Data Science and Artificial Intelligence

Machine Learning

Classification

Lecture No. 1

By- SIDDHARTH SABHARWAL SIR

Recap of Previous Lecture

Topic Ridge Reg.

Topic Cross Validation $\Rightarrow$ K-Fold Cross Validation

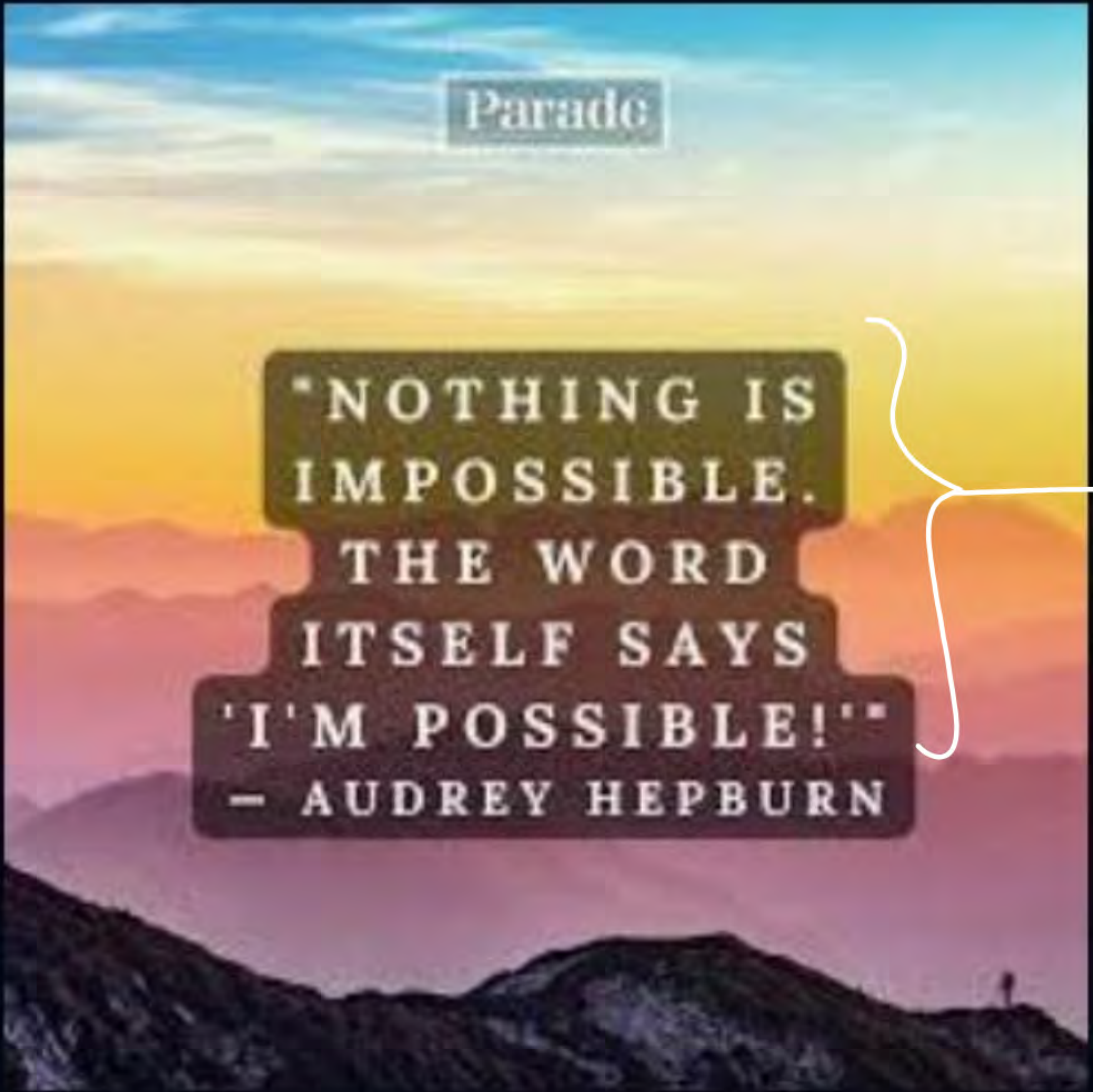
Topic effect of λ

Topic

Topic

Topics to be Covered

- **Topic** Lasso
- **Topic** Lasso vs RR
- **Topic** Classification
- **Topic** doubts
- **Topic**



Plan and work for it

Ridge Regression Final expression

$\rightarrow \min \frac{1}{2}\sum(y_i-\hat{y}_i)^2$

Const $\sum_{i=1}^{D}\beta_i^2 < C$

Ridge Regression Final expression

Linear Reg

$$L = \min \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

$$\frac{\partial L}{\partial \beta} = \begin{bmatrix} \partial L/\partial \beta_0 \\ \partial L/\partial \beta_1 \\ \vdots \\ \partial L/\partial \beta_D \end{bmatrix} \Rightarrow -2\left[X^T Y - (X^T X)\beta_{old} \right]$$

Ridge Reg

$$L = \min \frac{1}{2} \sum (y_i - \hat{y}_i)^2 + \frac{\lambda}{2} \sum_{i=1}^{D} \beta_i^2$$

$$\frac{\partial L}{\partial \beta} \Rightarrow -\left[X^T Y - X^T X \beta \right] + (\lambda I \beta)$$

$$\left(\beta^{new} = \beta^{old} - \eta \cdot \frac{\partial L}{\partial \beta} \right)$$

Ridge Regression – lets practise

Ridge Regression is a regularization technique used in linear regression to:

A) Increase model complexity.
B) Reduce model complexity and prevent overfitting.
C) Make the model fit the training data perfectly.
D) Enhance the interpretability of the model.

Ridge Regression – lets practise

In Ridge Regression, the penalty term added to the cost function is based on:

$\lambda/2 \sum_{i=1}^{p} \beta_i^2$ Siddharth Sir AI/ML.

A) The absolute values of the regression coefficients.
B) The square of the regression coefficients.
C) The number of features.
D) The dependent variable.

Ridge Regression – lets practise

What happens to the magnitude of regression coefficients in Ridge Regression compared to ordinary linear regression?

A) They become larger.
B) They become smaller.
C) They stay the same.
D) It depends on the dataset.

Ridge Regression – lets practise

Ridge Regression is particularly useful when:

RR is used

A) There is no multicollinearity among the independent variables.
B) There is a high degree of multicollinearity among the independent variables.
C) The model needs to fit the training data perfectly.
D) The dataset has very few observations.

Ridge Regression

Ridge Regression – lets practise

Which of the following values of λ (lambda) in Ridge Regression would lead to the strongest regularization effect?

Strongest Reg.

