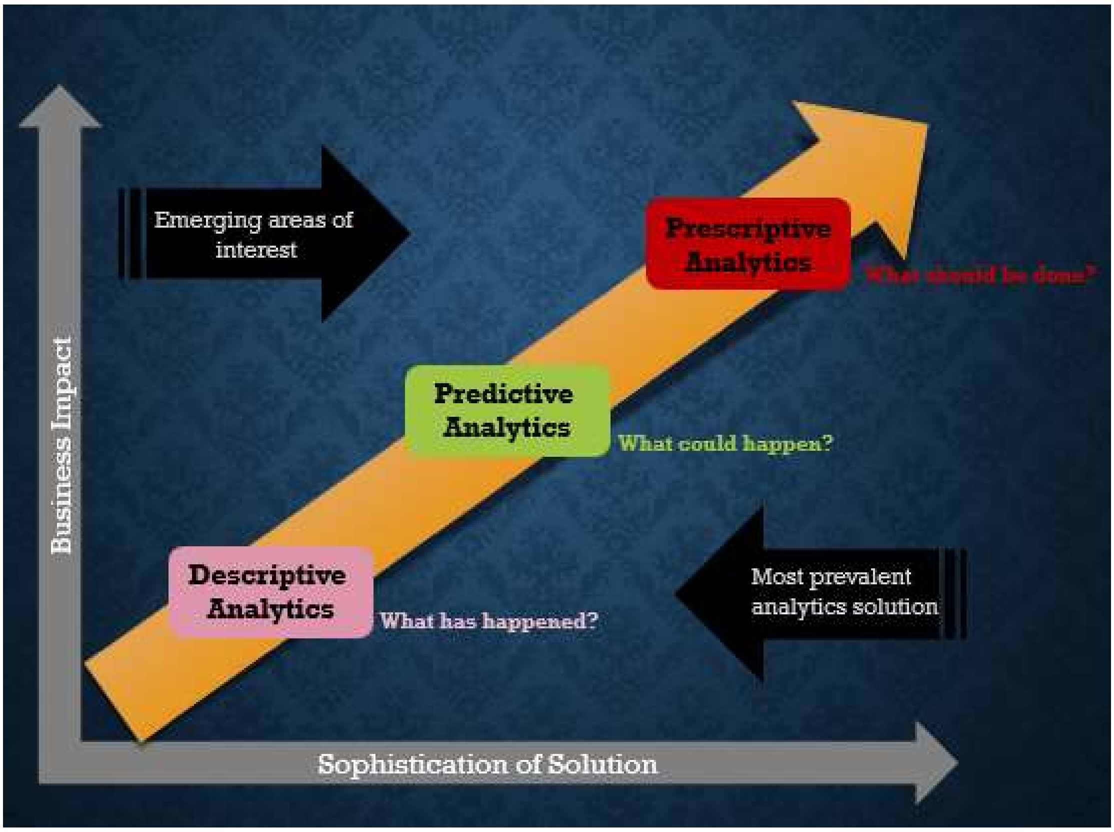
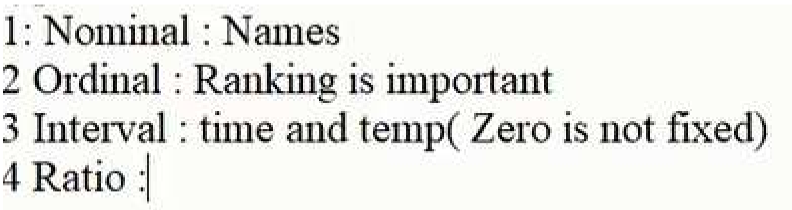
PRANJAL SIR LECTURE 29-DEC-2020

Different Types of analytics



Types of Measurement


1: Nominal : Names
2 Ordinal : Ranking is important
3 Interval : time and temp(Zero is not fixed)
4 Ratio :

What is measure of central tendency

Central tendency is a descriptive summary of a dataset through a single value that reflects the center of the data distribution. Along with the variability (dispersion) of a dataset, central tendency is a branch of descriptive statistics.

Mean, Median and Mode, S.D, Variance, Skewness is the measurement of Central

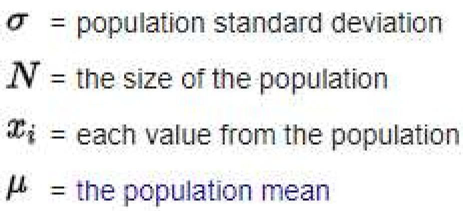
Tendency.

Formula for S.D

The standard deviation value tells us how much all data points deviate from the mean value. The **standard deviation is affected by the outliers because it uses the mean for its calculation.**

Variance is the square of standard deviation. In the case of outliers, the variance value becomes large and noticeable. Hence, it is also affected by outliers.Here is a code to calculate the variance

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$$



σ = population standard deviation

N = the size of the population

x_i = each value from the population

μ = the population mean

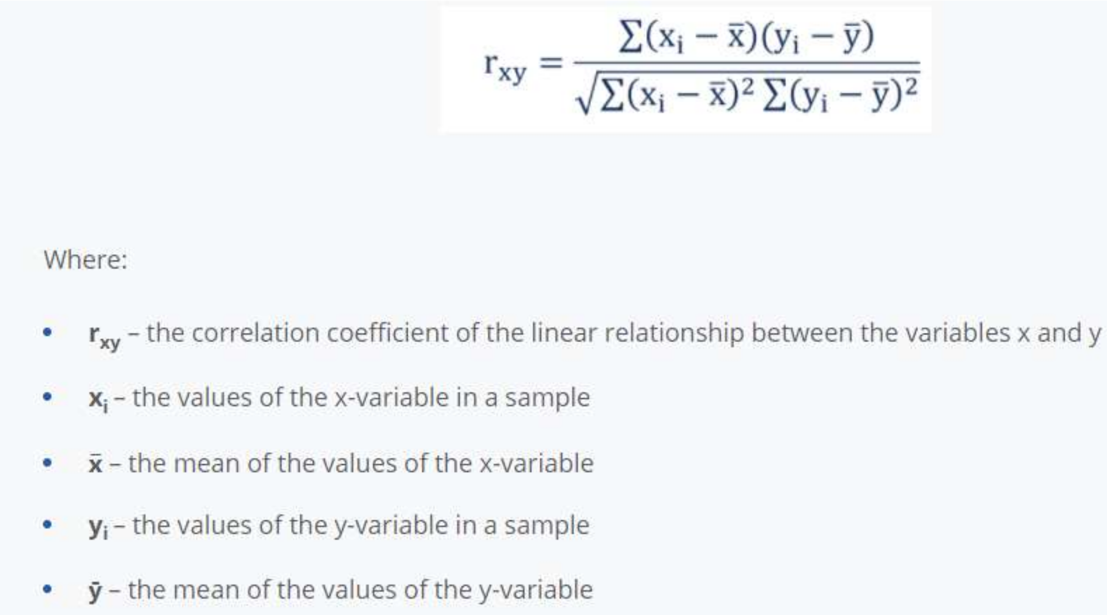
for sample

$$\sigma = \sqrt{\frac{\sum(x-\mu)^2}{n-1}}$$

Why S.D can not be negative

Beacuse SD is the deviation from the mean, which is nothing but a distance from mean, and distance can not be negative

Correlarion Furmula



$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Where:

- **r_{xy}** – the correlation coefficient of the linear relationship between the variables x and y
- **x_i** – the values of the x-variable in a sample
- **$\bar{x}$** – the mean of the values of the x-variable
- **y_i** – the values of the y-variable in a sample
- **$\bar{y}$** – the mean of the values of the y-variable

OR (covariance / s.d)

s.d can not be negative. so, covariance can be negative

DATA TANSFORMATION TECHNIQUES

Normalization (min-max scaling)

data convert between -1 to 1

$$X_{sc} = \frac{X - X_{min}}{X_{max} - X_{min}}.$$

standerdization

mean = 0 S.D =1

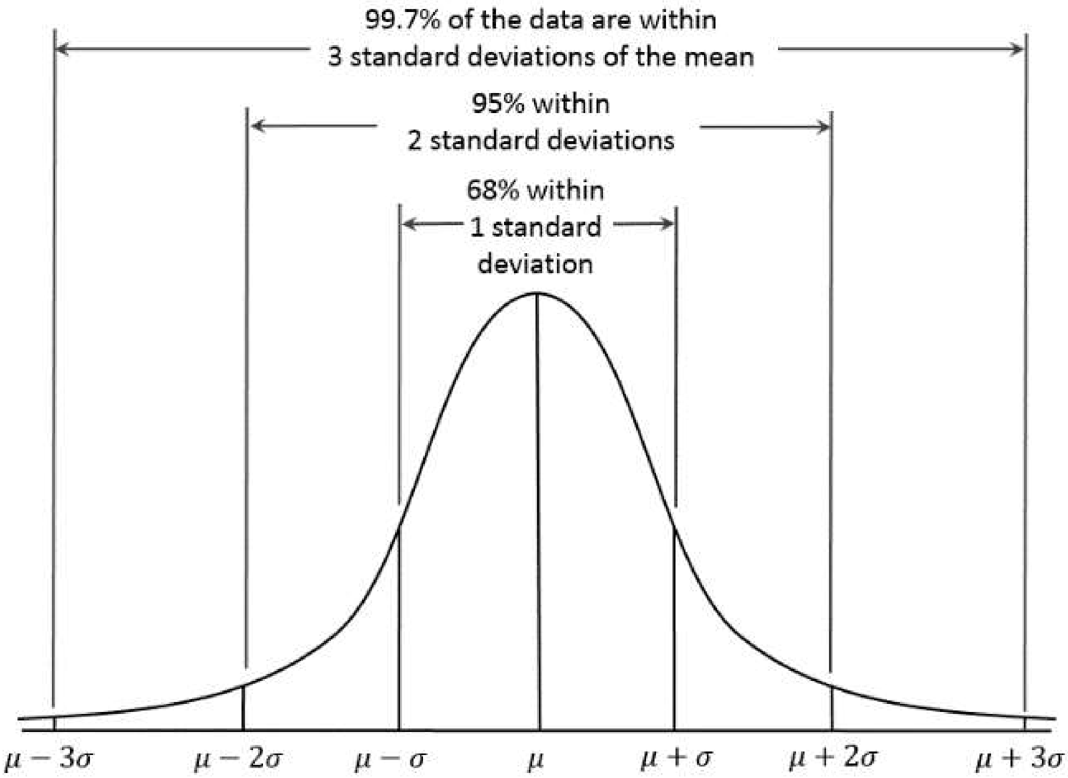
box-cox

The Box-Cox method is a data transform method that can perform a range of power transforms, including the log and the square root. The method is named for George Box and David Cox.

