Data Science and Artificial Intelligence

Machine Learning

Regression

Lecture No. -07

By- SIDDHARTH SABHARWAL SIR

Recap of Previous Lecture

- Topic: Multicollinearity
- Topic: Homoscedasticity
- Topic: Correlation Matrix
- Topic: VIF
- Topic

Topics to be Covered

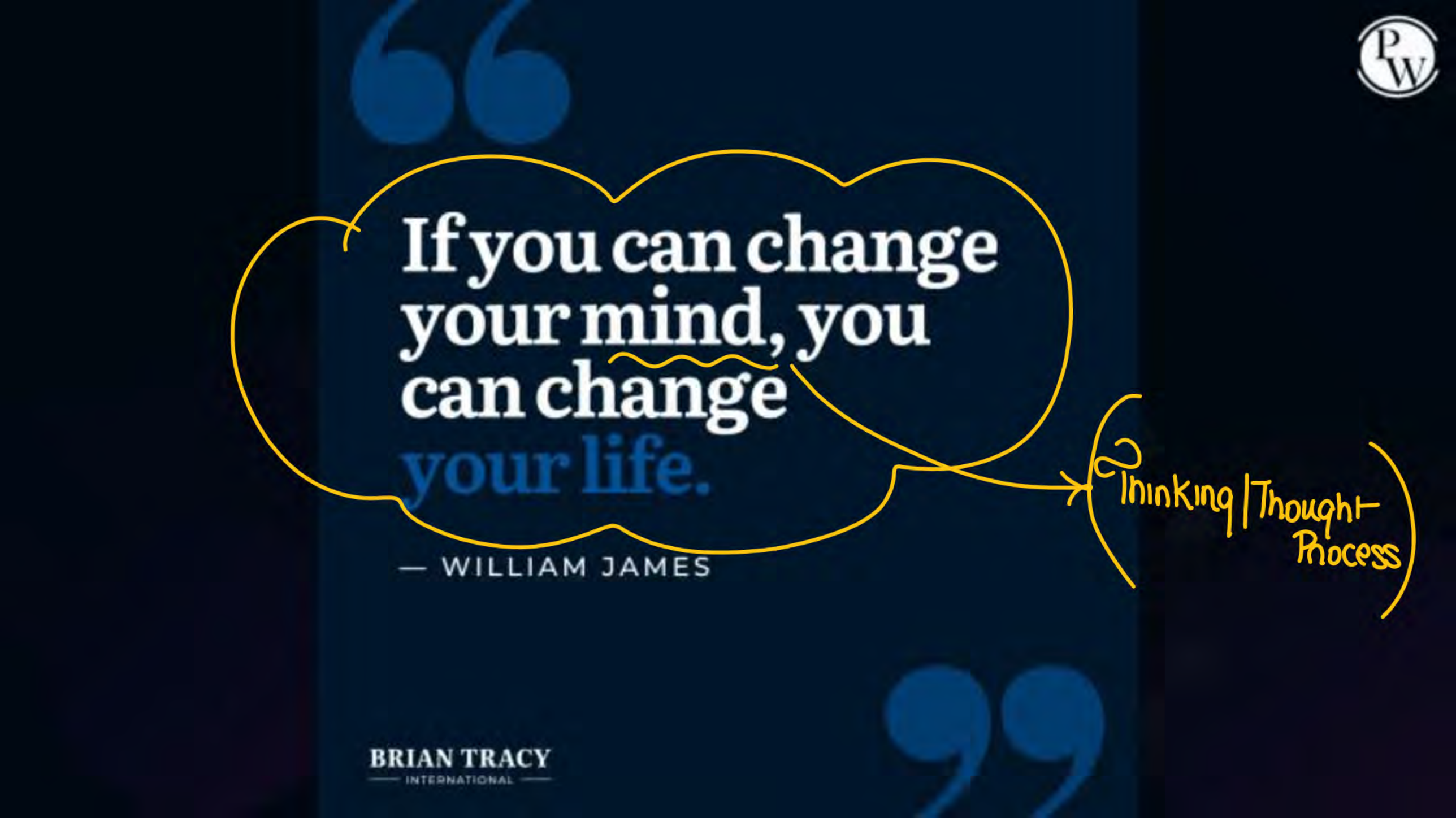
- Topic
- Topic
- Topic
- Topic
- Topic

Advantage - disadvantage of LR

Space and time Complexity of LR

Regularisation

Ridge Regression


If you can change your mind, you can change your life.
— WILLIAM JAMES
Thinking / Thought Process
BRIAN TRACY
INTERNATIONAL

REVISION

Multicollinearity

- Correlation matrix → Value ≈ 0 → Not multicollinear
  - ≈ Values close to 1 → Multicollinear
- the dimension are collinear | dependent on each other
- VIF → $\frac{1}{1-R^2}$, ∞ → 100% multicollinear
  - 1 → No multicollinearity

REVISION

Homo and Heteroscedasticity

→ Noise in data changes with data

→ The noise in data ≈ Same allover

⇒ (SSE ⇒ Squared Sum of Error ⇒ RSS)

⇒ (SST ⇒ Total Sum of Square ⇒ TSS)

Basics of Machine Learning

REVISION

VIF and Correlation matrix

done.