$\beta \rightarrow 0$

A) λ = 0
B) λ = 1
C) λ = 10
D) λ = ∞

Ridge Regression

Ridge Regression – lets practise

Ridge Regression can help prevent overfitting, but what is the trade-off?

A) Increased model interpretability.

B) Increased computational complexity.

C) Reduced accuracy on the training data.

D) Smaller training dataset size.

Regularisation best λ

overfitting ✗

generalise model ✓

The model give more error on training data

* This is not a problem bcoz our model is generalising better

Ridge Regression – lets practise

In Ridge Regression, what is the effect of increasing λ (lambda) on the bias and variance of the model?

→ Testing err.

as λ inc → model generalise better
model complexity reduce
→ training err inc / testing err dec
Training err

Bias inc | var dec

A) Increases bias, decreases variance.
B) Decreases bias, increases variance.
C) Increases both bias and variance.
D) Decreases both bias and variance.

Ridge Regression – lets practise

In Ridge Regression, the penalty term added to the cost function is based on the L2 norm (Euclidean norm) of the regression coefficients. If the sum of squared regression coefficients (L2 norm) is 50 and the value of λ (lambda) is 3, what is the modified penalty term in the Ridge Regression cost function?

a)150
b)135
c)123
d)578

Penalty term $= \lambda \sum \beta_i^2 / 2$

$\Rightarrow 3 \times 50 / 2$

$\Rightarrow 150/2 \Rightarrow 75$

Ridge Regression – lets practise

In Ridge Regression, the penalty term added to the cost function is based on the L2 norm (Euclidean norm) of the regression coefficients. If the sum of squared regression coefficients (L2 norm) is 50 and the value of λ (lambda) is 3, what is the modified penalty term in the Ridge Regression cost function?

a)150 done

b)135

c)123

d)578

Interpretability

So in L.R we get

$y = \beta_0 + \beta_1 x^1 + \beta_2 x^2 - - - -$

from this equation we can imagine our model i.e the Pattern of data

⇒ LR ⇒ high Interpretability

Ridge Regre.

generalise better, remove effect of few useless dimensions

Thus gives a better/easier model that Represent Pattern of data ⇒ More Interpretable model.

Ridge Regression – lets practise

In a ridge regression model, the original sum of squared residuals is 60. If the regularization parameter λ is set to 0.4, and the sum of squared residuals after ridge regression becomes 50, what is the proportion of variance explained by the model?

we are Comparing RR with LR

$* \, LR \Rightarrow RSS \Rightarrow 60$

$* \, RR \Rightarrow RSS \Rightarrow 50$

So $R^2 \Rightarrow 1 - \dfrac{\text{RSS of RR}}{\text{RSS of LR}} \Rightarrow 1 - \dfrac{50}{60} \Rightarrow \dfrac{1}{6}$

$\left(R^2 \Rightarrow \dfrac{1}{6}\right) \Rightarrow 16.66\%$

What is Lasso Regularisation

(L1 Reg)

$$\text{Loss fxn} \Rightarrow \left(\frac{1}{2} \sum (y_i - \hat{y}_i)^2 + \lambda \sum_{i=1}^{D} |\beta_i| \right)$$

$$\text{OR} \Rightarrow \min \frac{1}{2} \sum (y_i - \hat{y}_i)^2$$

Such that $\sum_{i=1}^{D} |\beta_i| < C$

$$\left\{ |x| \rightarrow \begin{array}{ll} x & \text{when } x \text{ +ve} \\ -x & \text{when } x \text{ -ve} \end{array} \right\}$$

Taking a 2D Case

$$|\beta_1| + |\beta_2| \leq C$$

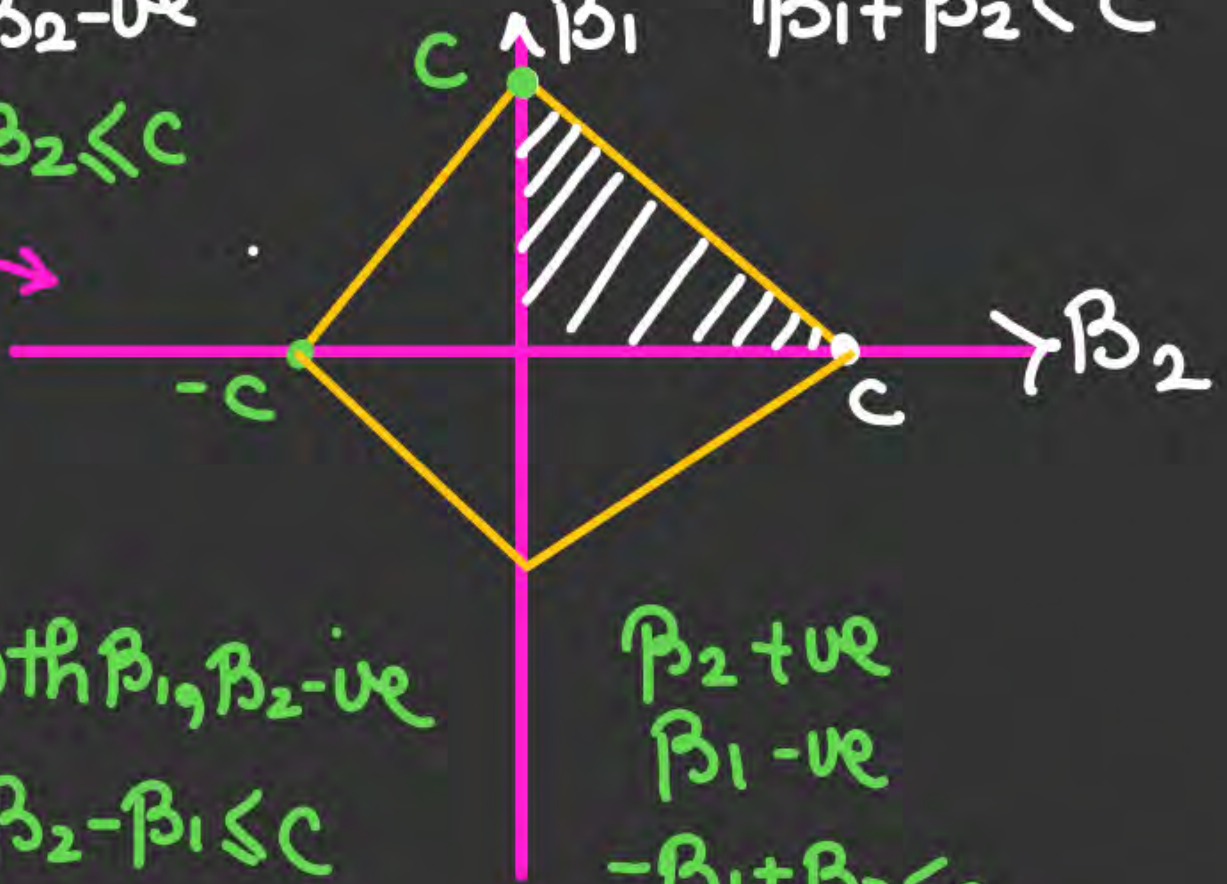
So Constraint Kite/square.

