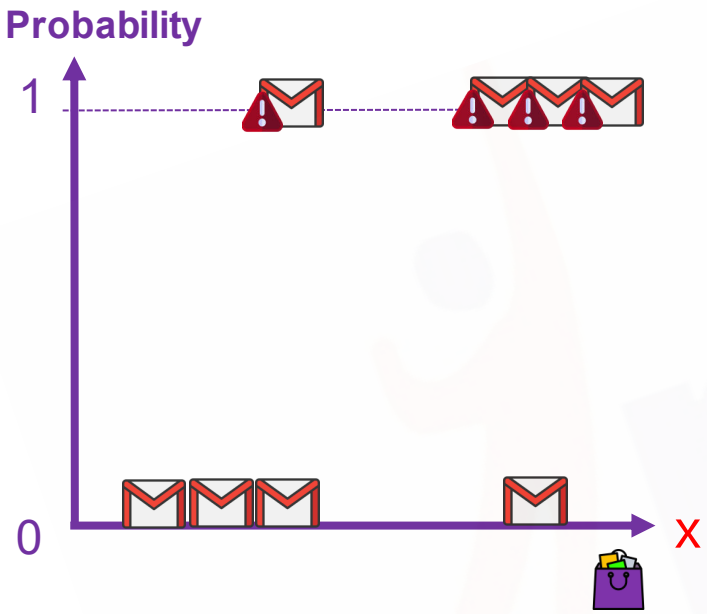
STEP:3

- Define the variable
- Plot labeled data
- Draw regression line
- Find out the best using MLE

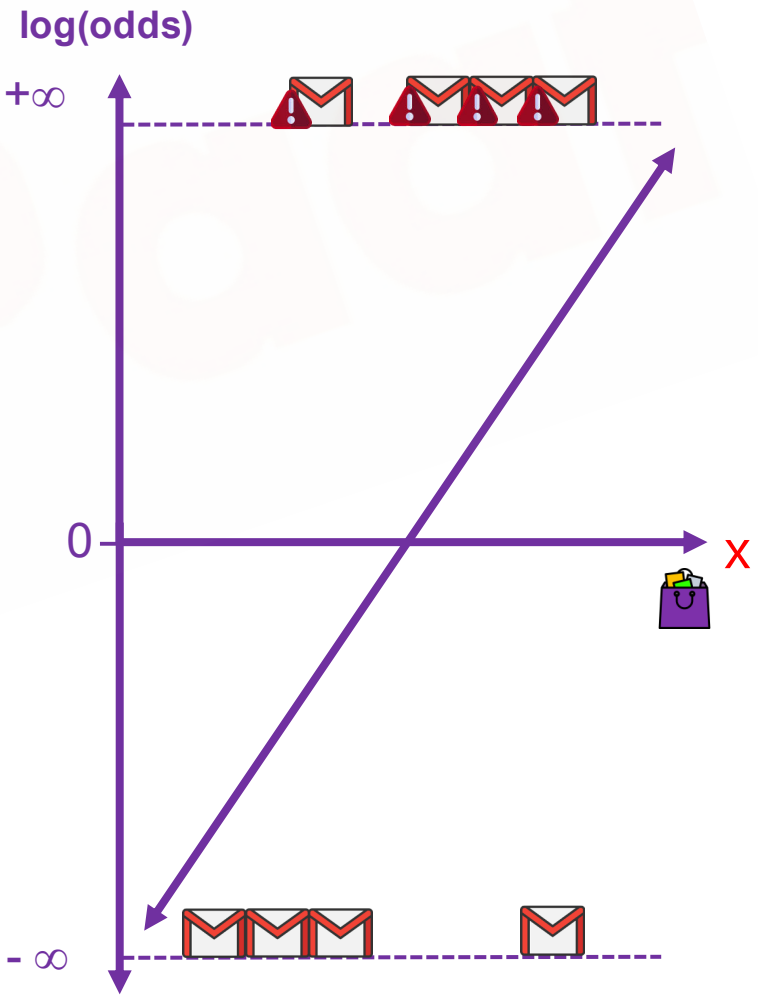
Log(odds) and log(odds ratio)

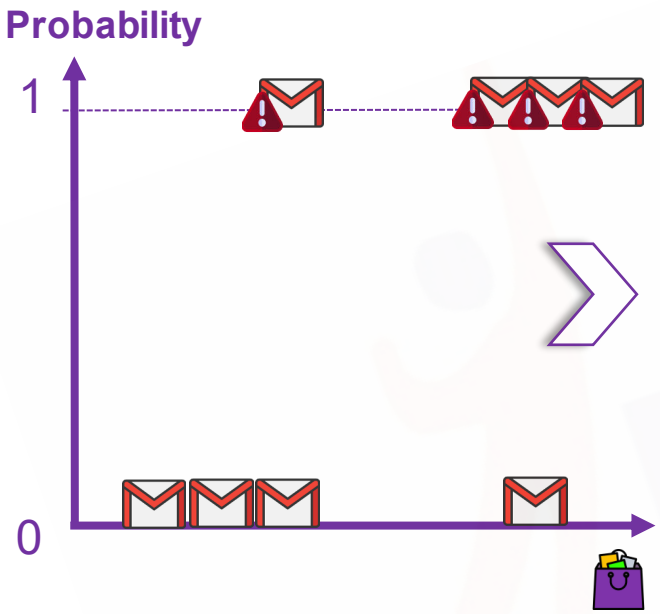


- Log(odds)=Logit Function
- *Note: Odds ≠ Odds Ratio*
- Odds of catching on sunny day= $\frac{2}{3}$
- Odds of catching on rainy day= $\frac{3}{2}$
- Log(odds of catching on sunny day) =log$\left(\frac{2}{3}\right)$
- Log(odds of catching on rainy day)=log$(\frac{3}{2})$
- Log(odds ratio)= Log$(\frac{odds\ on\ rainy\ day}{odds\ on\ sunny\ day})$=log $(\frac{\frac{2}{3}}{\frac{3}{2}})$ =log(0.44)



log(odds)





$$\log(\text{odds}) = \log\left(\frac{P(Spam)}{1-P(Spam)}\right)$$

$$\log(\text{odds}) = \log\left(\frac{1}{1-1}\right)$$

$$\log(\text{odds}) = \log\left(\frac{1}{0}\right) \rightarrow +\infty$$

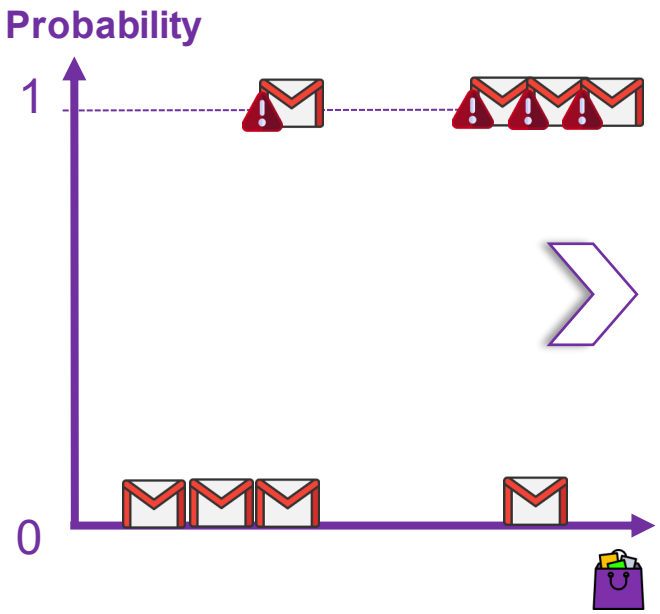
$\log_b 0 = c$

$\Rightarrow 0 = b^c$-------(1)

So, for equation(1) to be true.

If b<1 =>c → + ∞

If b>1 =>c → + ∞

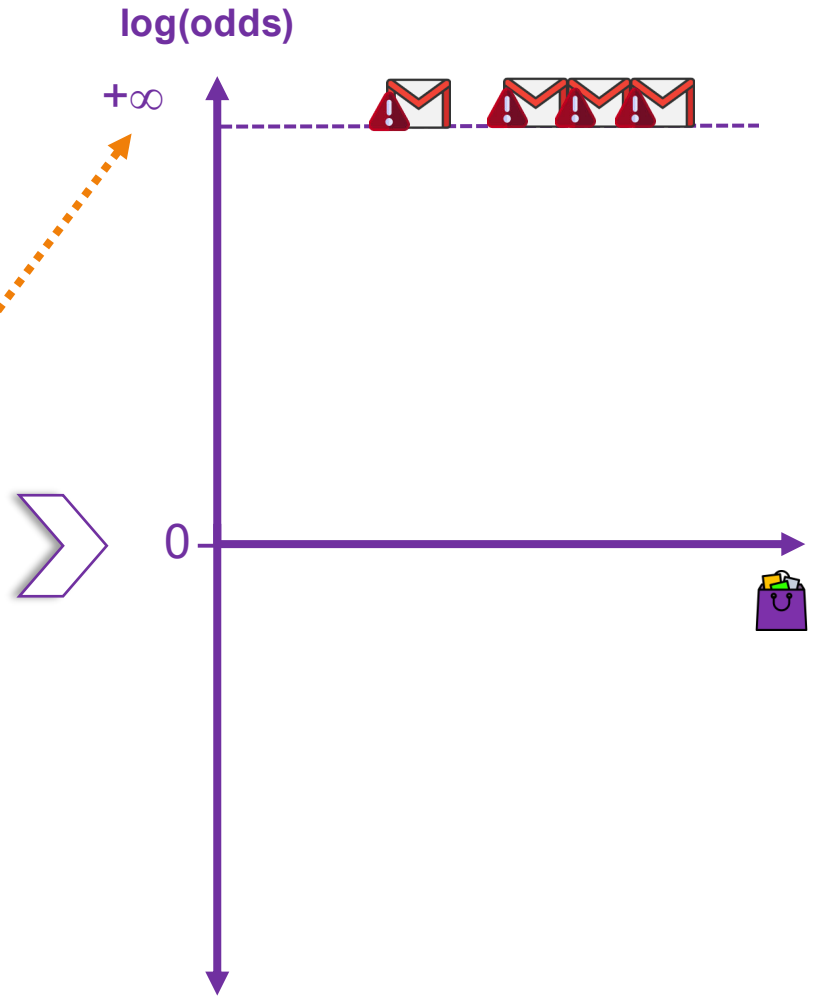


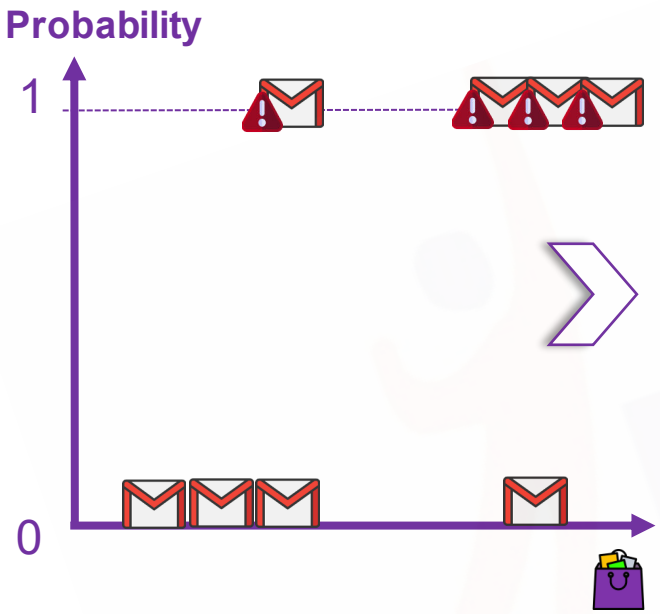
$$\log(odds) = \log\left(\frac{P(Spam)}{1-P(Spam)}\right)$$

