Data Science and Artificial Intelligence

Machine Learning

Regression

Lecture No. 10

By- SIDDHARTH SABHARWAL SIR

Recap of Previous Lecture

Topic

Topic

Topic

Topic

Topic

- Ridge Regression
  expression
  effect of λ

Topics to be Covered

- **Topic** How to find best λ
- **Topic** Constraint view of RR
  Questions.
- **Topic** Lasso $\rightarrow$ brief intro
- **Topic** Classification
- **Topic**


Parade
"NOTHING IS
IMPOSSIBLE.
THE WORD
ITSELF SAYS
'I'M POSSIBLE!'"
– AUDREY HEPBURN

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

Ridge Regression Final expression

Ridge Regression

Ridge Regression – lets practise

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

→ Prevent overfitting, → reduce model complexity

~~A)~~ Increase model complexity.

B) Reduce model complexity and prevent overfitting.

~~C)~~ Make the model fit the training data perfectly. → overfit

D) Enhance the interpretability of the model.

Ridge Regression

Ridge Regression – lets practise

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

Loss fxn = RSS + penalty term $\longrightarrow \sum \beta_i^2$

A) The absolute values of the regression coefficients.

B) The square of the regression coefficients.

C) The number of features.

D) The dependent variable.

absolute value $\Rightarrow |\beta_i|$

Ridge Regression – lets practise

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

β Ki values.

RR ⇒ try to reduce β's.

~~A) They become larger.~~

B) $\boxed{\text{They become smaller.}}$

C) They stay the same.

D) It depends on the dataset.

Ridge Regression

Ridge Regression – lets practise

Ridge Regression is particularly useful when: → β value small

Linear R = give stable model

dimensions.

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. → overfit

D) The dataset has very few observations.

→ LR give unstable model

RR = Solve Multi Collinearity

Ridge Regression – lets practise

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

- Strongest penalty term.
- → λ that give v. high importance to penalty term

A) λ = 0

B) λ = 1

C) λ = 10

D) λ = ∞

- $\min RSS + \lambda \sum \beta_i^2$

$\lambda = \infty$, β's → 0 underfit model.

Ridge Regression

RR → inc model interpretability. (adv)
inc Complexity Computation (dis).

Ridge Regression – lets practise

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

Benefit

RR 1 → 2

A) Increased model interpretability.

B) Increased computational complexity.

C) Reduced accuracy on the training data.

~~D) Smaller training dataset size.~~

overfit → Best fit

100 dimension Ka data.

LR ⇒ 100 β's Ka eq.
Jismain har dimension is given Importance

RR ⇒ 85 β's → 0 ⇒ less Imp dimension
$\boxed{15\ \beta\text{'s} \neq 0}$ ⇒ Imp dimension

So model will give more error on training data

Ridge Regression

Ridge Regression – lets practise

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

A) Increases bias, decreases variance.

B) Decreases bias, increases variance.

C) Increases both bias and variance.

D) Decreases both bias and variance.

0 — Bias inc, Var dec → | — Bias inc, Var dec → ∞

Overfit: Bias=0, Variance: high

Best Value: Best fit, Bias: low, Variance: low

Underfit: Bias high, Variance=0

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 $= \lambda \sum \beta^2$
$= 3 \times 50$
$= 150$

Ridge Regression

How to find the best hyerparameter

λ

How to find best λ

The method is called Cross Validation ⇒.