β_1 +ve

β_2 -ve

$\beta_1 - \beta_2 \leq C$

$\beta_1 + \beta_2 < C$



Both β_1, β_2 -ve

$-\beta_2 - \beta_1 \leq C$

β_2 +ve

β_1 -ve

$-\beta_1 + \beta_2 < C$

Important RR Vs lasso Reg

1. the algorithm of RR try to reduce β values but Lass Reg. has the algorithm that try to make β's $\to 0$.

2. lasso is more sensitive to 'λ' than Ridge Regression

Ridge Regression

Lasso Vs Ridge Regression

Interpret:

linear R
^
Ridge R
^
lasso R

Parameter	Ridge Regression	Lasso Regression
Regularization Type	L2 regularization: adds a penalty equal to the square of the magnitude of coefficients.	L1 regularization: adds a penalty equal to the absolute value of the magnitude of coefficients.
Primary Objective	To shrink the coefficients towards zero to reduce model complexity and multicollinearity.	To shrink some coefficients towards zero for both variable reduction and model simplification.
Feature Selection	Does not perform feature selection: all features are included in the model, but their impact is minimized.	Performs feature selection: can completely eliminate some features by setting their coefficients to zero.
Coefficient Shrinkage	Coefficients are shrunk towards zero but not exactly to zero.	Coefficients can be shrunk to exactly zero, effectively eliminating some variables.
Suitability	Suitable in situations where all features are relevant, and there is multicollinearity.	Suitable when the number of predictors is high and there is a need to identify the most significant features.
Bias and Variance	Introduces bias but reduces variance.	Introduces bias but reduces variance, potentially more than Ridge due to feature elimination.
Interpretability	Less interpretable in the presence of many features as none are eliminated.	More interpretable due to feature elimination, focusing on significant predictors only.
Sensitivity to λ	Gradual change in coefficients as the penalty parameter λ changes.	Sharp thresholding effect where coefficients can abruptly become zero as λ changes.
Model Complexity	Generally results in a more complex model compared to Lasso.	This leads to a simpler model, especially when irrelevant features are abundant.

Read

when # of feature is v.high

Interpretability LR < RR < lasso

Complexity LR > RR > lasso

generalise LR < RR < lasso

ω is the β

- β_{MLE} = the β of LR
- β_{RR}

Q7 Using the data $X=[-3,5,4]$ and $Y=[-10,20,20]$, ...uming a ridge penalty $\lambda = 50$, what ratio versus the Maximum Likelihood Estimate (MLE) estimate w_{mle} do you think the ridge regression L2 estimate estimate w_{ridge} estimate will be?

(A) 2

(B) 1

(C) 0.6

(D) 0.5

we will have (ω_0, ω_1)

$$\Rightarrow \text{Ratio of } \frac{\omega_{MLE}}{\omega_{RR}} = \frac{\omega_{LR}}{\omega_{RR}} = \frac{\omega_{1LR}}{\omega_{1RR}}$$

$RR \Rightarrow X \Rightarrow \begin{bmatrix} -3 \\ 5 \\ 4 \end{bmatrix} \qquad Y = \begin{bmatrix} -10 \\ 20 \\ 20 \end{bmatrix} \Rightarrow$ Centering

$X = \begin{bmatrix} \\ \\ \end{bmatrix} \qquad Y = \begin{bmatrix} \\ \\ \end{bmatrix}$

$\Rightarrow \beta_1 = (X^TX + \lambda I)^{-1}(X^TY)$

$\Rightarrow \beta_0 = \bar{y} - \beta_1 \bar{x}$

$\Rightarrow LR \Rightarrow \begin{bmatrix} 1 & -3 \\ 1 & 5 \\ 1 & 4 \end{bmatrix}$

$\beta = (X^TX)^{-1}(X^TY)$

Concept

$\Rightarrow$ $\|\beta\|^2$ $\|Y\|^2$

$A = \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix}$

$\|A\|^2 \Rightarrow (\text{norm})^2$

$\Rightarrow$ Sum of square of elements

$\|A\|^2 \Rightarrow a^2+b^2+c^2+d^2$

Q1 Consider the linear regression model Y = Xβ + ε with ε ~ N (0n, σε^2 Inn). This model (without intercept) is fitted to data using the ridge regression estimator β^(λ) = arg minβ ‖Y − Xβ‖ 2 2 + λ‖β‖ 2 2 with λ > 0.

The data are:

X^T =(−1 1 1 −1) and Y^T = (−1.5 2.9 −3.5 0.7)

What is the maximum likelihood/ordinary least squares estimator of the regression parameter for λ =0?

(A) [−0.3, 0.05] (B) [−0.5, 0.1]

(C) [0.1, −0.2] (D) [0.05, −0.3]

Q2 Suppose you are training a Ridge Regression model for a particular task and notice the following training error and validation RSS

Train: 57

Q5

$$\Rightarrow X = \begin{bmatrix} -1 \\ 1 \\ 1 \\ -1 \end{bmatrix}$$

$$Y = \begin{bmatrix} -1.5 \\ 2.9 \\ -3.5 \\ .7 \end{bmatrix}$$

$$X = \begin{bmatrix} 1 & -1 \\ 1 & 1 \\ 1 & 1 \\ 1 & -1 \end{bmatrix}$$

$$\beta = (X^TX)^{-1}(X^TY)$$

You have a dataset with 30 observations. After applying linear regression, you find that the residual standard error (RSE) is 5. If the coefficient of determination (R^2) is 0.8, what is the root mean square error (RMSE) for this model?

(A) 2

(B) 3

(C) 4

(D) 5

$\underbrace{RSE}$ = RMSE

5 = RMSE

Linear Classification

Linear classification

Classification vs Regression...

Classification:

- The X value can be anything, but the Y values are Categorical

Regression:

we try to create a model which predict Y based on X

→ Value of Y can be any real number.

label

Fever	Heart Rate	Sugar	Diabetic
105	67		Yes
102	73		No
100	72		Yes
103	⋮		No
⋮			⋮

It is a categorical data of 2 classes.

label is categorical in nature

So in this data ⇒ Task is to predict label of the new data point.

Linear Classification

Linear Regression of an Indicator Matrix

One hot Coding

2 classes

Yes ← Y_1 Y_2 → No

Fever	H.R	Sugar	Y_1	Y_2
–	–	–	1	0
–	–	–	0	1
			1	0
–	–	–	0	1
–	–	–	0	1

Let's consider a 2-class case

What is an Indicator Matrix

Linear Classification

Linear Regression of an Indicator Matrix

- So we have 2 classes and we create 2 Y values Y_1, Y_2
- For each data point only one Y value will be '1' rest will be '0'.