$$\log(odds) = \log\left(\frac{1}{0}\right)$$

$$\log(odds) = \log(1) - \log(0)$$

$$\log(odds) = 0 - (-\infty) \rightarrow +\infty$$



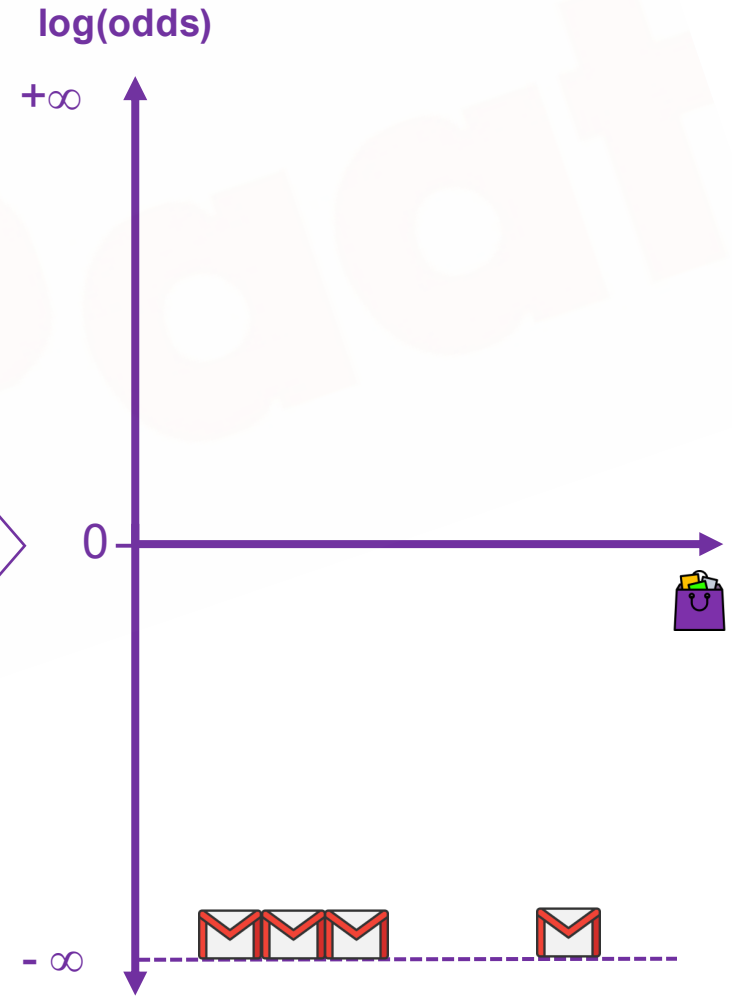


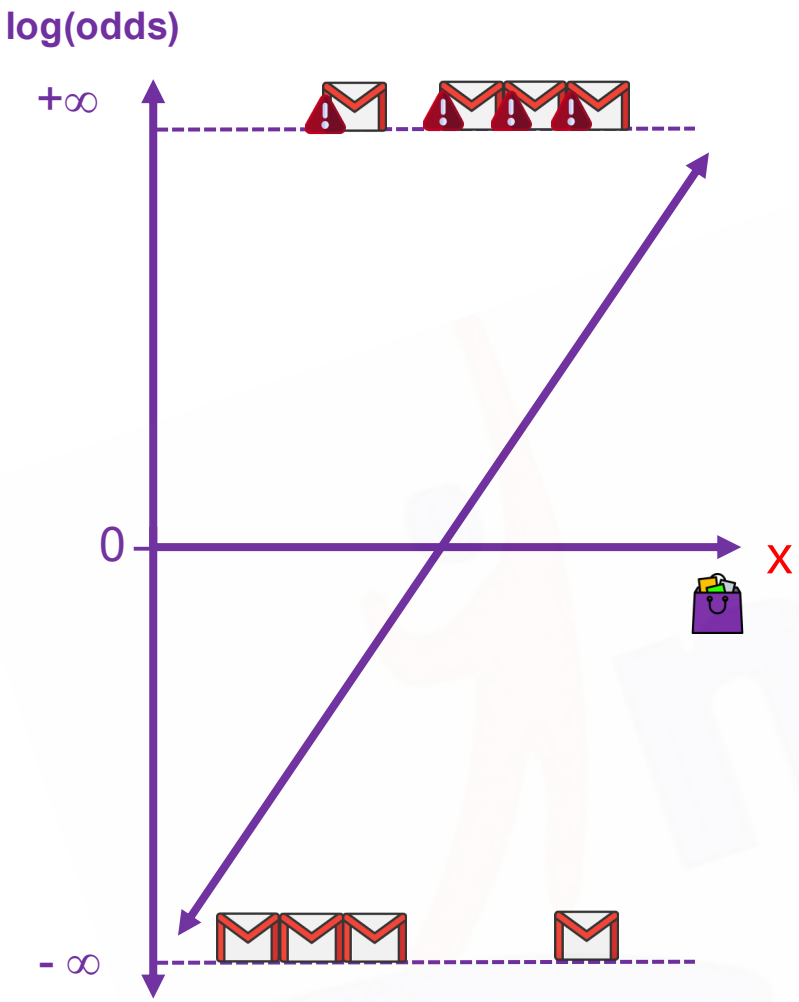
$$\log(odds) = \log\left(\frac{P(\cancel{Not}\ Spam)}{1-P(\cancel{Not}\ Spam)}\right)$$

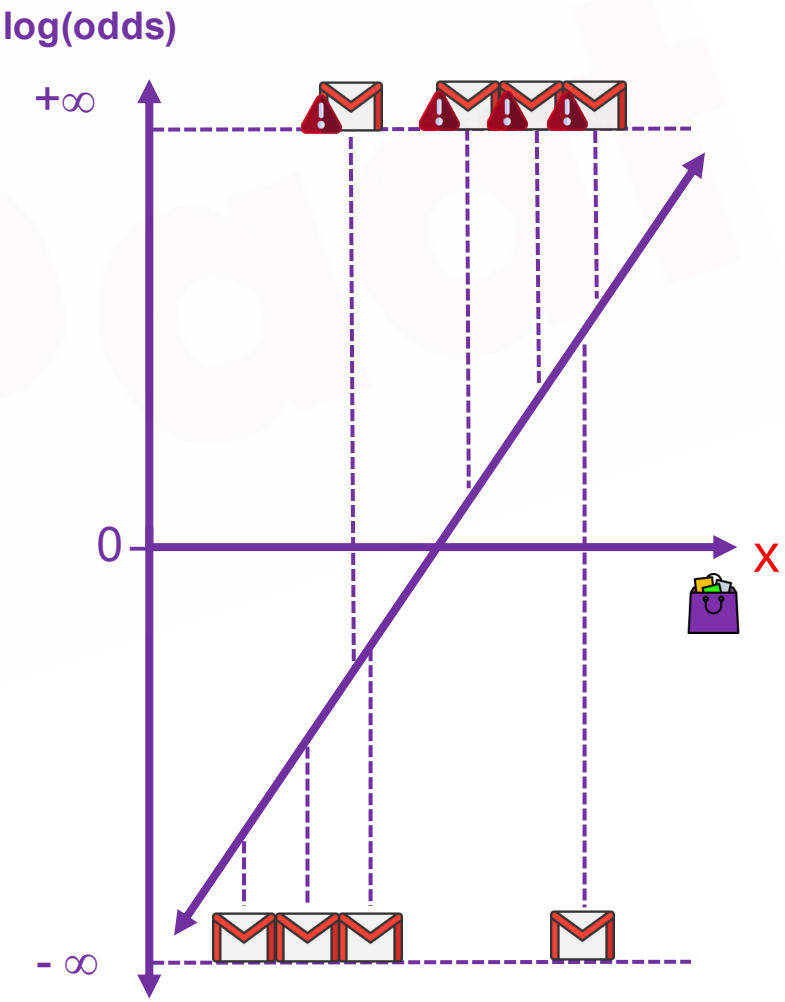
$$\log(odds) = \log\left(\frac{0}{1-0}\right)$$

$$\log(odds) = \log\left(\frac{0}{1}\right)$$

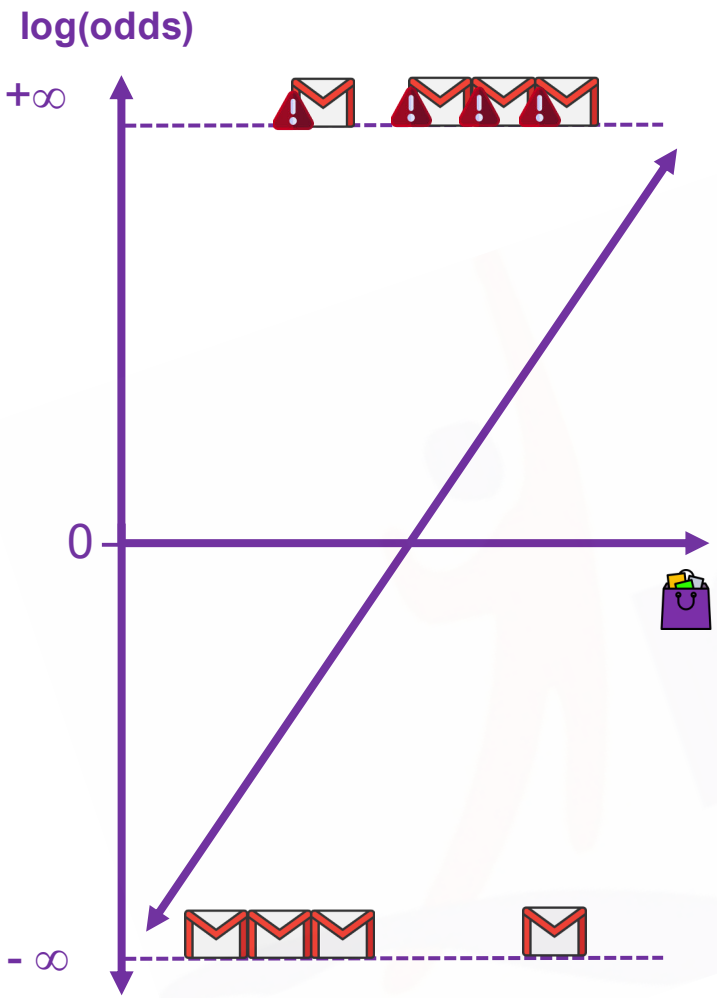
$$\log(odds) = \log(0) - \log(1) \rightarrow -\infty$$



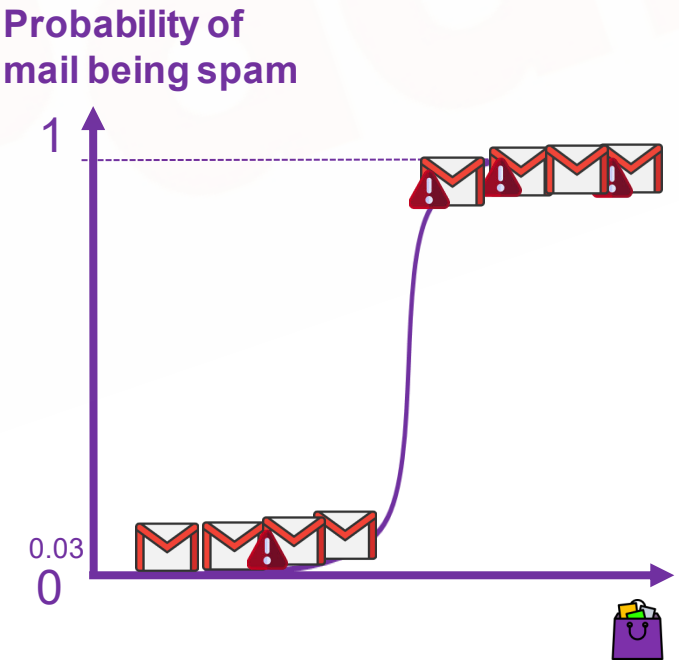

log(odds)
+∞
0
X
- ∞


log(odds)
+∞
0
X
- ∞

How?



