Data Science and Artificial Intelligence

Machine Learning

Regression

Lecture No. 9

By- SIDDHARTH SABHARWAL SIR

Recap of Previous Lecture

Topic

- OLS → Overfit
  → Multicollinearity → unstable model
  → means Min RSS
  → Solution Ridge Reg.

Regularisation

Topic

Topic

Topic

Topic

Topics to be Covered

Topic

Ridge regression

Topic

Topic

Topic

Topic

acha
Rank

Gate 2026
DSAI
DA
acha Rank.

Ridge Regression Final expression

Shrinkage Methods : Ridge Regression

- Ridge regression is a regularisation techniques...

updated loss fxn.

Hyperparameter

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

β_0 not included ??

model LR wala hai $\hat{y}_i = \beta_0 + \beta_1 x^1 + \beta_2 x^2 + \cdots$

we want to min the loss fxn, to find best β's.

Shrinkage Methods : Ridge Regression

- "In regularization technique, we reduce the magnitude of the features by keeping the same number of features.
- This helps in ....

(i) we are not reducing / limiting the No of dimension
for example if data has 100 dimensions
then we train model with 100 dimensions

(ii) In Regularisation we put limit on β's of the dimension.

Shrinkage Methods : Ridge Regression

- Ridge regression shrinks the regression coefficients by imposing a penalty on their size.
- The ridge coefficients minimize a penalized residual sum of squares of the weights.

we add $\frac{\lambda}{2}\sum_{i=1}^{D}\beta_i^2$ in the loss fxn.

The loss function are updated

Shrinkage Methods : Ridge Regression

done

The loss function are updated

Shrinkage Methods : Ridge Regression

The main reason for not regularizing the intercept term is that it represents the mean value of the target variable when all the features are zero. Regularizing the intercept can lead to shifting this mean value away from its natural value, which might not be desirable in many cases.

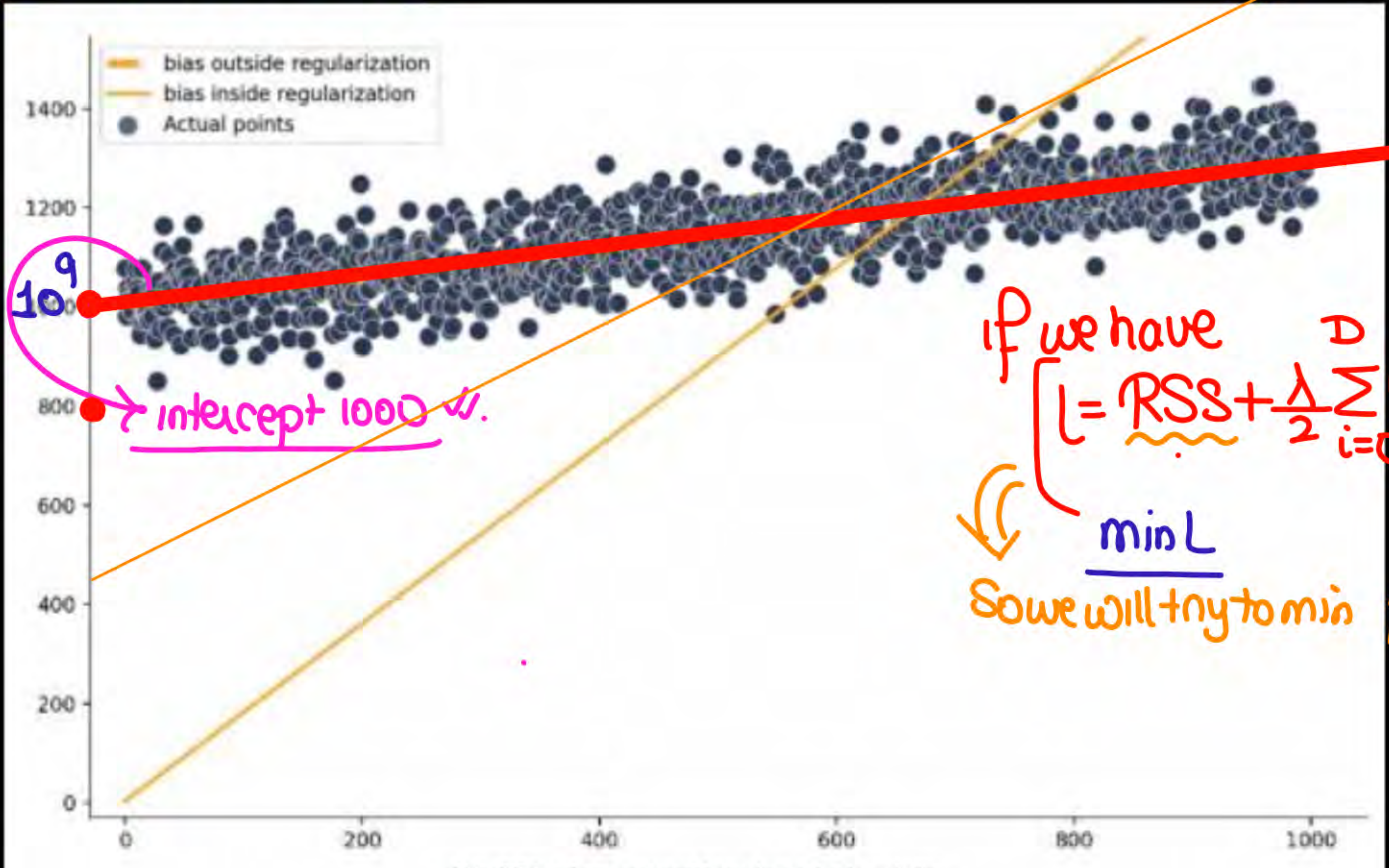
Why the ~~bias~~ term is not included in regularisation ..

