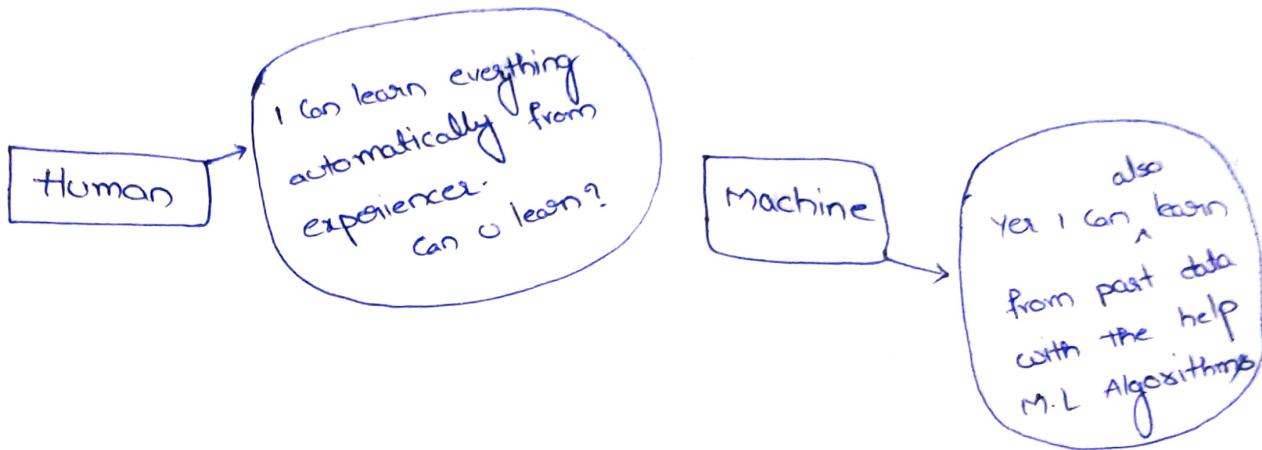


Machine learning

i) What is M.L

The machine that automatically learns from data that improves performance from the past experience and predict things without being explicitly programmed.



Classification of M.L

- Supervised learning
- Unsupervised learning
- Reinforcement learning

Supervised learning :- here the data is labelled. Ex:- students learn things in the supervision of the teacher

It is classified into 2 categories

- Classification
- Regression.

Unsupervised learning :- here the data is not labelled, classified or categorized and the algorithm needs to act on that data

Without any supervision

Unsupervised learning is classified into two categories

- clustering
- Association

Reinforcement learning :- It is an feedback-based learning method.

Machine learning life cycle

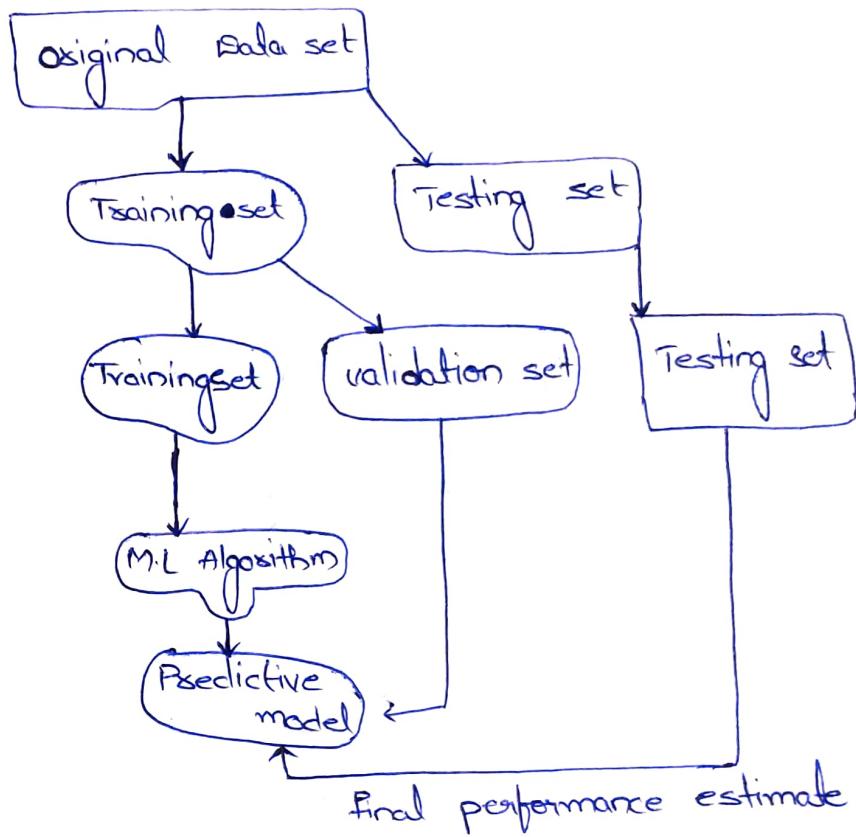
- * Gathering Data [Database, Files, Internet]
- * Data preparation [Data exploration, pre-process]
- * Data wrangling [Missing values, Invalid, Duplicates, Noise data]
- * Analyze Data [visualization]
- * Train the model [train data]
- * Test the model [test data]
- * Deployment

How to get Datasets for M.L

Numerical :- Data with numerical temp, house price etc

Categorical :- Yes/no ; Blue/green, True/False etc

ordinal :- Similar to categorical but can be measured on the basis of comparison.



Popular sources of M.L Dataset

- Kaggle
- UCI
- Dataset via AWS
- Google Dataset search Engine
- Microsoft Dataset
- Awesome public Dataset Collection
- Govt Dataset
- Computer vision Dataset
- Web scraping
- sklearn Dataset

Data Preprocessing

In real world data generally contains noise, missing values

so---etc

For that we follow the following steps

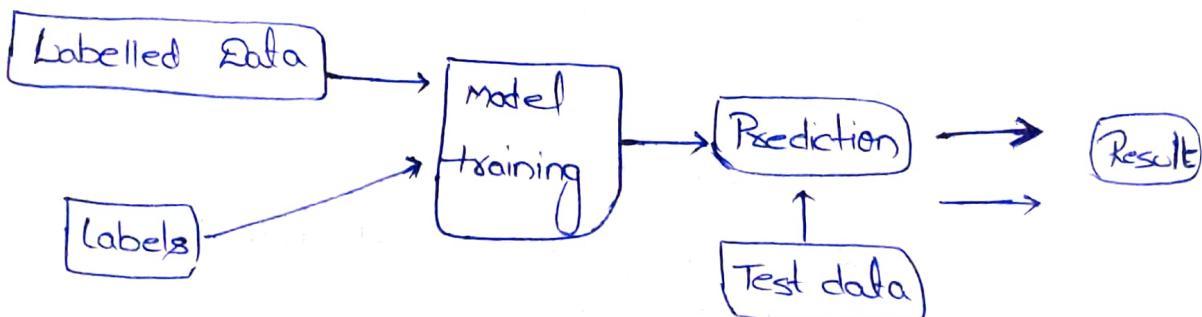
- Getting the dataset
- Importing libraries
- Importing dataset
- Finding missing data
- Encoding categorical data
- Splitting the data into train and test
- Feature scaling

Supervised Learning

The models are trained using labelled dataset, where the model learns about each type of data

once model is trained then it is tested on the basis of test data. and then it predicts the output

The working of supervised learning



Regression

Regression algorithms are used if there is a relationship between the input and output variable. It is used for the prediction of continuous variables.