Let's consider a 2-class case

What is an Indicator Matrix

So Now we use regression for classification ⇒

↳ So Now taking $\underline{Y_1}$ as Y we do L·R and we will get a line that will give '1' for points of 'Yes' Class '0' for points of 'No' Class

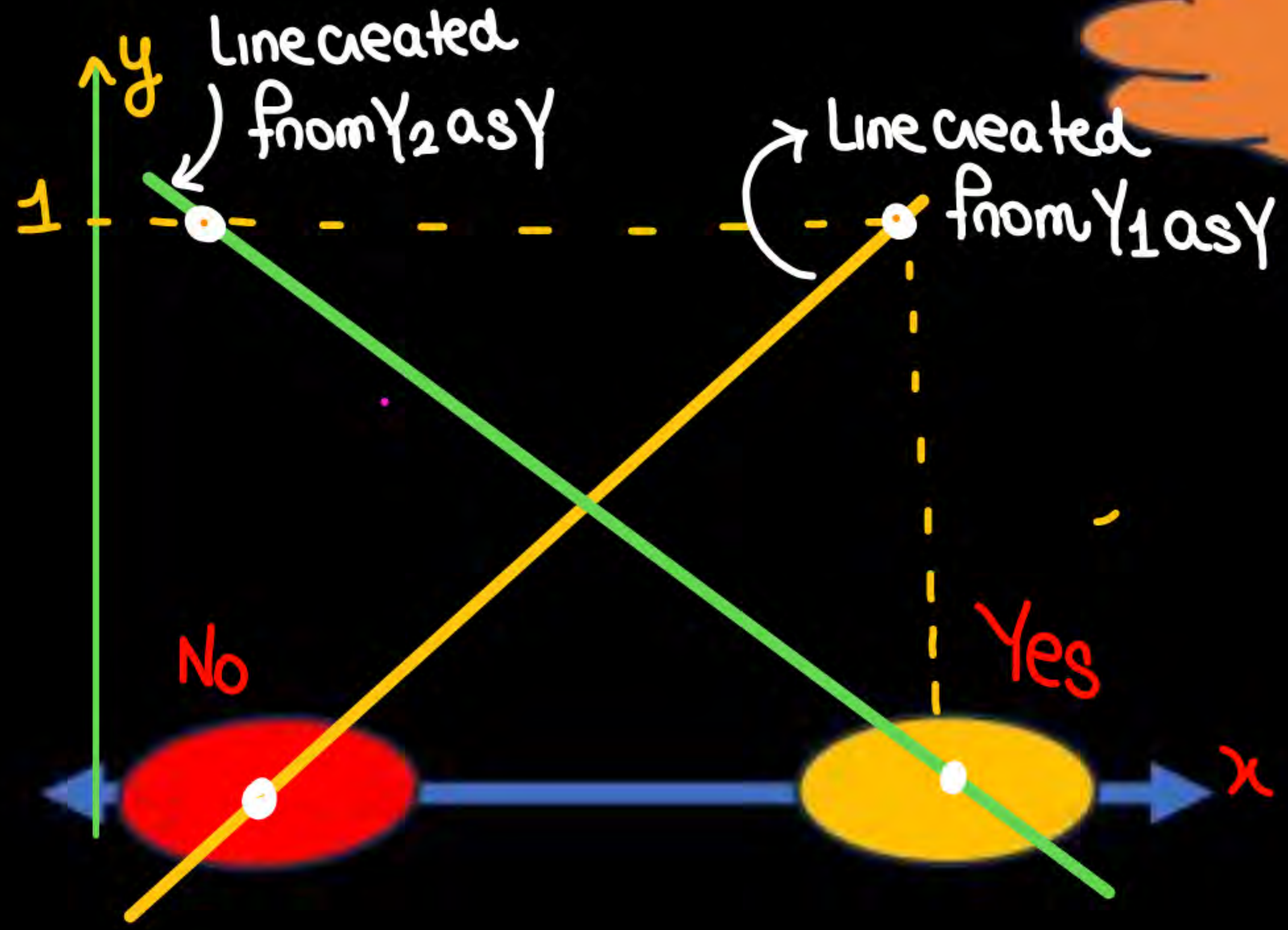
↳ Similarly taking $\underline{Y_2}$ as Y we do L·R and we will get a line that give '1' for 'No' Points and '0' for Yes points.

Linear Classification

Single dimension data

Linear Regression of an Indicator Matrix

Let's understand using figures



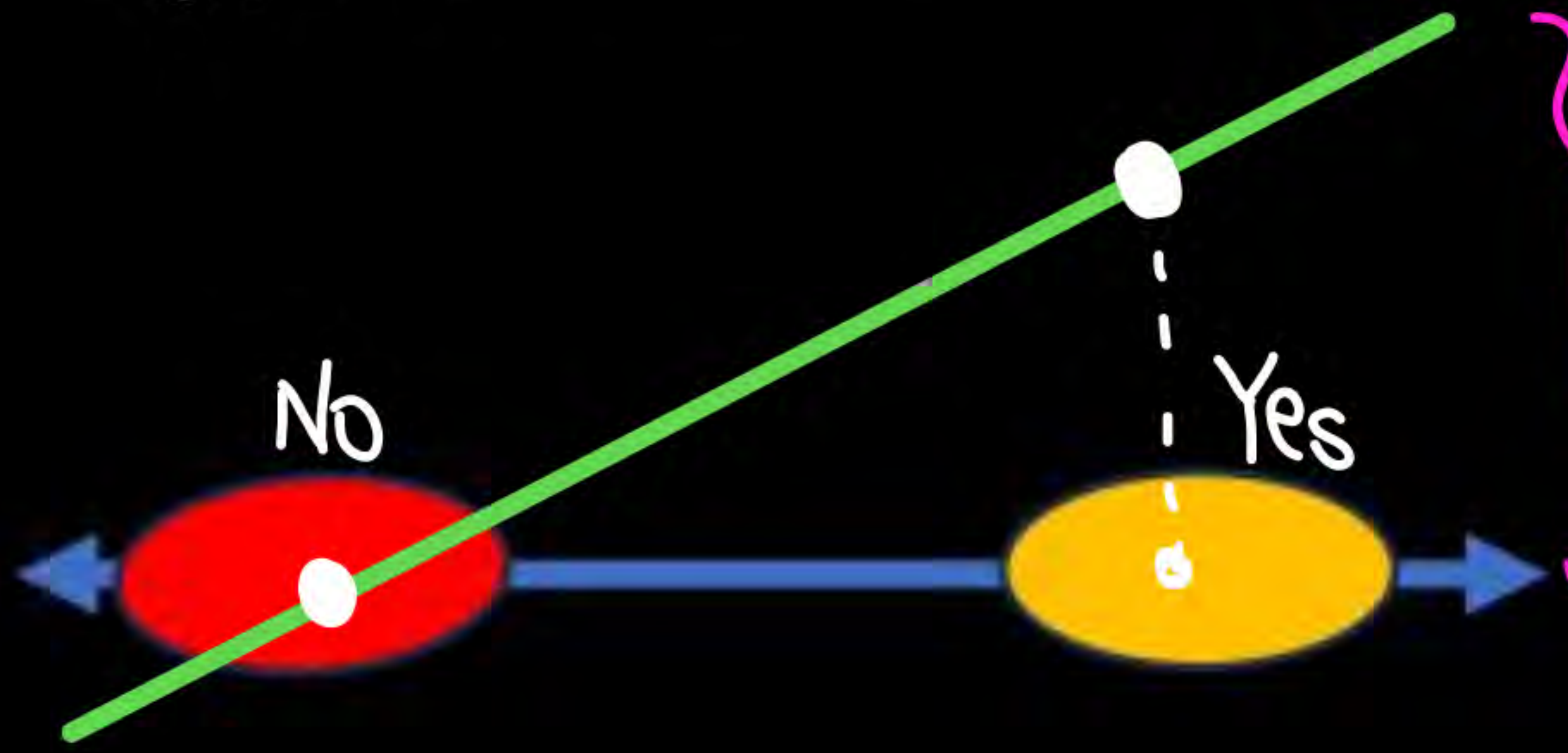
- So we have created 2 lines for 2 classes

