Data Science and Artificial Intelligence

Machine Learning

Regression

Lecture No. 9

By- SIDDHARTH SABHARWAL SIR

Recap of Previous Lecture

Topic — Ridge Regression

Topic — β's expression

Topic

Topic

Topic

Topics to be Covered

- **Topic** · What is 'λ'
- **Topic** · effect of λ, how to find best value of λ
- **Topic** Lasso Regularisation (only v. brief)
- **Topic** Questions.
- **Topic**

Think big

AIR 1

apne Teachers

Ridge Regression Final expression

- $\Rightarrow \beta = (X^T X + \lambda I)^{-1} X^T Y$ → β's obtained from the centred data is valid in original data.
  - → we use centred data here
  - → $y = y - \bar{y}$, $x^i = x^i - \overline{x^i}$
  - → $\beta_0 = (\bar{y} - \beta_1 \overline{x^1} - \beta_2 \overline{x^2} - - - -)$

REVISION

Ridge Regression Final expression

- β_0 is not included in Reg term
- L2 Reg. term
- $L \Rightarrow \min\left(\frac{1}{2}\sum(y_i-\hat{y}_i)^2 + \frac{\lambda}{2}\sum\beta_i^2\right)$

Constant

we can find best λ.

$X = [\;]_{n\times 1}$, n: Number of unknowns

- The basic linear Algebra ⇒

$AX = B$ → unique solution

→ No solution

→ ∞ many solution

⇒ Solution ⇒ $X = A^{-1}B$

- A is matrix of size $(n \times n)$

We create a Augmented matrix

$[A \mid B]$

1. Rank of Augmented matrix = Rank of A = n → unique Sol
2. " " " = " " < n → ∞ Sol
3. " " " ≠ " " A → No Solution.

Q. What will happen if $X^TX^{-1} \Rightarrow$ non inv

So in linear regression

$$\underbrace{(X^TX)}_{A}\beta = \underbrace{X^TY}_{B}$$

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

$$X^TY = \begin{bmatrix} \\ \\ \end{bmatrix}_{D+1 \times 1}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_D \end{bmatrix}_{D+1 \times 1}$$

$$(X^TX) = \begin{bmatrix} & \\ & \end{bmatrix}_{D+1 \times D+1}$$

- $(X^TX)^{-1}$ will have dimension of $(D+1 \times D+1)$

→ X^TX will be non inv if $|X^TX| = 0$

OR Rank of $X^TX < (D+1)$

Now Augmented matrix in this case

$$\left[\underbrace{(X^TX)}_{A} \;\middle|\; \underbrace{X^TY}_{B}\right]$$

If X^TX is non inv then Rank of A will be $< D+1$, and in that case we can have 2 situation

→ i.e No solution ✓

→ i.e ∞ many solutions ✓

$$\underbrace{\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}}_{A} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \underbrace{\begin{bmatrix} j \\ k \\ l \end{bmatrix}}_{B}$$

$$Aug = \left[\begin{array}{ccc|c} a & b & c & j \\ d & e & f & k \\ g & h & i & l \end{array}\right]_{\underline{3\times4}}$$

If Rank of $A = \underbrace{2}$, then Rank of Aug ⇒ Can be 3/2.

So the case of unique solution arise only when Rank of $X^T X \geq D+1$

Shrinkage Methods : Ridge Regression

- Solution to this ridge regression problem

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

- $\lambda=0 \Rightarrow$ this Become basic L.R $\Rightarrow$ Overfitting
  - $\rightarrow$ V. low training error
  - $\rightarrow$ i.e V. low Bias
  - $\rightarrow$ V. high testing error
  - $\rightarrow$ i.e V. high variance

- If $\lambda=\infty$ then Regularisation terms become dominant, the algorithm try to make all $\beta=0$ except β_0
  - $\rightarrow$ underfitting
  - $\rightarrow$ high training and testing error.

Shrinkage Methods : Ridge Regression

- **Solution to this ridge regression problem**

done

Ridge Regression

❖ How to choose the best λ.

Cross Validation

- In basic linear regression the complexity was very less
- we solve the algorithm just once to find β

- But in RR the algorithm has to run for various λ's and then we find the best λ to be used in the algorithm.

- So RR has much more complexity than L.R

Ridge Regression

❖ How to choose the best λ.

data ⇒ Training data / Testing data.