Eg:- Weather forecasting, market trends etc.

Some popular Regression Algorithms under supervised learning

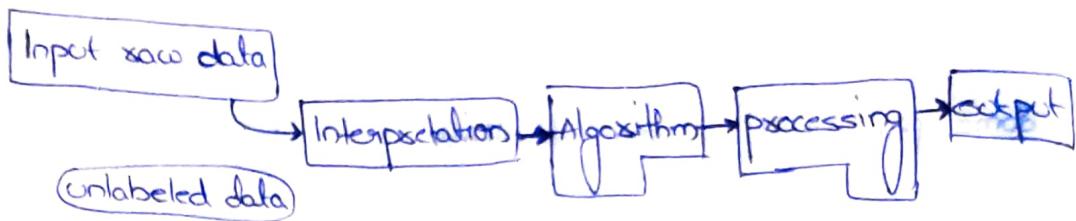
- Linear Regression
- Regression trees
- Non-Linear Regression
- Bayesian Linear Regression
- Polynomial Regression

Classification algorithms are used when the output variable is categorical.

- Random Forest
- Decision trees
- Logistic Regression
- Support vector machines

Unsupervised Learning

It works on the unlabelled data in which models are not supervised using training dataset. Model itself find the hidden patterns from the given data.



unsupervised learning can be classified into two types

- clustering
- Association

clustering :- It is method of grouping the object into clusters such that object with most similarities remain into a same group

Association :- used to find the relationship b/w variable in the large data set

Dimensionality reduction :- It is the transformation of data from higher - dimension to lower dimension to retain the meaningful properties of the original data.

Unsupervised M.L Algorithms

K-Means clustering	Neural Network	Singular value decomposition
KNN (k-nearest neighbour)	PCA (Principal Component Analysis)	
Hierarchical clustering	Independent component analysis	
Anomaly detection	Apriori Algorithm	

Regression :-

It is an stats model to find relationship b/w dependent (target) and Independent (features, predictors) with one or more independent variables.

Regression helps us to find the correlation b/w variables and enable us to predict the continuous output variable based on one or more predictor variables.

It is mainly used for prediction, forecasting, time series modeling and determining the causal-effect relationship b/w variables.

Terminologies Related to the Regression Analysis

Dependent :- Target

Independent :- Features

Outlier :- observation which contains very low value or very high value in observed values. Outliers should be avoided because these impact the results accuracy.

Multicollinearity :- If the independent variables are highly correlated with each other than other variables, then such condition is called multicollinearity.

It is should not present in the dataset.

Underfitting and Overfitting

low bias ; high variance

If algorithm works well with the training dataset but not well with test dataset then such problem is called overfitting.

If algorithm does not work performing well even with the training dataset then it is called under fitting.

high bias, low variance

Type of Regression

- Linear regression
- Logistic regression
- Polynomial regression
- Support vector regression
- Decision tree regression
- Random Forest regression
- Ridge Regression
- Lasso Regression

Type of classification

- Linear models
 - logistic
 - SVM
- Non-Linear models
 - K-nearest Neighbours
 - SVM Kernel
 - Naive Bayes
 - Decision tree classification
 - Random Forest classification

Linear Regression

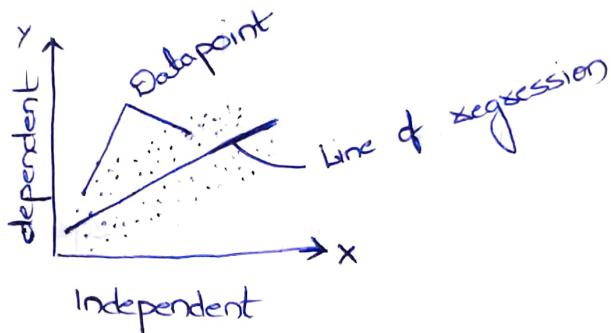
easiest and

it is one of the most popular ML model. It is an statistical method that is used to predictive analysis. It makes predictions for continuous/scale (as) numerical variables.

Ex:- sales; age; salary; product price etc.

Linear regression shows relationship b/w dependent (y) and independent (x) variables i.e how the value of the dependent variable

is changing according to the independent variable.



$$Y = a_0 + a_1 x + \epsilon$$

Y = Dependent variable (target variable)

X = Independent variable (features, predictor variable)

a_0 = Intercept of the line (An additional degree of freedom)

a_1 = Linear regression coefficient (scale factor to each input value)

ϵ = Random error

Type of Linear Regression

- Simple
- Multiple

} Linear Regression

Simple Linear Regression :-

If a single independent variable is used to predict the value of a numerical dependent variable.

Multiple Linear Regression :-

If more than one independent variable is used to predict the value of a numerical dependent variable.

Regression Line :-

A linear line showing the relationship b/w the dependent and independent variable is called regression line

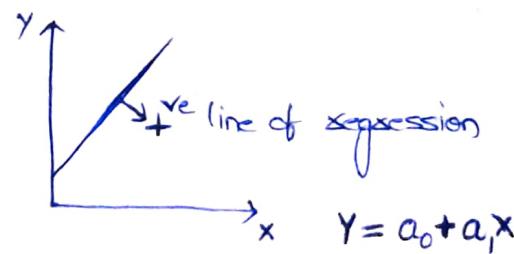
These are two types of regression line relationship

- positive linear relationship

- negative linear relationship

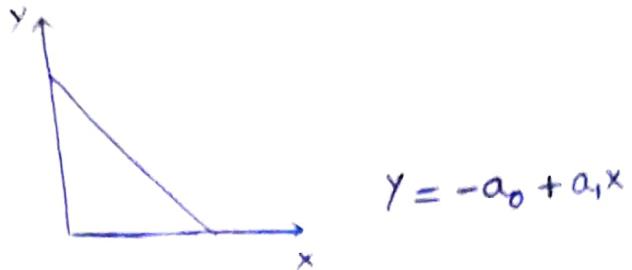
+ve linear relationship

If the dependent variable increases on the y-axis and the independent variable increases on x-axis



-ve linear relationship

If the dependent variable decreases on the y-axis and independent variable increases on the x-axis



$$y = -a_0 + a_1x$$

When we are working with the linear regression our main goal is to find the best fit line that means that errors b/w predicted values and actual values should be minimized.

