Data Science and Artificial Intelligence

Machine Learning

Regression

Lecture No. 8

By- SIDDHARTH SABHARWAL SIR

Recap of Previous Lecture

- Topic
- Topic
- Topic
- Topic
- Topic

Assumption in LR

H.W

Topics to be Covered

- Topic
- Topic
- Topic
- Topic
- Topic

Sol. uploaded

Adv and dis adv of LR

Space and time Complexity of LR

Ridge Regression


Be the change
that you wish to
see in the world.
— MAHATMA GHANDI
BRIAN TRACY
INTERNATIONAL

- advantage of LR } ⇒

(i) Simplest algorithm

(ii) High interpretability → $y = 3 + 2x^1 + 3x^2 + 10x^3$

- from the eq of model
- we can interpret
  - * Importance of each dimension

* It is possible to interpret the Plane created from model

REVISION

Problems in LR ...

Disadvantage in LR →

(i) It work only when data is linearly related

(ii) data must not have multicollinearity

(iii) data must not have Heteroscedasticity.

(iv) LR produces unstable model

(v) algorithm is severely effected by outlier

Linear Regression

Advantages of Simple Linear Regression:

- Simplicity and ease of interpretation.
- Transparent modeling with clear coefficient interpretations.
- Computational efficiency, suitable for large datasets.
- A baseline model for assessing feature significance.
- Effective when the relationship between variables is linear.

Disadvantages of Simple Linear Regression:

- Limited to linear relationships, may perform poorly for nonlinear data.
- Sensitive to outliers, leading to parameter influence.
- Prone to underfitting when facing complex relationships.
- Assumptions of independent and normally distributed errors are critical.
- Suitable only when one independent variable is involved in the analysis.

- Assumptions of independent and normally distributed errors are critical.
- Suitable only when one independent variable is involved in the analysis.

Homoscedasticity

Gaussian noise

Residue = due to noise

Residue = 0

- LR has one more assumption
  that data points must be independent to each other.

Linear Regression

Advantage & Disadvantage of Linear Regression

Advantages	Disadvantages
Linear Regression is simple to implement and easier to interpret the output coefficients.	On the other hand in linear regression technique outliers can have huge effects on the regression and boundaries are linear in this technique.
When you know the relationship between the independent and dependent variable have a linear relationship, this algorithm is the best to use because of it's less complexity compared to other algorithms.	Diversely, linear regression assumes a linear relationship between dependent and independent variables. That means it assumes that there is a straight-line relationship between them. It assumes independence between attributes.
Linear Regression is susceptible to over-fitting but it can be avoided using some dimensionality reduction techniques, regularization (L1 and L2) techniques and cross-validation.	But then linear regression also looks at a relationship between the mean of the dependent variables and the independent variables. Just as the mean is not a complete description of a single variable, linear regression is not a complete description of relationships among variables.

Linear Regression

Why Linear Regression is Important

The interpretability of linear regression is a notable strength.
The model's equation provides clear coefficients that elucidate the impact of each independent variable on the dependent variable, facilitating a deeper understanding of the underlying dynamics.
Its simplicity is a virtue, as linear regression is transparent, easy to implement, and serves as a foundational concept for more complex algorithms.

REVISION

Why LR is unstable model...

$|A| = .002$ $A = \begin{bmatrix} 1 & 2 \\ 1.000 & 2.002 \end{bmatrix}$ $\Rightarrow$ invertible A^vv.

→ $|A|$ = very small = close to 0

$A^{-1} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ So terms in A^{-1} matrix will be V. large.

$|A| = .001$ $\begin{bmatrix} 1 & 2 \\ 1.000 & 2.001 \end{bmatrix}$

$$X = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 2 & 4 \\ 1 & 3 & 6 \\ 1 & 4 & 8 \end{bmatrix} \qquad x^2 = 2x^1$$

If X has multicollinearity

$$X^TX = \begin{bmatrix} \text{2 or more than 2 rows equal} \end{bmatrix}$$

- $|X^TX| = 0$
- $X^TX \Rightarrow$ non Invertible

$$X = \begin{bmatrix} 1 & 1 & .9999 \\ 1 & 2 & 1.9889 \\ 1 & 3 & 3.0002 \\ 1 & 4 & 4.0001 \end{bmatrix}$$

✓ $(X^TX) = \begin{bmatrix} \cdots \\ \cdots \end{bmatrix}$ Row exactly Same nahi

$|X^TX| \neq 0$, $\beta = (X^TX)^{-1}X^TY$

Slight change in data

$$X = \begin{bmatrix} 1 & 1 & .9999 \\ 1 & 2 & 1.9889 \\ 1 & 3 & 3.0002 \\ 1 & 4 & 4.0001 \end{bmatrix}$$

$(X^TX)^{-1}X^TY$ we get large β's

$$X = \begin{bmatrix} 1 & 1 & .9989 \\ 1 & 2 & 1.9998 \\ 1 & 3 & 3.0001 \\ 1 & 4 & 4.0002 \end{bmatrix}$$

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

β's will be of large value

But

β's value will be totally different.