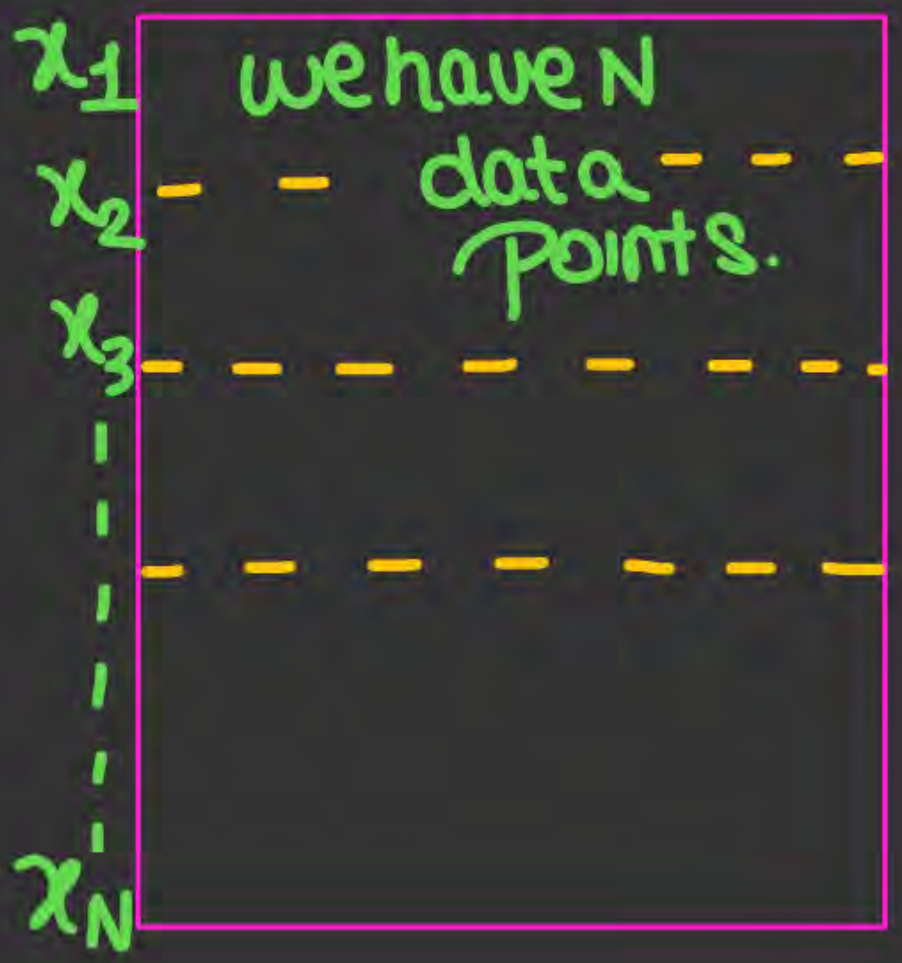
Question Bank data

Training data

Testing data

- model is prepared using training data
- then accuracy is tested on Testing data.

Training data.



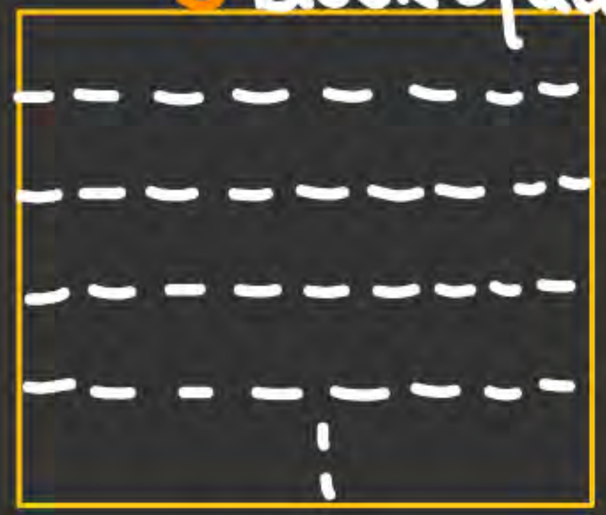
K fold cross validation

→ Step ① divide training data into K folds.

each block/fold has $\frac{N}{K}$ data points.

Step 2 : let $K = 5$

5 block of data



K=5, 5 fold CV.

we choose a λ

λ = ·1

	Train	Validate	% accuracy
1	1,2,3,4	5	a
2	2,3,4,5	1	b
3	1,3,4,5	2	c
4	1,2,4,5	3	D
5	1,2,3,5	4	E

for λ = ·1

accuracy $= \frac{a+b+c+d+e}{5}$.

So by this we check that model prepared by λ = ·1, is good for overall data or not

K=5, 5 fold CV.

Now

we choose a λ

$\lambda = \cdot 2$

	Train	Validate	% accuracy
1	1,2,3,4	5	a
2	2,3,4,5	1	b
3	1,3,4,5	2	c
4	1,2,4,5	3	D
5	1,2,3,5	4	E

for $\lambda = \cdot 1$

accuracy $= \dfrac{a+b+c+d+e}{5}$.

So by this we check that model prepared by $\lambda = \cdot 1$, is good for overall data or not

K=5, 5 fold CV.

