NAME: Dipesh Ramesh Limaje

INTERNSHIP BATCH : 33

SME : Mr. Shwetank Mishra

WORKSHEET NO : 1

Q1. Least Square Error

Q2. Linear regression is sensitive to outliers

Q3. Negative

Q4. Correlation

Q5. Low bias and high variance

Q6. Predictive Model

Q7. Regularization

Q8. SMOTE

Q9. TPR and FPR

Q10. False

Q11. Apply PCA to project high dimensional data

Q12. A) We don't have to choose the learning rate.
B) It becomes slow when number of features is very large.
C) We need to iterate.

Q13. Explain the term regularization?

Ans. If our model learns too fast we have to slow it down to penalize it, we have to make it learn slow as far as possible, so that it will give a better results.

When we use Regression model to train some data thee is good chance that the model will Overfit the given training dataset, regularization will help to sort this overfitting problem by restricting the Degree Of Freedom(DOF).

There are 3-types of Regularization they are: 1) LASSO (L1 Form)

2) RIDGE (L2 Form)

3) ELASTONET

Q14. Which particular algorithms are used for regularization?

Ans. We basically have 3-types of Regularization they are: 1) LASSO (L1 Form)

2) RIDGE (L2 Form)

3) ELASTONET

But we use only LASSO and RIDGE

For example : we have to predict car price and we have some feature

NAME	PINCODE	EMAIL	BRAND	COUNTRY	STATE	CAR PRICE
Dipesh	421503	123@gmail.com	BMW	INDIA	Maharashtra	50LAC

so here (NAME,PINCODE,EMAIL,BRAND ,COUNTRY ,STATE) are Features and CAR PRICE are label .

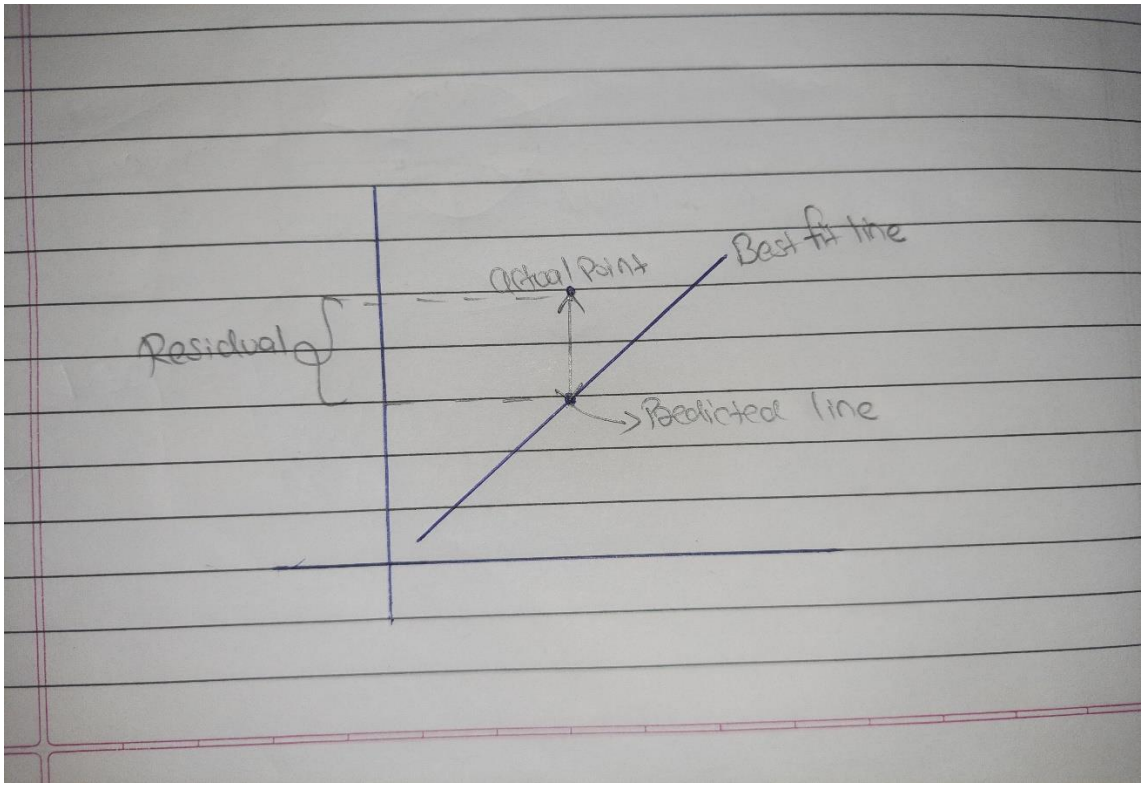
so what LASSO does to predict new car price (Label) it will first compare all features and label and find which one have relationship with it, from example: PINCODE,BRAND,COUNTRY,STATE only has relationship, so it will only give importance to those feature and not to NAME and EMAIL , in short it will act as feature Selection.it is also called as L1 FORM.