It can be configured to evaluate a suite of transforms automatically and select the best fit.

The resulting data sample may be more linear and will better represent the underlying non-power distribution, including Gaussian.

Properties of Normal Distribution (Gaussian Distribution)


99.7% of the data are within 3 standard deviations of the mean
95% within 2 standard deviations
68% within 1 standard deviation
$\mu - 3\sigma$
$\mu - 2\sigma$
$\mu - \sigma$
μ
$\mu + \sigma$
$\mu + 2\sigma$
$\mu + 3\sigma$

mean =median =mode

Symmetric around mean

skewness is zero and kurtosis is 3 (meso kurtic)

68% within 1 SD, 95% within 2 SD, 99.7% within 3 SD

Unipolar or Unimodal

How to analyse the normal distribution

Histogram

Q-Q plot

Statistical test (Hypothesis)

Hypothesis

A hypothesis is an educated guess about something in the world around you. It should be testable, either by experiment or observation. It considered to be true untill you prove it to be wrong.

lifebuoy claims that, it kills 99.9% of germs. So how can they say so? There has to be a testing technique to prove this claim right?? So hypothesis testing uses to prove a claim or any assumptions.

Like, if we make a statement that “Dhoni is the best Indian Captain ever.” This is an assumption that we are making based on the average wins and losses team had under his captaincy. We can test this statement based on all the match data.

Decide NULL hypothesis

we can understand the null hypothesis as already accepted statements, For example, Sky is blue. We already accept this statement.

Decide Alternate

The alternative hypothesis complements the Null hypothesis. It is the opposite of the null hypothesis such that both Alternate and null hypothesis together cover all the possible values of the population parameter.

Decide it left right or double tail

If the alternate hypothesis gives the alternate **in both directions** (less than and greater than) of the value of the parameter specified in the null hypothesis, it is called a Two-tailed test.

If the alternate hypothesis gives the alternate in **only one direction** (either less than or greater than) of the value of the parameter specified in the null hypothesis, it is called a One-tailed test.

do the sampling and calculation

compare the calculated with actual

if cal val falls in critical then reject the nullor else there is not enough evidence to reject the null hypothesis

The critical region is that region in the sample space in which if the calculated value lies then we reject the null hypothesis.

Suppose you are looking to rent an apartment. You listed out all the available apartments from different real state websites. You have a budget of Rs. 15000/ month. You cannot spend more than that. The list of apartments you have made has a price ranging from 7000/month to 30,000/month.

You select a random apartment from the list and assume below hypothesis:

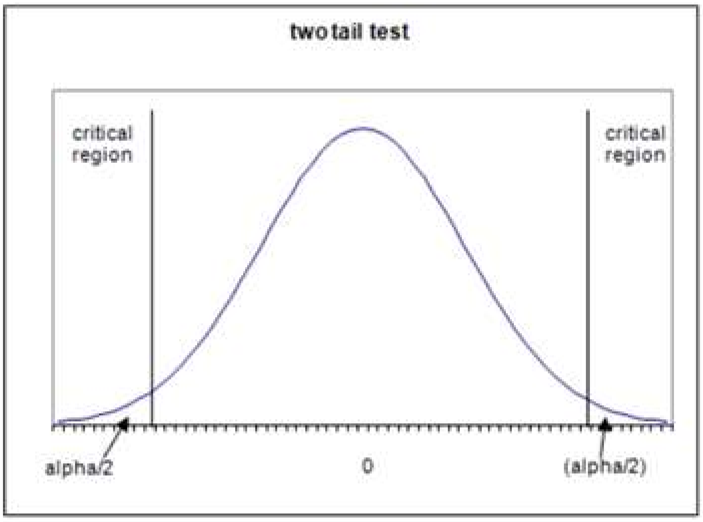
H0: You will rent the apartment.

H1: You won’t rent the apartment.

Now, since your budget is 15000, you have to reject all the apartments above that price.

Here all the Prices greater than 15000 become your critical region. If the random apartment’s price lies in this region, you have to reject your null hypothesis and if the random apartment’s price doesn’t lie in this region, you do not reject your null hypothesis.

$\mathbf{H_0}: \mu = \mu_0$ $\mathbf{H_1}: \mu \neq \mu_0;$



Level of significance(α)

The significance level, in the simplest of terms, is the threshold probability of incorrectly rejecting the null hypothesis when it is in fact true. This is also known as the type I error rate.

It is the probability of a type 1 error. It is also the size of the critical region.

Generally, strong control of α is desired and in tests, it is prefixed at very low levels like 0.05(5%) or 0.01(1%).

If H0 is not rejected at a significance level of 5%, then one can say that our null hypothesis is true with 95% assurance.

P value

the p-value is the smallest level of significance at which a null hypothesis can be rejected.

For right tailed test:

p-value = P[Test statistics >= observed value of the test statistic]

For left tailed test:

p-value = P[Test statistics <= observed value of the test statistic]

For two tailed test:

p-value = 2 * P[Test statistics >= |observed value of the test statistic|]

Decision making with p-value

We compare p-value to significance level(alpha) for taking a decision on Null Hypothesis.

If p-value is greater than alpha, we do not reject the null hypothesis.

If p-value is smaller than alpha, we reject the null hypothesis.

Which factor determine Power of Testing

By increasing sample size

By increasing SD

Significant level

CHI-SQUARE AND ANOVA

Null hypothesis in chi-square

Ho = There is no relation between two variable

H1= opposite of null

known as test of independence

Null hypothesis in ANOVA

Ho = mean is same for all the samples (population)

H1= opposite of null

Python List

A list in Python is a heterogeneous container for items.

languages=['C++',[1,2,3],(4,5),'Scratch']

A list is mutable

Mutability is the ability to be mutated, to be changed. A list is mutable, so it is possible to reassign and delete individual items as well.

Python Tuple

But the major difference between the two (tuple and list) is that a list is mutable, but a tuple is immutable. This means that while you can reassign or delete an entire tuple, you cannot do the same to a single item or a slice.

Python Set

A set, in Python, is just like the mathematical set. It does not hold duplicate values and is unordered. However, **it is not immutable**, unlike a tuple.

myset={3,1,2}

Python Dictionaries

It holds word-meaning pairs. Likewise, a Python dictionary holds key-value pairs.

mydict={1:2,2:4,3:6}

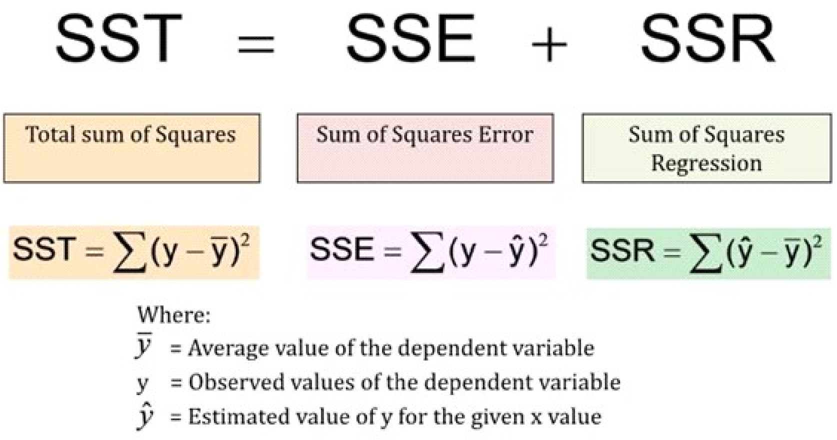
What is R-Square and Adj.R square

R-squared statistic or **coefficient of determination** is a scale invariant statistic that gives the proportion of variation in target variable explained by the linear regression model.