Basics of Machine Learning

Outlier and its effect

⇒ Linear Reg is effected by the Outlier Yes

Linear Regression

Considering data of P Dimensions

Lets Practice

Based on the data provided below, answer questions from (7-10). We consider a function we wish to minimize. $J(w) = \frac{1}{10}\Sigma_i^5 (y^{(i)} - w_1 x^{(i)} - w_0)^2$ where the constants $x^{(i)}$, $y^{(i)}$ are provided in the table below

$$J(w) = \frac{1}{10}\sum_{i=1}^{5}\left(y^{i} - w_1 x^{i} - w_0\right)^2$$

i	$x^{(i)}$	$y^{(i)}$
1	0	1.48[illegible]
2	0.25	1.81[illegible]
3	0.50	1.91[illegible]
4	0.75	2.39[illegible]
5	1.00	2.58[illegible]

Dataset

7) The dimension of w is _____.

Linear Regression

Considering data of P Dimensions

Lets Practice

8) Start with the initial guess of $[w_0, w_1] = [0, 0]$. Take the value of learning rate = 1. The value of w_0 after 1 iterations of gradient descent will be _____.

$$\left(\frac{\partial J}{\partial w}\right) = \begin{pmatrix} \partial J/\partial w_0 \\ \partial J/\partial w_1 \end{pmatrix} = \begin{pmatrix} -\frac{2}{10}\sum_{i=1}^{5}(y^i - w_1 x^i - w_0) \\ -\frac{2}{10}\sum_{i=1}^{5} x^i(y^i - w_1 x^i - w_0) \end{pmatrix}$$

$$w^{new} = w^{old} - 1 \cdot \left.\frac{\partial J}{\partial w}\right|_{w^{old}}$$

$$\begin{pmatrix} w_0 \\ w_1 \end{pmatrix}^{new} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} - 1 \begin{pmatrix} -\frac{1}{5}\sum y^i \\ -\frac{1}{5}\sum x^i y^i \end{pmatrix} \Rightarrow \begin{pmatrix} 0 \\ 0 \end{pmatrix} - \begin{pmatrix} -\frac{1}{5}\times 10.17 \\ -\frac{1}{5}\times 5.78 \end{pmatrix} \Rightarrow \begin{pmatrix} 2.034 \\ 1.156 \end{pmatrix}$$

Advantage of LR ⇒

- → easy / Simple Algorithm
- → The model created by LR ⇒ has high interpretability ...

General

Bias ⇒ Training error
Variance ⇒ Test error

Highly Complicated Model ⇒
- data pattern missed but data learned
- Rote learning

Overfitting
⇓
Bias = 0
Variance = V. high.

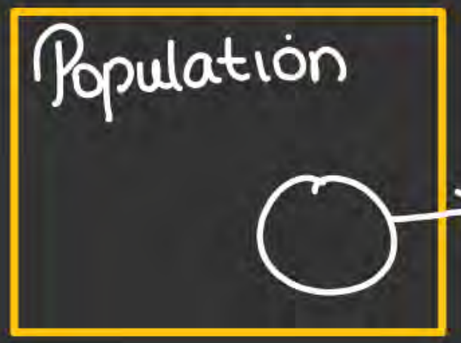
Training data
$y = \hat{y}$ ⇒ Zero error

Test data (new data)
↳ huge error

Complexity ↓↓
Training error ≠ 0 Bias ≠ 0
Testing error ⇒ less

Disadvantage of LR ⇒

1. it assume that data has a linear Pattern, thus L-R will not perform good in Case of Non linear data
2. L.R fails in case of multicollinearity
3. The sample data is used to predict the model for whole population ⇒
   - ↳ in LR the loss function ⇒ min (RSS) ⇒ the model try to min the RSS ⇒ that is LR try to make RSS=0 i.e L.R is prone to overfit the data..



Biased Sample, with y values dependent on few dimensions

4. LR may create unstable model ⇒

↳ So if the data/sample is incorrect and the sample/data is such that the label or y values are highly dependent on on few dimensions then in these cases the L-R will give us a model where β's of few dimensions is very large and β's of other dimension is V. Smalled.

Such model are unstable ⇒ $y = (10000x^1 + 2x^2 + 2.5x^3 - - -)$

" So in this model where few β's are v. large then a little change in x^1 will hugely effect y"

" Unstable models "

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. → B's show significance of any dimension
- 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. → When data is NL, LR fails
- Assumptions of independent and normally distributed errors are critical.
- Suitable only when one independent variable is involved in the analysis.

LR work when variables are independent.