* Why β_0 is not included
* Why we donot want β_0 to be minimize ??

we always want values of β's shd be within some limit

bcoz if any dimension get large β, then that dimension dominate model ⇒ unstable model.

Ridge Regression



This GIF has been sourced from the author's website

10^9

intercept 1000 w.

if we have $L = RSS + \frac{\lambda}{2}\sum_{i=0}^{D}\beta_i^2$

min L

So we will try to min β_0

$$y = \beta_0 + \beta_1 x$$

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

Similarly

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

$$\beta_0 = \left(\bar{y} - \beta_1 \overline{x^1} - \beta_2 \overline{x^2} \dots\right)$$

* Value of β_0 depend on mean location of data

* β_0 shd not be minimized

Complete analysis of RR

we have the data $\Rightarrow$ Step1 Remove the β_0 from analysis.

x^1	x^2	y

$\hookrightarrow \overline{x^1}$ $\hookrightarrow \overline{x^2}$ $\hookrightarrow \overline{y}$

So Create Centred data

Now x^1. x^2. y.

$x^1-\overline{x^1}$	$x^2-\overline{x^2}$	$y-\overline{y}$

Now apply RR $\Rightarrow$ $\boxed{\text{2D data.}}$

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

N data points and 2 dimension

$$\left(\hat{y}_i = \beta_1 x_i^1 + \beta_2 x_i^2\right)$$

$$\mathcal{L} = \sum_{i=1}^{N} \left(y_i - \left(\beta_1 x_i^1 + \beta_2 x_i^2\right)\right)^2 + \frac{\lambda}{2} \sum_{i=1}^{2} \beta_i^2$$

min L $\Rightarrow$

$$\frac{\partial L}{\partial \beta_1} = -2 \sum_{i=1}^{N} x_i^1 \left(y_i - \beta_1 x_i^1 - \beta_2 x_i^2\right) + \lambda \beta_1 = 0$$

$$\frac{\partial L}{\partial \beta_2} = -2 \sum_{i=1}^{N} x_i^2 \left(y_i - \beta_1 x_i^1 - \beta_2 x_i^2\right) + \lambda \beta_2 = 0$$