- $(X^TX)^{-1}X^TY$

Since $(X^TX)^{-1}$, all the terms will be vv. large.

$\beta \Rightarrow$ we will get very large β values.

→ So in LR due to multicollinearity we may get large β's.

- LR give unstable model.

⇓

Slight change in data

Produces Completely new model

Linear Regression

Space and Time Complexity of Linear Regression

Assumptions:

n = number of training examples, **m** = number of features, **n'** = number of support vectors,

k = number of neighbors, **k'** = number of trees

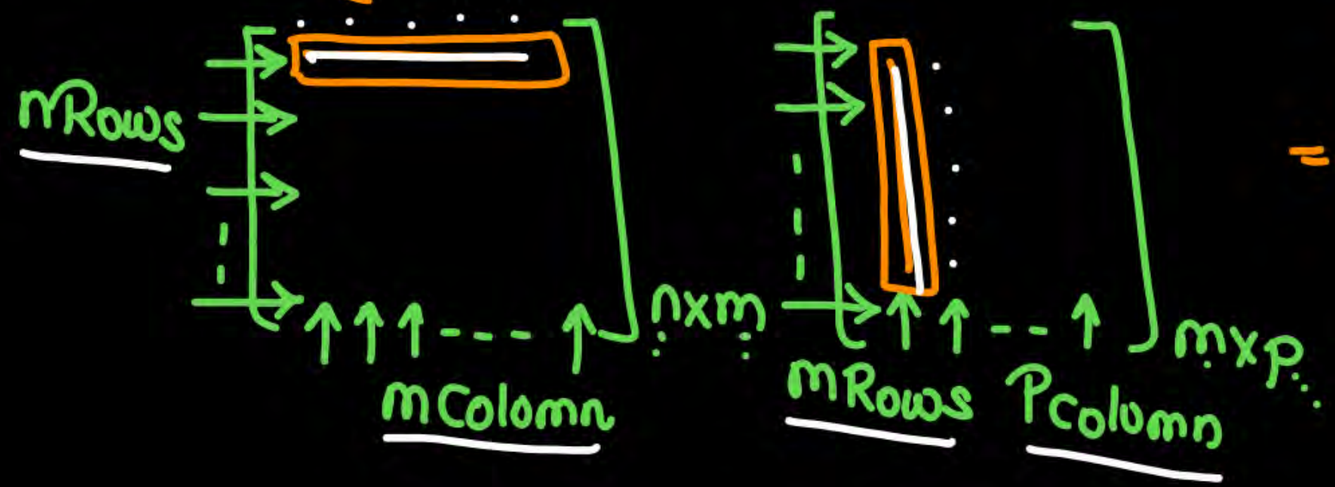
- **Linear Regression**
  - Train Time Complexity=O(n*m^2 + m^3)
  - Test Time Complexity=O(m)
  - Space Complexity = O(m)

Linear Regression

Space and Time Complexity of Linear Regression

- Matrix multiplication

$(A)_{n \times m} (B)_{m \times p} = (C)_{n \times p}$



→ For each value m times mults
(m-1) times addition

nxp number of values

(nxpxm) number of mult

nxpx(m-1) number of add.

Inverse of a matrix

$$\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}_{3\times3}$$

$$\begin{bmatrix} C_{11} & C_{12} & C_{13} \\ C_{21} & C_{22} & C_{23} \\ C_{31} & C_{32} & C_{33} \end{bmatrix}$$

$$A^{-1} = \frac{1}{|A|}\left[\text{Adj}A\right]$$

$$\text{Adj}A \rightarrow (\text{Cof Matrix})^{T}$$

$$C_{11} = (ei - fh)(-1)^{1+1}$$

$$C_{12} = (-1)^{1+2}(di - fg)$$

To find Cofactor = 3 Mult

9 Cofactors $\Rightarrow$ 27 Mult

* So for $(3\times3)^{-1}$ we need 27 operation

N: No of data points
D: dimension of data

KxK matrix Ka inv ⇒ Number of Calculation ⇒ $O(K^3)$

In LR ⇒ $\beta = (X^TX)^{-1}X^TY$

$(X^T)_{(D+1)\times N}(Y)_{N\times 1} \Rightarrow$ To create X^TY $(D+1)N\times 1$