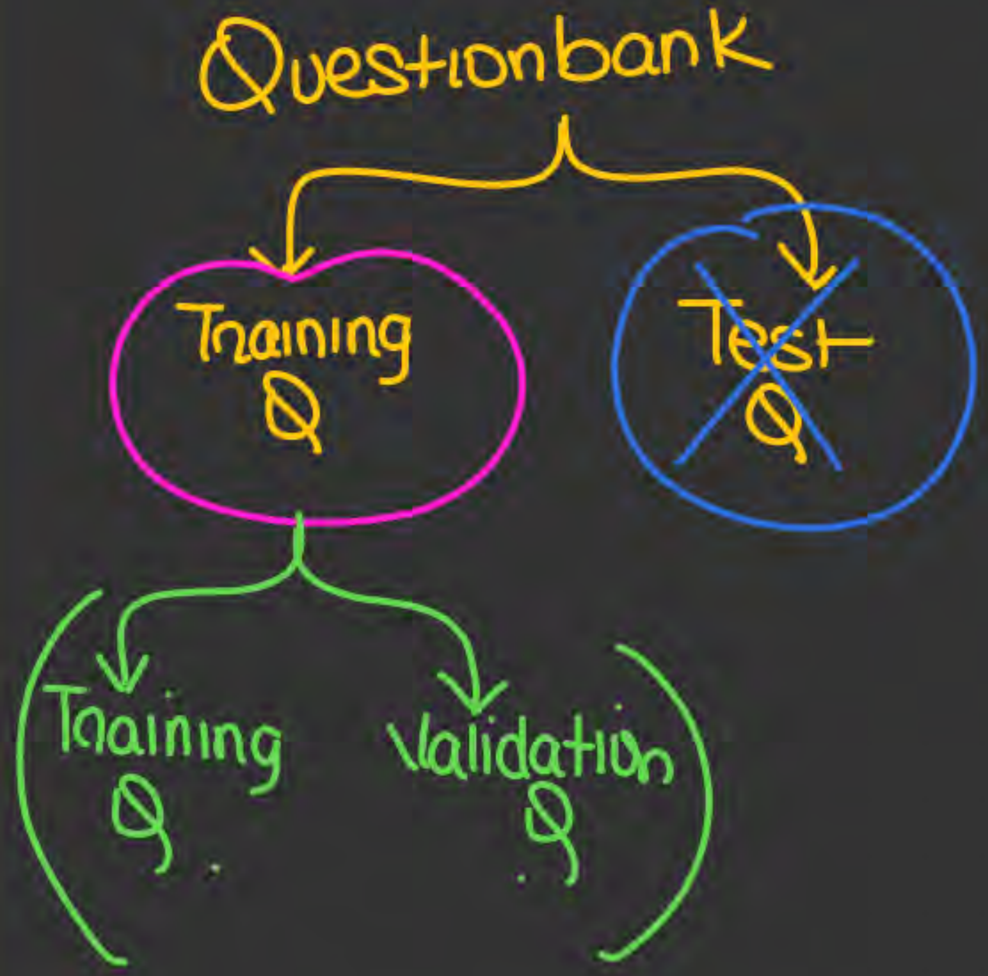
- So data is broken into training & testing data
- Testing data is not available to us
- we only have the training data.