Linear Classification

Linear Regression of an Indicator Matrix

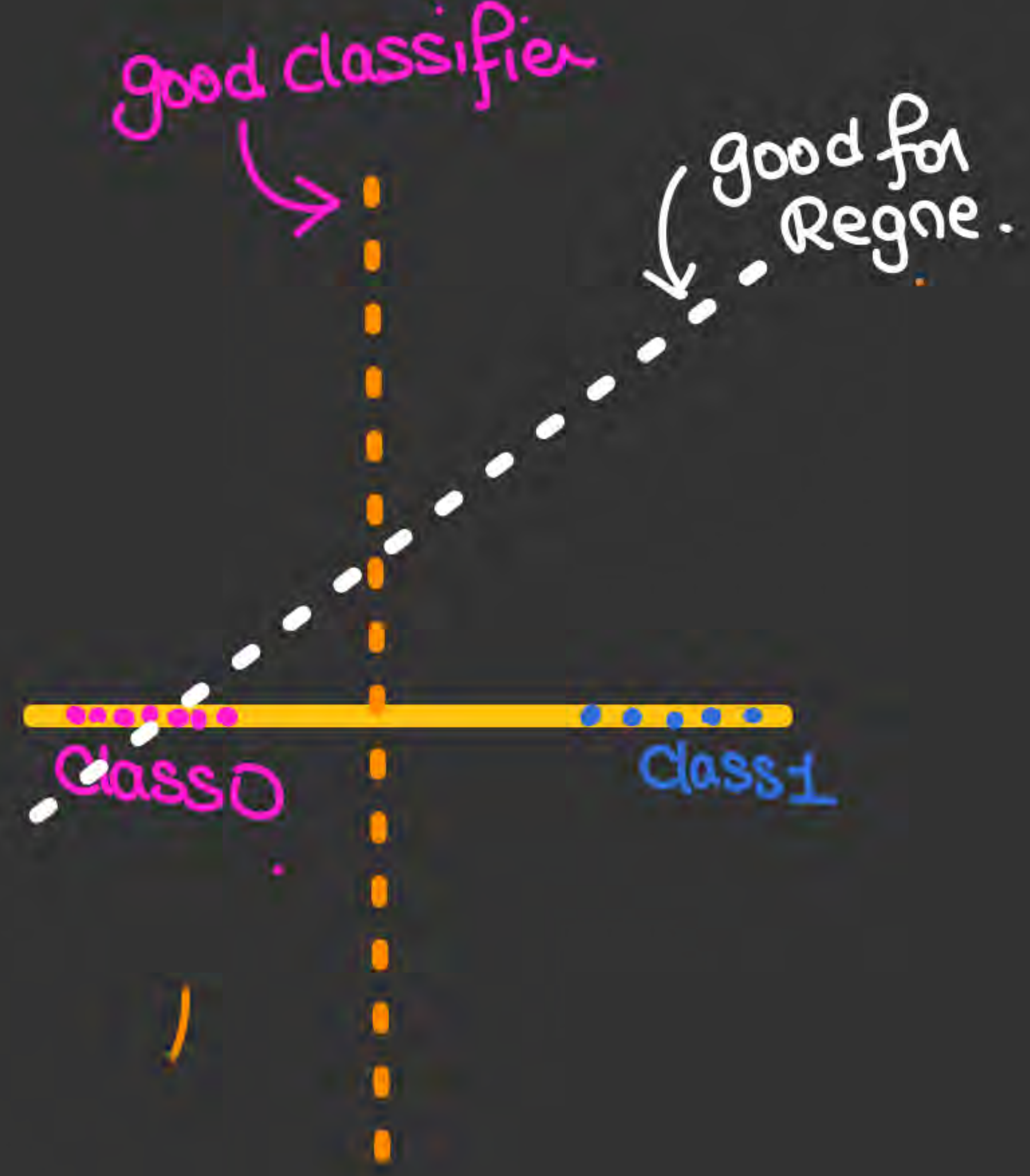
Let's understand using figures

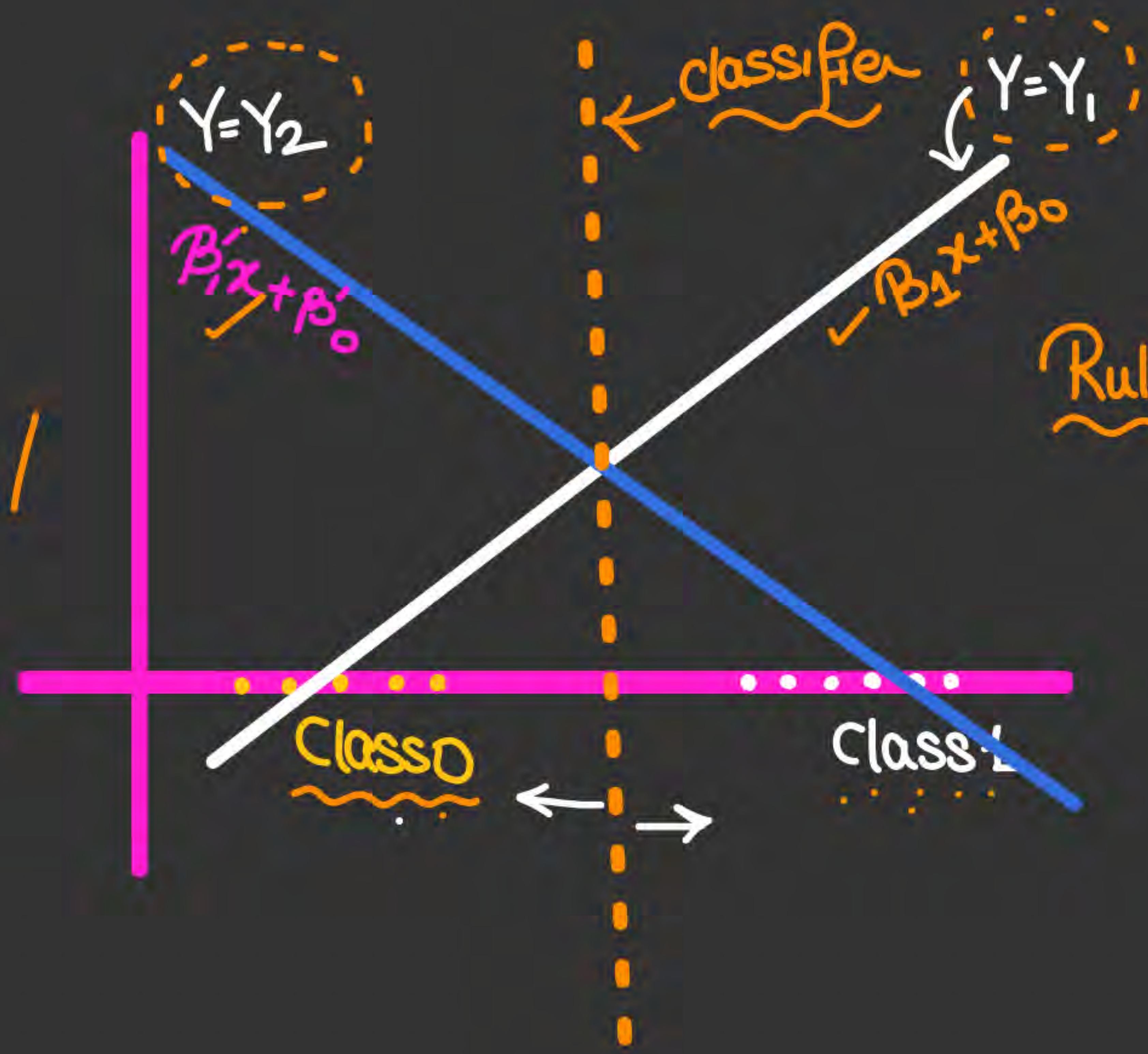
doubt: We can work with a single line such that if value of y from line at any point is close to 1 ⇒ Yes
y from " " " " " " 0 ⇒ No



Purely Reg. Activity not a seperator.

- But actually in classification our motive is not to find y value from line. Rather we need a line that seperate the Points Belonging to diff classes





Rule ⇒ Classification Rule

Class 0

$\hookrightarrow (\beta_1' x + \beta_0') > \beta_1 x + \beta_0$