Now in RIDGE from upper example to predict new car price (Label) it will first compare all feature and label that it has relationship with it, and from that it give more importance which contribute to label and which does not contribute it will also give some or little importance to those feature also. It is only major difference between LASSO and RIDGE. It is also called as L2 FORM.

Q15. Explain the term error present in linear regression equation?

Ans. The term Error Present in Linear Regression is called as Residual.

Residual is the distance between the actual line and prediction line.



NOTE : (IN FIGURE : instead of predicted line it is predicted point by model)

The residual should always be less so that the prediction become good. The residual can be positive or may be negative. By adding all the residual we will get negative data or zero, so we are going to square the residual to avoid that negativity and to avoid zero.

Mathematically,

$R=y-(mx+c)$

Where R = residual

y= prediction

m = coefficient or slope

x=user define data

c = intercept

#########PYTHON WORKSHEET#########

Q1. %

Q2. 0

Q3. 24

Q4. 2

Q5. 6

Q6. the finally block will be executed no matter if the try block raises an error or not.

Q7. It is used to raised an ecception

Q8. In defining a generator

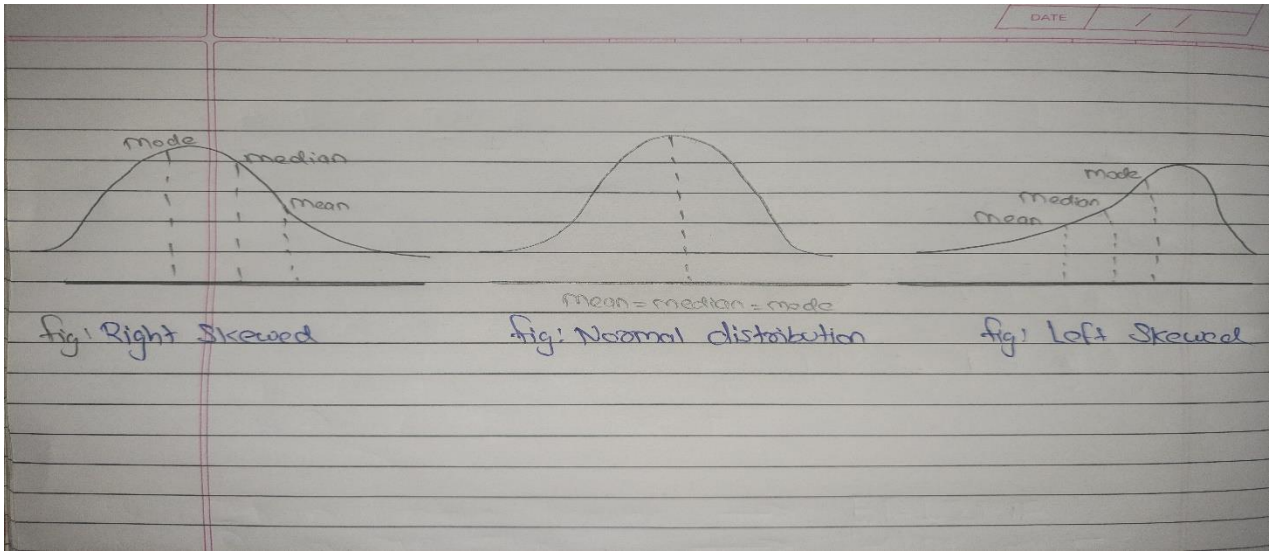
Q9. A) _abc
C)abc2

Q10. A)yield
B)raise

STATISTICS WORKSHEET-1

Q1. True
Q2. Central limit theorem
Q3. Modeling bounded count data
Q4. All of the mentioned
Q5. Poisson
Q6. False
Q7. Hypothesis
Q8. 0 (Zero)
Q9. Outliers cannot conform to the regression

Q10. What do you understand by the term Normal Distribution?
Ans . A normal distribution is the continuous probability distribution with a probability density function that gives you a Symmetrical Bell Curve. That means mean is equal to zero.The Distribution of data can be Right Skewed, Normal Distribution , Left Skewed. If the data is Right or Left Skewed then it is our duty to get it back to Normal Distribution.



Q11. How do you handle missing data? What imputation techniques do you recommend?

Ans. So basically some common issue is to handle missing data or null values , so we commonly treat this missing values by mean , median , mode etc but we use some advance technique that is impute technique to impute the null values.

Some Imputation Techniques are :

1) Simple Inputer
2) KNN Imputer
3) Iterative Imputer

So I will recommend Simple Imputer and Iterative Imputer because in Simple Imputer we have to pass directly the column name and it will fill the null value by internally taking mean of that column

And in iterative imputer it basically treat null column as label and remaining column as feature , this will going to predict the null values depend upon the feature and fill the null value. It is just like solving the regression problem, here null column is label.