Sigmoid Function



IntelliPaat

STEP:3

- Define the variable
- Plot labeled data
- Draw regression line
- Find out the best using MLE

What does Sigmoid Function mean?

IntelliPaat

STEP:3

- Define the variable
- Plot labeled data
- Draw regression line
- Find out the best using MLE

Sigmoid Function:

Sigmoid function is the standard logistic function

$$\text{Logistic function} = \frac{L(e^{k(x-x_0)})}{1+e^{k(x-x_0)}}$$

Here,

L – Curve's maximum value
k – Steepness of the curve
x0 – x value of Sigmoid midpoint

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

Here,

k=1
x0=0
L=1

IntelliPaat

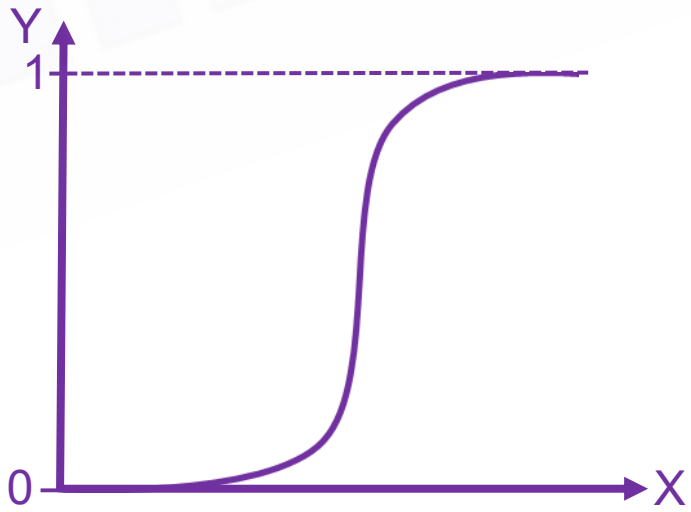
STEP:3

- Define the variable
- Plot labeled data
- **Draw regression line**
- Find out the best using MLE

Sigmoid Function:

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

- S' shaped curve.
- Sigmoid curve has a finite limit of:
  - '0' as x approaches $-\infty$
  - '1' as x approaches $+\infty$

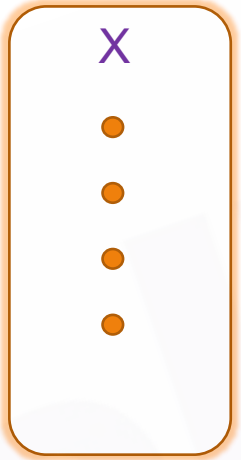


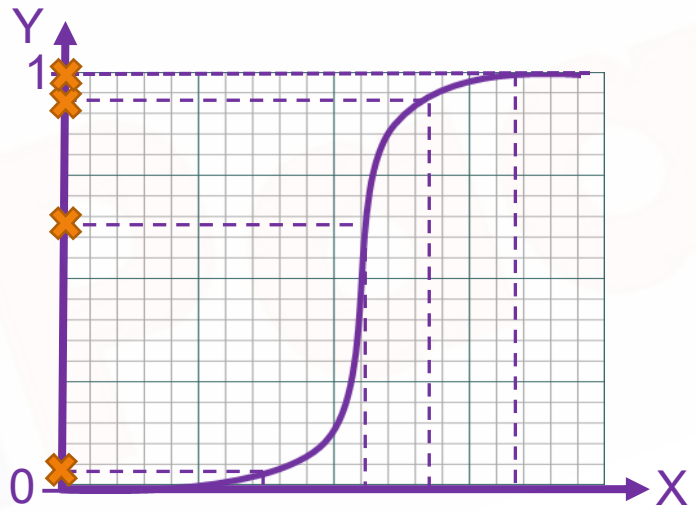
IntelliPaat

STEP:3

- Define the variable
- Plot labeled data
- Draw regression line
- Find out the best using MLE

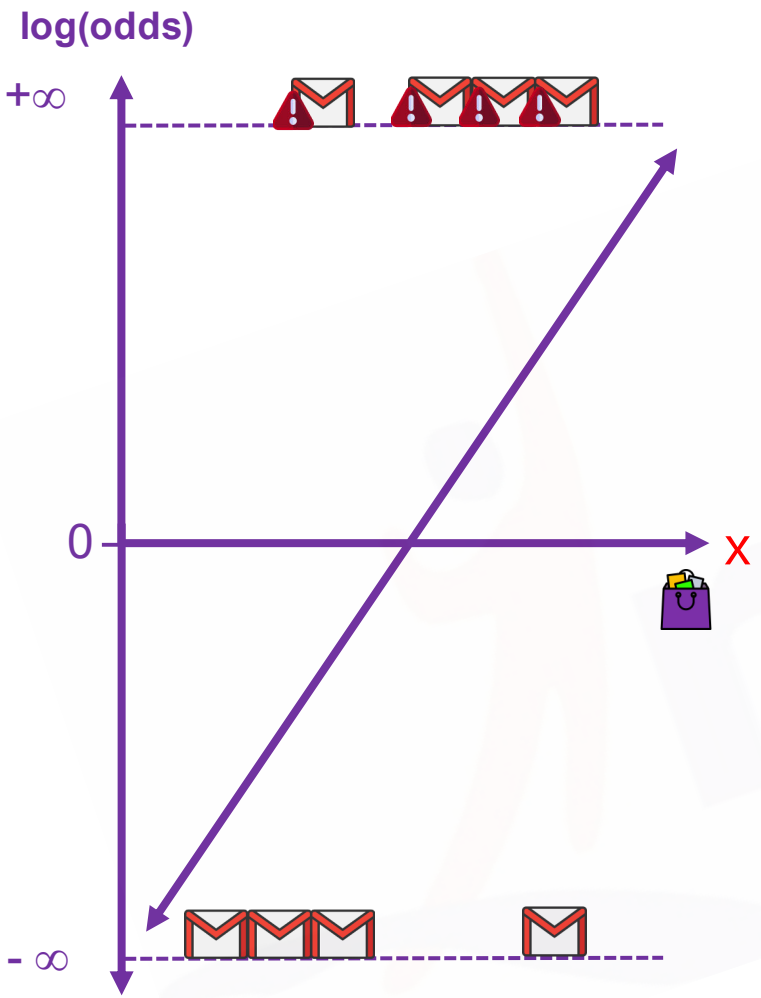
Sigmoid Function:



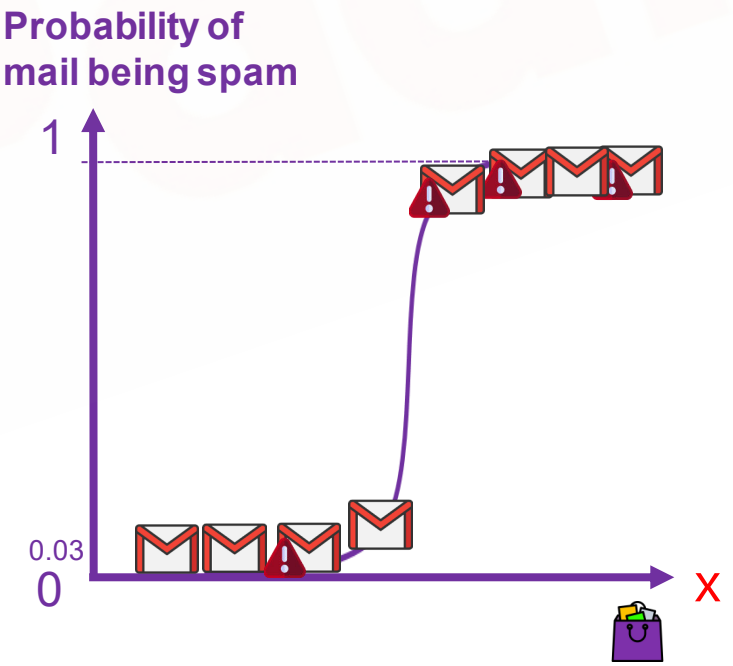


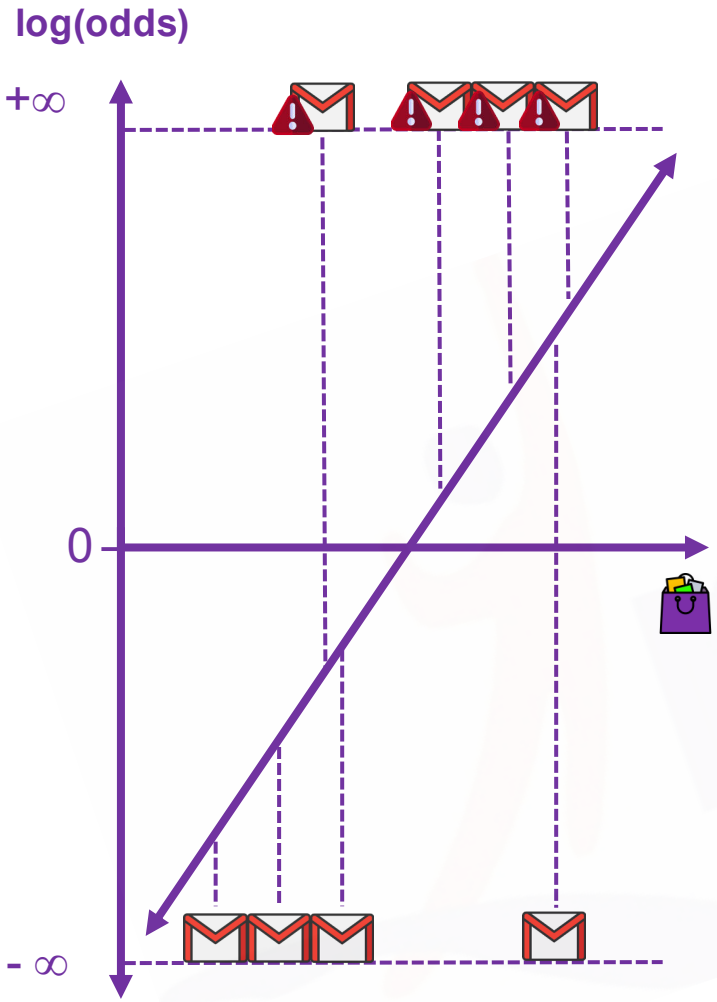
- Takes any real-valued number and maps into a value between 0 and 1.
- Helpful while solving classification problems

How?