$\underline{\min L} \Rightarrow \dfrac{\partial L}{\partial \beta_1} = -2\sum_{i=1}^{N} x_i^1 (y_i - \beta_1 x_i^1 - \beta_2 x_i^2) + \lambda \beta_1 = 0$

$$\frac{\partial L}{\partial \beta_2} = -2\sum_{i=1}^{N} x_i^2 (y_i - \beta_1 x_i^1 - \beta_2 x_i^2) + \lambda \beta_2 = 0$$

$$\frac{\partial L}{\partial \beta} = \begin{bmatrix} \partial L/\partial \beta_1 \\ \partial L/\partial \beta_2 \end{bmatrix} = -2\begin{bmatrix} \sum x_i^1 y_i - \beta_1 \sum (x_i^1)^2 - \beta_2 \sum x_i^1 x_i^2 \\ \sum x_i^2 y_i - \beta_1 \sum x_i^2 x_i^1 - \beta_2 \sum (x_i^2)^2 \end{bmatrix} + \lambda \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = 0$$

$$= -2\left[\begin{bmatrix} \sum x_i^1 y_i \\ \sum x_i^2 y_i \end{bmatrix} - \begin{bmatrix} \sum (x_i^1)^2 & \sum x_i^1 x_i^2 \\ \sum x_i^1 x_i^2 & \sum (x_i^2)^2 \end{bmatrix}\begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}\right] + \lambda \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = 0$$

$$\frac{\partial L}{\partial \beta} = \begin{bmatrix} \partial L/\partial \beta_1 \\ \partial L/\partial \beta_2 \end{bmatrix} = -2 \begin{bmatrix} \sum x_i^1 y_i - \beta_1 \sum (x_i^1)^2 - \beta_2 \sum x_i^1 x_i^2 \\ \sum x_i^2 y_i - \beta_1 \sum x_i^2 x_i^1 - \beta_2 \sum (x_i^2)^2 \end{bmatrix} + \lambda \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = 0$$

$$= -2 \left[\underbrace{\begin{bmatrix} \sum x_i^1 y_i \\ \sum x_i^2 y_i \end{bmatrix}}_{X^T Y} - \underbrace{\begin{bmatrix} \sum (x_i^1)^2 & \sum x_i^1 x_i^2 \\ \sum x_i^1 x_i^2 & \sum (x_i^2)^2 \end{bmatrix}}_{X^T X} \underbrace{\begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}}_{\beta} \right] + \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = 0$$

$$X = \begin{bmatrix} x_1^1 & x_1^2 \\ x_2^1 & x_2^2 \\ x_3^1 & x_3^2 \\ \vdots & \vdots \end{bmatrix} \quad X^T = \begin{bmatrix} x_1^1 & x_2^1 & x_3^1 & \cdots \\ x_1^2 & x_2^2 & x_3^2 & \cdots \end{bmatrix}$$

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \end{bmatrix}$$

$$\frac{\partial L}{\partial \beta} = -2\left[X^T Y - (X^T X)\beta\right] + \lambda I \beta = 0.$$

$$= -2X^T Y + 2(X^T X)\beta + \lambda I \beta = 0$$

$$\Rightarrow (X^T X)\beta + \frac{\lambda}{2} I \beta = X^T Y$$

$$\Rightarrow \left(X^T X + \frac{\lambda}{2} I\right)\beta = X^T Y$$

$$\boxed{\beta = \left(X^T X + \frac{\lambda}{2} I\right)^{-1} X^T Y.}$$

$\Downarrow$

(I) Centre the data

(II) we will get β_1, β_2

Centred data ka model

$\underline{y} = \beta_1 \underline{x^1} + \beta_2 \underline{x^2}$

Centred

Original model will be

$$\underline{y} = \beta_0 + \beta_1 \underline{x^1} + \beta_2 \underline{x^2}$$

Original Values

$$\beta_0 = \bar{y} - \beta_1 \overline{x^1} - \beta_2 \overline{x^2}$$

2D data

x^1	x^2	y
5	8	16
7	10	26
9	12	30
7	14	40

$\overline{x^1} = 7 \quad \overline{x^2} = 11 \quad \overline{y} = 28$

Find $\beta_0, \beta_1, \beta_2$.