The best fit line will have the least error.

Cost function

The cost function is the avg error of n-samples in the data (for the whole training data)

Loss function

The loss function is the error for individual data point
 (for one training example)

The cost function of linear regression is mean squared error (or) root mean squared error.

MSE is the avg of squared errors occurred b/w the Predicted values and actual values

Residuals : the distance b/w the actual values and predicted value is called Residual. If the observed point are far from the Regression line then the residual will be high so that the Cost function will be high. If the scatter points are close to the regression then residual and cost function is low.

Gradient Descent

It is an optimization algorithm used to find the values of parameters (coefficients) of a function (f) that minimizes a cost function (cost)

It is used when the parameters cannot be calculated analytically (e.g. using linear algebra) and must be searched for by an optimization algorithm

$$R^2 \text{ squared} = \frac{\text{Explained variation}}{\text{total variation}}$$

Assumptions of Linear Regression

- Linear relationship b/w features and target
- Small (as) no multicollinearity b/w the features
- Homoscedasticity (a) Homogeneity refers to a condition in which the variance of the residual is same for any value of X
- The linear regression model assumes no autocorrelation in error terms.
- Linear regression assume that the error terms should follow the normal distribution patterns. If the error terms are not normally distributed then the confidence interval will become either too wide (as) too narrow which is difficult to find coefficients

Simple linear regression

It is a type of regression algorithm that models the relationship b/w a dependent variable and a single independent variable.

The relationship shown by a simple linear Regression model is linear or a sloped straight line hence it is called simple linear regression

The key point in simple linear regression is that the dependent variable must be a ~~scal~~ continuous value. However the independent variable can be measured on continuous (or) categorical values

S.L.R has mainly two objectives

→ Relationship b/w the two variables

* Relation b/w Income and expenditure, experience and salary etc

→ Forecasting New observations

* Weather forecasting according to temperature

* Revenue of a company according to the investment in a year etc.

$$Y = a_0 + a_1 x + \epsilon$$

a_0 = it is the Intercept of the Regression line (can be obtained putting $x=0$)

a_1 = it is the slope of the regression line, which tells whether the line is increasing (or) decreasing.

ϵ = the Error term. (For good model it will be negligible)

Multiple Linear Regression

In multiple linear regression is one of the most important algorithm which models the linear relationship b/w a single dependent continuous variable and more than one independent variable.

- For MLR (y) must be continuous / Real but (x) may be categorical (as) continuous
- Each feature must be linear relationship with the dependent variable
- MLR tries to fit a regression line through a multidimensional space of data points.

$Y = \text{output} / \text{Response variable}$

► Assumptions for MLR

- A linear relationship should exist b/w the target and predictor variables
- Regression residuals must be normally distributed
- MLR assumes little (as) multicollinearity in the data.



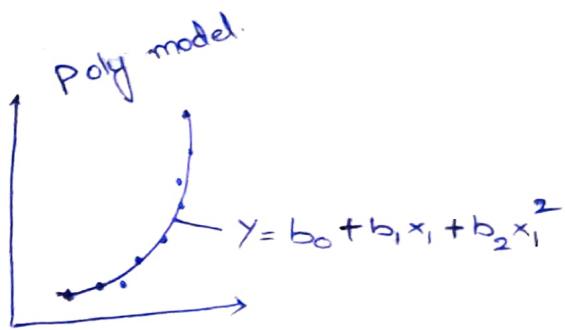
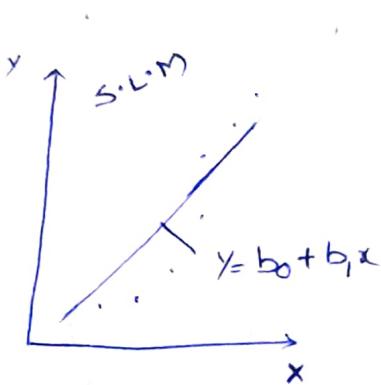
(Correlation b/w Independent variable)

ML polynomial Regression

It is the relationship b/w a dependent (y) and Independent Variable (x) at nth degreee polynomial.

The polynomial Regression equation is given below

$$Y = b_0 + b_1 x_1 + b_2 x_1^2 + b_3 x_1^3 + \dots + b_n x_1^n$$



It is a special case of Multiple Linear Regression. Because we add some polynomial features to the M.L.R to convert into Polynomial Regression

It is also linear model with some modification in order to increase accuracy

The data set used into polynomial regression for training is non-linear nature

We use polynomial regression when the linear Regression is unable to capture the points.

so this leads to underfit to over come these we increase the complexity of model by increasing the higher order (or) dimensions.

In polynomial regression the original features are converted into

Polynomial features of required degree (2, 3, ..., n) and then modeled using a linear model

Simple Linear Regression Equation $y = b_0 + b_1 x$

Multiple Linear Regression Equation $y = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_n x_n$

Polynomial Regression Equation $y = b_0 + b_1 x + b_2 x^2 + b_3 x^3 + \dots + b_n x^n$

Backward Elimination

Backward elimination is a feature selection technique while building ML model. It is used to remove those feature that do not have a significant effect on the dependent variable or prediction of output.

These are various ways to build a model in ML

All-in

Backward elimination

Forward selection

Bidirectional elimination

F-Score comparison

Suppose we have included all the independent variables and don't know which independent model is most affecting and which one is the least affecting for the prediction.

Unnecessary features increase the complexity of the model.

In order to optimize the performance of the model, so we will use the Backward elimination method. This process is used to ~~remove~~ the feature which is least affecting feature.

Forward selection:-

It is an iterative method in which we start with having no feature in the model. In each iteration we keep adding the feature which best improves our model till an addition of a new variable does not improve the performance of the model.

Once the model no longer improves with adding more variables, then the process stops.

Bidirectional elimination:- It is the combination of both forward selection and backward elimination testing each step for variables to be included or excluded.

Lasso & Ridge Regression & elastic Regression

Lasso shrinkage	comes towards mean	one should not increase
-----------------	--------------------	-------------------------

Lasso (least absolute shrinkage and selection operator)

Lasso & Ridge helps us to overcome the overfitting model.