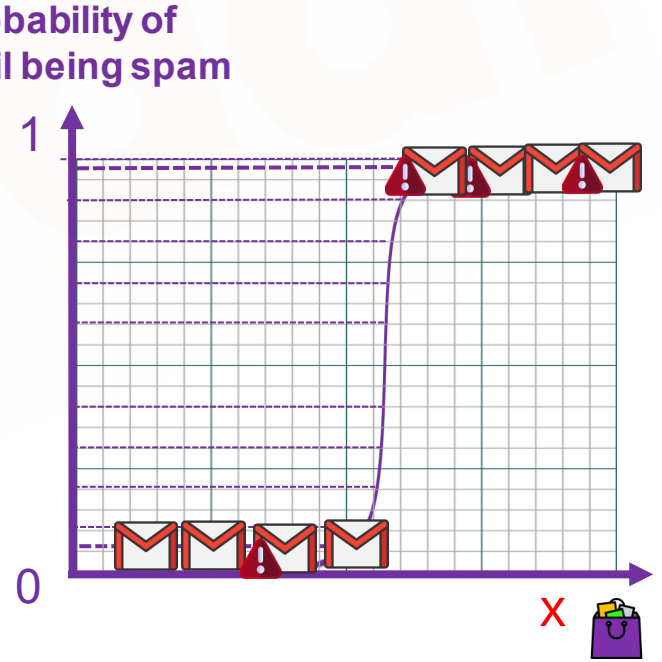
Sigmoid Function

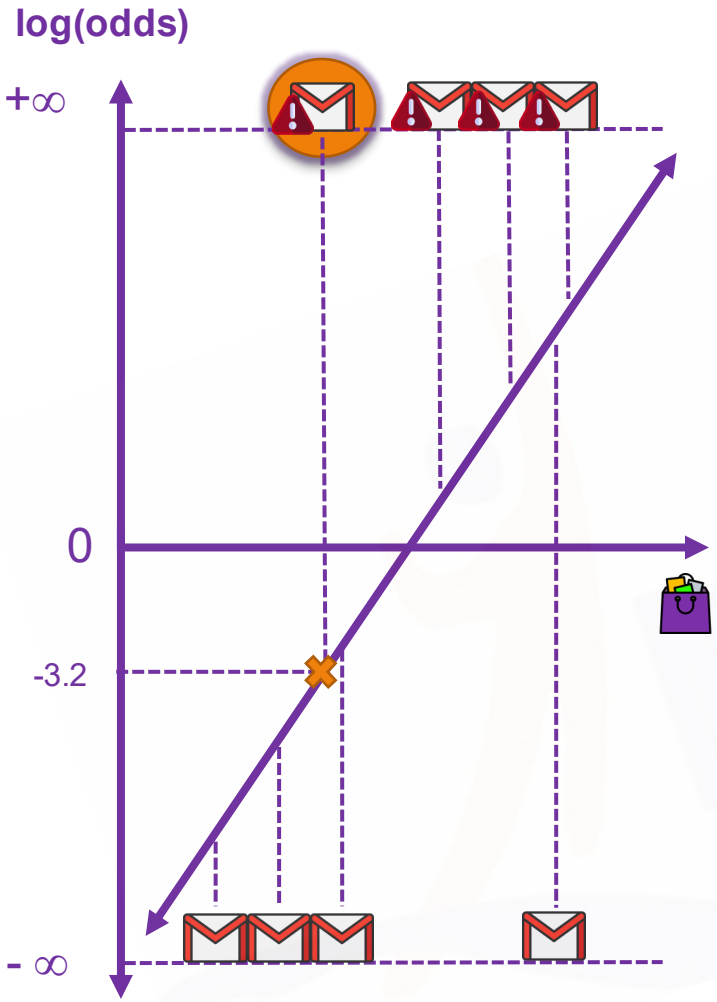




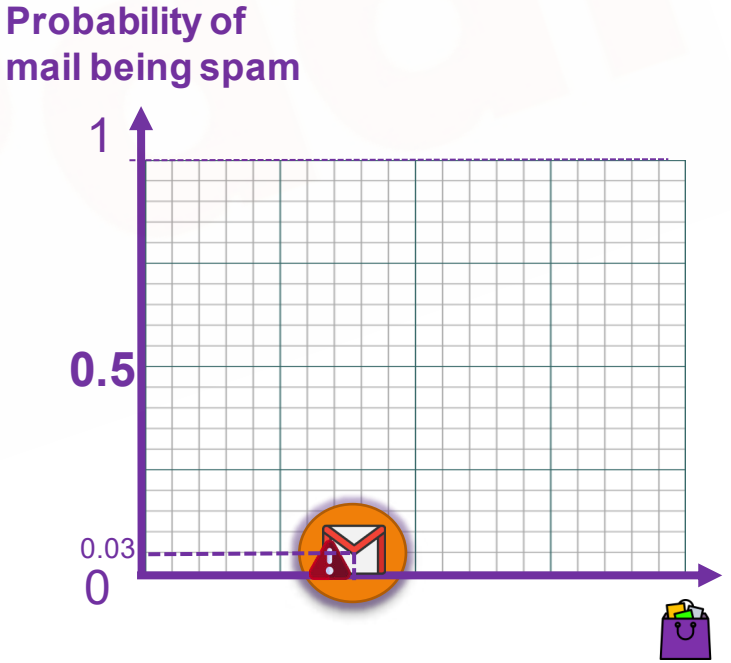
$$P = \frac{e^{\log(odds)}}{1 + e^{\log(odds)}}$$

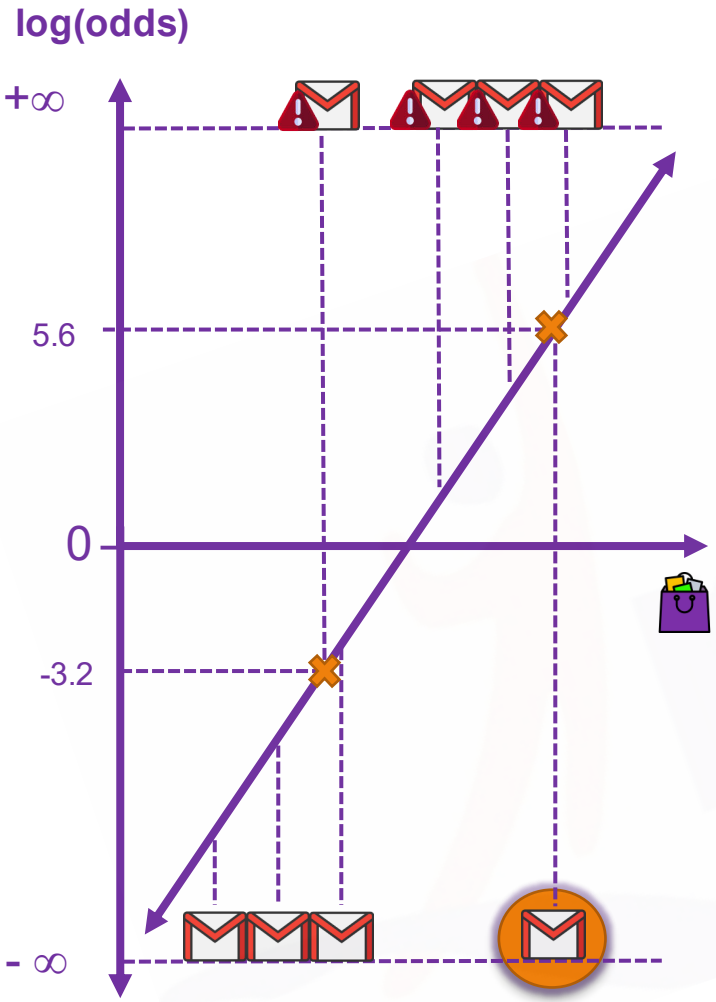
$$\frac{1}{1+e^{-x}} \quad \text{or} \quad \frac{e^{x}}{1+e^{x}}$$



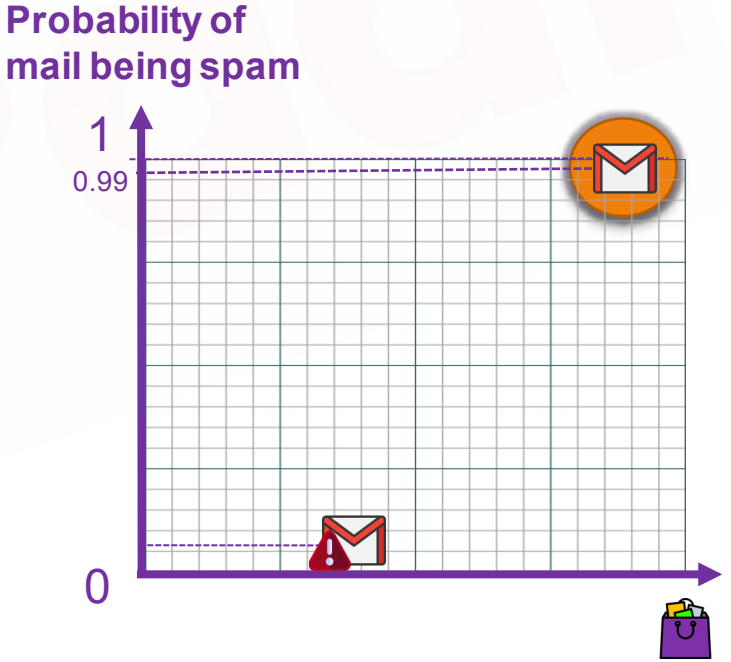


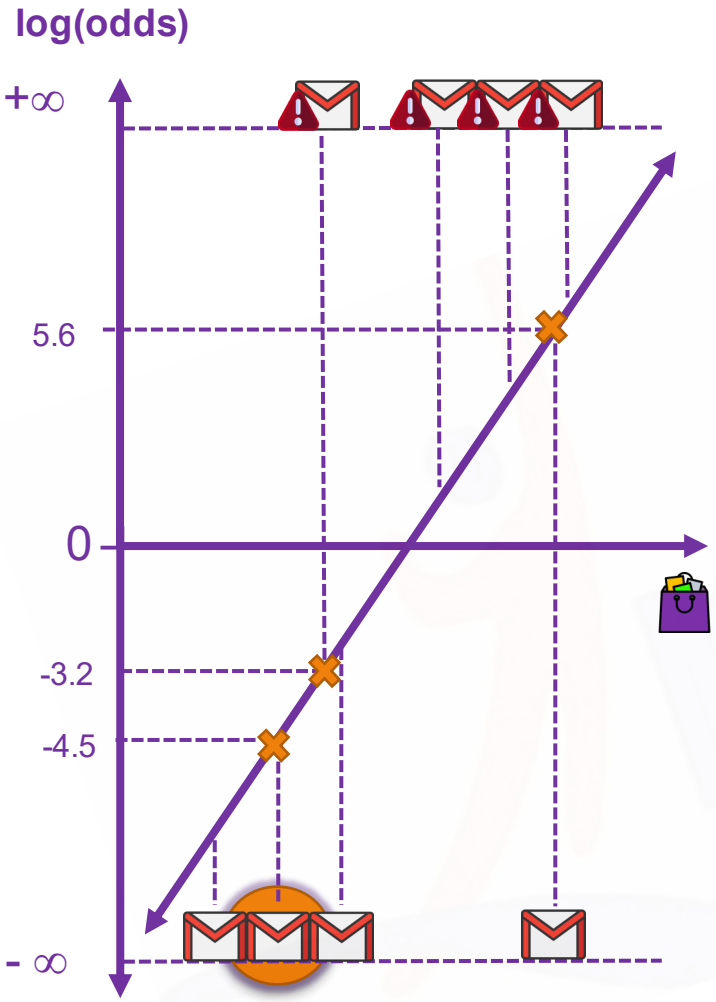
$$P = \frac{e^{\log(-3.2)}}{1+e^{\log(-3.2)}} = 0.03$$