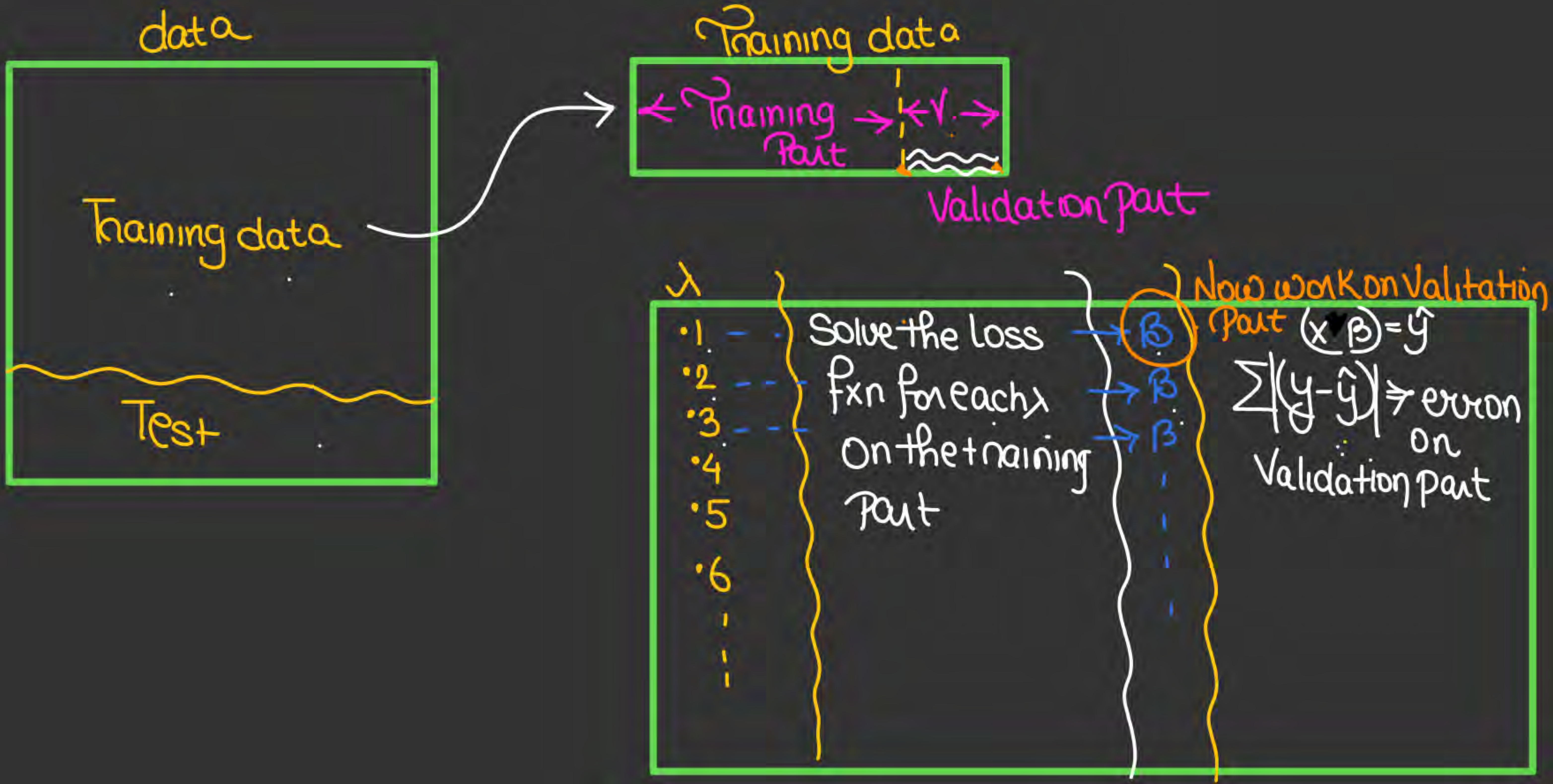
Cross Validation

To find best λ ⇒

So we break the Training data into 2 parts ⇒ Training and Validation parts.

- So we traing the model on training part for various λ's and then test the model on the Validation part.
- The λ that give the best performance on Validation part will be choosen

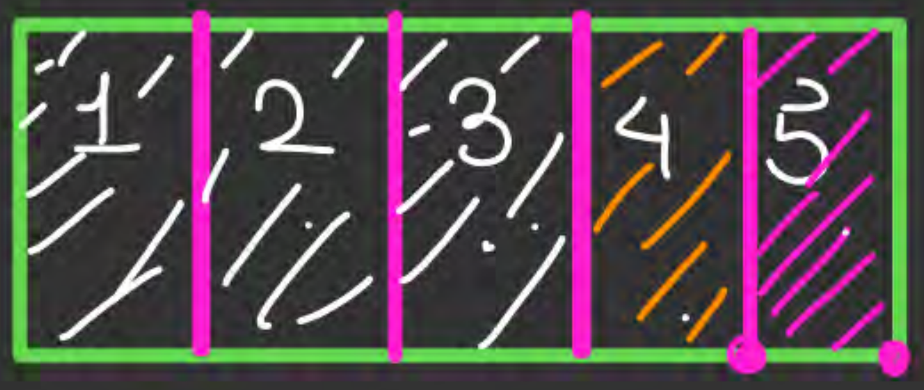



data
Training data
Test
Training data
Training Part
V.
Validation Part
λ
1
2
3
4
5
6
Solve the loss fxn for each λ On the training Part
β
β
β
Now work on Validation Part $(X\beta) = \hat{y}$
$\sum |(y-\hat{y})| \Rightarrow$ error on Validation part

- So for various λ's we train the model and then test it on the validation and then simply find best λ

- But in the above process we are actually finding best λ that suits our validation part.