$\beta_1 x + \beta_0 > \beta_1' x + \beta_0'$

Class 1

Linear Regression of an Indicator Matrix

So, now the analysis is as follows :

So if we have data with 2 Classes

- → one hot Coding ⇒ Indicator matrix
- → two times L·R ⇒ 2 lines
- → Condition $\beta_1' x + \beta_0' > \beta_1 x + \beta_0 \Rightarrow$ class 0

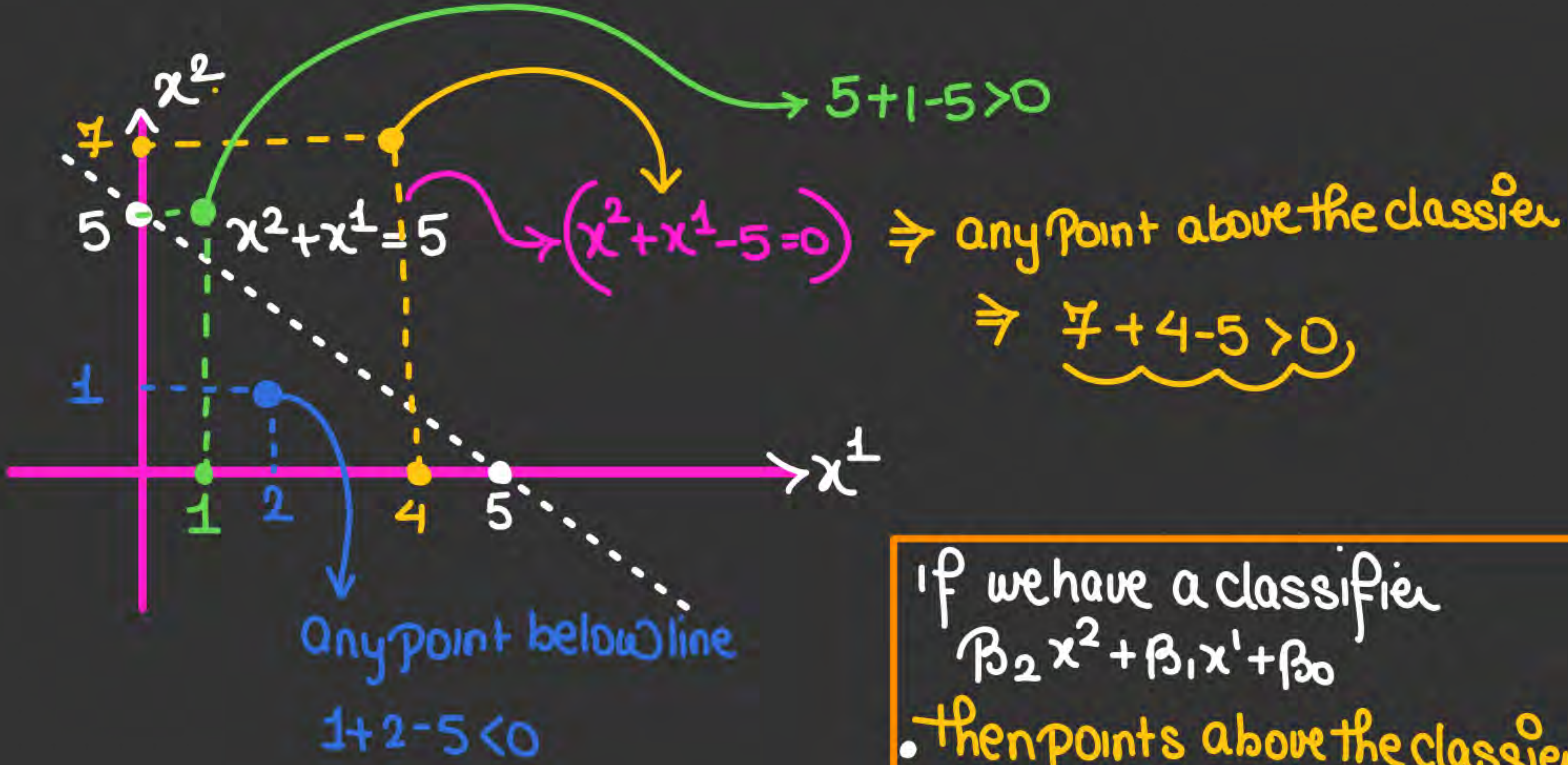
  $\beta_1 x + \beta_0 > \beta_1' x + \beta_0' \Rightarrow$ class 1

Linear Regression of an Indicator Matrix

Lets extend the case for K classes

From these 2 conditions we actually get a single classifier line b/w 2 classes.