R-squared **gives the degree of variability** in the target variable that is explained by the model or the independent variables. **If this value is 0.7, then it means that the independent variables explain 70% of the variation in the target variable.**

OR

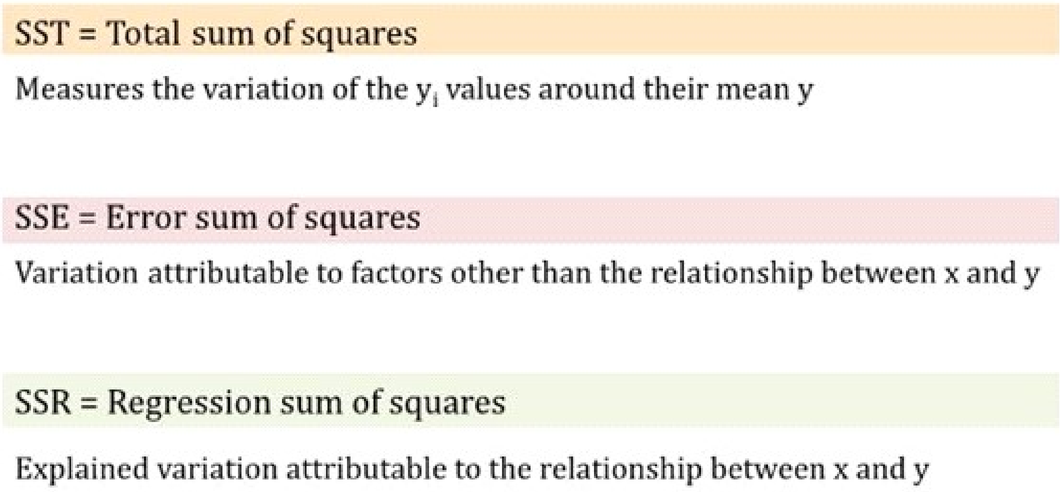
The R-squared, also called the coefficient of determination, is used to explain the degree to which input variables (predictor variables) explain the variation of output variables (predicted variables). It ranges from 0 to 1. For example, if the R-squared is 0.9, it indicates that 90% of the variation in the output variables are explained by the input variables.



$$SST = SSE + SSR$$

$SST = \sum(y - \bar{y})^2$ $SSE = \sum(y - \hat{y})^2$ $SSR = \sum(\hat{y} - \bar{y})^2$

Where:
$\bar{y}$ = Average value of the dependent variable
y = Observed values of the dependent variable
$\hat{y}$ = Estimated value of y for the given x value



SST = Total sum of squares

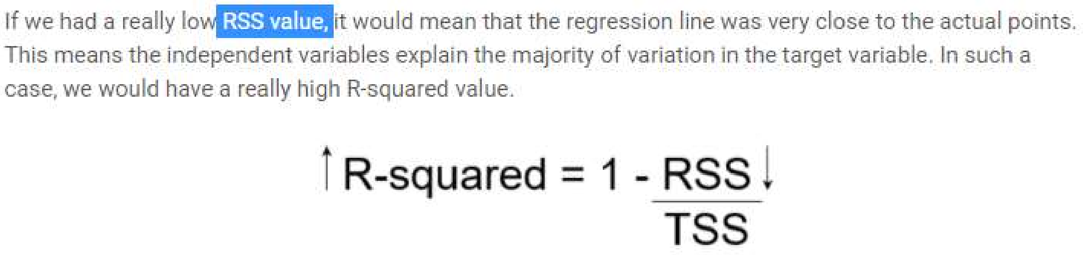
Measures the variation of the y_i values around their mean y

SSE = Error sum of squares

Variation attributable to factors other than the relationship between x and y

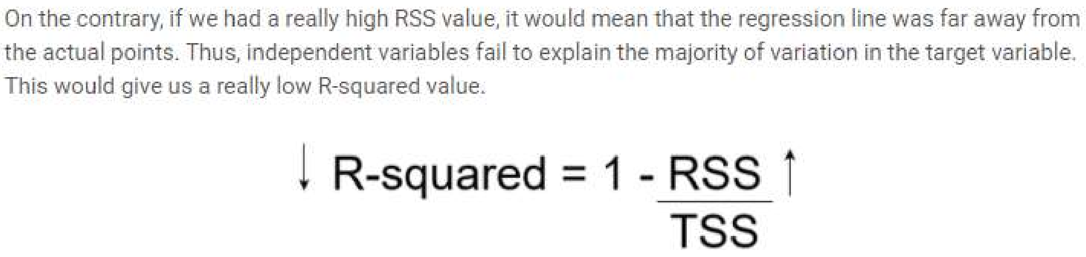
SSR = Regression sum of squares

Explained variation attributable to the relationship between x and y



If we had a really low RSS value, it would mean that the regression line was very close to the actual points. This means the independent variables explain the majority of variation in the target variable. In such a case, we would have a really high R-squared value.

$$\uparrow \text{R-squared} = 1 - \frac{\text{RSS}}{\text{TSS}} \downarrow$$



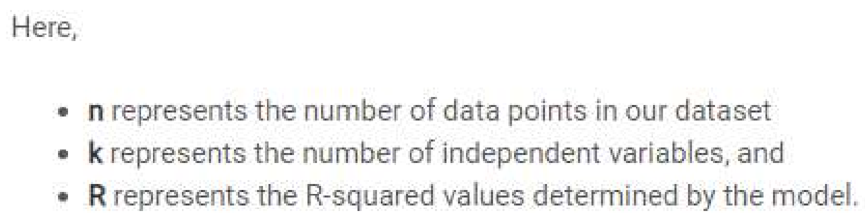
On the contrary, if we had a really high RSS value, it would mean that the regression line was far away from the actual points. Thus, independent variables fail to explain the majority of variation in the target variable. This would give us a really low R-squared value.

$$\downarrow \text{R-squared} = 1 - \frac{\text{RSS}}{\text{TSS}} \uparrow$$

Adjusted R-squared statistic

The Adjusted R-squared takes into account the number of independent variables used for predicting the target variable. In doing so, we can determine whether adding new variables to the model actually increases the model fit.

$$Adjusted\, R^2 = \{1 - [\frac{(1-R^2)(n-1)}{(n-k-1)}]\}$$



Here,

- **n** represents the number of data points in our dataset
- **k** represents the number of independent variables, and
- **R** represents the R-squared values determined by the model.

Range if R square is between 0 to 1, where Adj R square can be more

How to check the assumptions in Linear Regression

Histogram

Q-Q plot

Collinearity chart

Scatter plot of (actual - predicted) error terms

What are the evolution metrics

MSE

RMSE

MAE - (sumation of |Actual - predicated|) / n

R SQUARE

ADJT. R SQUARE

Advantages of MSE

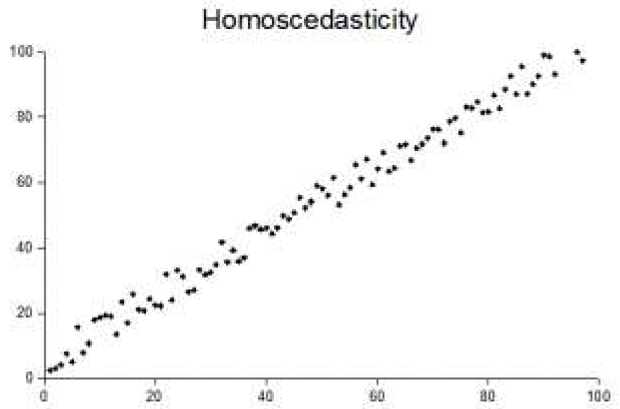
1. Negative and Positiove error should not be cancel out

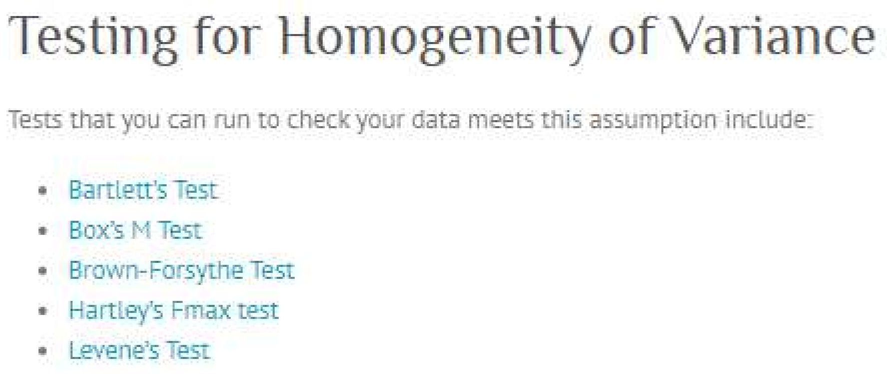
2. If your errors are more then MSE will be more and punish the model for more errors

Homoscedasticity

homoscedasticity means “having the same scatter.”

the points must be about the same distance from the line.





Testing for Homogeneity of Variance

Tests that you can run to check your data meets this assumption include:

- Bartlett's Test
- Box's M Test
- Brown-Forsythe Test
- Hartley's Fmax test
- Levene's Test

Q-Q chart

A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution.

F-Test

F-test in regression compares the fits of different linear models. Unlike t-tests that can assess only one regression coefficient at a time, the F-test can assess multiple coefficients simultaneously.

Curse of Dimentionality

When you analyze and organize data in high-dimensional spaces (usually in thousands), various situations can arise.

<u>Example</u>

All kids love to eat chocolates. Now, you bring a truckload of chocolates in front of the kid. These chocolates come in different colors, shapes, tastes, and price. Consider the following scenario.

The kid has to choose one chocolate from the truck depending on the following factors.

Only taste – There are usually four tastes, sweet, salty, sour, and bitter. Hence, the child will have to try out only four chocolates before choosing one to its liking.

Taste and Color – Assume there are only four colors. Hence, the child will now have to taste a minimum of 16 (4 X 4) before making the right choice.

Taste, color, and shape – Let us assume that there are five shapes. Therefore, the child will now have to eat a minimum of 80 chocolates (4 X 4 X 5).