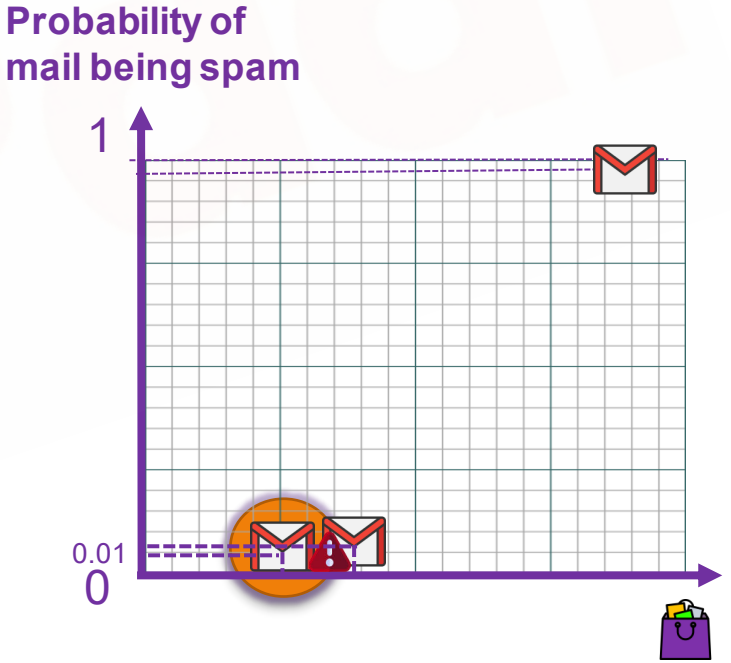


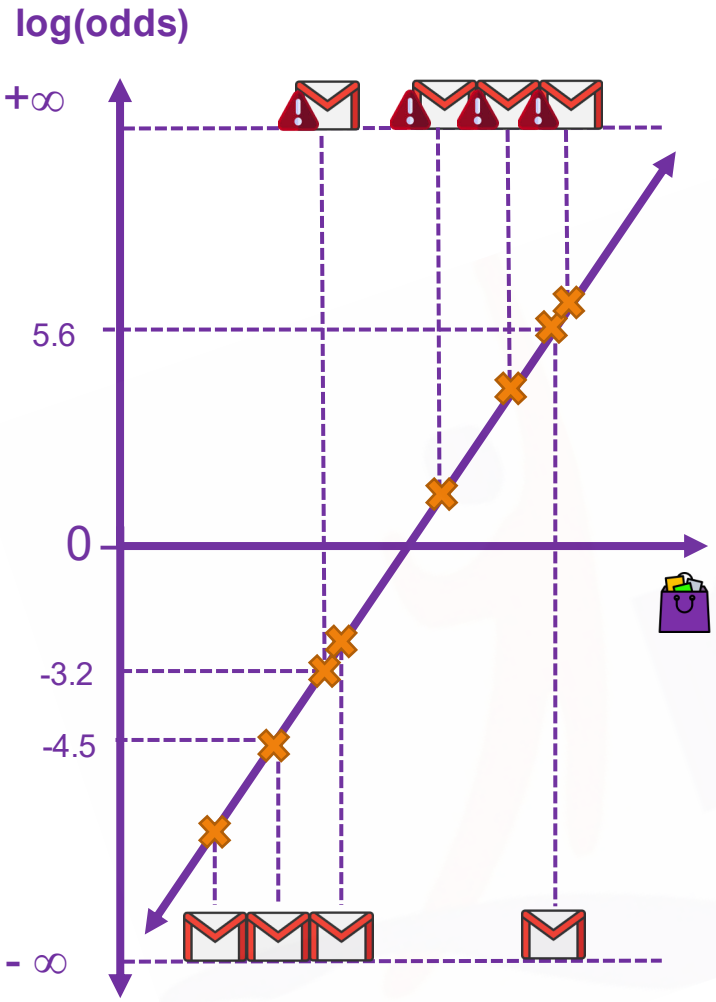
$$P = \frac{e^{\log(5.6)}}{1+e^{\log(5.6)}} = 0.99$$



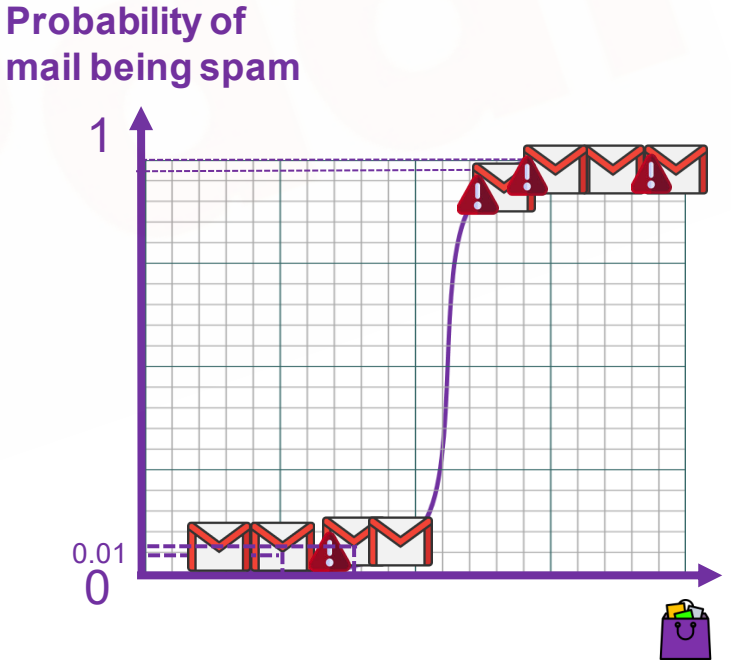


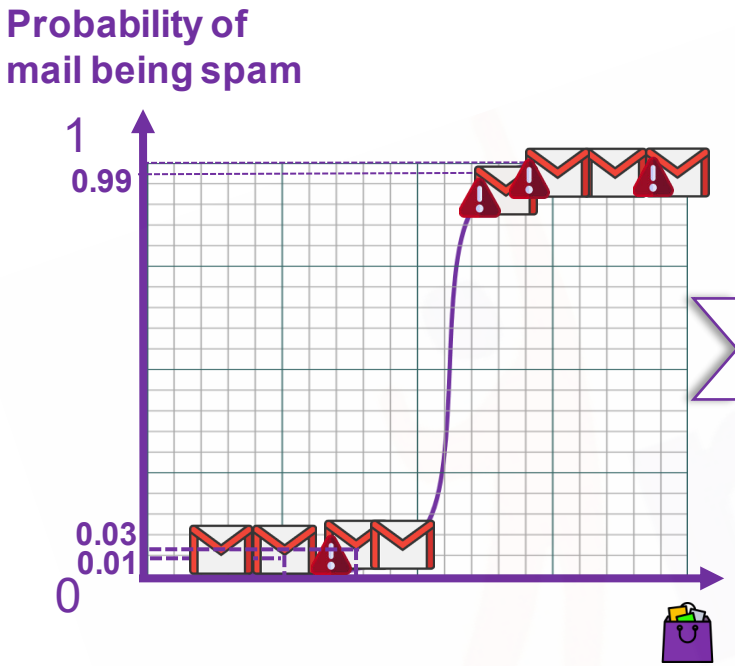
$$P = \frac{e^{\log(-4.5)}}{1+e^{\log(-4.5)}} = 0.01$$





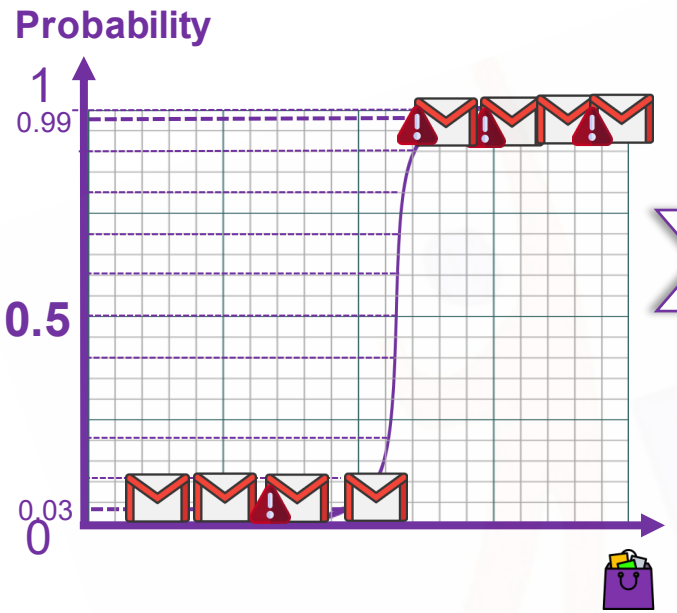
$$P = \frac{e^{\log(odds)}}{1+e^{\log(odds)}}$$





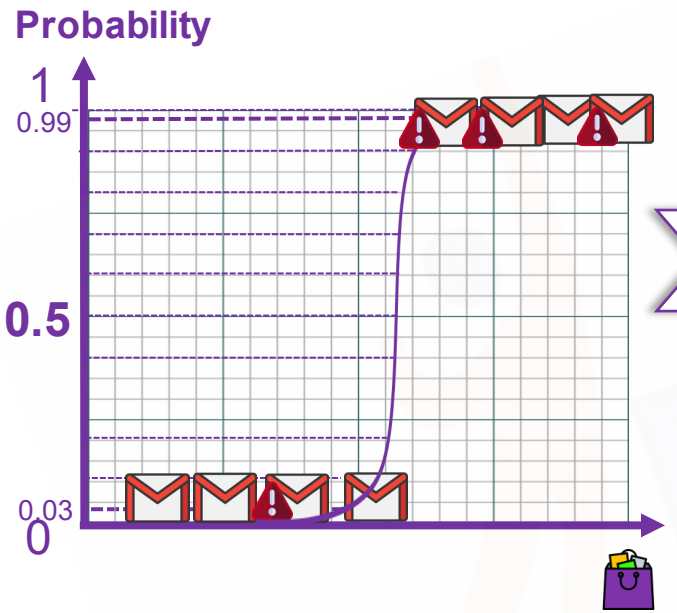
Now let us find out the individual log likelihood of each mail

- Likelihood of 1st mail being spam = 0.01
- Likelihood of 2nd mail being spam = 0.01
- Likelihood of 3rd mail being spam = 0.03
- Likelihood of 4th mail being spam = 0.05
- Likelihood of 5th mail being spam = 0.97
- Likelihood of 6th mail being spam = 0.99
- Likelihood of 7th mail being spam = 0.99
- Likelihood of 8th mail being spam = 0.99