So we do cross validation



5-fold cross validation

→ Training data is broken into 5 parts

Step 1 — 1,2,3,4 ⇒ Training Part
5 ⇒ validation part
↳ find best λ.

Step 2 ⇒ 1,2,3,5 ⇒ Training Part
4 ⇒ Validation
↳ find best

Step 3 ⇒

Step 4 ⇒

Step 5 ⇒

The best λ for overall data → avg of all

↳ This process is called K fold cross validation.

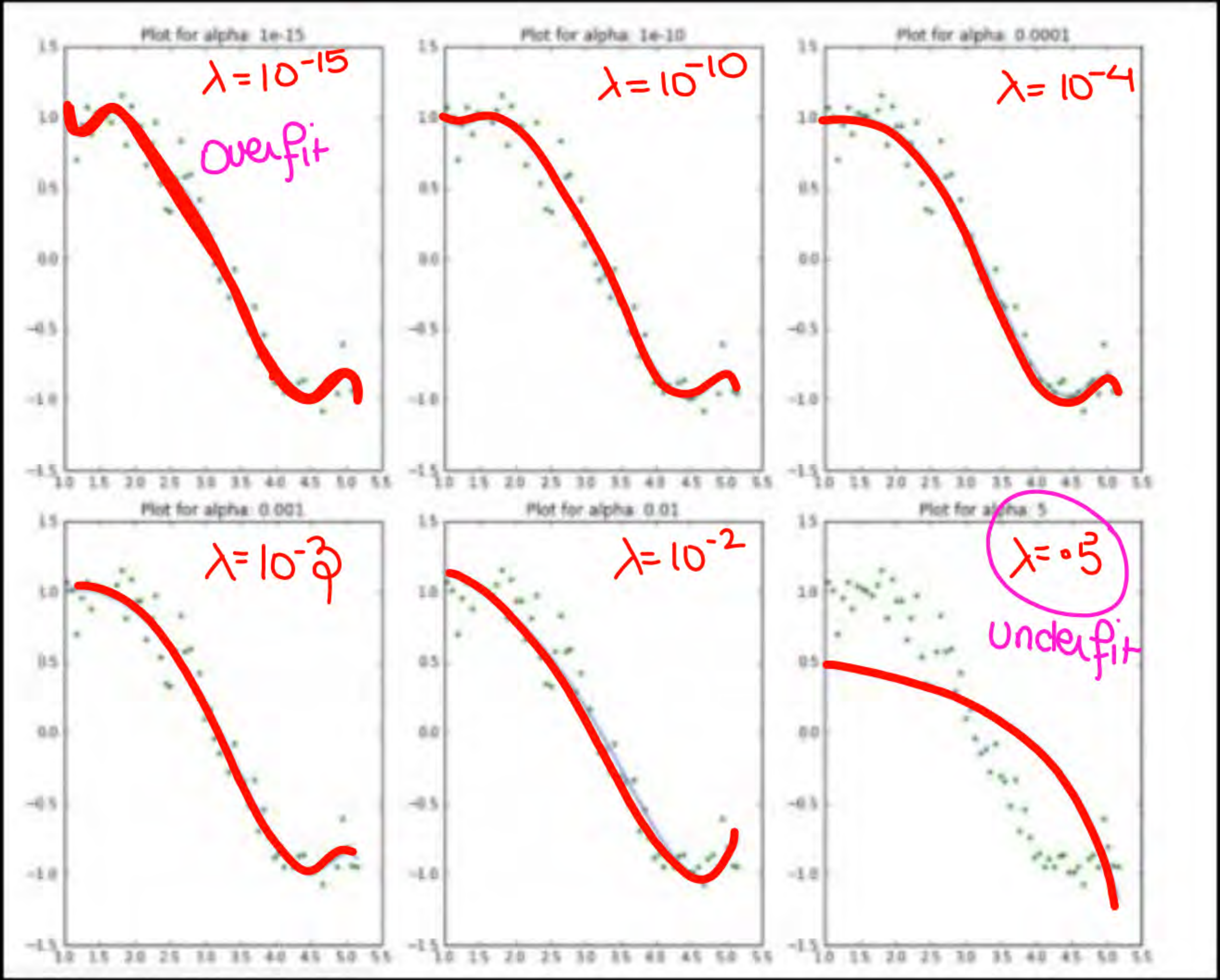
Ridge Regression

❖ Can λ be negative.

$$L \Rightarrow \min\left[\frac{1}{2}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2 + \frac{\lambda}{2}\sum_{i=1}^{D}\beta_i^2\right]$$

* If λ is negative ⇒ then $\beta \Rightarrow \infty$

* So λ will never be -ve.