Now

we choose a λ

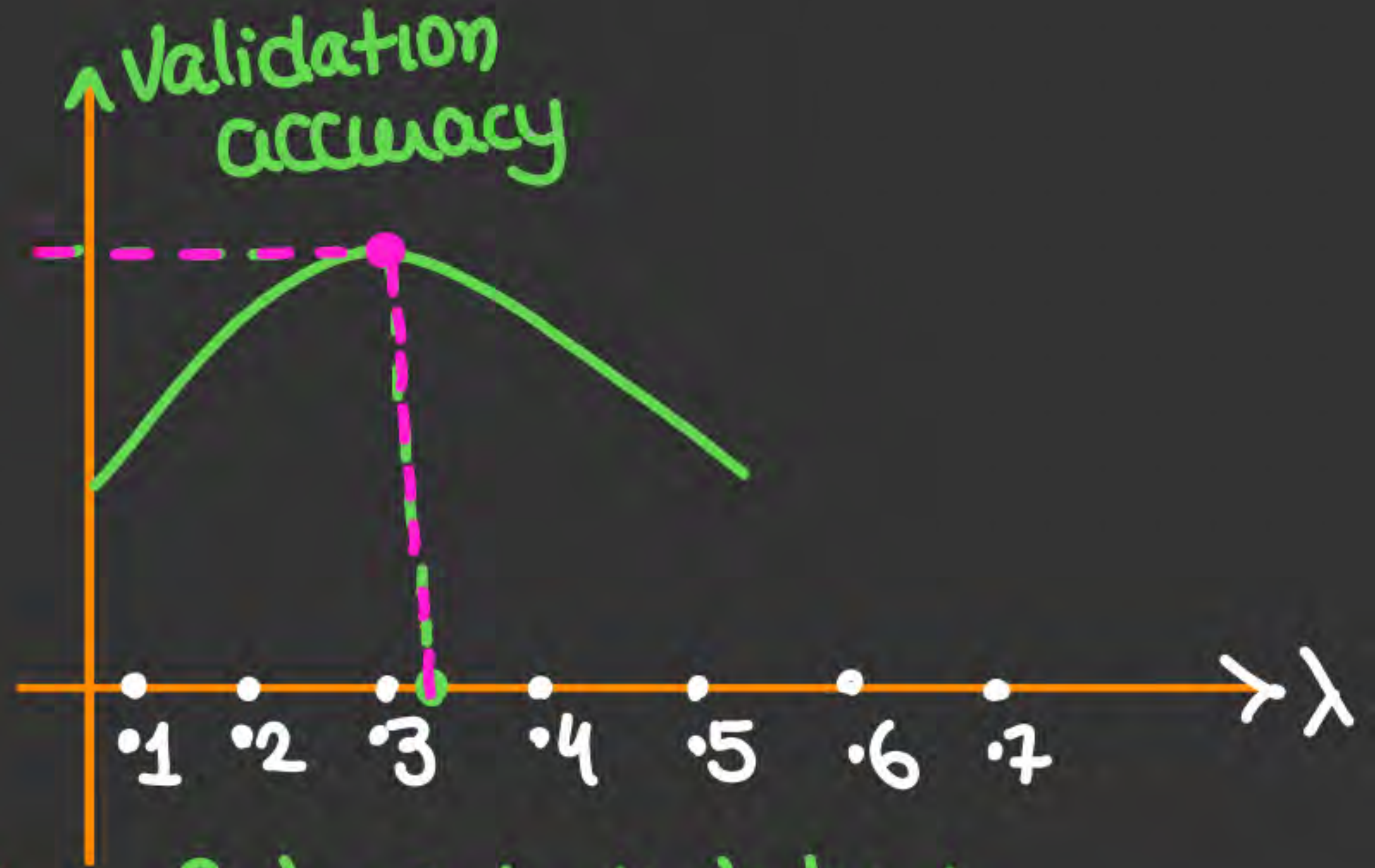
$\lambda = \cdot 3$

	Train	Validate	% accuracy
1	1,2,3,4	5	a
2	2,3,4,5	1	b
3	1,3,4,5	2	c
4	1,2,4,5	3	D
5	1,2,3,5	4	ε

for $\lambda = \cdot 1$

$$\text{accuracy} = \frac{a+b+c+d+e}{5}$$

So by this we check that model prepared by $\lambda = \cdot 1$, is good for overall data or not



So $\lambda = .3$ to $.4$ is best

Repeat process for $\lambda = .31$
$\lambda = .32$
$\lambda = .33$
⋮

To find best λ

finding best λ is really very lengthy.