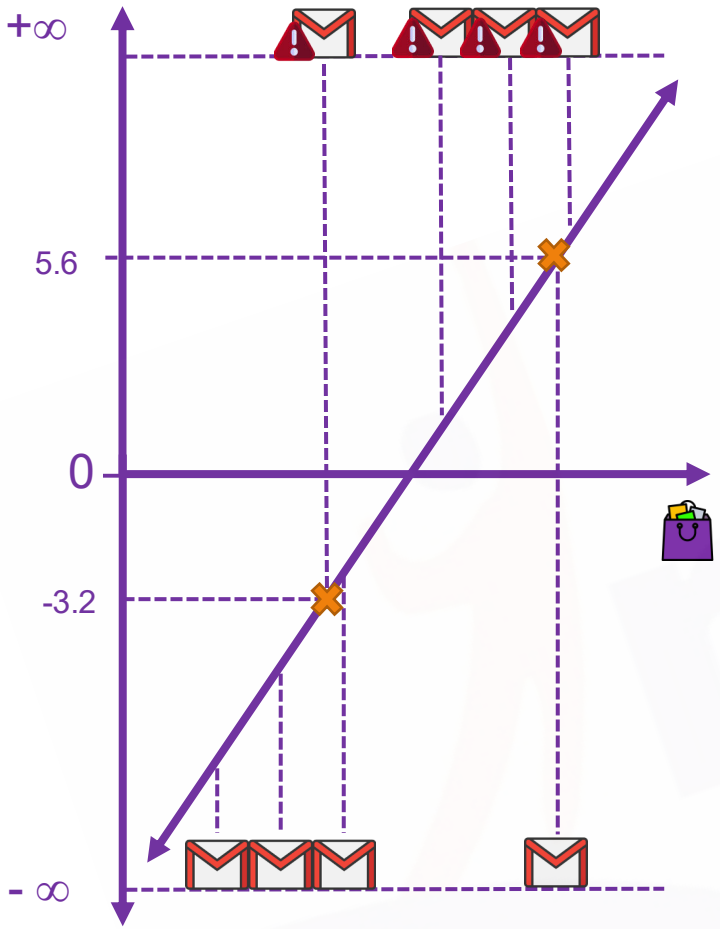
Likelihood of data given this curve =

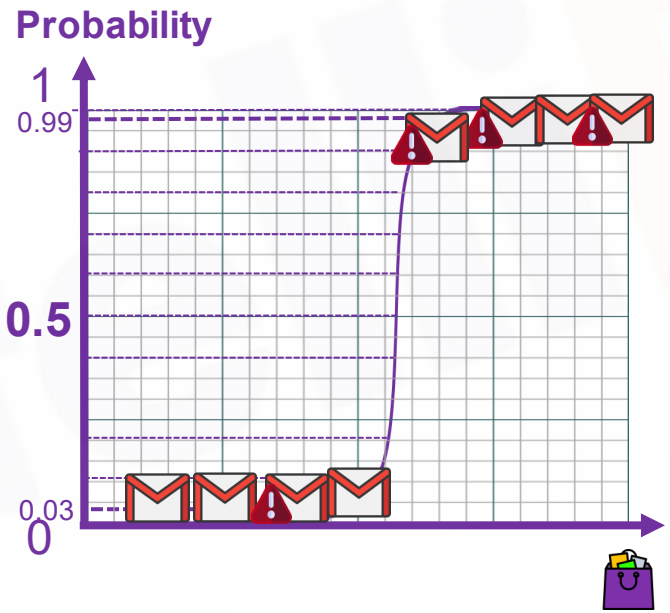
(1-0.01) x (1-0.01) x (1-0.03) x (1-0.05) x 0.97 x 0.99 x 0.99 x 0.99



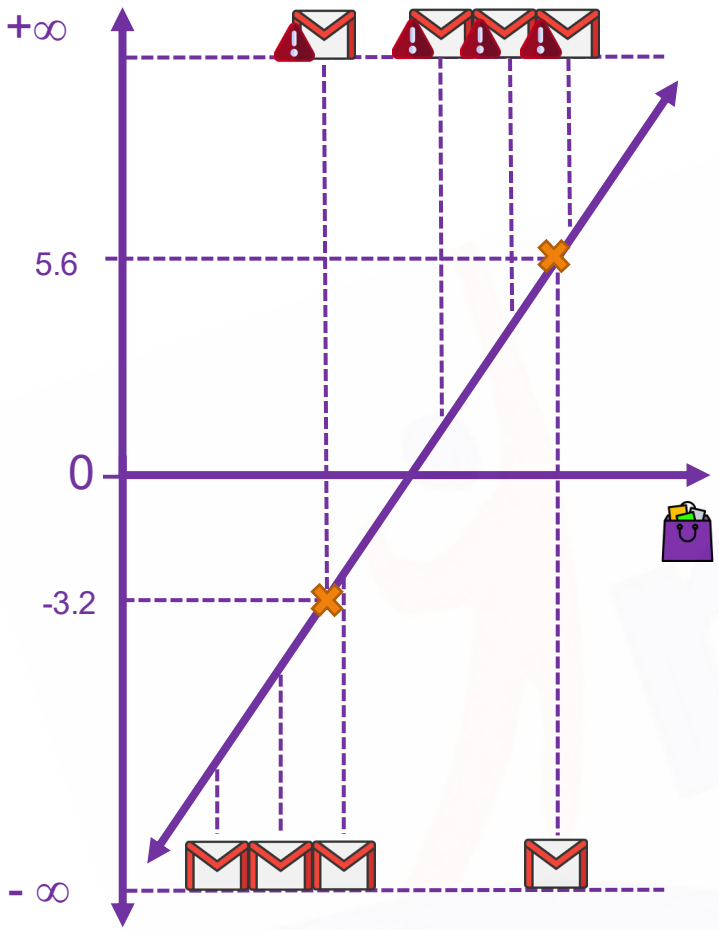
log(Likelihood of data given this curve) =

log(1-0.01) + log(1-0.02)
+ log(1-0.05) + log(1-0.05)
-0.084
+ log(0.97) + log(0.99)
+ log(0.99) + log(0.99)

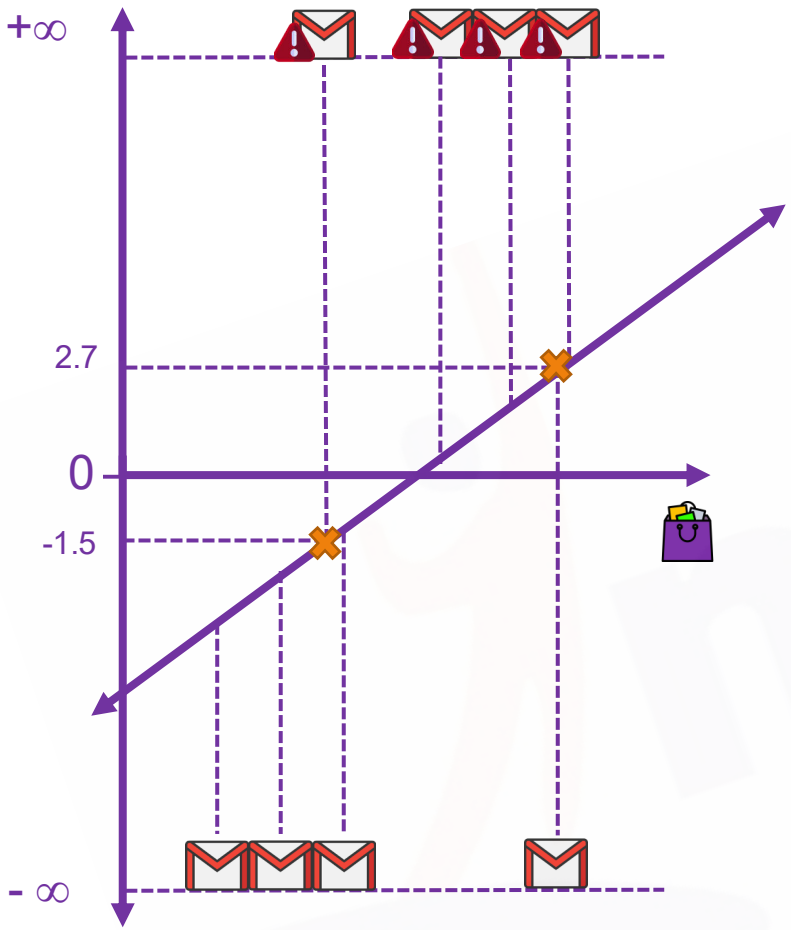


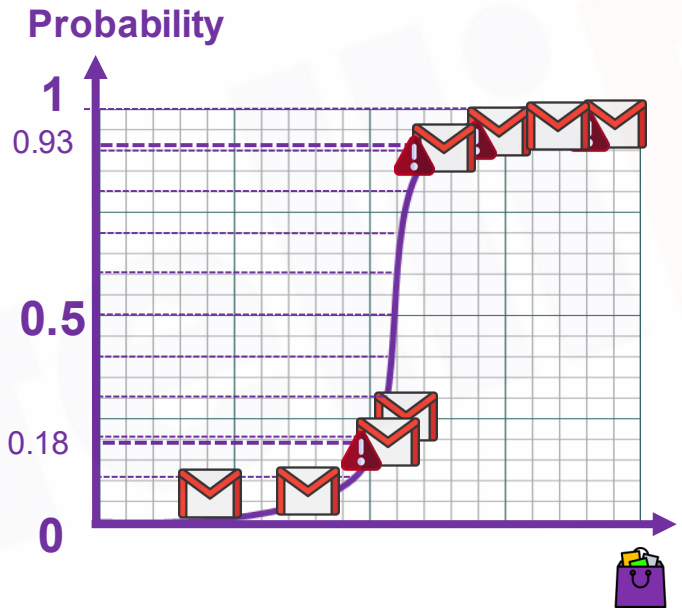


This means that log likelihood of this line is -0.084

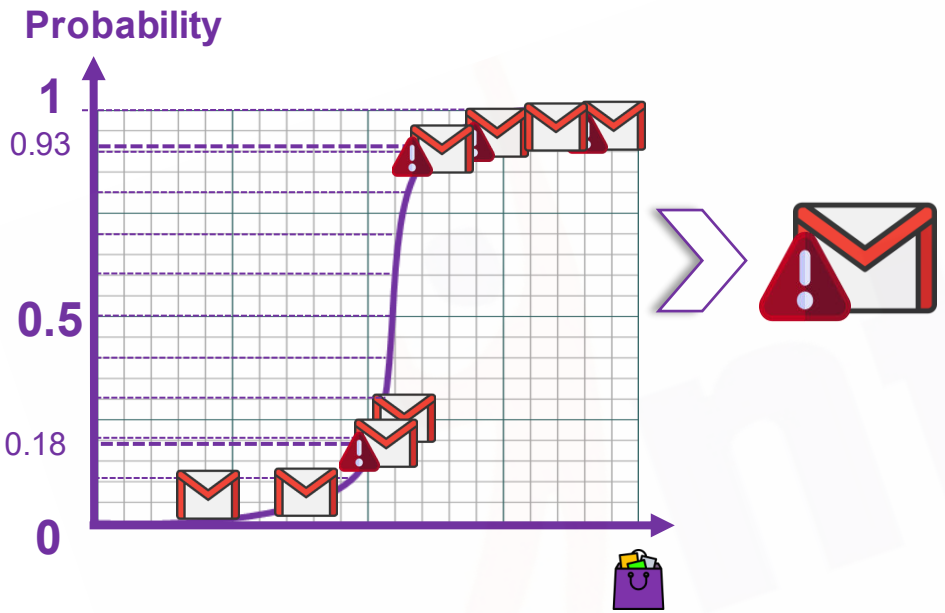


Now let us rotate this line to find out the best fitted regression line.