Ridge and Lasso

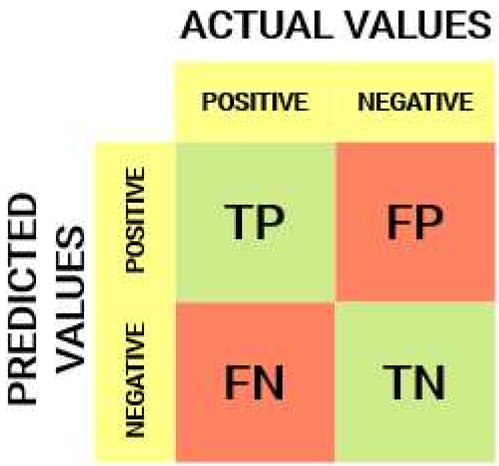
cost function- MSE + Lambda (slop)^2

Overcome the Over-fitting problem

Overcome multicollinearity of the data

Logistic Regression

Confusion Matrics



		ACTUAL VALUES	
		POSITIVE	NEGATIVE
PREDICTED VALUES	POSITIVE	TP	FP
	NEGATIVE	FN	TN

False Positive (FP) – Type 1 error

False Negative (FN) – Type 2 error

Precision vs. Recall

Precision tells us how many of the correctly predicted cases actually turned out to be positive.

$$Precision = \frac{TP}{TP + FP}$$

Recall tells us how many of the actual positive cases we were able to predict correctly with our model. Also, known as TPR or sensitivity.

$$Recall = \frac{TP}{TP + FN}$$

False Negative Rate

$$FNR = \frac{FN}{TP + FN}$$

0

Specificity / True Negative Rate

$$Specificity = \frac{TN}{TN + FP}$$

False Positive Rate

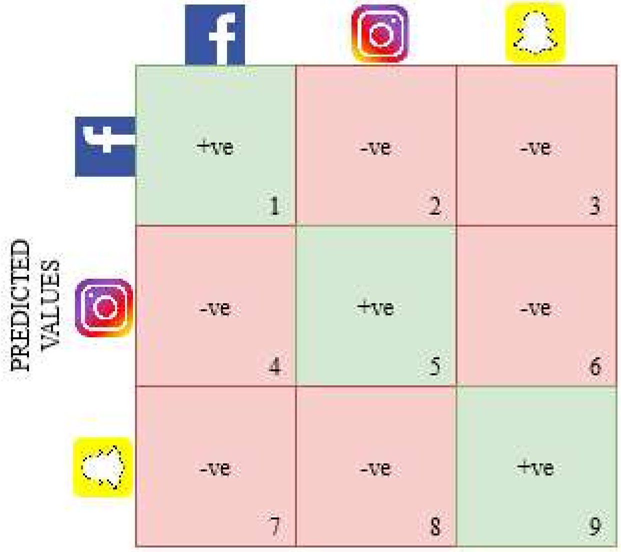
$$FPR = \frac{FP}{TN + FP} = 1 - Specificity$$

F1-Score

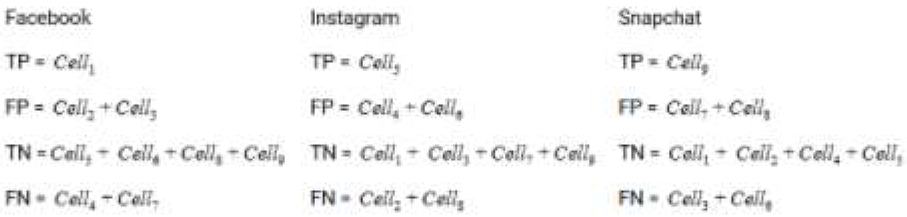
$$F1 - score = \frac{2}{\frac{1}{Recall} + \frac{1}{Precision}}$$

TN is not available in the formula of F1- Score

For multi class



	f		
f	+ve 1	-ve 2	-ve 3
	-ve 4	+ve 5	-ve 6
	-ve 7	-ve 8	+ve 9



Facebook	Instagram	Snapchat
TP = $Cell_1$	TP = $Cell_5$	TP = $Cell_9$
FP = $Cell_2 + Cell_3$	FP = $Cell_4 + Cell_6$	FP = $Cell_7 + Cell_8$
TN = $Cell_5 + Cell_6 + Cell_8 + Cell_9$	TN = $Cell_1 + Cell_3 + Cell_7 + Cell_9$	TN = $Cell_1 + Cell_2 + Cell_4 + Cell_5$
FN = $Cell_4 + Cell_7$	FN = $Cell_2 + Cell_8$	FN = $Cell_3 + Cell_6$

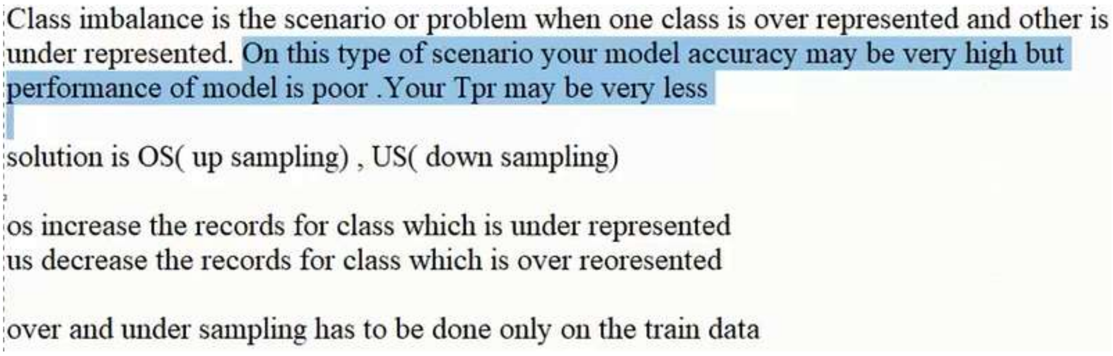
AUC-ROC curve

AUC-ROC curve helps us visualize how well our machine learning classifier is performing.

F1 score = 2 * Precision * Recall / (Precision + Recall)

F1 score does not TN score , when have highly imbalace data may get good auroc but not f-1 score.

Class Imbalance



Class imbalance is the scenario or problem when one class is over represented and other is under represented. On this type of scenario your model accuracy may be very high but performance of model is poor .Your Tpr may be very less

solution is OS(up sampling) , US(down sampling)

os increase the records for class which is under represented
us decrease the records for class which is over reoresented

over and under sampling has to be done only on the train data

STRATIFY WHILE TRAIN_TEST_SPLIT

Stratified splits are desirable in some cases, like when you're classifying an imbalanced dataset, a dataset with a significant difference in the number of samples that belong to distinct classes.

Decision Tree

2 Critearea

Entropy (purity of the data)

Gini Score

Hyper parameters (which are not part of the data Parameter, User can change it)

depth of tree

no. of leafs

class weight

min_split

Parameters (which can be not changed)

mean

median

mode

number of rows and columns

etc.

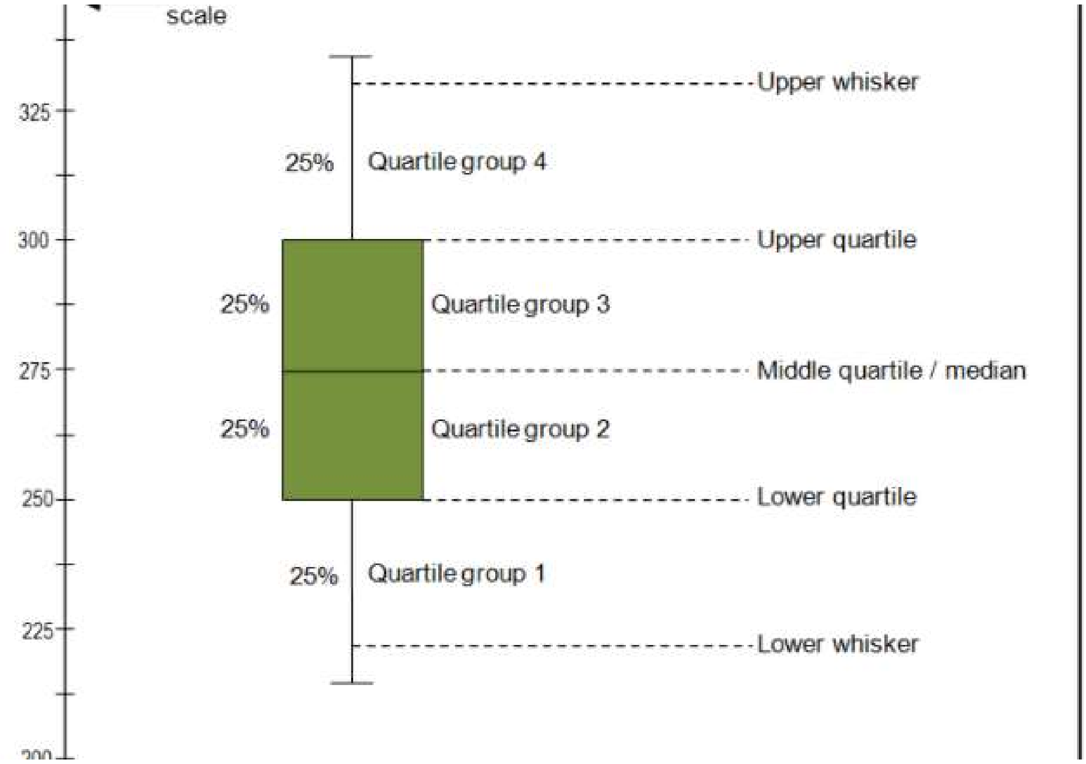
Outliers finding Techniques

Z-Score

Data lies far away from the mean

Boxplot graph

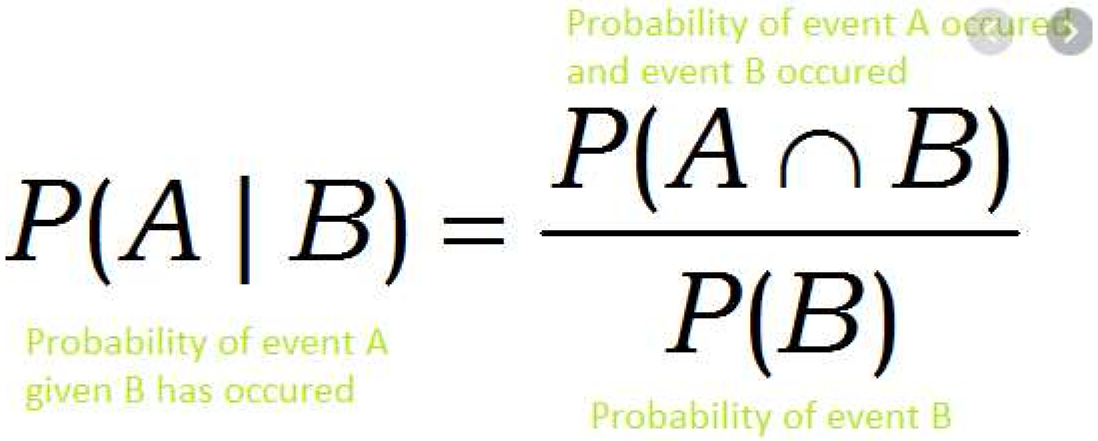
IQR = Interquartile range (Q3-Q1)