Linear Regression

Space and Time Complexity of Linear Regression

⇒ Space and time complexity

$\begin{bmatrix} a & b & c \\ d & e & f \end{bmatrix}_{2\times 3} \begin{bmatrix} h & K \\ i & L \\ j & m \end{bmatrix}_{3\times 2}$

⇒ (12 operations)

$\begin{bmatrix} ah+bj+cj & ak+bl+mc \\ ah+ei+fj & dk+le+fm \end{bmatrix}$ (3,2) (3,2)

(12 ⇒ mult ~~& addition~~)

In general

$\begin{bmatrix} A \end{bmatrix}_{n\times m} \begin{bmatrix} B \end{bmatrix}_{m\times p}$ ⇒ # of mult. op ⇒ $n\times m\times p$

So if each multiplication take 1 unit time then time to multiply 2 matrix $\Rightarrow (n \times m \times p) \times 1$

$\Rightarrow$ Time complexity of this operation $\Rightarrow (n \times m \times p)$

$$\Rightarrow \left(C_{ij} = (-1)^{i+j} M_{ij}\right)$$

- Inverse of a matrix $\Rightarrow$

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

Here each minor will have 4 elements

$$\begin{vmatrix} e & f \\ h & i \end{vmatrix} = (ei - hf) \Rightarrow 2 \text{ multiplication}$$

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

$$\text{Adj } A \Rightarrow [\text{Cof. Mat}]^T$$

So 3 multiplication for each Cofactor $\Rightarrow$ Total 27 multiplication in inv of 3x3 matrix.

So in any mxm matrix the order of multiplication = m^3

So the time complexity $\Rightarrow$ is of order of m^3.

Ex $\begin{bmatrix} \quad \end{bmatrix}_{a \times b} \begin{bmatrix} \quad \end{bmatrix}_{b \times c} \Rightarrow (abc) = \text{No of mult.}$

(D+1 ⇒ K)

So intraining

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

$X = \begin{bmatrix} & \\ & \end{bmatrix}_{N \times K}$

$X^T = \begin{bmatrix} & \\ & \end{bmatrix}_{K \times N}$

$X^TX \Rightarrow \begin{bmatrix} & \\ & \end{bmatrix}_{K \times N} \begin{bmatrix} & \\ & \end{bmatrix}_{N \times K} \Rightarrow KNK \Rightarrow NK^2$

$(X^TX)_{K \times K} \Rightarrow$ inv of $(X^TX)_{K \times K} \Rightarrow$ Multiplic. K^3

$X^TY \Rightarrow \begin{bmatrix} X^T \end{bmatrix}_{K \times N} \begin{bmatrix} \end{bmatrix}_{N \times 1} \Rightarrow NK \Rightarrow$ mult.

$(X^TX)^{-1}_{K \times K} (X^TY)_{K \times 1} \Rightarrow K^2 \Rightarrow$ mult.

⇒ Total number of mult ⇒

$(NK^2 + K^3 + NK + K^2)$

So in training

Time complexity ⇒ order of $(K^3 + NK^2 + NK + K^2)$.

after training we get Betas ⇒ $\beta_0, \beta_1, \ldots, \beta_D$ ⇒ Soweget K Betas

So new test point $(1\ x^1\ x^2\ x^3 \ldots x^D)$

$$y = X\beta \Rightarrow (1\ x^1\ x^2 \ldots x^D) \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_D \end{bmatrix}$$

⇒ Only D+1 ⇒ K number of mult.

Time complexity of testing ⇒ Order of K.

So in L.R, after training we can delete the data only Store D+1 Beta values ⇒

So in memory we need only (D+1) ⇒ K locations

Space Complexity of LR ⇒ K.

Question 12: In simple linear regression, which variable is considered the independent variable?

© PW

A. The variable being predicted
B. The response variable
C. The predictor variable
D. There is no independent variable in simple linear regression

Question 19: If the R-squared value in simple linear regression is 0.75, what does it indicate? (a)

A. A strong linear relationship between the variables
B. A weak linear relationship between the variables
C. No linear relationship between the variables
D. The model is overfitting

Question 20: Which of the following statements is true regarding the residual plot in simple linear regression?

(b) HW.

A. Residuals should exhibit a clear linear pattern.
B. Residuals should be randomly scattered around the horizontal line.
C. Residuals should be negatively correlated with the predictor variable.
D. Residuals should have a positive correlation with the dependent variable.

5.FOr a give N independent input variables (X1,X2... Xn) and dependent (target) variable Y a linear regression is fitted for the best fit line using least square error on this data. The correlation coefficient for one of it's variable(Say X1) with Y is -0.97. Which of the following is true for X1? ?

- A) Relation between the X1 and Y is weak
- B) Relation between the X1 and Y is strong
- C) Relation between the X1 and Y is neutral
- D) Correlation does not imply relationship

6.Given below characteristics which of the following option is the correct for Pearson correlation between V1 and V2? If you are given the two variables V1 and V2 and they are following below two characteristics. 1. If V1 increases then V2 also increases 2. If V1 decreases then V2 behavior is unknown ? H.W

- A) Pearson correlation will be close to 1
- B) Pearson correlation will be close to -1
- C) Pearson correlation will be close to 0
- D) None of these

1) A regression analysis is inappropriate when;
   a) you have two variables that are measured on an interval or ratio scale.
   b) you want to make predictions for one variable based on information about another variable.
   c) the pattern of data points forms a reasonably straight line.
   d) **there is heteroscedasticity in the scatter plot.**

2) In regression analysis, the variable that is being predicted is;

a) the independent variable

b) the dependent variable

c) usually denoted by x

d) usually denoted by r

3) In the regression equation $y = b_o + b_1x$, b_o is the;

a) slope of the line

b) independent variable

c) **y intercept**

d) coefficient of determination

6) Least square method calculates the best-fitting line for the observed data by minimizing the sum of the squares of the _____ deviations.

a) **Vertical**
b) Horizontal
c) Both of these
d) None of these

7) Which one is the least square method formula;

a) $\min \sum(y_i - \hat{y}_i)^2$

b) $\min \sum(\hat{y}_i - y_i)$

c) **$\min \sum(y_i - \hat{y}_i)^2$**

d) $\min \sum(y_i - \hat{y}_i)$

13) Below you are given a summary of the output from a simple linear regression analysis from a sample of size 15, SSR=100, SST = 152. The coefficient of determination is;

a) 0.5200

b) **0.6579**

c) 0.8111

d) 1.52

10) **A residual is defined as**

a) The difference between the actual Y values and the mean of Y.

b) **The difference between the actual Y values and the predicted Y values.**

c) The predicted value of Y for the average X value.

d) The square root of the slope.