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

To create $X^TX \Rightarrow (D+1)^2 \times N$

$(X^TX) = [X^T]_{(D+1)\times N}(X)_{N\times(D+1)} = (X^TX)_{(D+1)\times(D+1)}$ → Inv ⇒ $(D+1)^3$

$(X^TX)^{-1}_{(D+1)\times(D+1)}(X^TY)_{(D+1)\times 1} \Rightarrow$ Total op. $(D+1)^3$.

N: No of data points
D: dimension of data

In LR ⇒ $\beta = (X^T X)^{-1} X^T Y$

$(X^T)_{(D+1)\times N} (Y)_{N\times 1} \Rightarrow$ To create $X^T Y$ $(D+1)N\times 1$

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

To create $X^T X \Rightarrow (D+1)^2 \times N$

$(X^T X) = [X^T]_{(D+1)\times N} (X)_{N\times(D+1)} = (X^T X)_{(D+1)\times(D+1)}$ → Inv $\Rightarrow (D+1)^3$

$(X^T X)^{-1}_{(D+1)\times(D+1)} (X^T Y)_{(D+1)\times 1} \Rightarrow$ Total op. $(D+1)^2$.

Total No of Calculation $\Rightarrow (D+1)^2 \times N + (D+1)^3 + N(D+1) + (D+1)^2$

This is order of No of Calci° $\Rightarrow$ $N(D+1)^2 + (D+1)^3$

Not exact No of Calci

$K = D+1$

- Training time ⇒

data → LR machine → @end of training we get. model, β values

Time taken to produce β_0.

if 1 unit time is needed for a multiplication

we need $O(NK^2+K^3)$ number of mult$_0$

So training time ⇒ $O(NK^2+K^3)$

Training Time Complexity of LR

⇒ $O(NK^2+K^3)$

$D+1 = K$

after training

⇓

we only need to store ⇒ only β's.

→ $(D+1)$ Beta values

Space complexity.

→ $(D+1)$ spaces.

How testing is done

we get a new x value

$$\begin{bmatrix} x_t^1 & x_t^2 & x_t^3 & x_t^4 & -- & x_t^D \end{bmatrix}$$

@ Testing $(D+1)$ multiplication

Order of Complexity in Testing $= O(K)$

To find $\hat{y}$, testing

$$\begin{bmatrix} 1 & x_t^1 & x_t^2 & --- & x_t^D \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_D \end{bmatrix} \Rightarrow \hat{y}$$

Ridge Regression

Problem with the least square error ??

Problem in OLS

min RSS

(i) Multicollinear data ⇒ produce unstable model

Very large β.

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

- we min RSS ⇒ thus algo want to make RSS = 0.
- So algo has overfitting tendency.

The method to solve the problem

Regularisation

Ridge Regression

we change the loss function

Updated loss fxn.

$$\text{loss fxn} = \left[\underbrace{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}_{RSS} + \lambda \sum_{i=1}^{D} \beta_i^2 \right]$$

= we will minimize this loss fxn

Ex.

Same data → LR

LR

$x^2 = 100x^1$

x^1	x^2	y

$3x^2 + 1x^1 + 100 = 0$

Best model

$301x^1 + 0x^2 + 100 = 0$

$3 \cdot 01x^2 + 0x^1 + 100 = 0$

all 3 perform Similar

Similar MSE on data

Ex.

Same data → LR

LR

$x^2 = 100x^1$

x^1	x^2	y

$3x^2 + 1x^1 + 100 = 0$

Best model

$3.01x^1 + 0x^2 + 100 = 0$

$3.01x^2 + 0x^1 + 100 = 0$

(i) Solution of Ridge Regression remove problem of large B

(ii) The algorithm will make B's of irrelevant dimension = 0 Or close to zero.

Ridge Regression

Problem with the least square error ??

- Not all the dimensions are equally usefull

Ridge Regression

Problem with the least square error ??

- OLS may lead to unstable model

minRSS

Ridge Regression

Problem with the least square error ??

Let's Summarize : The problem in OLS or minimizing the RSS are as follows

1. It may lead to Overfitting.
2. No boundation on values of Betas may lead to unstable model.
3. The problem of multicollinearity.