Plot for alpha: 1e-15
λ=10^-15
Overfit
Plot for alpha: 1e-10
λ=10^-10
Plot for alpha: 0.0001
λ=10^-4
Plot for alpha: 0.001
λ=10^-3
Plot for alpha: 0.01
λ=10^-2
Plot for alpha: 5
λ=5
Underfit

So as λ inc

→ model Complexity reduce

→ Training error inc / becoz overfitting reduce

→ model generalise better

→ Testing error reduce

But large λ ⇒ underfitting

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:

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:

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 – lets practise

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

A) $\lambda = 0$
B) $\lambda = 1$
C) $\lambda = 10$
D) $\lambda = \infty$

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.

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.

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

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

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?

Ridge Regression – The constraint representation of the equation...

→ The loss function of RR ⇒ $\mathcal{L}$ ⇒ $\min\left(\frac{\sum_{i=1}^{N}(y_i-\hat{y}_i)^2}{2} + \frac{\lambda}{2}\sum_{i=1}^{D}\beta_i^2\right)$

Can be written as a Constraint minimization problem

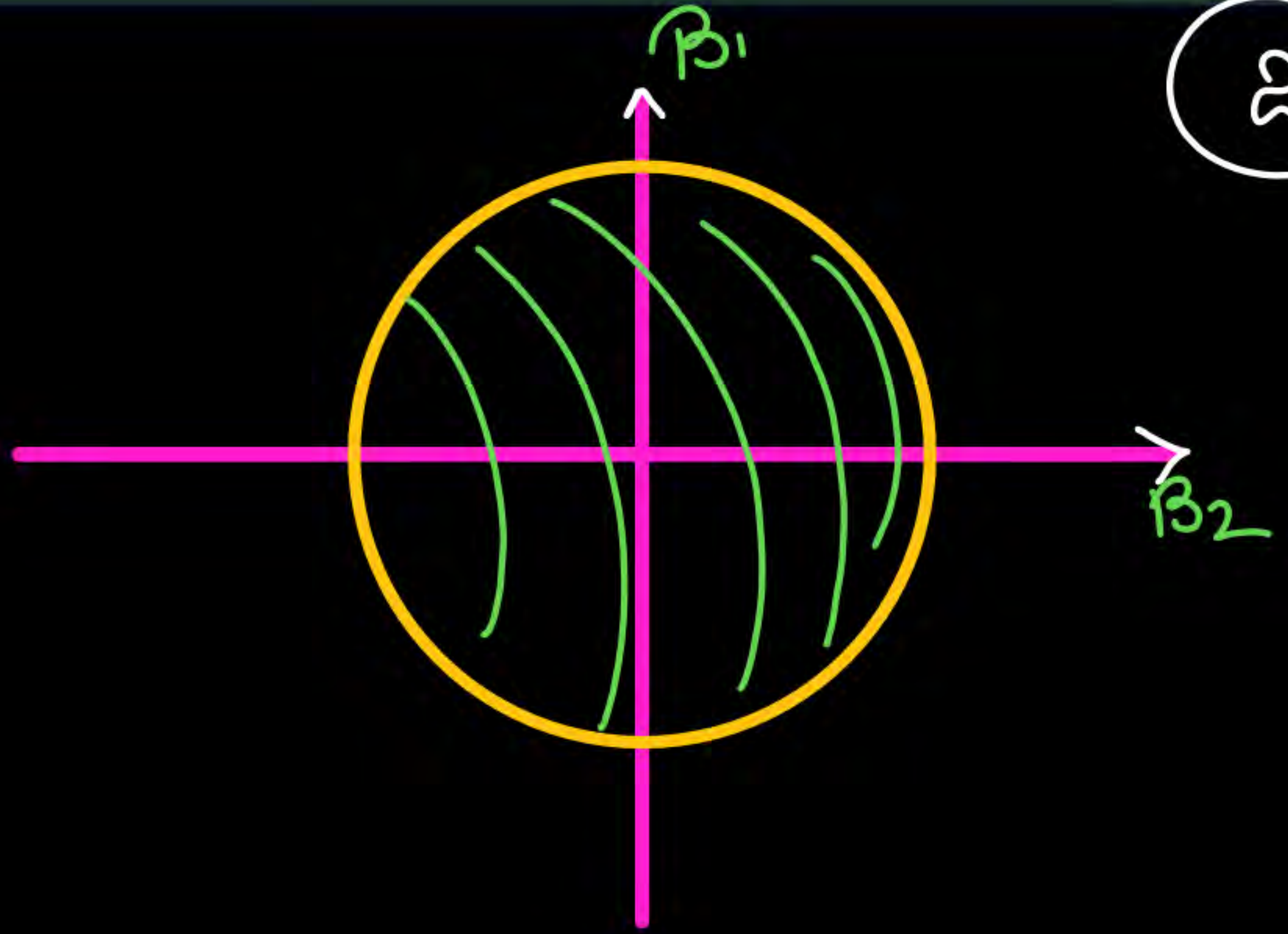
⇒ $\min\left(\frac{1}{2}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2\right)$