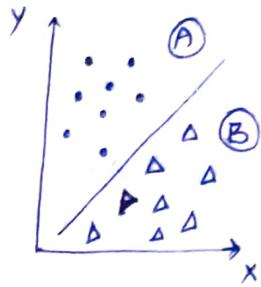
Lasso Regression which penalizes the sum of absolute value of the coefficients.

Ridge Regression which penalizes sum of squared coefficients (multicollinearity)

Classification Algorithm

It is a supervised M.L Algorithm that is used to identify the category of new observations on the basis of training data.

classes can be called as targets / labels (or) categories



Binary classifier :- If the classification problem has only two possible outcomes then it is called Binary classifier.

Ex :- Yes (or) no, female(or) male, spam(or) Not Spam

Multi class classifier :- It is also classification problem which has more than two classes.

Ex types of Animals , types of Crops

Learners in classification problem

These are two types of learners

1) Lazy learners

2) Eager learners

Lazy learners :- Firstly stores the training dataset and wait until it receives the test dataset. In lazy learners classification is done on the most relevant data stored in training dataset.
• less time for training but more time for the prediction.

Lazy learning examples :- K-NN algorithm

Eager learning :- It develops a classification model based on a training dataset before receiving a test dataset. Eager learning takes more time in learning and less time in prediction.

Example :- Decision trees, Naive Bayes

Logistic Regression in machine learning

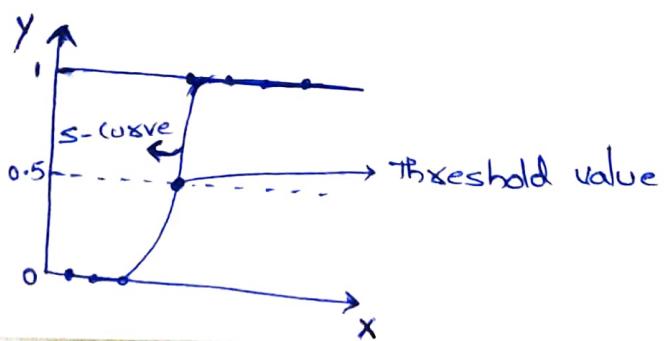
It is used to predict the categorical dependent variable using a given set of independent variables.

It predicts the output variables are categorical in the dependent variables.

It gives probabilistic values which lie b/w 0 to 1

Logistic Regression is similar to Linear Regression. But Linear Regression is used for solving regression problems and Logistic Regression is used for solving classification problems.

Instead of fitting a best fit line we use S-shaped logistic function. which predict max values (0 and 1)



Logistic function (Sigmoid function)

It is math function used to map the predicted values to the Probabilities

It maps any real value to another value within range 0 and 1.

The value of logistic Regression must be b/w 0 and 1

which cannot be beyond the limit so this form a S shaped curve.

It is called sigmoid function (or) logistic function

In Logistic Regression we use concept of threshold value.

The dependent values must be in categorical in nature.

The independent values ~~---~~ should not have any multi-collinearity.

Types of Logistic Regression

Binomial \rightarrow only two possible types

Multinomial \rightarrow There are 3 (or) more possible unordered types

Ordinal \rightarrow There are 3 (or) more possible ordered types

Naive Bayes classifier

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Naive Bayes is a supervised learning algorithm which is based on Bayes theorem and used for solving classification problems. It is mainly used in text classification that includes a high-dimensional training dataset.

It is one of the simple and most effective classification algorithms, which helps in building the fast ML model and make quick predictions.

Naive Bayes algorithm is composed of two words Naive and Bayes.

Naive → It assumes that each input variable is independent.

Bayes → It is because of depends on the principle of Bayes' theorem.

Working of Naive Bayes Algorithm

- Convert the given dataset into frequency table
- Generate the likelihood table by finding the probabilities of the given features
- Now use the Bayes' theorem to calculate the posterior probability

Type of Naive Baye's model

The are three types of Naive Baye's model

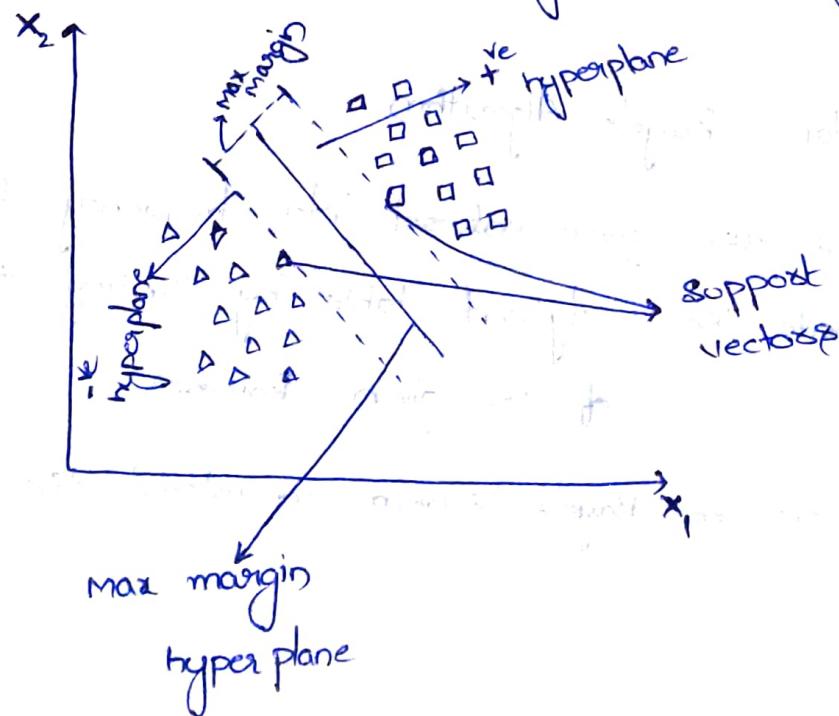
- Gaussian \rightarrow feature follows a normal distribution.
- Multinomial \rightarrow When the data is multinomial distributed. (Doc)
- Bernoulli \rightarrow It is similar to multinomial classifier. but the predictor variables are independent Boolean variables.

Support vector machine Algorithm

SVM is the one of the most popular supervised learning algorithm

which is used for classification as well as regression problems.

The goal of SVM algorithm is to create the best line (or) decision boundary that can segregate n-dimensional space into classes so that we can put the new data easily in the correct category in the feature. The best decision boundary is called hyperplane



SVM can be classified into two types

Linear SVM

Non-linear SVM

Linear SVM :- If the data can be easily or linearly separated by a single straight line can be divided into classes.

Non-linear SVM :- If the cannot be easily classified by using a straight line then it is known as Non-linear SVM.

* Hyperplane / Decision boundary

The max distance b/w datapoints that helps to divide the data points.