• K fold CV

• This process is used to find best λ

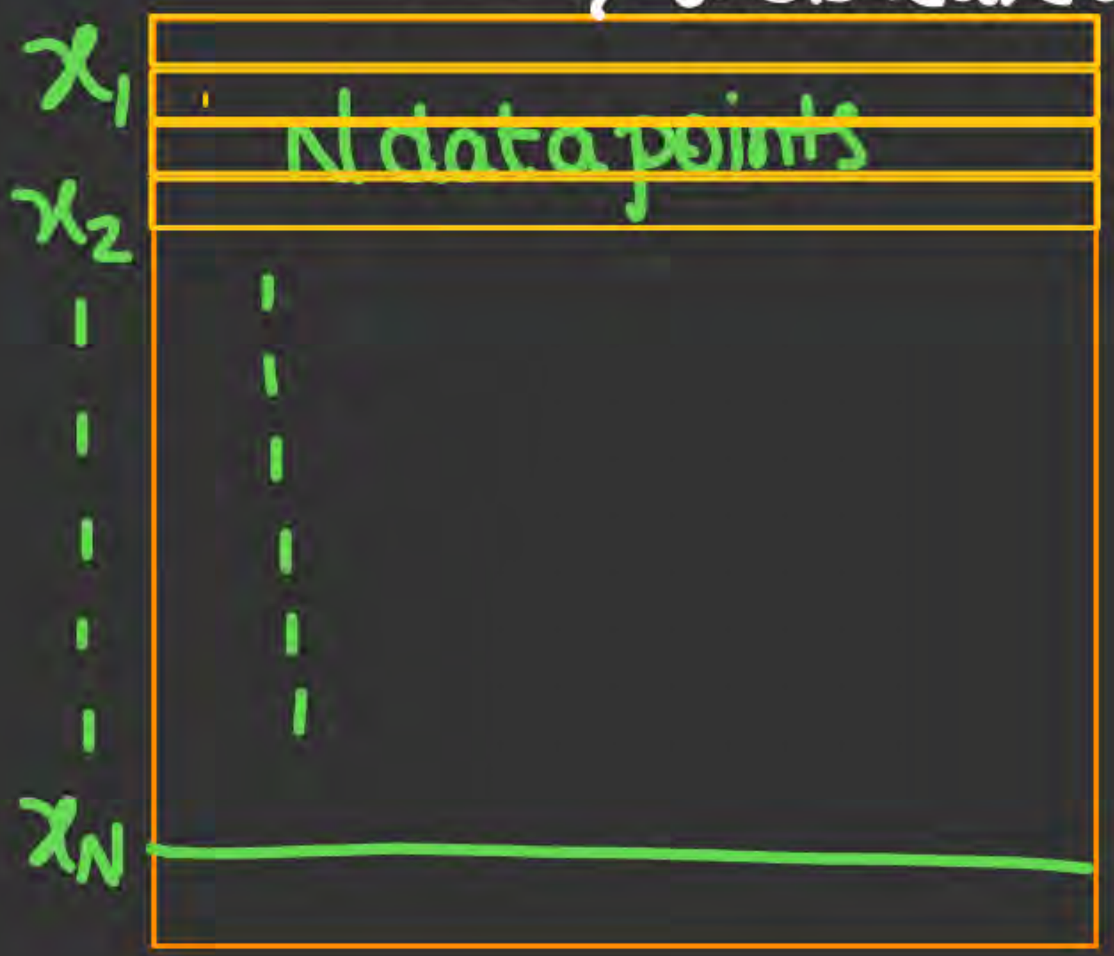
(i) divide training data into K fold

(ii) 1 fold for validation, (K-1) fold for training

Repeat process K times.

- LOOCV ⇒ leave one out Cross Validation.

→ we break data into N folds.



each fold has 1 data point

@each step (N-1) fold Training
1 fold Testing

→ Process Repeat "N" No of times.

Ex 1000 datapoints

↓

LOOCV

↓

1000 folds

↳ 1000 times process Repeat

1 fold Validation

999 fold training.

Ridge Regression

Constraint representation of Ridge Regression

loss fxn $\Rightarrow \left\{ \sum_{i=1}^{N}(y_i - \hat{y}_i)^2 + \lambda \sum_{i=1}^{D} \beta_i^2 \right\}$ $\Rightarrow$

minimize loss function

- another way $\Rightarrow$.

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

Such that

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

↳ Constant.