Q12. What is A/B testing?

Ans. A/B testing is a user experience research methodology. A/B test consist of a randomized experiment with two variants, A and B. it includes application of statistical hypothesis testing or two sample hypothesis testing as used in the filed of statistics. A/B testing I a way to compare two version of a single variable, typically by testing a subject response to variant A against variant B, and determining which of the two variable is more effective.

Q13. Is mean imputation of missing data acceptable practice?

Ans. The process of replacing null values in a data collection with the data mean is known as mean imputation.

The mean imputation is typically considered as bad practice since it ignore feature correlation. Consider we have a table with age and fitness score and a 80 year old has a missing value for fitness score (nan). If we take average the fitness score of the people in between the age of 20 to 80, the 80 year old will appear to have a significantly greater fitness level than other. Which is not possible by 80 year old human.

Second , the mean imputation decrease the variance of our data while increasing bias. As a result of the reducing variance, the model is less accurate and the confidence interval is narrowed.

Q14. What is linear regression in statistics?

Ans. In statistics Linear Regression is a linear approach for modelling the relationship between scalar response and one or more explanatory variable (Dependent Variable and Independent Variable)

When the case of one explanatory variable is called Simple Linear Regression for more than one, the process is called as Multiple Linear Regression. In linear regression the relationship are modelled using linear prediction function whose unknown model parameter are estimated form the data such model are called as linear model.

Matematically ,

The equation of the Linear Regression is

$Y=mx+c$

Where Y = prediction

m = coefficient

x = user define data (investment)

c = intercept

Q15. What are the various branches of statistics?

Ans. Branches Of Statistics Are:

1) Descriptive Statistics
2) Inferential Statistics

1) Descriptive Statistics: It is a term given to the analysis of data that help to describe, show and summarize data in a meaningful way. It is a simple way to describe data. Descriptive statistics is very important to present raw data in effective/meaningful way using numerical calculation

Example: Mean,Median,Mode,Range,Standard Deviation etc.

2) Inferential Statistics: In Inferential Statistics prediction are made by taking any group of data in which you are intrested. It can be defined as a random sample of data taken from population to describe and make inferential about the population. It basically allow you to make prediction by taking a small sample instead of working on whole set of population.

Example: Election of India (Exit Pole).