* Support vectors

Support vector is the datapoint which are closest to the hyperplane since the vector support the hyperplane hence these is known as support vectors

$$f(x) > +1 \rightarrow +^{\text{ve}} \text{ region}$$

$$f(x) < -1 \rightarrow -^{\text{ve}} \text{ region}$$

Type of margins

Soft margin ; Hard margin



It allows the
data points in b/w
the margin



It doesn't allow the
datapoint in b/w
the margin

K-Nearest Neighbor (KNN) Algorithm

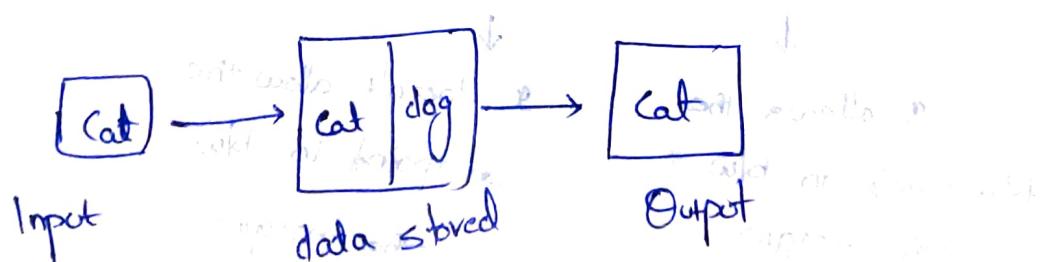
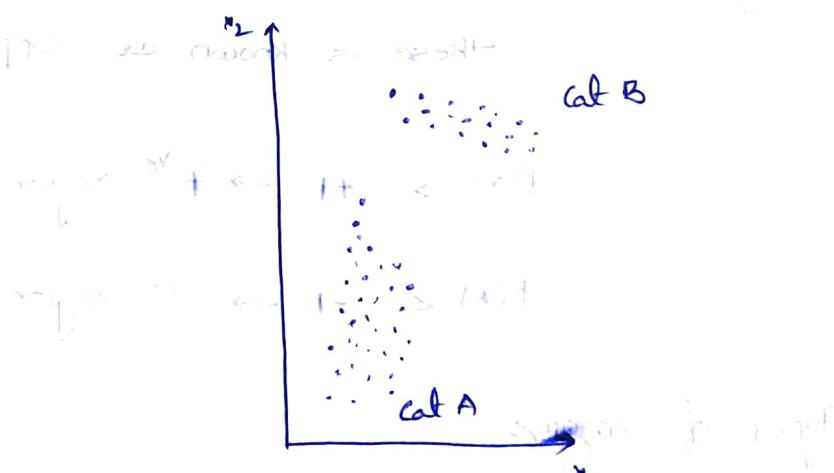
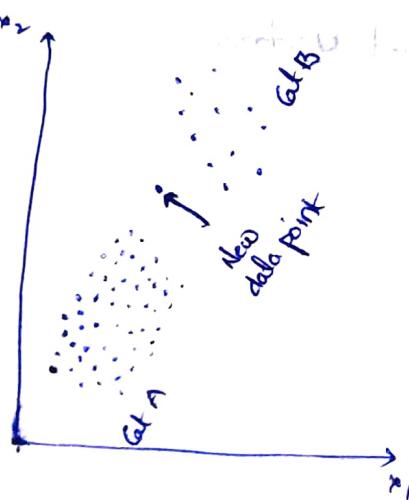
K-NN is a non-parametric and lazy learning algorithm. In K-NN there is no assumption for underlying data distribution.

Lazy learning means it does not need any training data point for model generation. All training is done in the testing phase which makes training faster and testing phase slower and costlier.

K-NN is a simple algorithm that stores all the available cases and classifier the new data (or) case based on a similarity measure.

Based on the nearest neighbor the new data point is classified if $K=1$ i.e. positive integer then is subjected to some class

K is an hyperparameter in machine learning



Decision Trees

It is a supervised learning technique that can be used for both classification and Regression.

It is an tree structured classifier

In Decision Tree there are two nodes

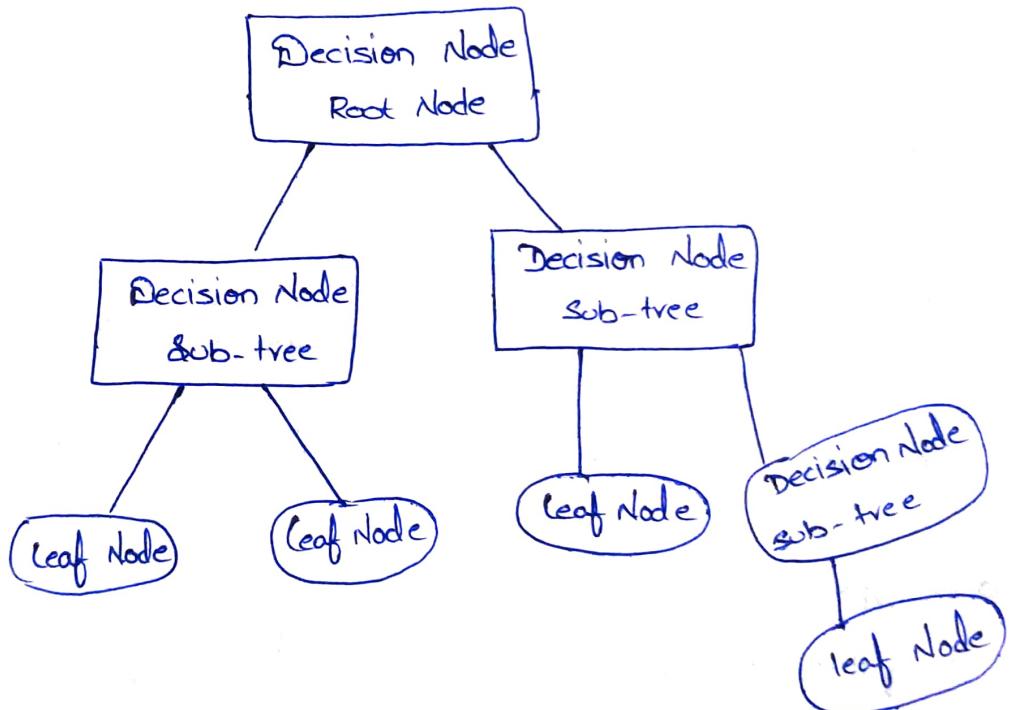
- Decision Node
- Leaf Node

Decision Nodes are used to make any decision and have multiple decision branches

Leaf Nodes are the output of those decisions and do not contain any further branches

Further branches

In order to build a tree we use the CART Algorithm i.e Classification and Regression tree Algorithm



Decision trees usually mimic human thinking ability while making a decision, so it is easy to understand

The logic behind the decision tree can be easily understood because it shows tree-like structure.