Can we find this classifier directly ?? ⇒ Yes by linear Classifier



If we have a classifier
$\beta_2 x^2 + \beta_1 x^1 + \beta_0$

- then points above the classier
$\beta_2 x^2 + \beta_1 x^1 + \beta_0 > 0$
- Similarly points below classifier
$\beta_2 x^2 + \beta_1 x^1 + \beta_0 < 0$

Linear Regression of an Indicator Matrix

Lets extend the case for K classes

11 am

Linear Regression of an Indicator Matrix

How to find the variables for the linear regression

Linear Classification

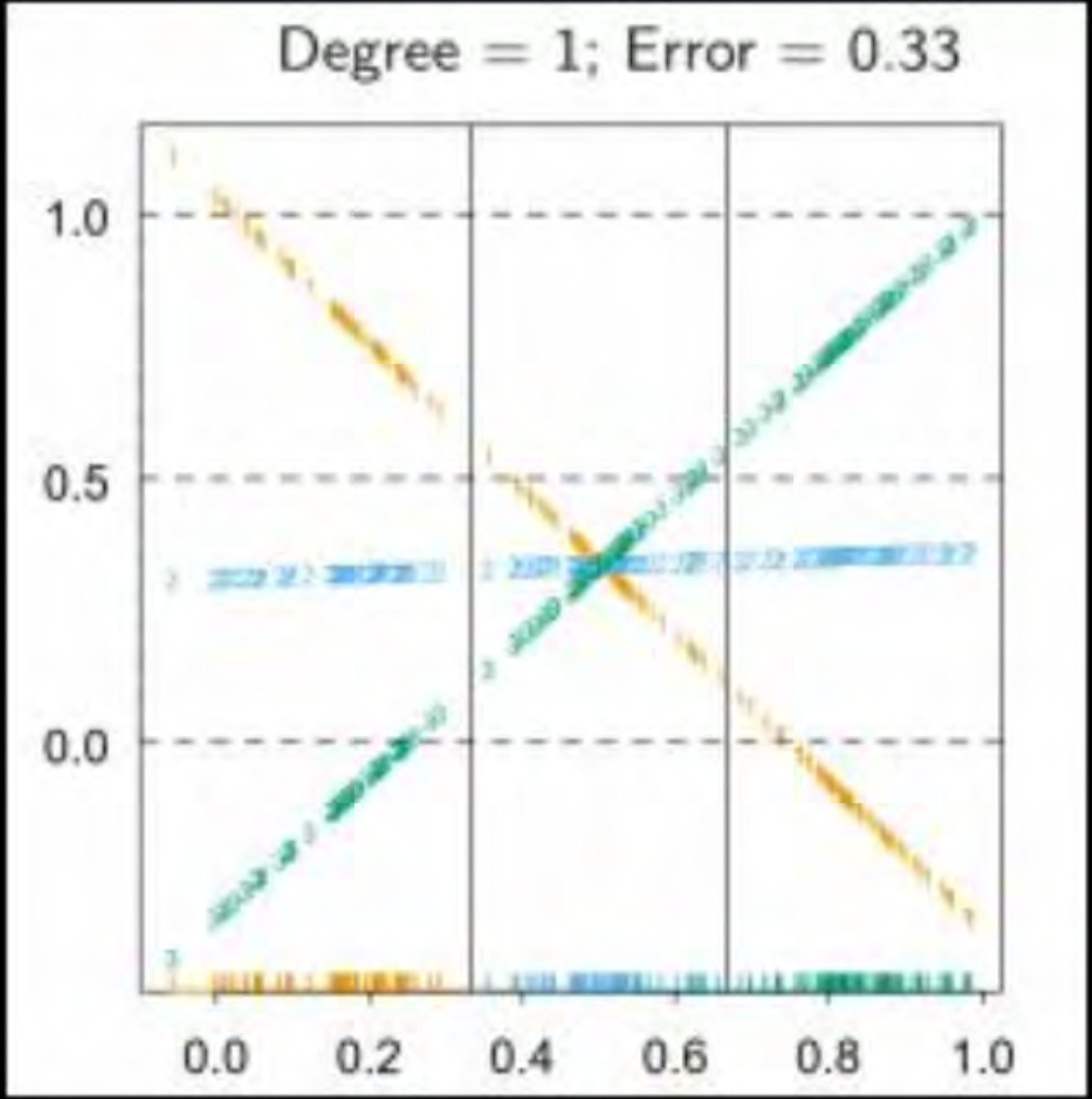
So linear regression can be used for classification also