Ridge Reg $\lambda = 2$.

x^1	x^2	y
-2	-3	-12
0	-1	-2
2	1	2
0	3	12
$x^1 - \overline{x^1}$	$x^2 - \overline{x^2}$	$y - \overline{y}$

$$X = \begin{bmatrix} -2 & -3 \\ 0 & -1 \\ 2 & 1 \\ 0 & 3 \end{bmatrix} \quad X^T = \begin{bmatrix} -2 & 0 & 2 & 0 \\ -3 & -1 & 1 & 3 \end{bmatrix}$$

$$(X^T X) = \begin{bmatrix} 8 & 8 \\ 8 & 20 \end{bmatrix}$$

$$X^T Y = \begin{bmatrix} 28 \\ 76 \end{bmatrix}$$

$$\left(X^T X + \frac{\lambda}{2} I\right) = \begin{bmatrix} 9 & 8 \\ 8 & 21 \end{bmatrix}$$

$$\beta = \left(X^T X + \frac{\lambda}{2} I\right)^{-1} X^T Y$$

$$= \frac{1}{125} \begin{bmatrix} 21 & -8 \\ -8 & 9 \end{bmatrix} \begin{bmatrix} 28 \\ 76 \end{bmatrix} \Rightarrow \begin{bmatrix} 20/125 \\ 460/125 \end{bmatrix}$$

$\checkmark\ \beta_1 = -.16$

$\checkmark\ \beta_2 = 3.68$

$$\bar{y} - \beta_1 \bar{x}_1 - \beta_2 \bar{x}_2 = \beta_0$$

$\checkmark\ \beta_0 = -11.36$

- if RR
- $L = RSS + \lambda \sum_{i=1}^{D} \beta_i^2$

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

Ridge Regression

Shrinkage Methods : Ridge Regression

- Here $\lambda \geq 0$ is a complexity parameter that controls the amount of shrinkage:

- we always keep $\lambda \geq 0$

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

Loss fxn has 2 terms

RSS + Penalty.

- $\lambda = 0 \Rightarrow$ Same as L.R
- $\lambda =$ Bahut Badi

λ: hyperparameter

λ ~~~~~~~~

Control Karta hai.

- λ Control that which term is more Imp
- λ large Penalty term imp

- we always keep $\lambda \geq 0$

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

Loss fxn has 2 terms

RSS + Penalty.

- $\lambda = 0 \Rightarrow$ Same as L.R
- $\lambda =$ Bahut Badi

λ: hyper parameter

λ ~~~~~~~~ Control karta hai.

- λ is very small then RSS shd be zero or very small even if $\sum \beta^2$ is large

- λ Control that which term is more Imp
- λ large Penalty term imp
  - ↳ matlab chahe RSS is large, we need $\sum \beta^2$ to be v. small

λ Ko badhane se kya hoga.

λ = small

RSS Pyara hai

RSS $\rightarrow 0$

$\Sigma\beta^2$ Can be large

Same as Linear Reg.

Overfit model.

How to find best β

hyperparameter tuning

Cross Validation

λ large

RSS kuch ho Jae

$\Sigma\beta^2$ very very small

Underfit model

- we always keep $\lambda \geq 0$

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

- $\lambda = 0 \Rightarrow$ Same as L.R
- $\lambda =$ Bahut Badi

λ: hyperparameter

λ ~~~~~~

Control Karta hai.