Naive Bayes

p(A/B)- probability of occurance of A, when given probabitliy of B

p(B/A) p(A) / p(B)



$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Probability of event A occured and event B occured
Probability of event A given B has occured
Probability of event B

Why naive bayes is so naive

NB takes parameter at a time

KNN

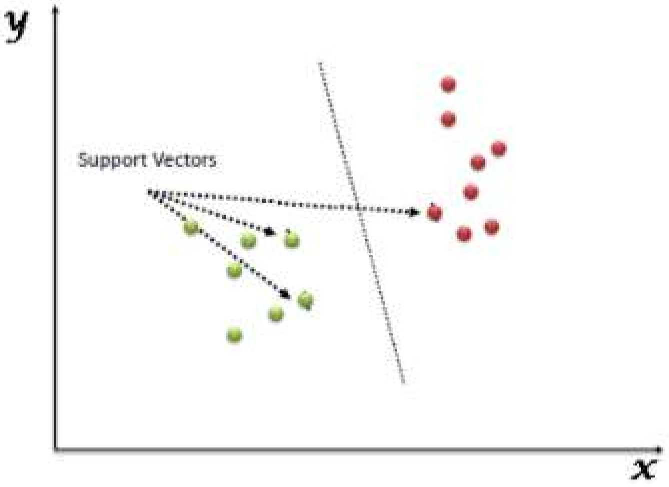
K-nn algorithm is basically algorithm in which nearest neighbours are calculated by using uclaidian distance(for p1(X1,Y1) and p2(X2,Y2) points distance formula = [(X2-X1)^2 + (Y2-Y1)^2]^1/2.

To check the best value of k in k-nn

We find the accuracy rate and (error rate = 1- accuracy rate) with value of k ranging from 1 to 40. and the value of k, after which graph shows constant value of error rate, that point can be taken as k value.

SVM

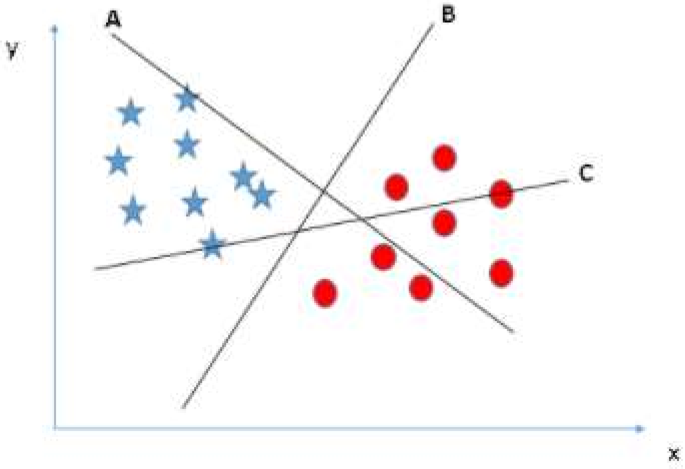
In SVM, we perform classification by finding the hyper-plane that differentiates the two classes very well.



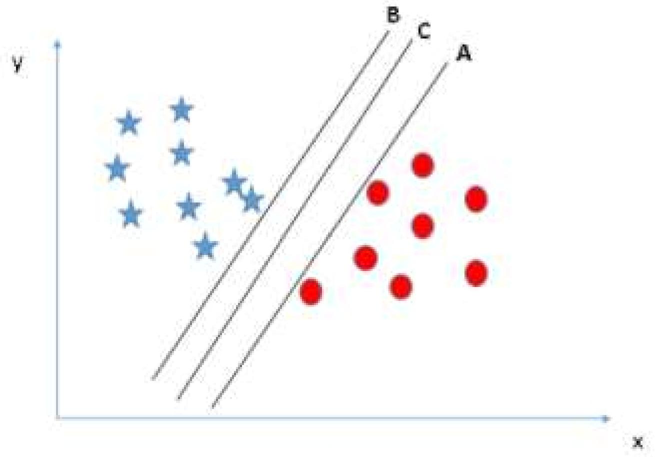
Identify the right hyper-plane (Scenario-1):

“Select the hyper-plane which segregates the two classes better”

here we can choose B as hyper plane

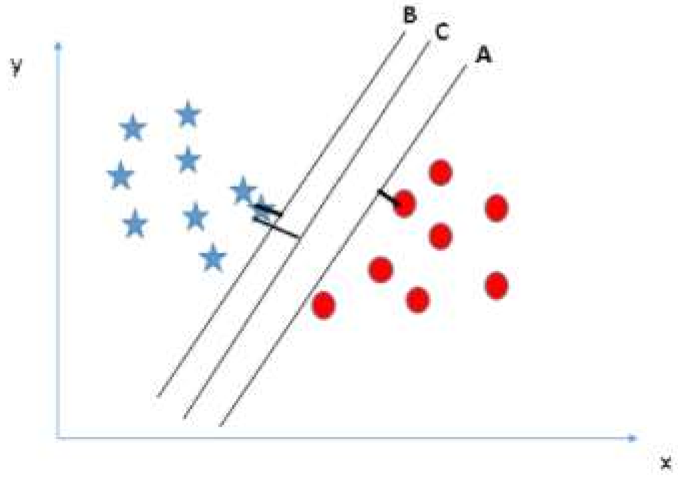


BUT



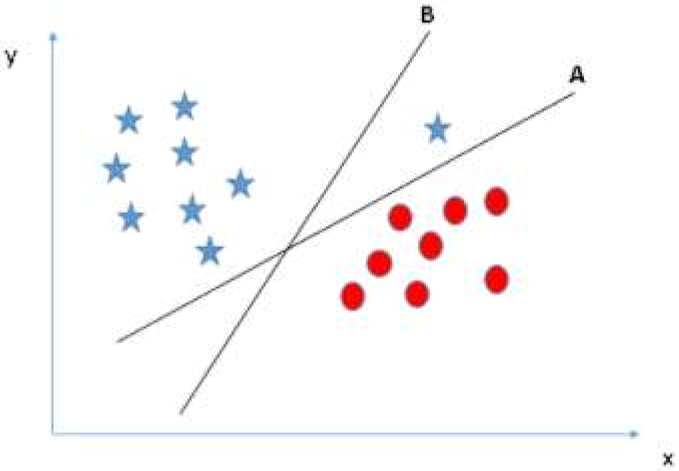
in this kind of conditions,

maximizing the distances between nearest data point (either class) and hyper-plane will help us to decide the right hyper-plane. This distance **is called as Margin**



here C hyper plane has high margin as compare to others.

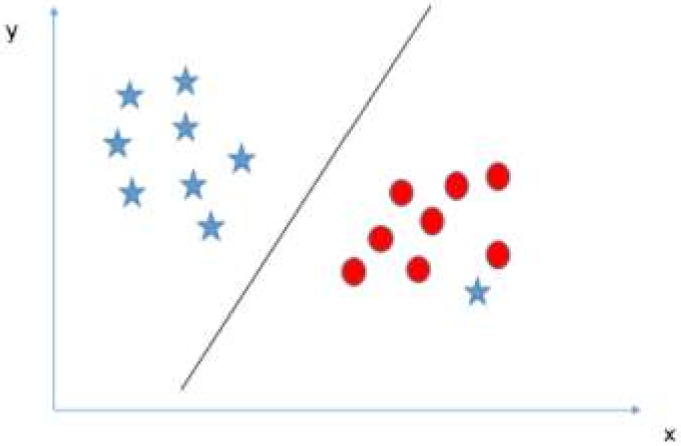
BUT



SVM selects the hyper-plane which classifies the classes accurately prior to maximizing margin.