- C will be a hyper parameter, which will be calculated by CV

Ridge Regression

Constraint representation of Ridge Regression

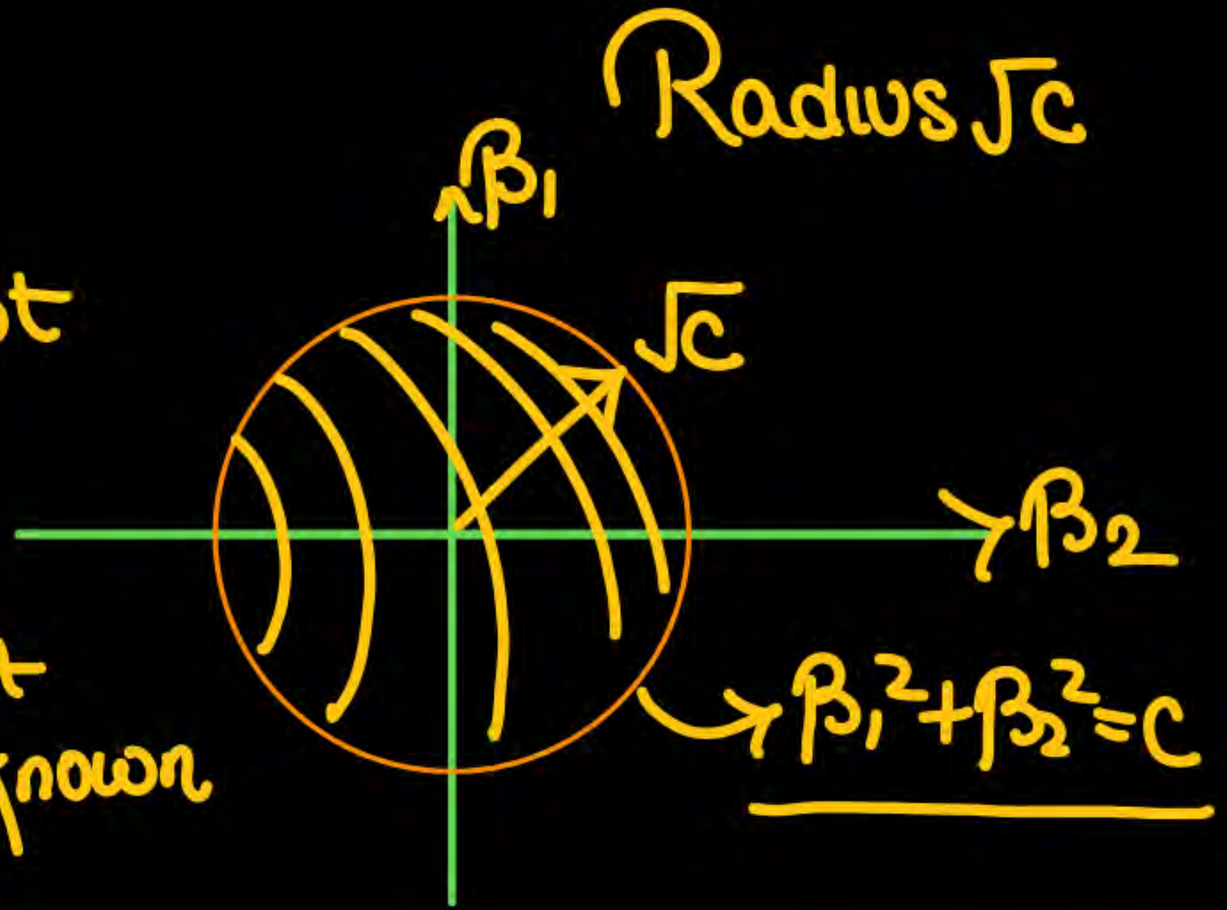
let data is of 2 dimension

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

St. $\beta_1^2 + \beta_2^2 \leq c$ ⇒ Plot

c : Constant

β_1, β_2 unknown



What is Lasso Regularisation

Basic (not in syllabus).

$$\text{loss function} = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2 + \lambda\sum_{i=1}^{D}|\beta_i|$$

The penalty term has absolute value of β.

we minimize the loss function

Arjun

$f(x) = x^2 \rightarrow x = \sqrt{.001} = .031$

Karan

$f(x) = x \rightarrow x = .001$

- I want min $f(x)$, $f(x)$ Ki value Zero nahi ho sakti $f(x) = .001$ Kardo.

When $x = .031$ then $x^2 = .001$

When $x = .001$ then $x = .001$

$$RR \Rightarrow \min RSS + \lambda \sum_{i=1}^{D} \beta_i^2$$

(RR Keha hai $\beta_i = .031$ then $\beta_i^2 \Rightarrow .001$)

$$\text{Lasso} \Rightarrow \min RSS + \lambda \sum |\beta_i|$$

$\rightarrow \beta_i = .001$

- lasso make β's closer to 0 than RR.