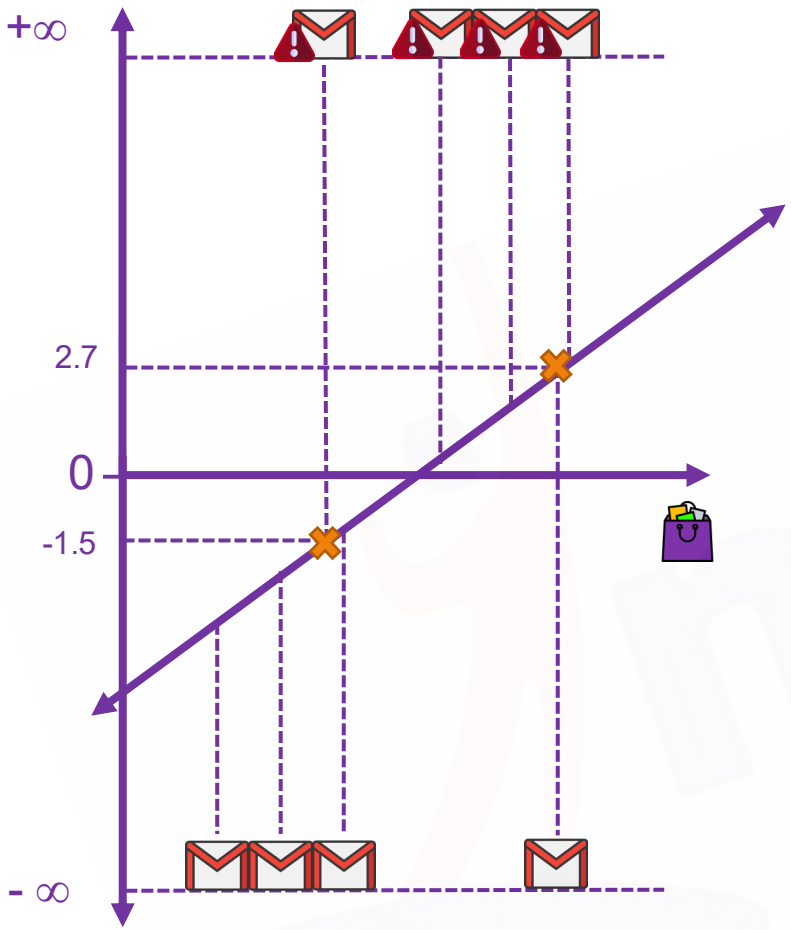


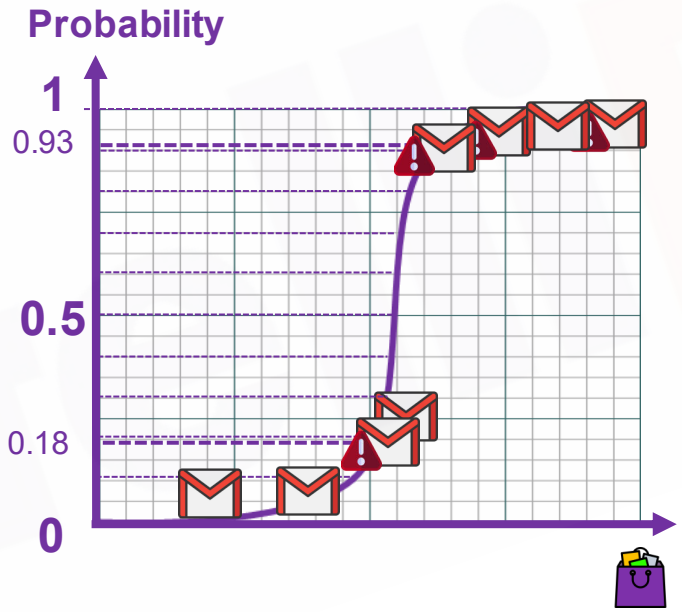
Again we will calculate individual log likelihood of each mail



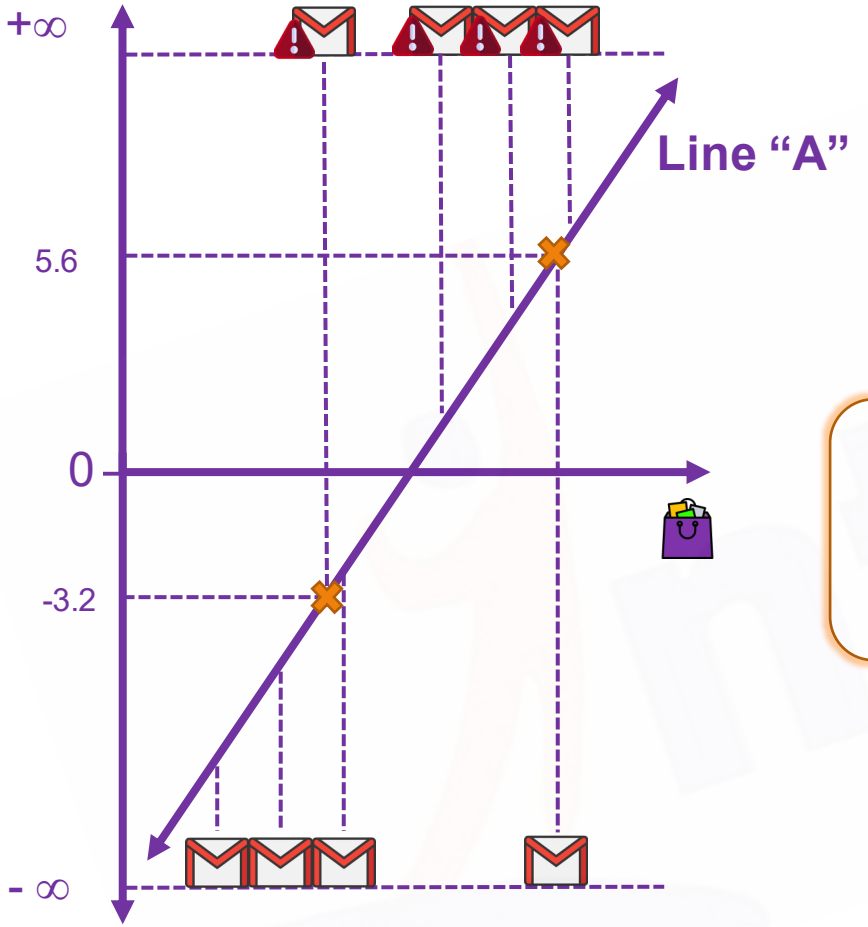
log(Likelihood of data given this curve) = log(0.93) + log(0.97) + log(0.99) + log(0.99) + log(1-0.18) + log(1-0.2) + log(1-0.03) + log(1-0.03)

= -0.207



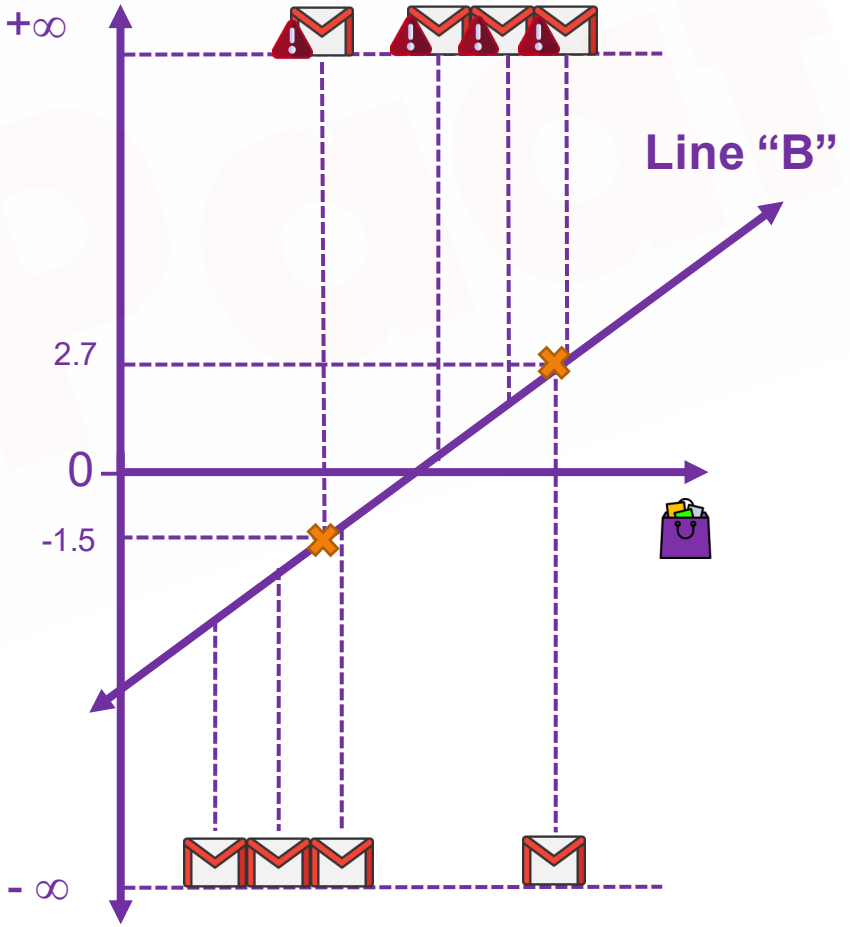


This means that log likelihood of this line is -0.207

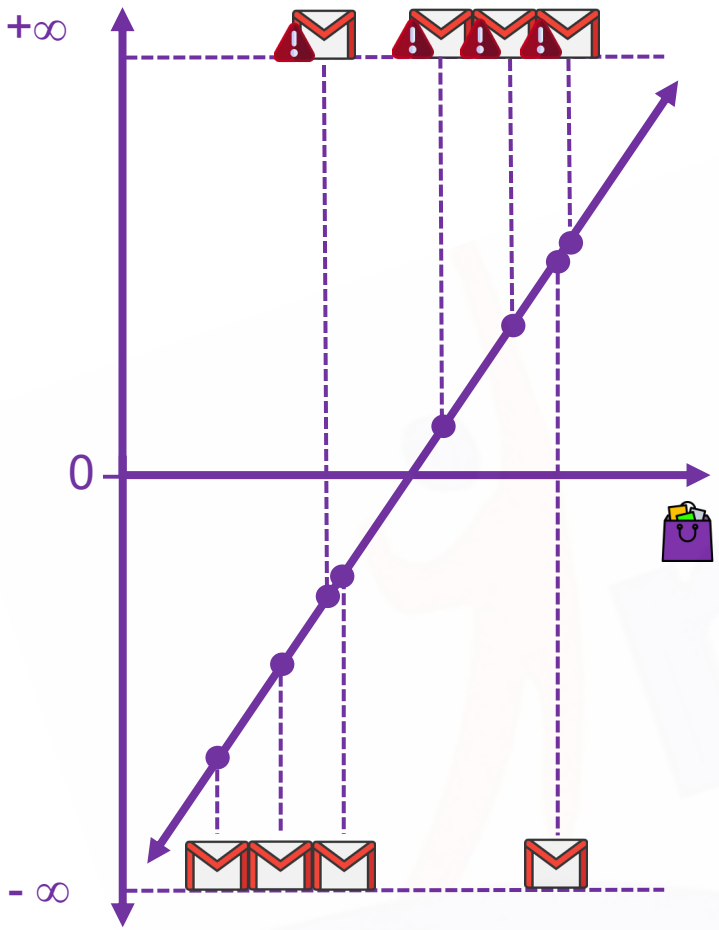


log(likelihood)= - 0.084

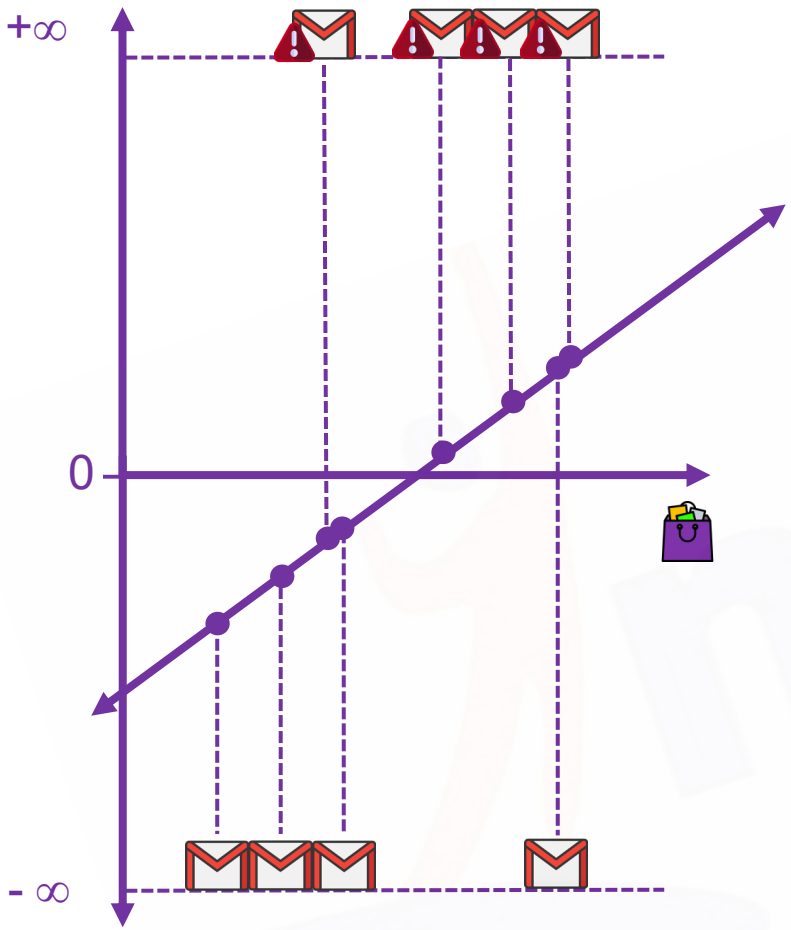
Line “A” has a better likelihood value than line “B”