11) If the regression equation is equal to $y=23.6-54.2x$, then 23.6 is the _____ while -54.2 is the _____ of the regression line.

a) Slope, intercept

b) Slope, regression coefficient

c) **Intercept, slope**

d) Radius, intercept

Q8. Suppose we have N independent variables (X1, X2... Xn) and Y's dependent variable.

Now Imagine that you are applying linear regression by fitting the best-fit line using the least square error on this data. You found that the correlation coefficient for one of its variables (Say X1) with Y is -0.95.

Which of the following is true for X1?

A) Relation between the X1 and Y is weak

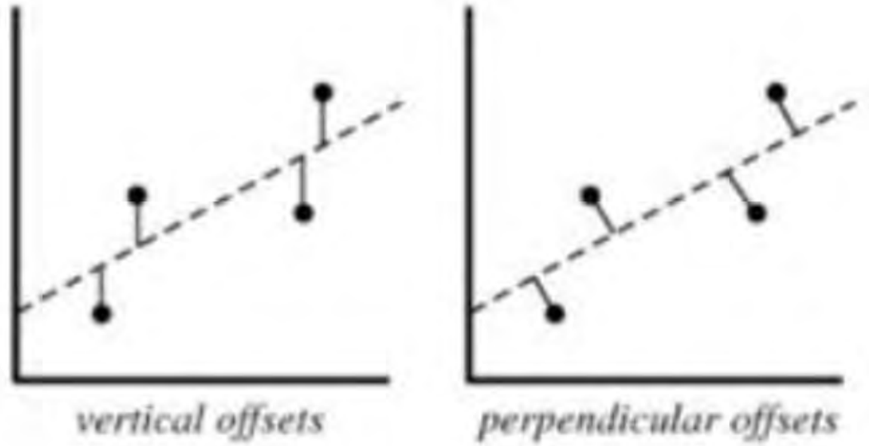
B) Relation between the X1 and Y is strong

C) Relation between the X1 and Y is neutral

D) Correlation can't judge the relationship

Solution: (B)

Q11. Suppose the horizontal axis is an independent variable and the vertical axis is a dependent variable. Which of the following offsets do we use in linear regression's least square line fit?



B) Perpendicular offset

C) Both, depending on the situation

D) None of above

Q12. True- False: Overfitting is more likely when you have a huge amount of data to train.

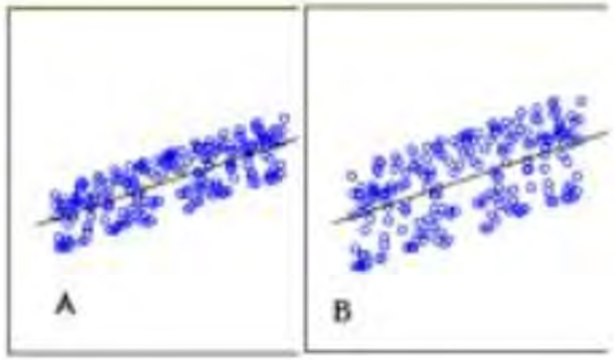
A) TRUE

B) FALSE

Solution: (B)

Q14. Which of the following statement is true about the sum of residuals of A and B?

Below graphs show two fitted regression lines (A & B) on randomly generated data. Now, I want to find the sum of residuals in both cases, A and B.



A) A has a higher sum of residuals than B

B) A has a lower sum of residual than B

C) Both have the same sum of residuals

D) None of these

Q18. Which of the following statement is true about outliers in Linear regression?

A) Linear regression is sensitive to outliers

B) Linear regression is not sensitive to outliers

C) Can't say

D) None of these

Q19. Suppose you plotted a scatter plot between the residuals and predicted values in linear regression and found a relationship between them. Which of the following conclusion do you make about this situation?

A) Since there is a relationship means our model is not good

B) Since there is a relationship means our model is good

C) Can't say

D) None of these

Suppose that you have a dataset D1 and you design a linear model of degree 3 polynomial and find that the training and testing error is "0" or, in other words, it perfectly fits the data.

Q20. What will happen when you fit a degree 4 polynomial in linear regression?