→ isiliye lasso model mein से zyada β's = 0 hote hai,

Zyada dimension gayab from model.

Sparse model

lasso give more sparse model.

Ridge Regression

Constraint representation of Lasso Regularisation

- Lasso

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

Such that (2D Case)

$$\sum_{i=1}^{2} |\beta_i| \leq C \rightarrow \text{hyperparameter}$$

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

$\beta_1 > 0$

$\beta_2 < 0$

$\beta_1 - \beta_2 < C$

Both +ve

$\beta_1 + \beta_2 < C$

$\beta_1, \beta_2 < 0$

$-\beta_1 - \beta_2 < C$

$\beta_2 > 0$

$\beta_1 < 0$

$-\beta_1 + \beta_2 < C$

we have a relation
to find β's directly.

- LR → Closed form sol. $\beta = (X^T X)^{-1} X^T Y$

RR → " → $\beta = (X^T X + \lambda I)^{-1} X^T Y$

Lasso → no closed form sol

Model Complexity.

increasing order of Complexity

Lasso

Bahut Saare β's $\Rightarrow$ 0

Ridge

Try to reduce β's of many dimension making them ineffective in model.

LR

$\Downarrow$

include all dimension in model

high.

low

Interpretability

- Complexity high
- Interpretability low

Penalty term $|\beta|$

Lasso is called L1 Regularisation

Ridge is called L2 Regularisation

Penalty term β^2

dPP

* H.W
* 800 drive
*

Ridge Regression

Lasso Vs Ridge Regression

Read

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.

• Linear Classification

• Linear classification

Classification vs Regression...

Linear Regression of an Indicator Matrix

Let's consider a 2-class case

What is an Indicator Matrix

Linear Classification

Linear Regression of an Indicator Matrix

Let's understand
using figures

Linear Regression of an Indicator Matrix

So, now the analysis is as follows :

Linear Regression of an Indicator Matrix

Lets extend the case for K classes

Linear Regression of an Indicator Matrix

How to find the variables for the linear regression

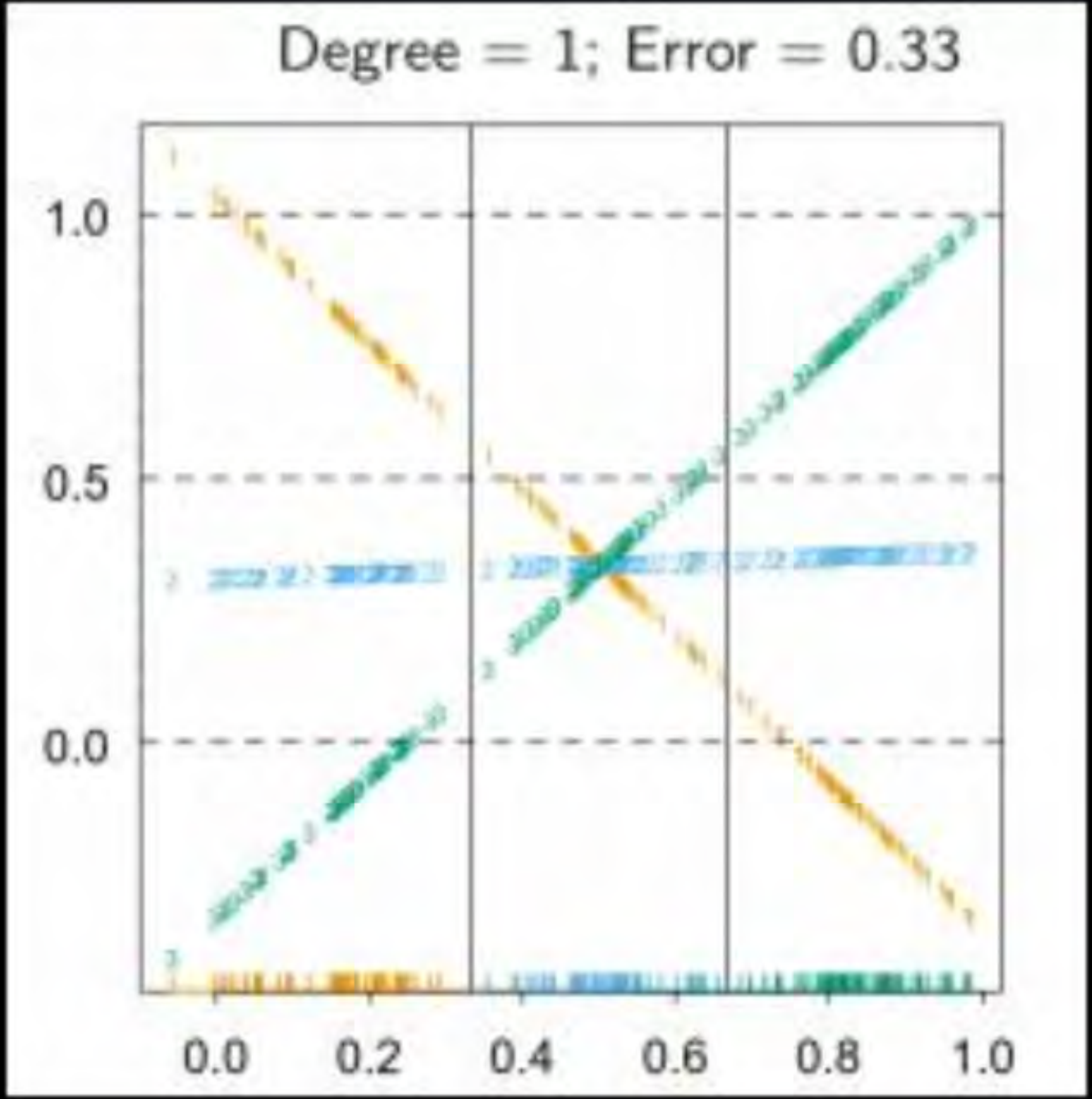
So linear regression can be used for classification also