Root Node :- It is from where the decision tree starts. It represents the entire dataset which further gets divided into two more homogeneous sets.

Leaf Node :- Leaf nodes are the final output node and the tree cannot be separated.

Splitting :- It is the process of dividing root / decision node into

subnodes according to the given conditions (a) conditions

Branch / sub Tree :- A tree formed by splitting the tree

Pruning :- It is the process of removing the unwanted branches from the tree

Parent / child node :- The root node of the tree is called parent

and other nodes are called child node

Advantages

- Simple to understand
- Little effort required for data preparation
- Can handle both numerical and categorical data
- Non-linear parameter does not effect the performance

Disadvantages

overfitting occurs when the algorithms capture noise data

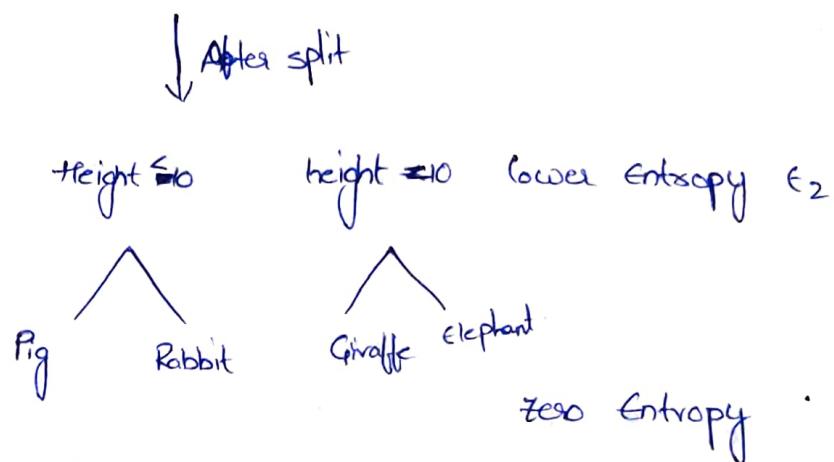
High variance model can be unstable due to variation in data

low bias it is difficult for the model to work with the

New data when it has highly complicated tree

Entropy :- It is the Randomness (or) unpredictability in the dataset

Group of Animals (High Entropy) E_1



Information Gain

It is the measure of decrease in Entropy after the Dataset is split

$$\text{Gain} = E_1 - E_2$$

Leaf Node carries the classification (or) the Decision

How to split the data:-

- Information Gain must be high

Gini Index, gini Impurity

It is a standard decision tree splitting metric

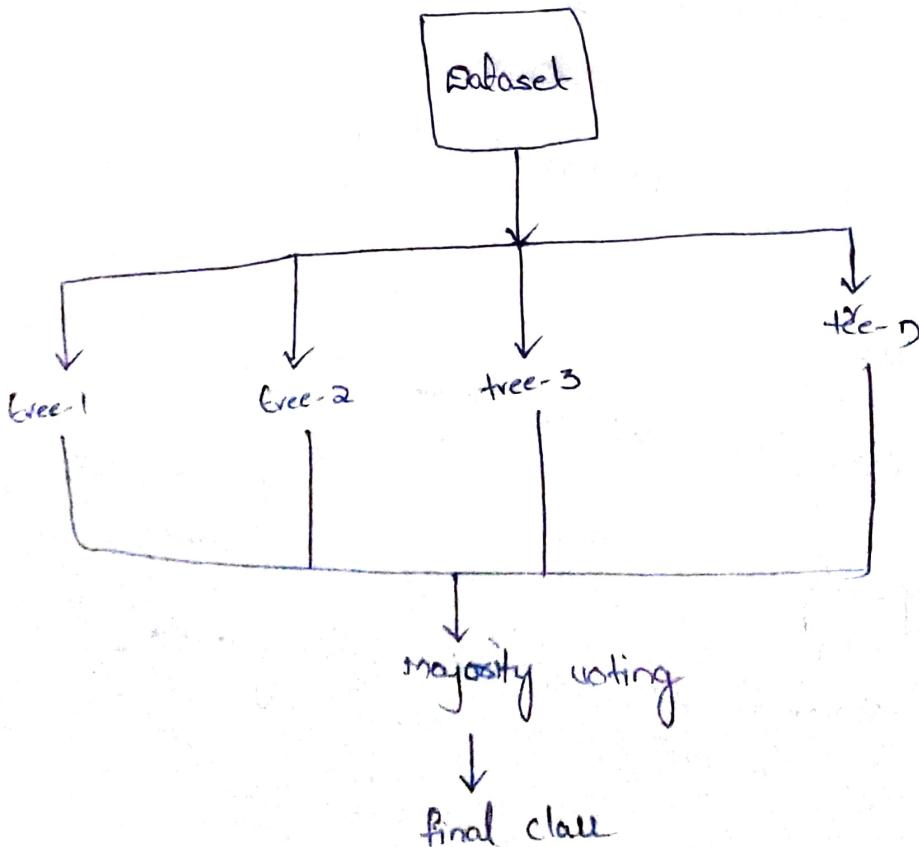
Random forest

It is an popular ML Algorithm that belongs to supervised learning technique. It is used for both classification and Regression.

It is derived from Decision tree to overcome the overfitting problem.

Random forest contains a number of decision trees. Various subsets of the given dataset and takes the average to improve the predictive accuracy of the data.

The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.



What is Bagging and Boosting

Bagging :- Bagging is the way to decrease the variance in the prediction by generating addition data for training from dataset using combination with repetitions to produce the multiset of the original data

Boosting :- It is an iterative technique which adjusts the weight of an observation based on the last classification

Difference b/w Boosting and Bagging

Bagging is a method for merging the same type of prediction

decrease ~~bias~~ variance, Not bias and solve over-fitting issues in a model

The weights are equal for each model

It is used for unstable with high variance.

Model built independently

Boosting is a method of merging different types of prediction

Boosting decreases Bias not variance

Here the weights are weighed based on their performance

It is used high bias and stable

models are affected by ~~the~~ previous built model.