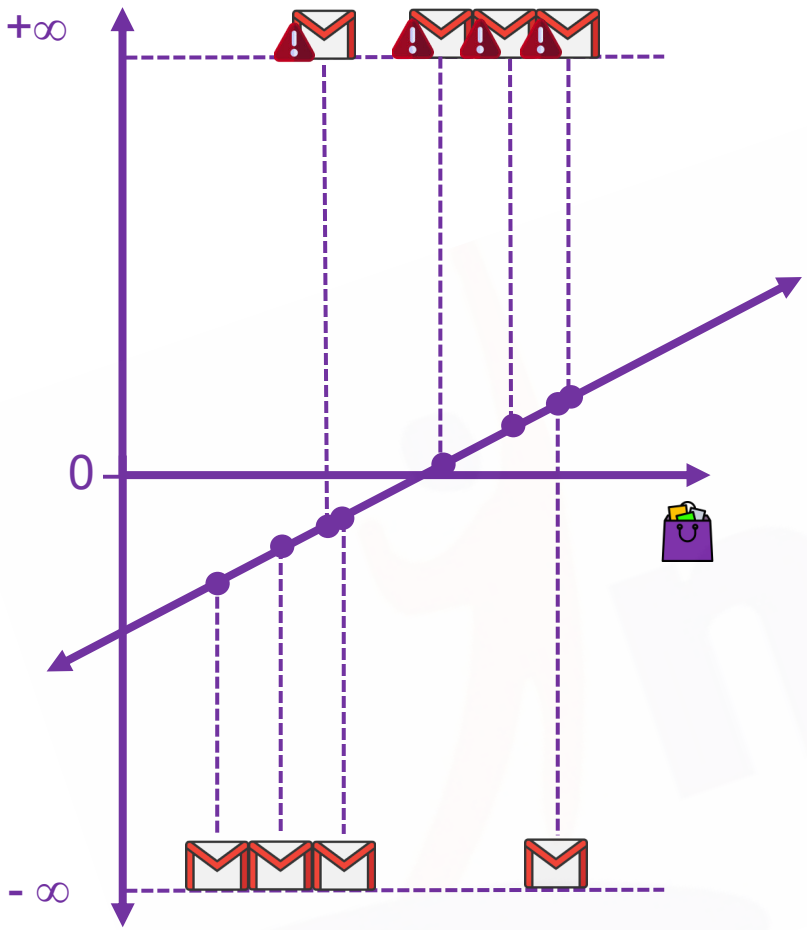
log(likelihood)= - 0.207



Again we will rotate the line

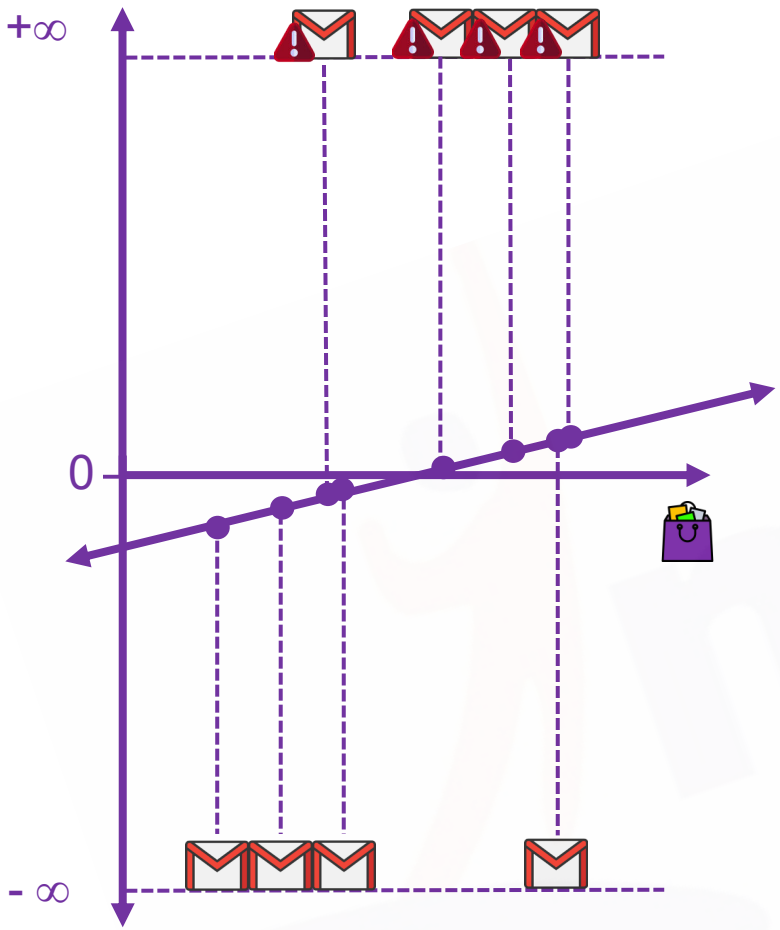


We will keep on rotating the line



We will keep on rotating the line

Until we get the maximum log likelihood



We will keep on rotating the line until we get the maximum log likelihood

That would be the best fitted regression line

Summary

📈 Logistic Regression Table

Model equations:

- Linear score:

$$z = \beta_0 + \beta_1 x = -3.75 + 0.88x$$

- Probability:

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

x (Spam Words)	Linear Score `z = β₀ + β₁·x` (with computed z)	Log Odds (z)	Probability `p = 1 / (1 + e^(-z))`
1	z = -3.75 + 0.88·1 = -2.87	-2.87	0.0536
5	z = -3.75 + 0.88·5 = 0.65	0.65	0.6574
3	z = -3.75 + 0.88·3 = -1.11	-1.11	0.2477
2	z = -3.75 + 0.88·2 = -1.99	-1.99	0.1199
7	z = -3.75 + 0.88·7 = 2.41	2.41	0.9178
4	z = -3.75 + 0.88·4 = -0.23	-0.23	0.4427
9	z = -3.75 + 0.88·9 = 4.17	4.17	0.9849
8	z = -3.75 + 0.88·8 = 3.29	3.29	0.9643

Summary

A

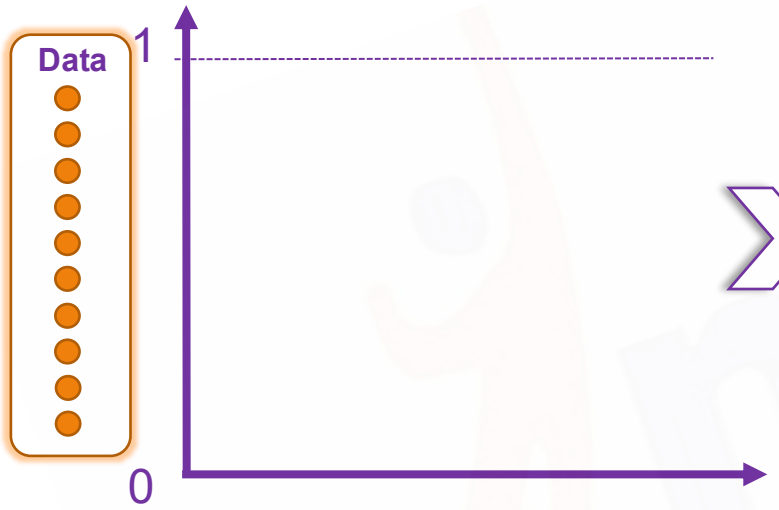
Plot the labeled data

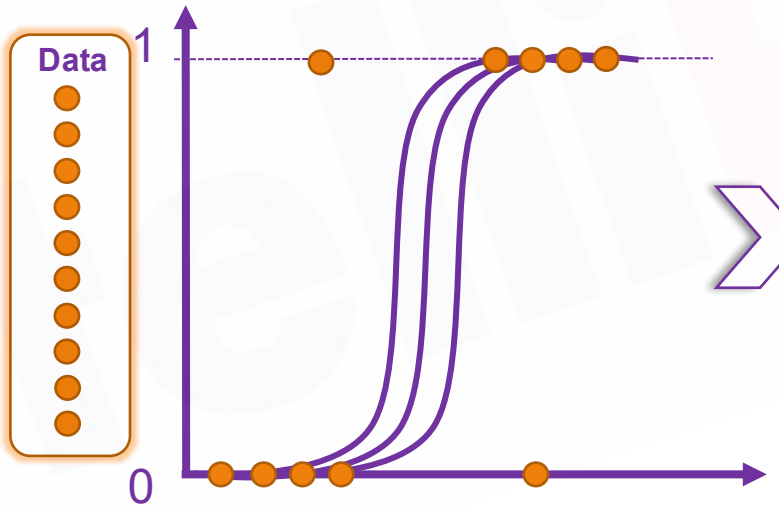
B

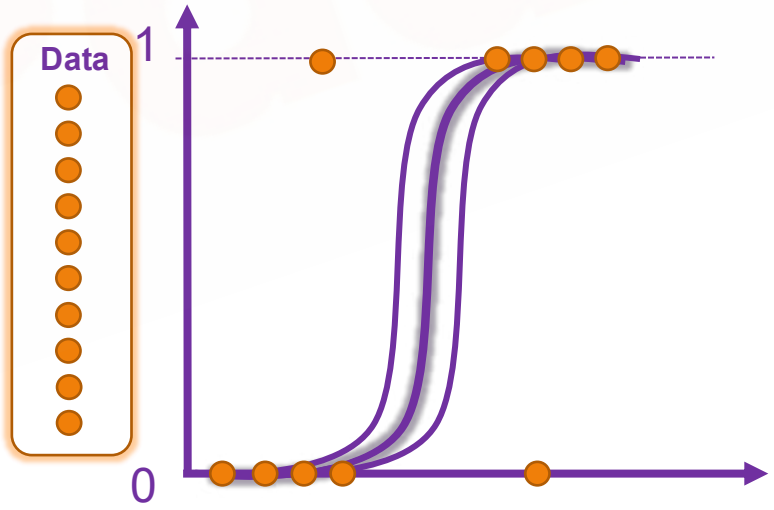
Draw a regression line and keep rotating

C

Use MLE to find out the best fitted regression line







Log(odds)

Sigmoid Function

MLE

IntelliPaat

www.intellipaat.com

CALL US NOW

India : +91-7847955955

US : 1-800-216-8930 (TOLL FREE)

sales@intellipaat.com