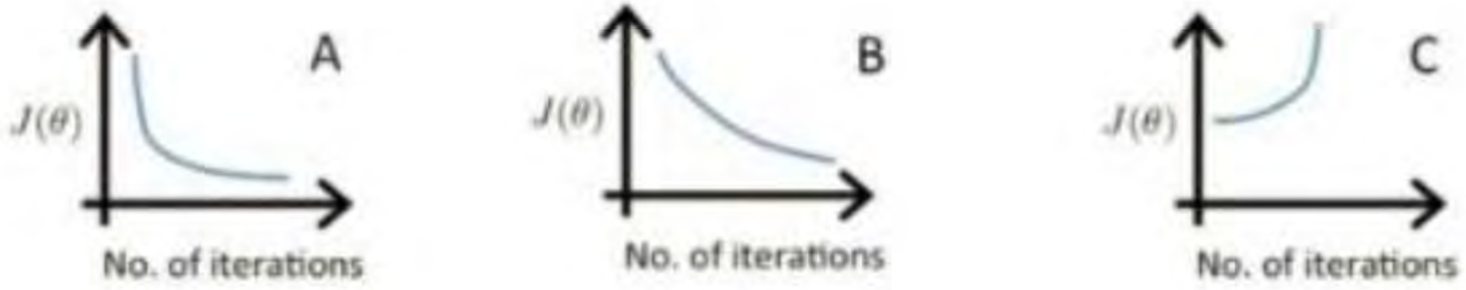
A) There is a high chance that degree 4 polynomial will overfit the data

B) There is a high chance that degree 4 polynomial will underfit the data

C) Can't say

D) None of these

Below are three graphs, A, B, and C, between the cost function and the number of iterations, I1, I2, and I3, respectively.



Q23. Suppose I1, I2, and I3 are the three learning rates for A, B, and C, respectively. Which of the following is true about I1,I2, and I3?

A) $I2 < I1 < I3$

B) $I1 > I2 > I3$

C) $I1 = I2 = I3$

D) None of these

QUESTION 1

How many coefficients do you need to estimate in a simple linear regression model (One independent variable)?

- 1
- 2
- 3
- 4

QUESTION 2

In a linear regression model, which technique can find the coefficients?

- ○ Ordinary Least Squares
- ○ Gradient Descent
- ○ Regularization
- ○ All of the above

Which one is the disadvantage of Linear Regression?

- The assumption of linearity between the dependent variable and the independent variables. In the real world, the data is not always linearly separable.
- Linear regression is very sensitive to outliers
- Before applying Linear regression, multicollinearity should be removed because it assumes that there is no relationship among independent variables.
- All of the above

QUESTION 4

Which parameter determines the size of the improvement step to take on each iteration of Gradient Descent?

- learning rate
- epoch
- batch size
- regularization parameter

QUESTION 5

5 marks

For a linear regression model, start with random values for each coefficient. The sum of the squared errors is calculated for each pair of input and output values. A learning rate is used as a scale factor and the coefficients are updated in the direction towards minimizing the error. The process is repeated until a minimum sum squared error is achieved or no further improvement is possible. This technique is called _____?

- ○ Gradient Descent
- ○ Ordinary Least Squares
- ○ Homoscedasticity
- ○ Regularization

QUESTION 6

In a linear regression model, which technique cannot find the coefficients?

- ○ Ordinary Least Squares
- ○ Gradient Descent
- ○ Regularization
- ○ Normalization

QUESTION 8

What is predicting y for a value of x that is within the interval of points that we saw in the original data called?

- Regression
- Extrapolation
- Intrapolation
- Polation

QUESTION 9

5 marks

The correlation coefficient between the age of a person and their IQ test score is found to be -1.0087. What can you conclude from this?

- ○ Age is not a good predictor of IQ.
- ○ Age is a good predictor of IQ.
- ○ None of the above

QUESTION 10

5 marks

In order to determine whether the coefficient in a simple linear regression model is significant or not, which Null Hypothesis do we propose?

- $\beta 0 \neq 0$
- $\beta 1 = 0$
- $\beta 0 = 0$
- $\beta 1 \neq 1$

QUESTION 4

A term used to describe the case when the independent variables in a multiple regression model are correlated is

- regression
- correlation
- multicollinearity
- none of the above

QUESTION 5

A multiple regression model has the form: $y = 2 + 3x1 + 4x2$. As x1 increases by 1 unit (holding x2 constant), y will

- ○ increase by 3 units
- ○ decrease by 3 units
- ○ increase by 4 units
- ○ decrease by 4 units

QUESTION 6

5 marks

The adjusted multiple coefficient of determination accounts for

- ○ the number of dependent variables in the model
- ○ the number of independent variables in the model
- ○ unusually large predictors
- ○ none of the above

QUESTION 7

A multiple regression model has

- ○ only one independent variable
- ○ more than one dependent variable
- ○ more than one independent variable
- ○ none of the above

Lets Practise...

Questions

26. In a linear regression model, if the sum of squared residuals (SSE) is 100 and the total sum of squares (SST) is 200, what is the coefficient of determination (R-squared)?

a) 0.5
b) 1
c) 0
d) -1

Lets Practise...

Questions

29. You are using the mean squared error (MSE) as an evaluation metric for a regression model. The predicted values are [3, 4, 5, 6], and the actual values are [2, 3, 4, 7]. What is the MSE?

a) 0.5
b) 1.0
c) 1.5
d) 2.0

Lets Practise...

Questions

32. You are performing linear regression with the following data points:

X: [1, 2, 3, 4]
Y: [4, 3, 6, 5]

What is the intercept (b) of the regression line, assuming a simple linear model $Y = aX + b$?

a) 1.5
b) 2
c) 2.5
d) 3

Ridge Regression

Problem with the least square error ??