So what is the solution

Regularisation

Ridge Regression

Shrinkage Methods : Ridge Regression

- Ridge regression is a regularisation techniques...

Ridge Regression

Shrinkage Methods : Ridge Regression

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

Ridge Regression

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.

The loss function are updated

Ridge Regression

Shrinkage Methods : Ridge Regression

The loss function are updated

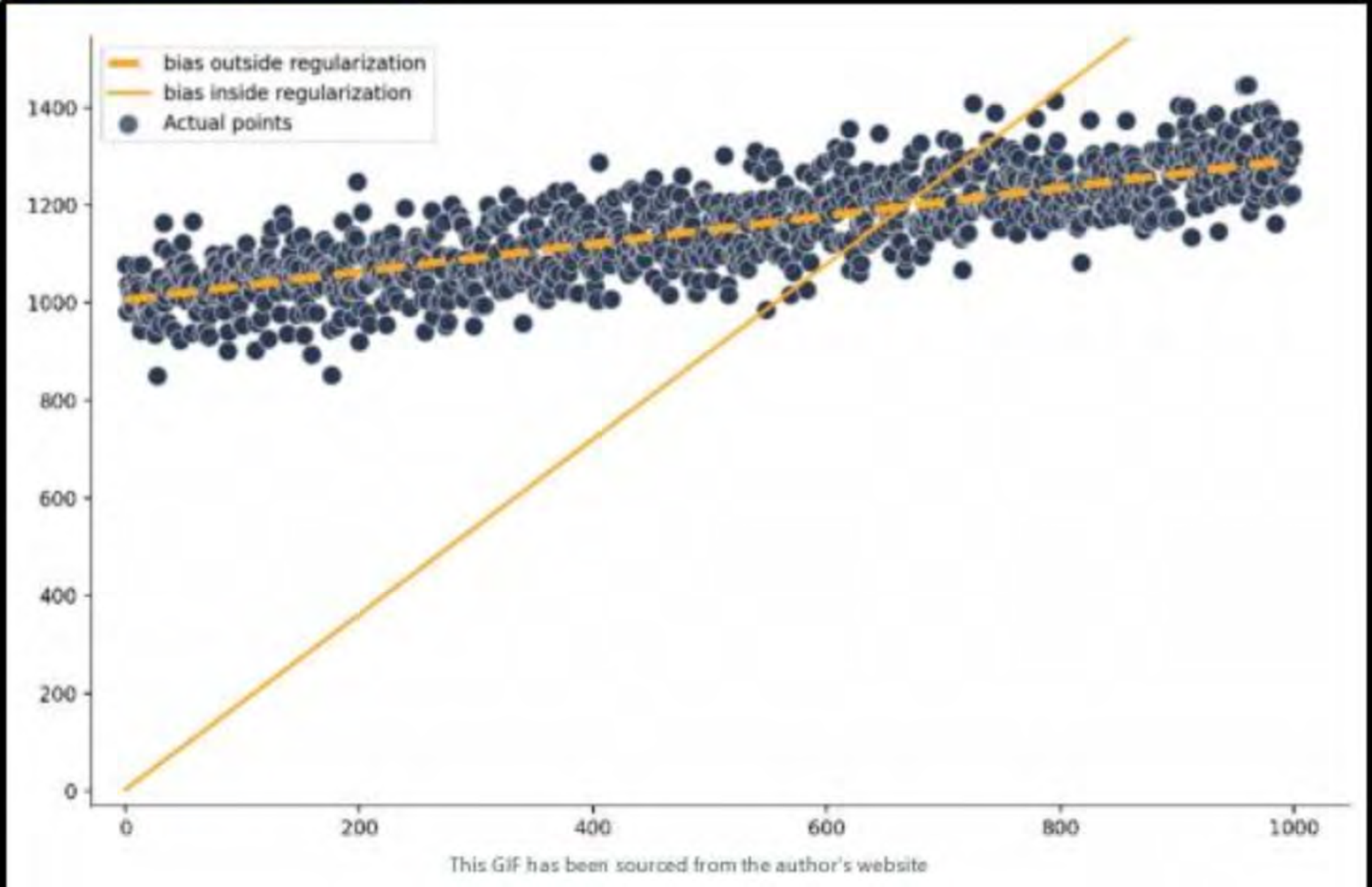
Ridge Regression

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 ..