A would be the perfect hyper plan in this case

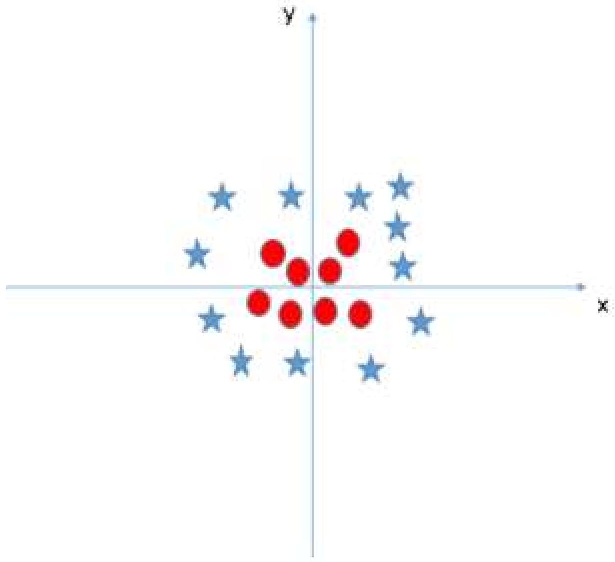
BUT

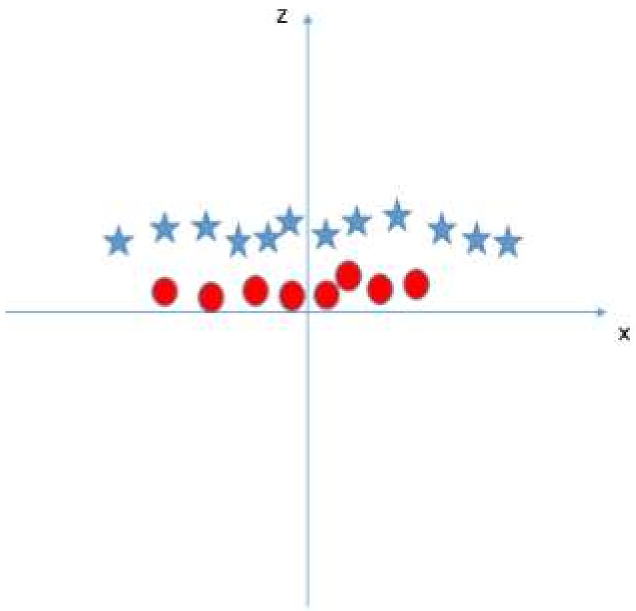


SVM algorithm has a feature to ignore outliers and find the hyper-plane that has the maximum margin.

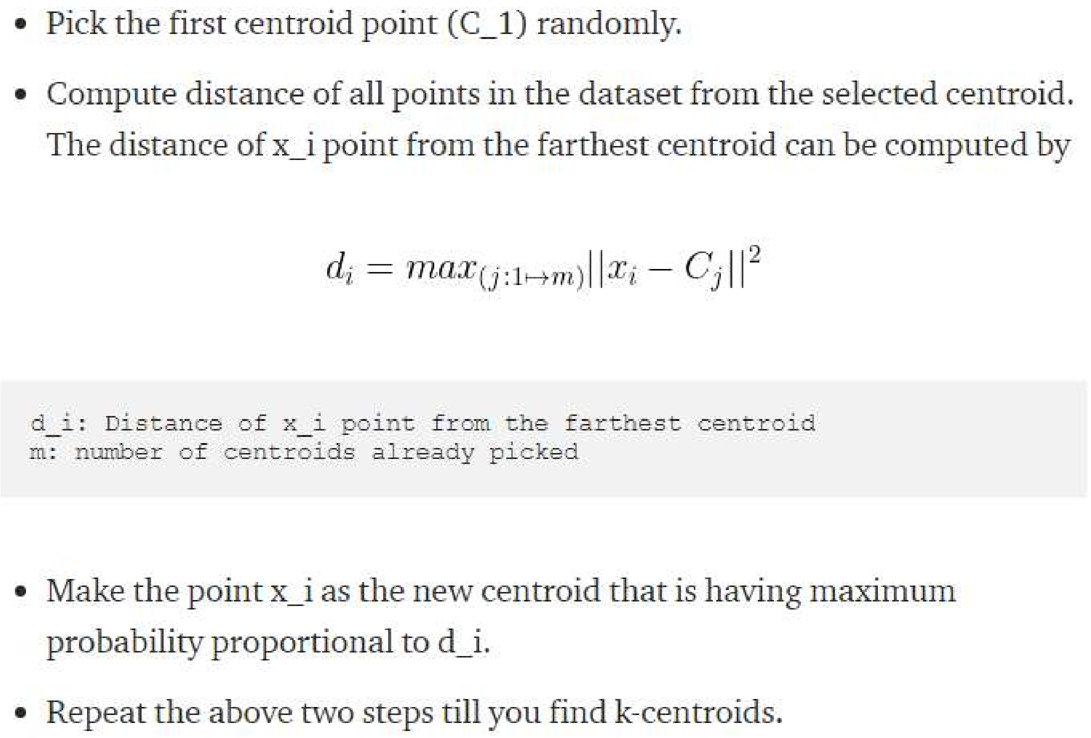
the SVM algorithm has a technique called the kernel trick.

The SVM kernel is a function that takes low dimensional input space and transforms it to a higher dimensional space


y
x


z
x

K means ++



- Pick the first centroid point (C_1) randomly.
- Compute distance of all points in the dataset from the selected centroid. The distance of x_i point from the farthest centroid can be computed by

$$d_i = max_{(j:1 \mapsto m)} ||x_i - C_j||^2$$

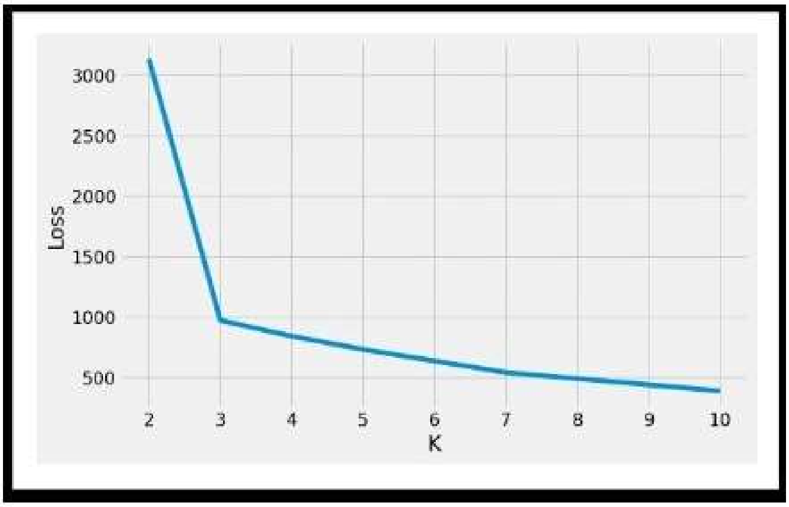
```
d_i: Distance of x_i point from the farthest centroid
m: number of centroids already picked
```

- Make the point x_i as the new centroid that is having maximum probability proportional to d_i.
- Repeat the above two steps till you find k-centroids.

Determine K in K-means

$$loss = argmin \sum_{i=1}^{k} \sum_{x \epsilon S_i} ||x - C_i||^2$$

To determine the right 'K', draw a plot between loss vs k.



Plot for Loss vs K, (Image 10)

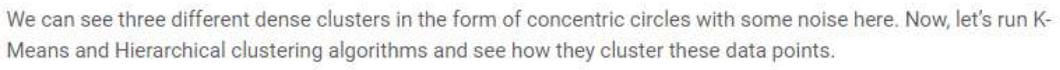
DBSCAN

Two hyperparameter

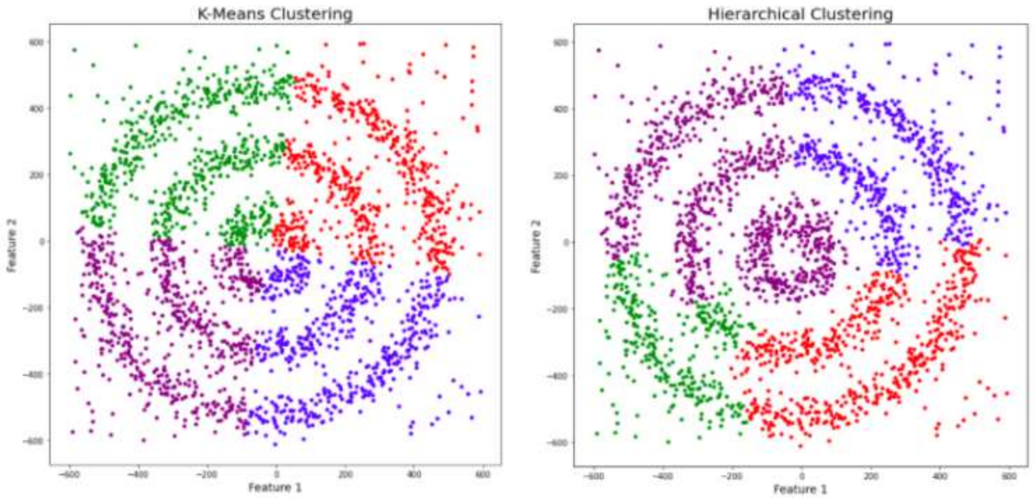
Min distance

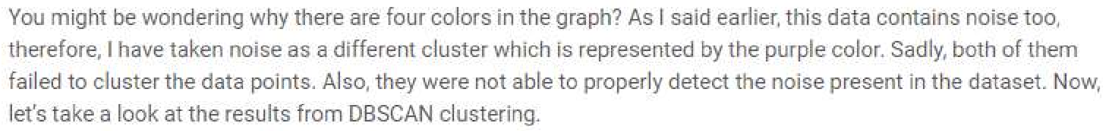
Min point

It is used

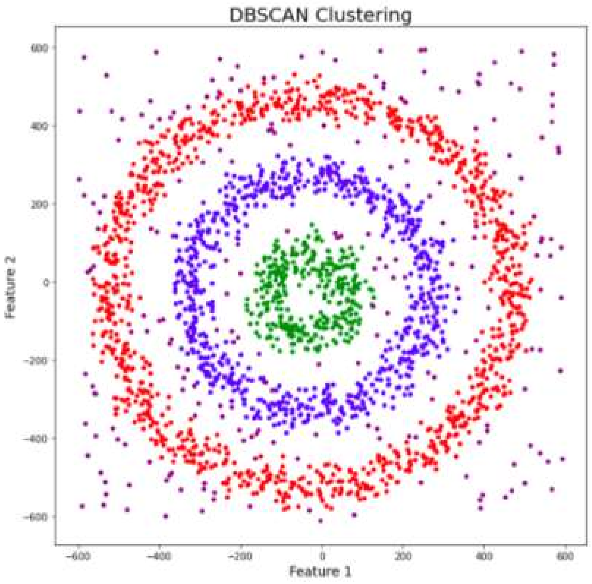


We can see three different dense clusters in the form of concentric circles with some noise here. Now, let's run K-Means and Hierarchical clustering algorithms and see how they cluster these data points.