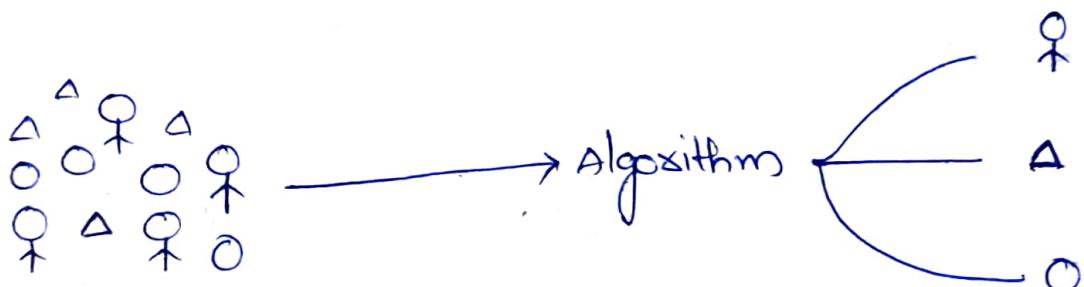
Clustering

* is an unsupervised M.L ~~for~~ technique when data is unlabelled.

A way of grouping the data points into different clusters, consisting of similar data points. The object with possible similarity remains in a group that has less (or) no similarity with another group.

It is done by identifying some similar patterns in the unlabelled dataset such as size, shape, color, behavior etc.

The clustering technique is commonly used in the statistical data analysis.



Type of clustering

- Exclusive
- Overlapping
- Hierarchical

Exclusive clustering :- (hard clustering)
Data point / item belong exclusively to one cluster

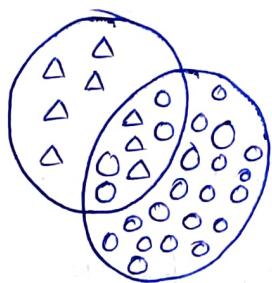
For example k-means clustering



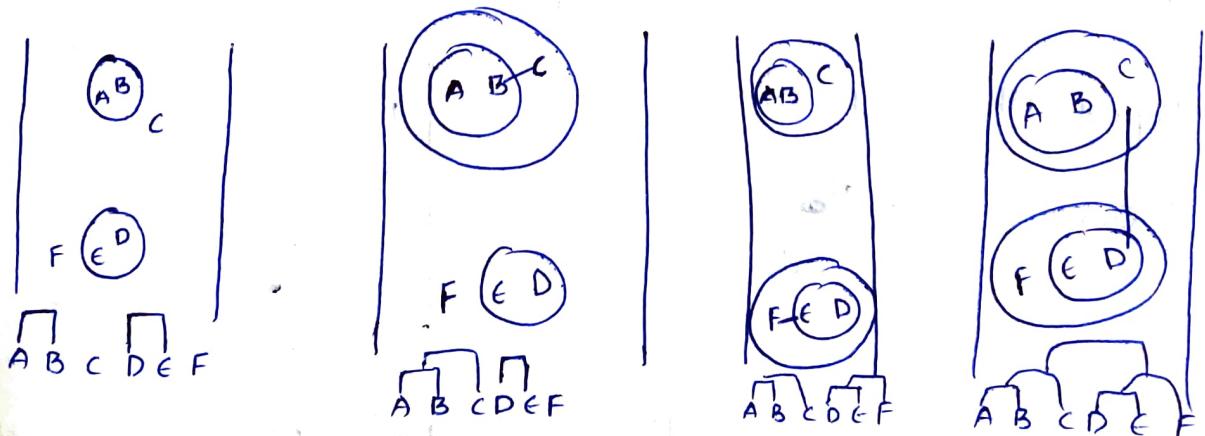
Overlap clustering :- (soft clustering)

Data point/item belong to multiple clusters

for example Fuzzy c-means clustering



Hierarchical clustering

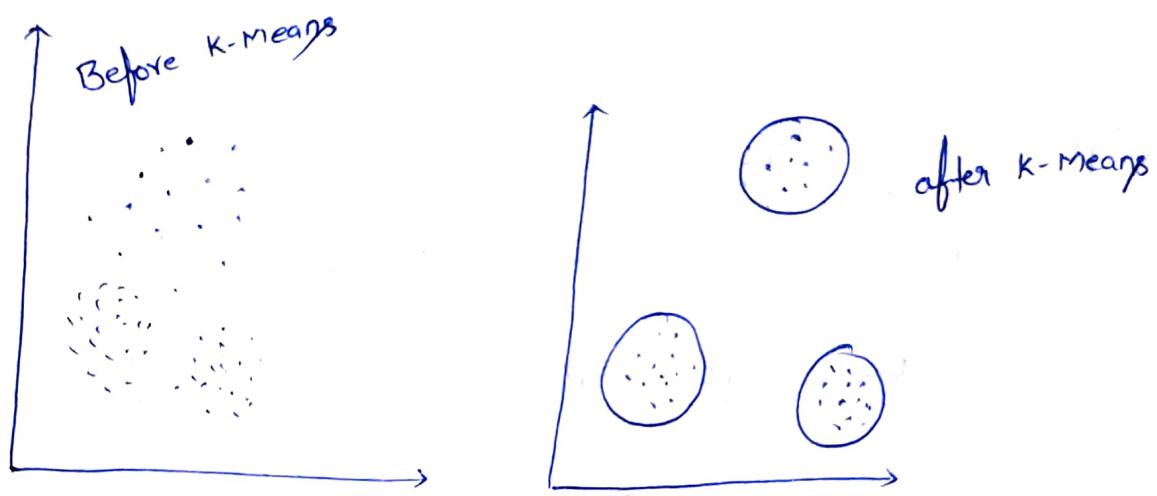


Ex Dendrogram

K-Means clustering

The main goal is to group similarities of elements (or data points) into a cluster.

The number of clusters (k) groups are represented by \textcircled{k}



steps

- ① select Number of clusters
- ② select Random Centroids for \textcircled{k}
- ③ Assign each data point close to the centroids
- ④ Calculate the variance and place a New centroid the cluster
- ⑤ Repeat the ③ point
- ⑥ Model is Ready



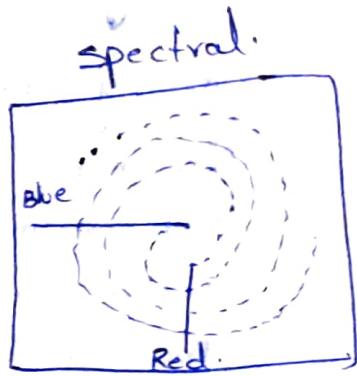
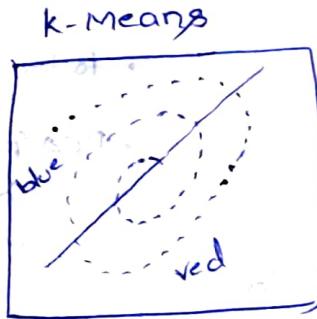
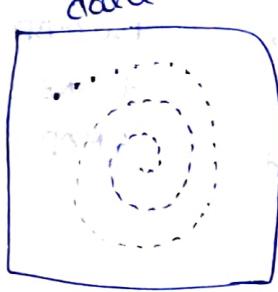
→ By using the elbow method we choose no of clusters

Spectral clustering

spectral clustering is a technique with roots in graph theory, where the approach is used to identify communities of nodes in a graph based on the edges connecting them.