- **The OLS try to minimize the RSS, We know that when RSS = 0 then it means ...**

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

So LR try to make RSS=0 $\Rightarrow$ (Overfit)

Ridge Regression

Problem with the least square error ??

- The OLS has a problem of multicollinearity ...

already pata है|

Ridge Regression

Problem with the least square error ??

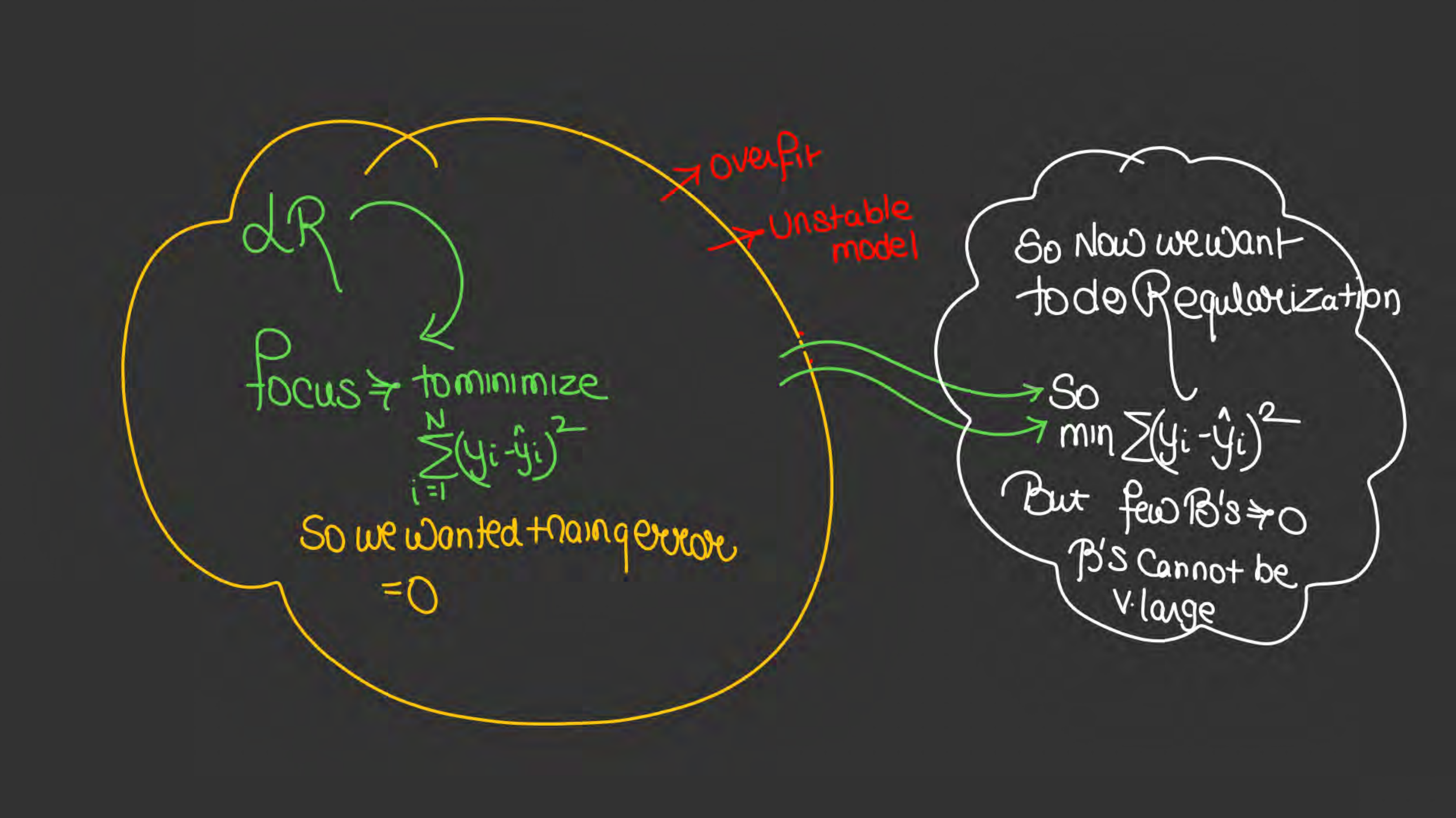
- **Not all the dimensions are equally usefull**
  - So in any exam instead of focusing on 100% of Syllabus it is better to focus on only 75-85% of Syllabus.

So in data Some of features (dimension) are not Imp.

↳ Some of features (dimension) are dependent on each other.

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

So to make contribution of any dimension = 0 ⇒ Simply make it $\beta = 0$.


LR
Focus → to minimize
$\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$
So we wanted train error =0
overfit
Unstable model
So Now we want to do Regularization
So
$\min \sum(y_i - \hat{y}_i)^2$
But few B's → 0
B's Cannot be V. large

Ridge Regression

Problem with the least square error ??

- OLS may lead to unstable model

done

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...