You might be wondering why there are four colors in the graph? As I said earlier, this data contains noise too, therefore, I have taken noise as a different cluster which is represented by the purple color. Sadly, both of them failed to cluster the data points. Also, they were not able to properly detect the noise present in the dataset. Now, let's take a look at the results from DBSCAN clustering.



It groups 'densely grouped' data points into a single cluster. It can identify clusters in large spatial datasets by looking at the local density of the data points. **The most exciting feature of DBSCAN clustering is that it is robust to outliers.**

It also does not require the number of clusters to be told beforehand, unlike K-Means, where we have to specify the number of centroids.

DBSCAN requires only two parameters: epsilon and minPoints. Epsilon is the radius of the circle to be created around each data point to check the density and minPoints is the minimum number of data points required inside that circle for that data point to be classified as a Core point.

NLP

Stemming - mapping to the root word

Stemming is the process of reducing the words(generally modified or derived) to their word stem or root form. The objective of stemming is to reduce related words.

beautiful and beautifully are stemmed to beauti

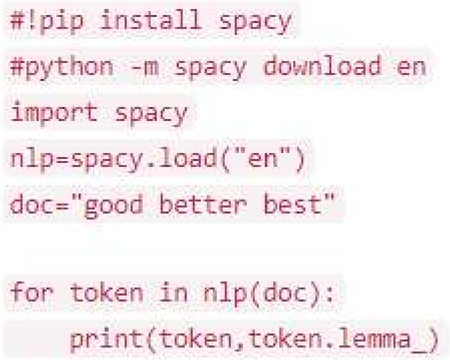
good, better and best are stemmed to good, better and best respectively

Lemmatisation

Lemmatisation is the process of reducing a group of words into their lemma or dictionary form. It takes into account things like POS(Parts of Speech), the meaning of the word in the sentence, the meaning of the word in the nearby sentences etc. before reducing the word to its lemma. For example, in the English Language-

beautiful and beautifully are lemmatised to beautiful and beautifully respectively.

good, better and best are lemmatised to good, good and good respectively.



```
#!pip install spacy
#python -m spacy download en
import spacy
nlp=spacy.load("en")
doc="good better best"

for token in nlp(doc):
    print(token,token.lemma_)
```

What is Word Embeddings?: Word Embeddings is the name of the techniques which are used to represent Natural Language in vector form of real numbers. They are useful because of computers' inability to process Natural Language. So these Word Embeddings capture the essence and relationship between words in a Natural Language using real numbers. In Word Embeddings, a word or a phrase is represented in a fixed dimension vector of length say 100.**Word Embeddings**

So for example-

A word “man” might be represented in a 5-dimension vector as where each of these numbers is the magnitude of the word in a particular direction.

STOP WORD

A stop word is a commonly used word but do not any meaning (such as “the”, “a”, “an”, “in”) that a search engine has been programmed to ignore, both when indexing entries for searching and when retrieving them as the result of a search query.

IDF (Invert Document Frequency)

log (total number of doc/ num of doc in that word appear)

TDM (term document matrics)

each unique word become a feature or seperate column

0 indicate a absence and 1 or more indicate number of time word has repeated

sparse matrix

Most of the elements are zero or haivng very less non-zero element

Name Entity Recognation

It recognise the object. rg. dollar - Money, Surat - Place

Hyper parameter in Nueral Network

Activation Function

Optimizer

Number of Hidden Layer

Number of nuerons

Learning Rate

Batch Size

Drop out ratio

Validation Size

Number of epochs

Disadvantages of nueral network

Black Box

Time consuming

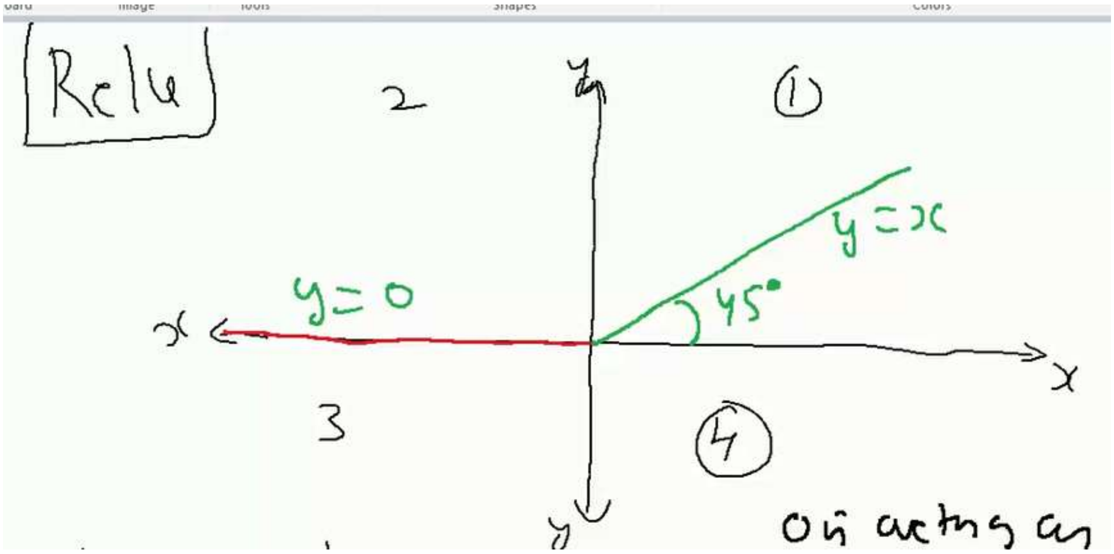
Huge computation

We can not do feature selection and other things

Relu Activation Function

It is an activation function where, for x <=0, y = 0 and x >0, y = X

here 0 acting as cutoff or threshold



convert linear input to non-linear output

if the input to the activation function is greater than a threshold, then the neuron is activated, else it is deactivated, i.e. its output is not considered for the next hidden layer.

The main advantage of using the ReLU function over other activation functions is that it does not activate all the neurons at the same time.

Linear Models

A linear model is one that outputs a weighted sum of the inputs, plus a bias (intercept) term. Where there is a single input feature, X, and a single target variable, Y, this is of the form:

$$f(X) = \beta_0 + \beta_1 X$$

Graphically, a linear model produces:

a point in 1-dimension (no features)

a line in 2-dimensions (one feature)

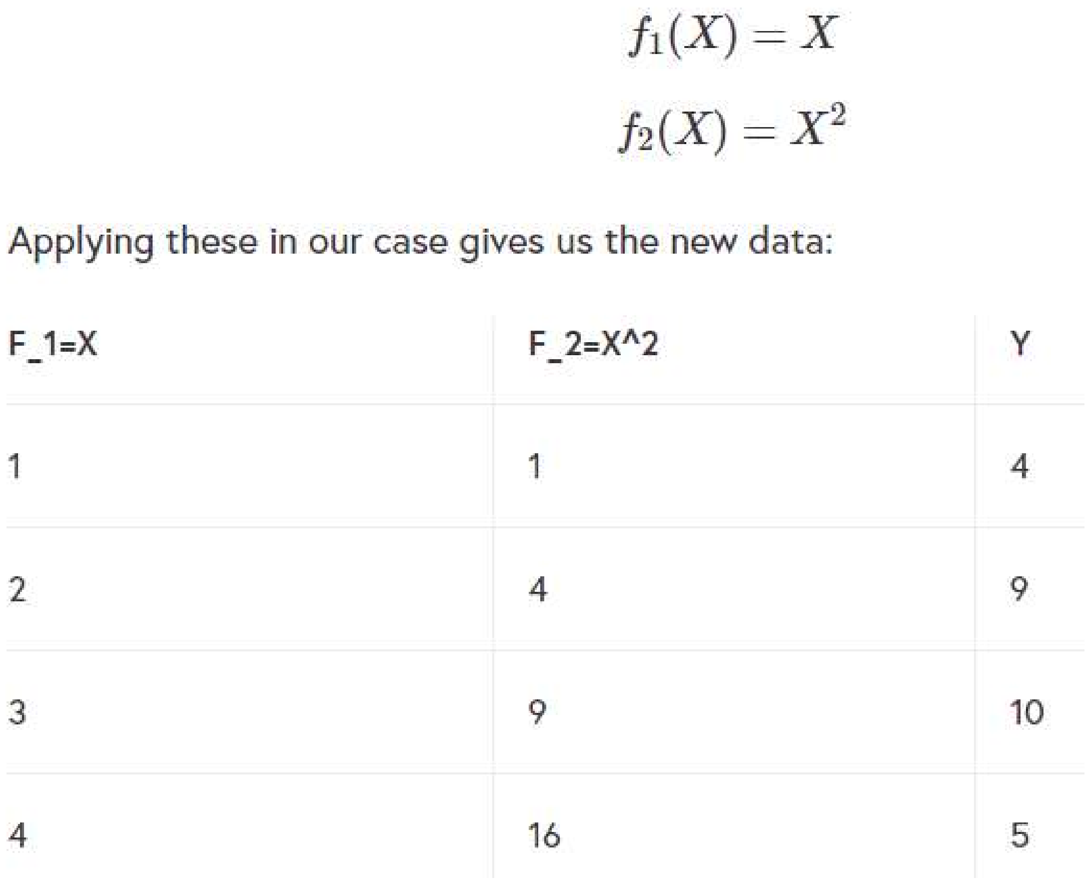
a plane in 3-dimensions (two features)

a hyperplane in n-dimensions (n-1 features)

Non-Linear Models

Let us assume we have the data given below. We wish to generate a model that estimates the value of Y given X.