Ridge Regression



This GIF has been sourced from the author's website

Ridge Regression

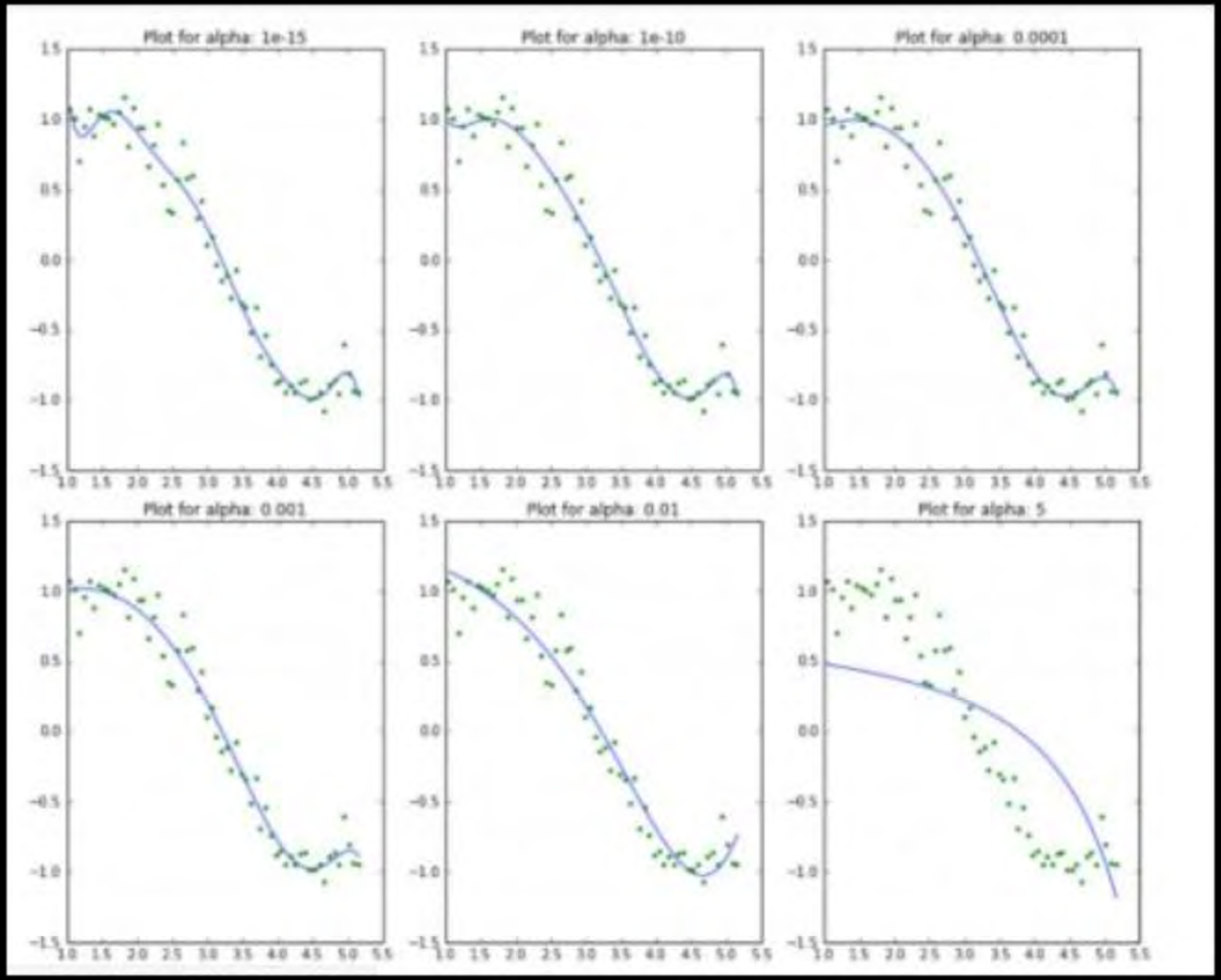
Shrinkage Methods : Ridge Regression

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

Ridge Regression

Shrinkage Methods : Ridge Regression

- Here λ is very important control parameter:


Plot for alpha: 1e-15
Plot for alpha: 1e-10
Plot for alpha: 0.0001
Plot for alpha: 0.001
Plot for alpha: 0.01
Plot for alpha: 5

Ridge Regression

Shrinkage Methods : Ridge Regression

- Lets find the solution to this ridge regression problem

Ridge Regression

Shrinkage Methods : Ridge Regression

- How to find λ (can this be negative?)

Linear Regression

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.

Linear Regression

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.

Linear Regression

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.

THANK - YOU