Where the loss function →

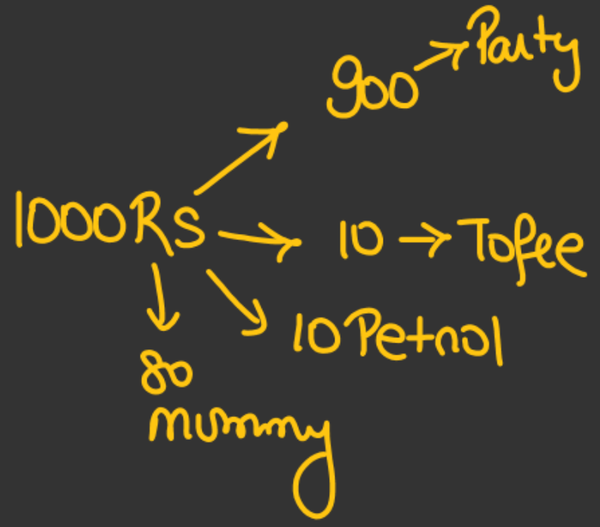
$$\min \left[\sum_{i=1}^{N} (y_{i0} - \hat{y}_i)^2 + \frac{\lambda}{2} \sum_{i=1}^{D} \beta_i^2 \right]$$

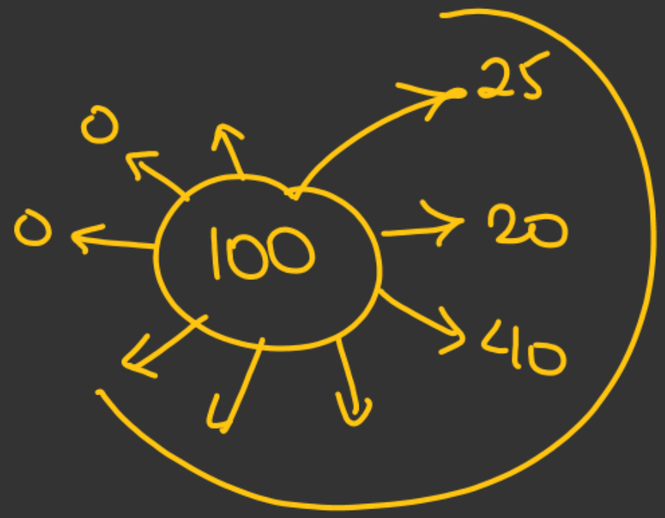
RSS

Regularisation Term

This term
→ make few β's ~ 0

→ no β of any feature can be v. large


900 → Party
1000 Rs
10 → Toffee
10 Petrol
80
Mummy


0
0
100
25
20