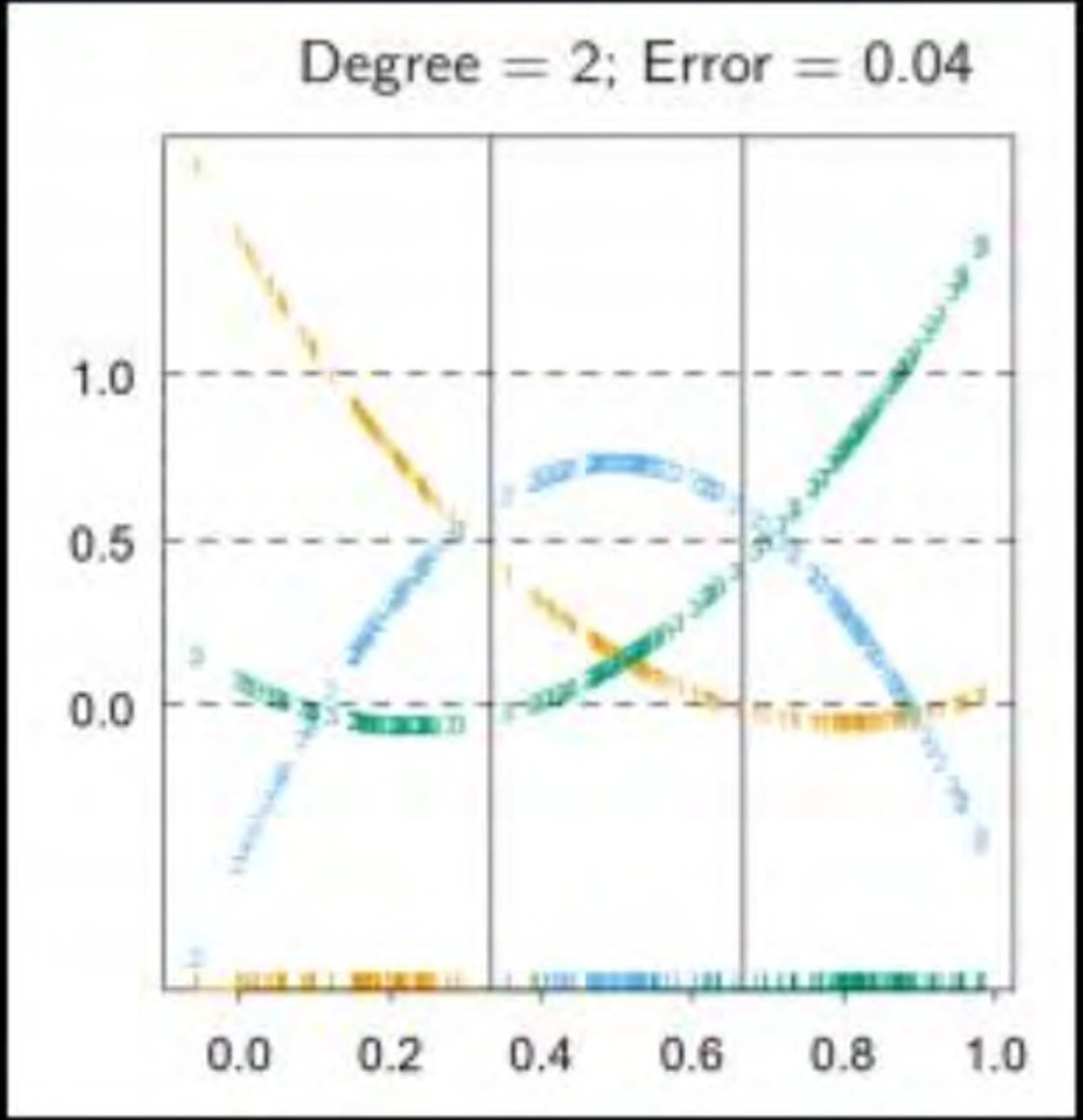
Linear Classification

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 $K \geq 3$ classes are lined up, polynomial terms up to degree $K - 1$ might be needed to resolve them.

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 Regression of an Indicator Matrix