Let us consider two such transformation functions of X:



$$f_1(X) = X$$

$$f_2(X) = X^2$$

Applying these in our case gives us the new data:

F_1=X	F_2=X^2	Y
1	1	4
2	4	9
3	9	10
4	16	5

The

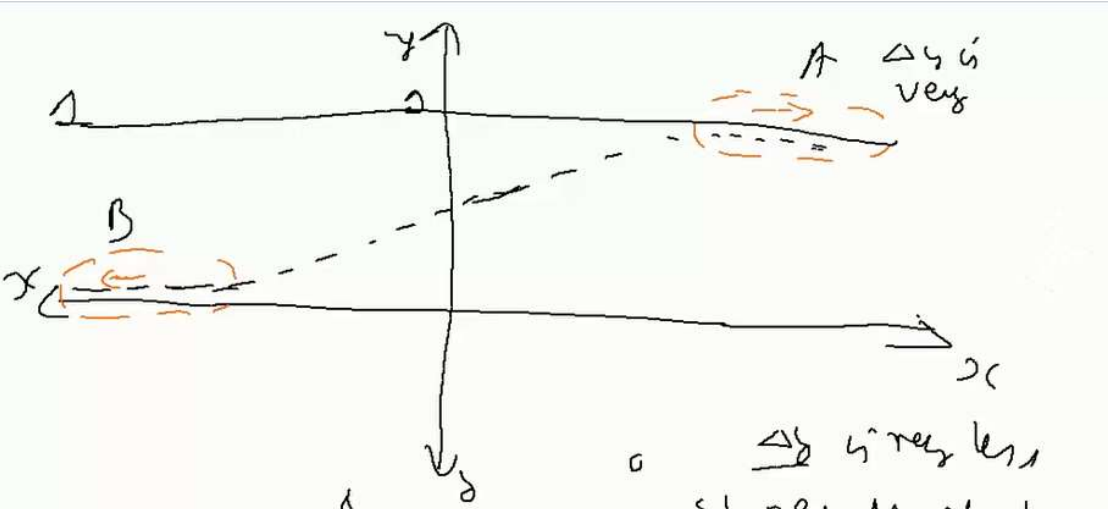
and

columns of this new data set are latent variables (a function of a random variable, or set of random variables, is itself a random variable). We can now consider the relationship between our target variable and these latent variables. Using OLS to model this relationship we generate the following model:

$$\hat{Y} = -6.5 + 12.9X - 2.5X^2$$

Problem with sigmoid

In sigmoid curve at point or area A, delta Y or change in slop is very slow, and if change in slop is slow then model would take huge time to converge or it will not converge



In other words, in sigmoid slop will not change near posite or negative infinity so it won't be able to change the weight

change at h3 is least as compare to h2 and h1, maximum change should have been at H1, but change at H1 is least to, H2 has max change(see sigmoid curve), it consume huge time to converge beacuse H1 has least change in weights

ADVANTAGE OF RELU OVER SIGMOID

As we discussed problem with sogmoid function, relu can overcome that as according to the relu function, for minimum value of x, y=x

WHAT WOULD BE THE PROLEM IF HAVING MANY HIDDEN LAYERS

rate of change of weight decayed down due to high number of hidden layers

model take huge time to converge

Vanishing Gradient Decent

we know that Weight updation formula is W (new weight) = W(old weight) - (Learning rate) * (derivative of loss function respect to old weights)

in some activation fuction such as sigmoid, the value of derivative of optimizer with respect to weight is very less (between 0 to 0.25 and it get reduce with increasing number of layers.

Also, the value of learning rate is too small. Therefore, at some point the value of old and new weights becomes almost same. Due to this our model face difficulty to trined. and this is known as vanishing gradient decent.

in case of tanh, we face a problem of gradient decent.

Exploding Gradient

main reason the exploding gradient decent problem happens, it is beacuse of WEIGHTS.

Basically, when weights are higher then the old weights and new weights varries a lot. Therefore, it would not reach at global minima ever. This problem is knoen as exploding gradient decent.

Advantage of Rectified Linear Unit (Relu) over Relu

In relu function sometime if we get 0 output for any of the derivative in chain then all , we get>> W (old) = W (new). It will create dead neuron or dead activation function.

for that Reky relu used.

Stochastic GD Vs GD Vs Mini Batch SGD

SGD comes into picture when 1 single recored is passed one by one

Minibatch comes into picture when k number of batch passes one by one into nueral network

Gradient Decent comes into picture when whole record passes to neural network at once.

CNN

in convolutional, it passes image through a filter of (x * x)

so, if we have 6 * 6 matrix and 3 * 3 convolutional filter, now if we pass it main matrix through this filter then it converted into 4 *4 accordign to the formula == (main - filter) + 1

but this formula will change for padding.

PADDING IN CNN

padding is just adding extra layer arround the martix so, we dont loose the info.

zero padding technique is most common where you add zero to all boxes

so, in order to get 6* 6 out in above example, we have to padding in main matrix then applt filter 3*3 so we get output of 6*6 matrix.

if we apply padding then formula changes = (n + 2p -f) +1 = (6+2(1)-3) + 1 = 6, where p is number of padding layer added.

Operation of CNN

In CNN filter is getting updated by back propogation in CNN.

when it run for first time, the values present in filters are randomly selected, and I get matrix 4*4(according to above example), then we apply relu function to each feed.

Then back propogation occurs to update that number

MAX POOLING

After convolutin operation, max pooling can be applied.

eg. we have multiple cat images, this faces should more precisely captured at higher levels.

in max pooling, we apply another filter, and always remember that jump in max poling is two

In maxpooling it picks maximum value from matrix area as output, therefore, high intensity image can be detected.

Moreover, this max pooling or min pooling matrix also get updated by back-propogation.

min pooling and max pooling, in min pooling lower number from matrix has been selected

STIDE

When we have a stride of one we move across and down a single pixel. With higher stride values, we move large number of pixels at a time and hence produce smaller output volumes.

We can apply a simple formula to calculate the output dimensions. The spatial size of the output image can be calculated as([W-F+2P]/S)+1. Here, W is the input volume size, F is the size of the filter, P is the number of padding applied and S is the number of strides. Suppose we have an input image of size 32,32,3, we apply 10 filters of size 3,3,3, with single stride and no zero padding.

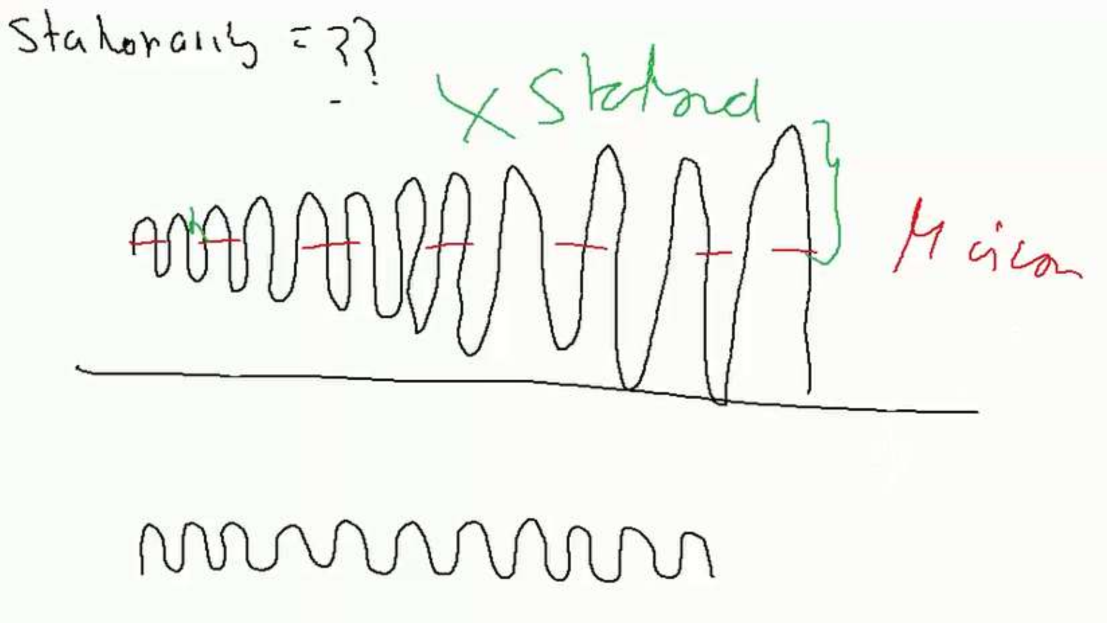
DATA AUGMENTATION CNN (Imp)

when we pass some image to CNN, data augmentation helps to create different diffrent images of that same image by flliping, horizantal shifting, vertical shifting, x% of zoom, or add some noise.

eg. if we take a example of cat, if image of cat is inverse or cat is on left corner, or in right, although our CNN model should predict that image as cat. So, we have to train our model something like that so, it can able to predict the given image.

What is Stationarity in time series

where mean and variance are stationary


Stationarity = ??
X Stationd
Mean