40

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

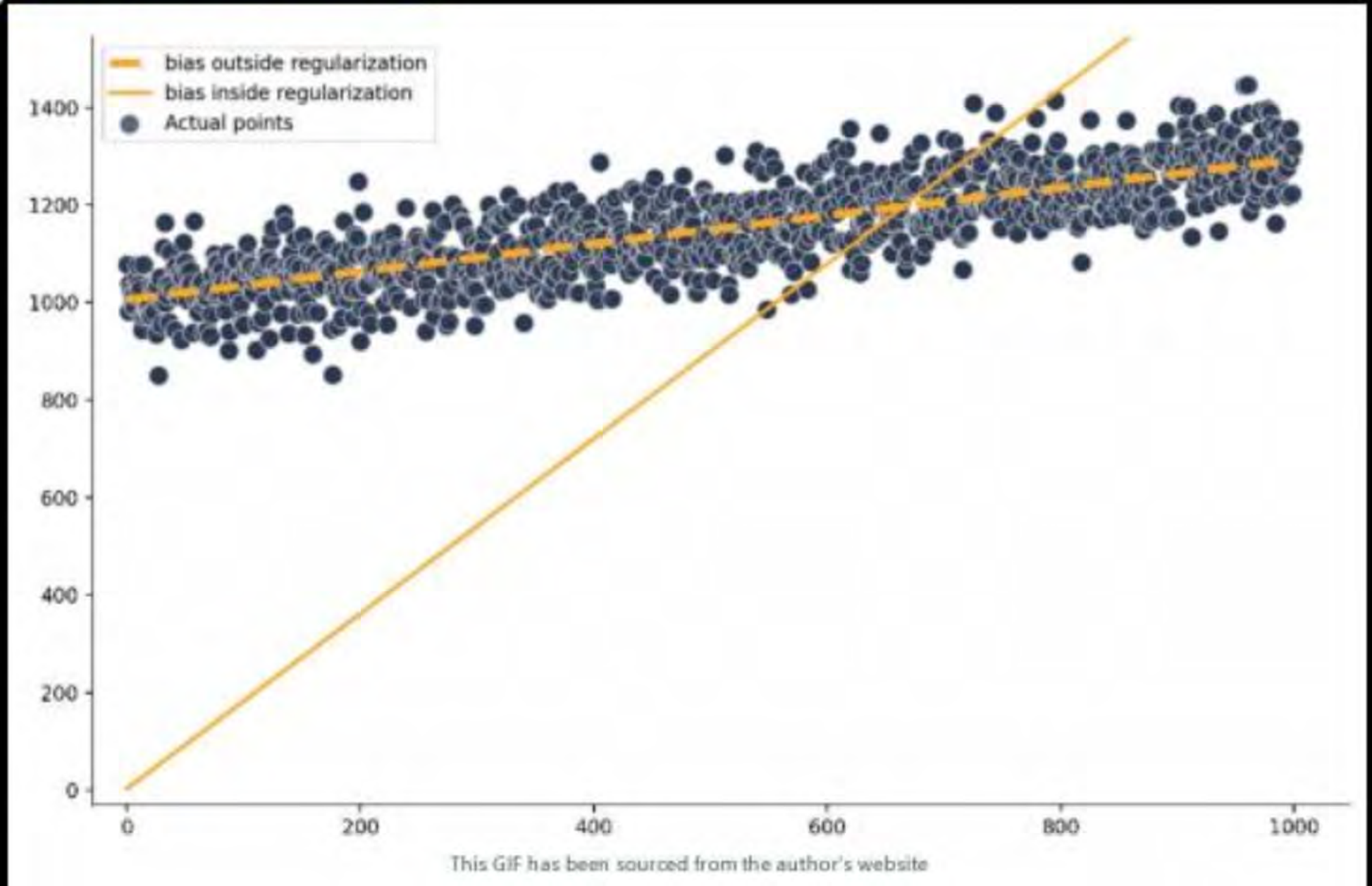
Ridge Regression

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

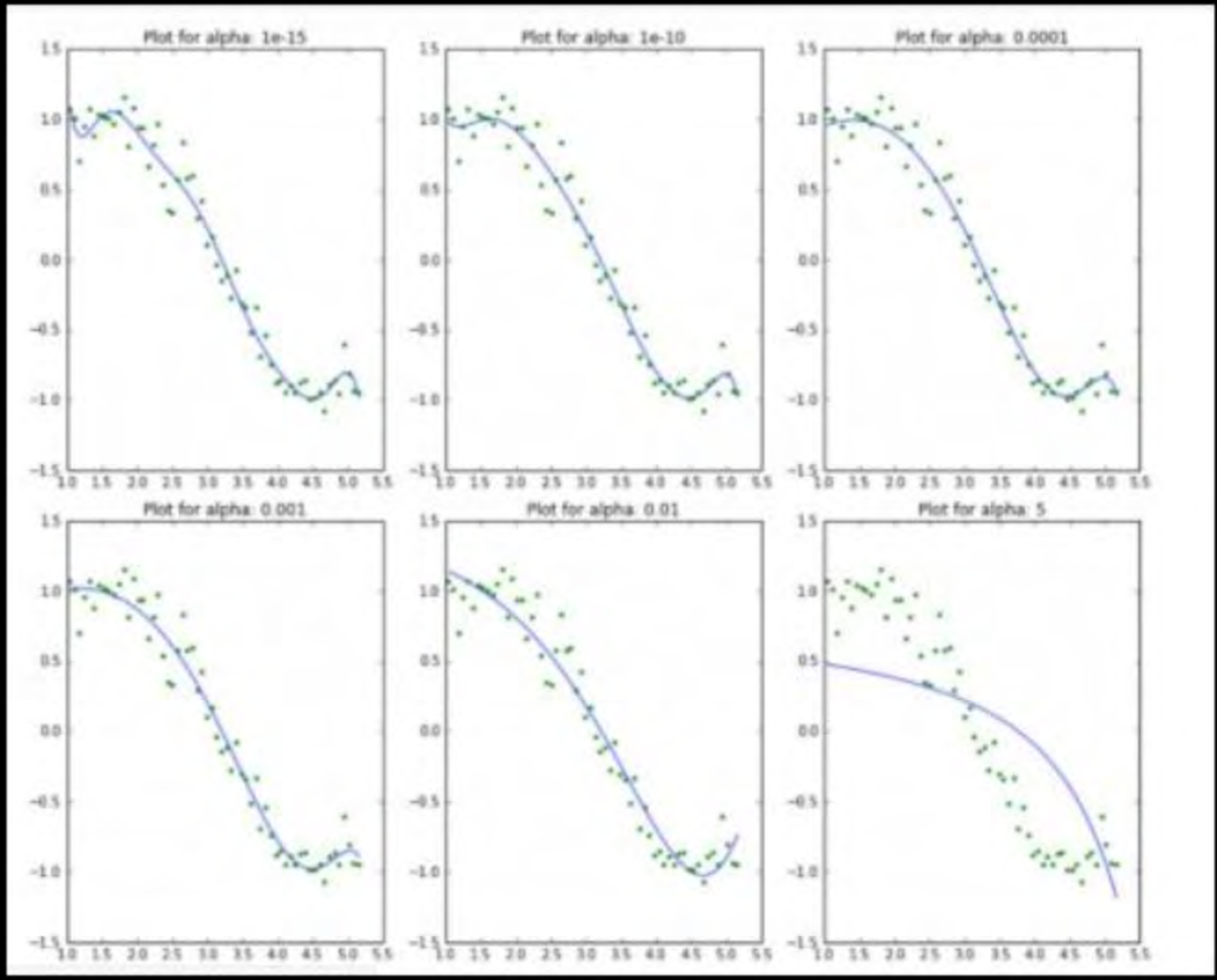
Shrinkage Methods : Ridge Regression

- 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.

2 mins Summary

- Topic
- Topic
- Topic
- Topic
- Topic

THANK - YOU