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

- How to find best 'C'

⇒ By Cross Validation

Ridge Regression – Shape of the contraint



2D data

Constraint

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

$$\Rightarrow \beta_1^2 + \beta_2^2 \leq C$$

Circular region of radius $\sqrt{c}$

Ridge Regression

Ridge Regression – lets practise

Numenical

2D data

x^1	x^2	y
5	6	10
8	9	15
15	10	20
4	5	9

Centering of data

$\bar{y} \Rightarrow 13.5$

$\bar{x^1} \Rightarrow 8$

$\bar{x^2} \Rightarrow 7.5$

x^1	x^2	y
-3	-1.5	-3.5
0	1.5	1.5
7	2.5	6.5
-4	-2.5	-4.5

Ridge Regression – lets practise

$$X \Rightarrow \begin{bmatrix} -3 & -1.5 \\ 0 & 1.5 \\ 7 & 2.5 \\ -4 & -2.5 \end{bmatrix}$$

$$(X^T X) = \begin{bmatrix} -3 & 0 & 7 & -4 \\ -1.5 & 1.5 & 2.5 & -2.5 \end{bmatrix} \begin{bmatrix} -3 & -1.5 \\ 0 & 1.5 \\ 7 & 2.5 \\ -4 & -2.5 \end{bmatrix}$$

$$(X^T X) = \begin{pmatrix} 74 & 32 \\ 32 & 17 \end{pmatrix}, \quad X^T Y \Rightarrow \begin{bmatrix} -3 & 0 & 7 & -4 \\ -1.5 & 1.5 & 2.5 & -2.5 \end{bmatrix} \begin{bmatrix} -3.5 \\ 1.5 \\ 6.5 \\ -4.5 \end{bmatrix}$$

$$= \begin{bmatrix} 74 \\ 35 \end{bmatrix}$$

So let $\lambda = 1$

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

$X^T X + \lambda I \Rightarrow \begin{pmatrix} 74 & 32 \\ 32 & 17 \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$

$\Rightarrow \begin{pmatrix} 75 & 32 \\ 32 & 18 \end{pmatrix}$

$\Rightarrow \frac{1}{326} \begin{pmatrix} 18 & -32 \\ -32 & 75 \end{pmatrix}$

$\Rightarrow \frac{1}{326} \begin{pmatrix} 18 & -32 \\ -32 & 75 \end{pmatrix} \begin{pmatrix} 74 \\ 35 \end{pmatrix}$

$\beta \Rightarrow \frac{1}{326} \begin{pmatrix} 212 \\ 257 \end{pmatrix}$

So $\beta_1 = \frac{212}{326}$, $\beta_2 = \frac{257}{326}$

$$\beta_0 = \left(\bar{y} - \beta_1 \overline{x1} - \beta_2 \overline{x^2}\right)$$

$$\Rightarrow 13.5 - 8 \times \frac{212}{326} - 7.5 \times \frac{257}{326}$$

$$\beta_0 \Rightarrow 2.384$$

What is Lasso Regularisation

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

- Square term in RR → L2 Regularisation
- In lasso → Single power term
  - → L1 Regularisation.

Ridge Regression

Lasso Vs Ridge Regression

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.

2 mins Summary

Topic

Topic

Topic

Topic

Topic

THANK - YOU