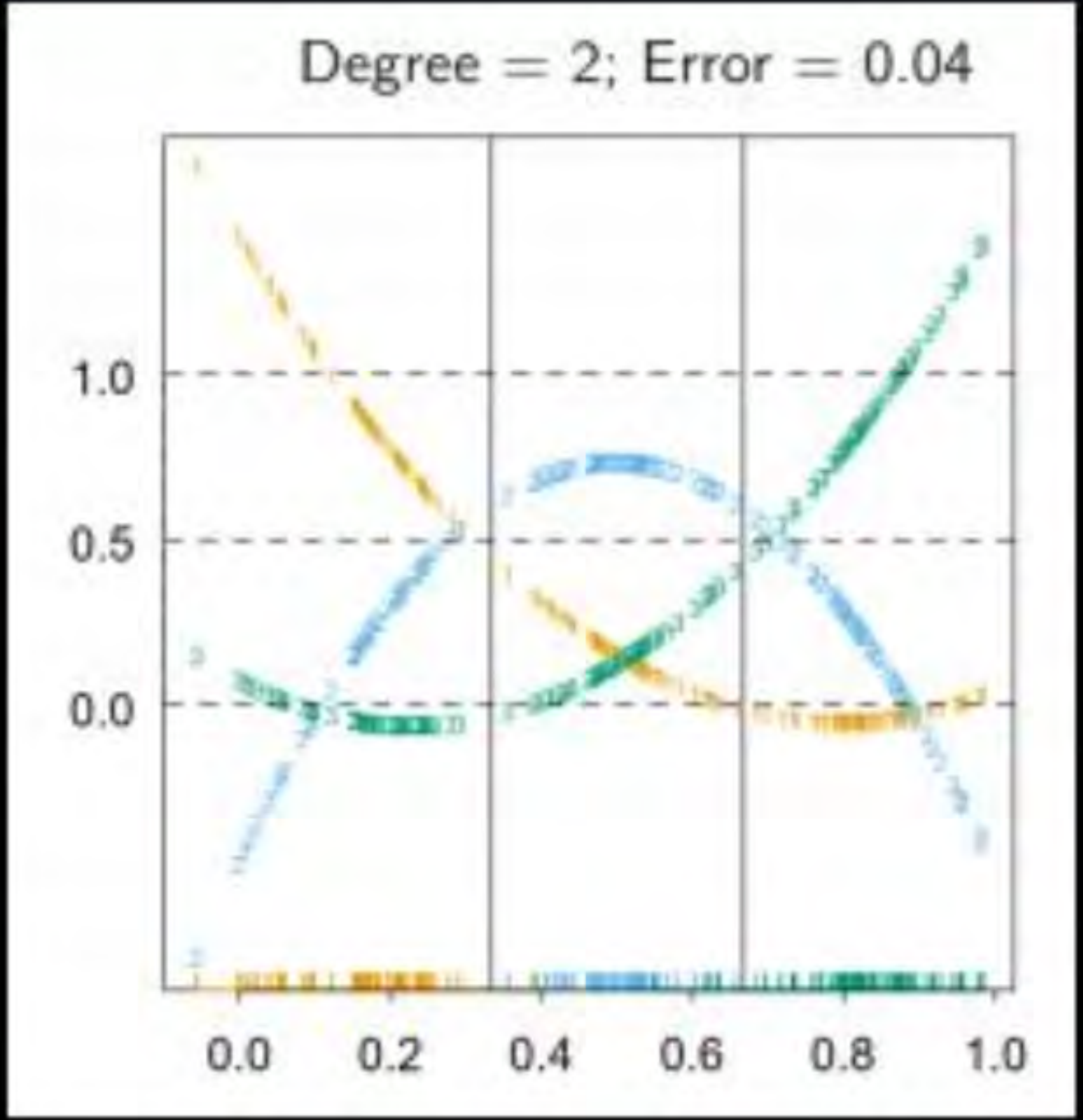
Linear Classification

Here we will have the error of 1/3, hence the linear regression fails to classify even the seperable points.

Linear Classification

Linear Regression of an Indicator Matrix





The three classes are perfectly separated by linear decision boundaries, yet linear regression misses the middle class completely.

But we can classify if we use the quadratic curves.

A loose but general rule is that if K ≥ 3 classes are lined up, polynomial terms up to degree K - 1 might be needed to resolve them.

Linear Classification

Linear Regression of an Indicator Matrix

In general p-dimensional input space, one would need general polynomial terms and cross-products of total degree $K - 1$, $O(p^{K-1})$ terms in all, to resolve such worst-case scenarios.

Linear Classification

Linear Regression of an Indicator Matrix

Lets consider a 2 class problem... We can have a single classifier for a 2 class problem...

Linear Classification

Linear Regression of an Indicator Matrix

The loss function for a 2 class case...

Linear Regression of an Indicator Matrix

But this loss function has 2 problems 1. outlier and 2. value of predicted Y

• Linear Classification

Problem of outliers

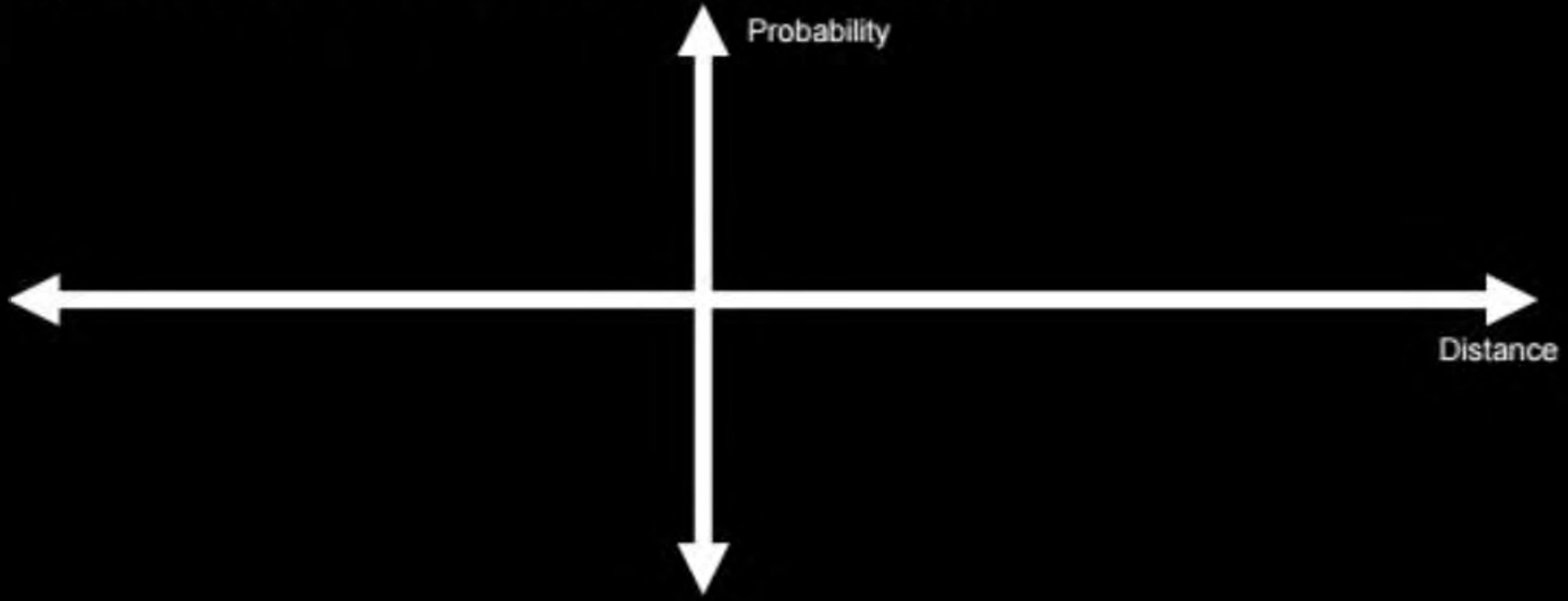
• Linear Classification

• Linear Classification

To solve the problem of outlier we will not use the distance in the analysis rather we will use the probability.

• Linear Classification

To solve the problem of outlier we will not use the distance in the analysis rather we will use the probability.



2 mins Summary

Topic

Topic

Topic

Topic

Topic

THANK - YOU