The loss function for a
2 class case...

Linear Classification

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

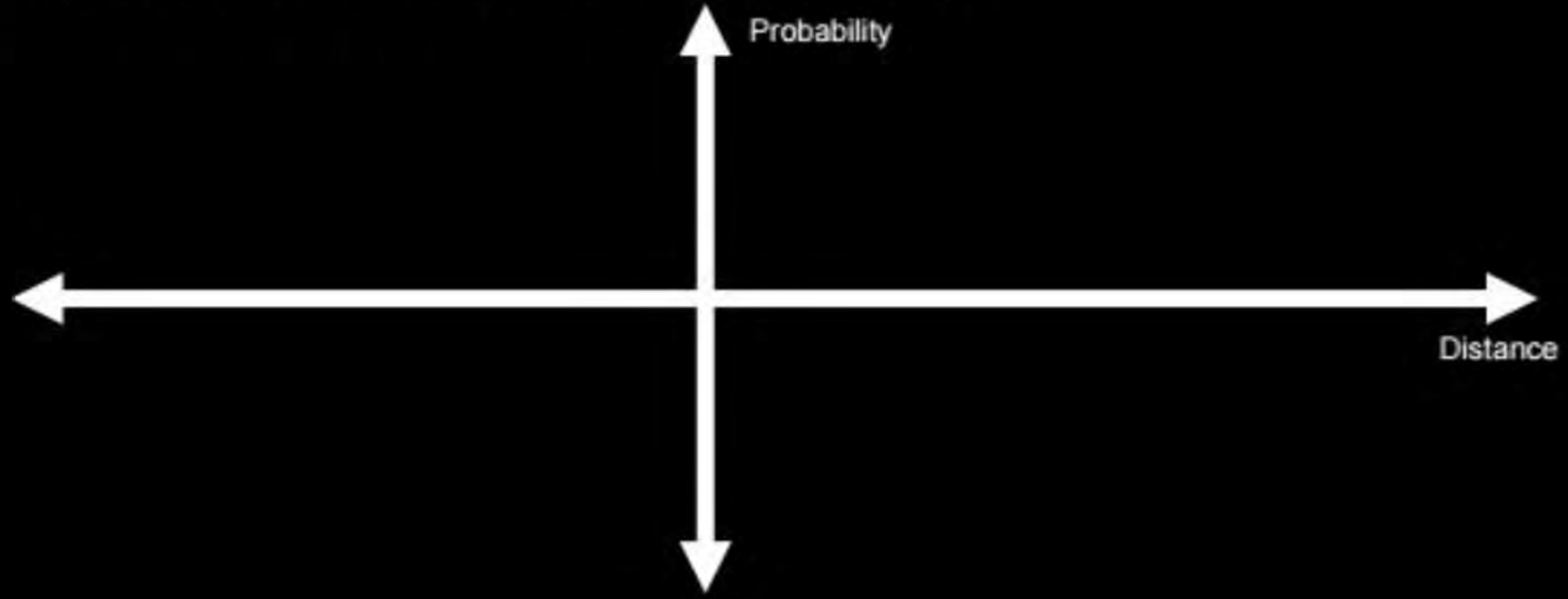
• Linear Classification

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

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.



THANK - YOU