What if λ is -ve

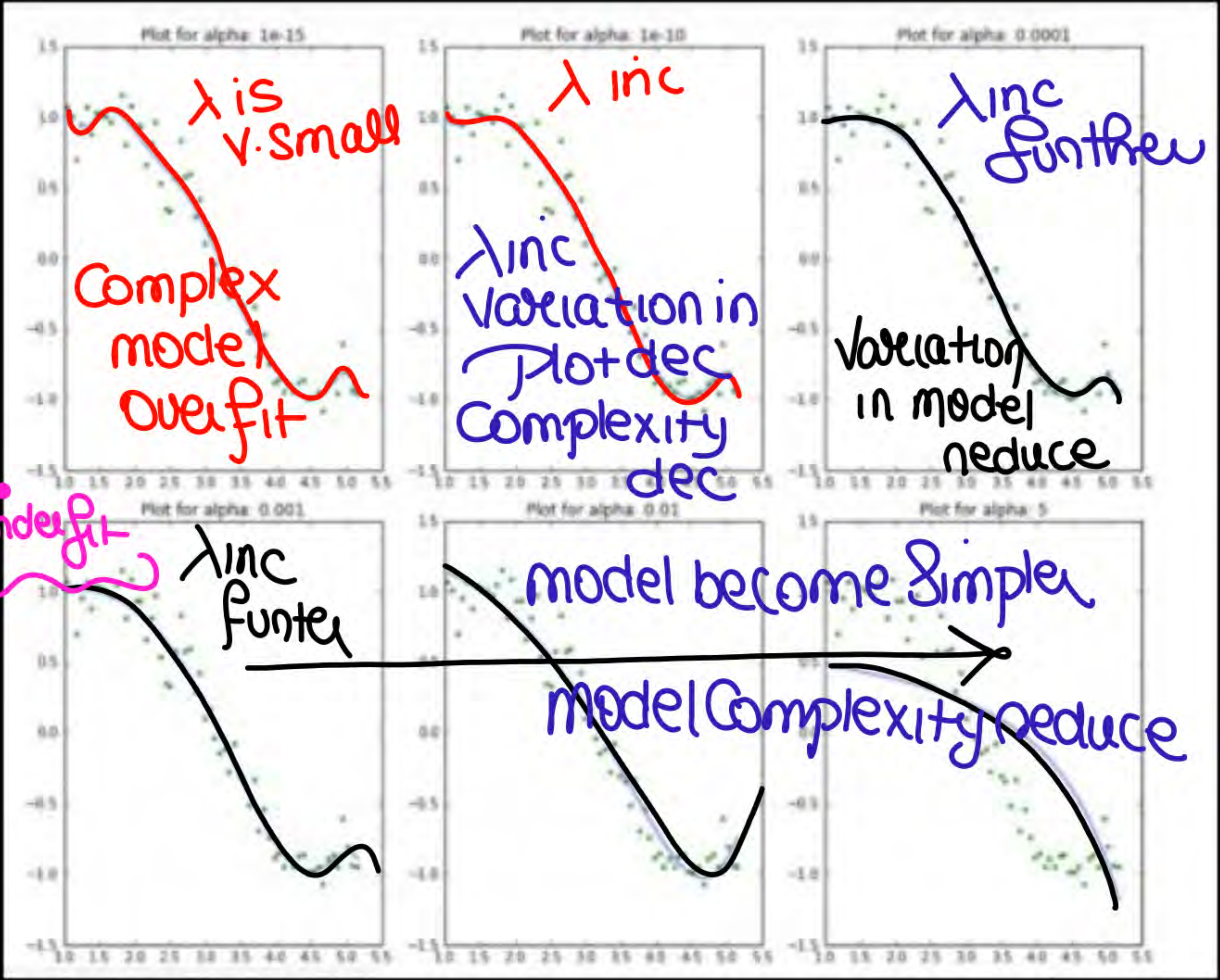
- If λ is -ve
- minimize loss fxn.
- Underfit model

-ve λ never used.

$\boxed{\lambda > 0}$

λ value

overfit → → → underfit



How to find best λ

In many algo we have → hyperparameter ⇒ they are used while training the model.

→ inki value ko hit and trail karke best value choose karte hai.

→ Parameters ⇒ Jo hamari final model ki equation ka part bante hein

→ Values of these are found after training process ends.

THANK - YOU