The 3 major steps involved in spectral clustering

- Constructing a similarity graph
- Projecting the data into lower dimensional space
- clustering the data.



DBSCAN (Density Based spatial clustering of Application with noise)

This clustering is done by separate clusters of high density from clusters of low density

Principle of DBSCAN is to find the neighborhoods of data points exceeds certain density threshold

Dimensionality Reduction

The number of input features, variables or columns present in a given dataset is known as Dimensionality and the process of reduce these features is called Dimensionality reduction.

It is a way of converting the higher dimensional dataset into lesser dimensional dataset ensuring that it provides similar information.

Dimensionality Reduction Techniques

Feature selection



missing value ratio

low variance filters

high correlation filters

Random forest

Backward feature selection

Forward feature selection

Dimensionality reduction



- Factor analysis ISO MAP
- PCA (Principal Component Analysis) t-SNE UMAP
- Independent Component Analysis

Approaches of Dimension Reduction

Feature selection

i) filter methods

- Correlation
- Chi-square test
- Anova
- Information Gain etc

ii) wrapper methods

- forward selection
- Backward selection
- Bi-directional elimination

iii) Embedded methods

- LASSO
- Elastic Net
- Ridge Regression etc

Feature Extraction

• PCA (Principal Component Analysis)

• Linear Discriminant Analysis

• Kernel PCA

• Quadratic Discriminant Analysis

Common techniques of Dimensionality Reduction

• PCA [Principal Component Analysis]

• Backward elimination

• forward selection

• Score comparison

• missing value Ratio

• low variance filter

• high correlation filter

• Random Forest

• Factor Analysis

• Auto - Encoder



- Handling the high-dimensional data is very difficult in practice, commonly known as the curse of dimensionality
- If the dimensionality of the input data increases any ML model becomes complex
- If the number of features increases then there is a chance of overfitting also increases

Filter method :- The dataset is filtered and a subset that contains only the relevant features is taken.

Wrappers method :- The wrapper method has the same goal as the filter method but in these methods some features are fed to the model and evaluate the performance. This performance decides whether to add features or to remove the features to increase the accuracy of the model.

This method is more accurate than filter method.

Embedded method :- They check the different training iterations of the machine learning model and evaluate the importance of each feature.

PCA (Principal Component Analysis)

It is a statistical process that converts the observations of correlated features into a set of linearly uncorrelated features with the help of orthogonal transformation. The new transformation features called principal components. It is one of the popular tools that is used for EDA (Exploratory Data Analysis) and predictive modeling.

PCA works by considering the variance of each attribute because the high attribute shows the good split b/w the classes and hence it reduces the dimensionality.

Missing value Ratio :- If the dataset has too many missing values then we drop those variables as they do not carry much useful information. To drop (or) perform this we can set a threshold level, and if a missing value more than the threshold we will drop that variable. The higher the threshold value the more efficient the reduction. variable has

Encoder :- The function encoder is to compress the input to form the latent space representation.

Decoder :- The function decoder is to recreate the output from the latent space representation.

Factor Analysis :- It is a technique in which each variable is kept within a group according to the correlation with other variables.

Variables within a group can have a high correlation b/w themselves but they have a low correlation with variables of other groups.

Confusion Matrix :-

It is a matrix used to determine the performance of the classification model for a given set of test data. It can be determined if the true values for test data are known.

The errors in the model performance in the form of a matrix is also known as Error matrix.

The matrix is divided into two dimensions

- predicted value
- actual value

n = total predictions

Actual : No

Actual : Yes

Predicted : No

True Negative

False Positive

Predicted : Yes

False Negative

True Positive

True Negative :- Model has given prediction No, and the real (or) actual value was also No

True positive :- Model has prediction Yes, and the actual was also True

False positive :- Model has predicted No, but the actual value was Yes. It is called type-II error

False Negative :- Model has predicted Yes, but the actual value was No. It is called type-I error.

Classification Accuracy :- one of the important parameter to determine the accuracy of the classification problem. It defines how often the model predict the correct output.

It is defined by calculating the ratio of the number of correct predictions made by the classifier to all number of predictions made by the classifier.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$$

misclassification Rate :- It is also termed as error rate. It defines how the model gives the wrong predictions. The value of error rate can be calculated as number of incorrect predictions to all the number of the predictions made by the classifier.

$$\text{Error rate} = \frac{FP + FN}{TP + FP + FN + TN}$$

Precision :- It is defined as the number of correct outputs provided by the model (as) out of all positive classes that have predicted correctly by the model.

$$\text{Precision} = \frac{TP}{TP + FN}$$

Recall :- It is defined as the out of total positive class how our model predicted correctly

The recall must be high as possible

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

F-Measure :- If two model have low precision and high recall (or) vice versa it is difficult to compare these model. F-score . This score help us to evaluate the recall and precision at the same time

F-score is max if the recall is equal to the precision.

$$F\text{-Measure} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}$$

Null Error rate :- It defines how often our model would be incorrect if it always predicted the majority class.

As per the accuracy paradox, it is said that

"The best classifier has a higher error rate than the null error."

Roc Curve :- The Roc is a graph displaying a classifier's performance for all possible thresholds. The graph is plotted b/w the true positive rate (on the y-axis) and the false positive rate (on the x-axis)

ROC
AUC

Cross-validation in ML

Cross-validation is a technique for validating the model efficiency by training it on the subset of input data and testing on previously unseen subset of the input data.

We can also say that it is a technique to check how a statistical model generalizes to an independent dataset.

Common methods are used for cross-validation

- 1) validation set approach
- 2) leave-p out cross-validation
- 3) leave one out cross-validation
- 4) k-fold cross-validation
- 5) stratified k-fold cross-validation

Validation set approach

We divide the input dataset into a training set and test set. Validation set in the validation set approach. Both the subsets are given 50% of the dataset.

One major disadvantage is we are using 50% for training so this may lead to miss some of the important information.

K-fold Cross validation

That divides the dataset into k groups of samples of equal size.
These samples are called folds. For each learning set the prediction function uses $k-1$ folds and the rest of the folds are used for the test set.

This approach is a very popular CV approach because it is easy to understand and the output is less biased than other methods.

Train / test split :- The input data is divided into two parts that are training set and testing set, on a ratio of

70 : 30, 80 : 20

It provides a high variance, which is one of the biggest disadvantages

Train data :- data which is used for training the model and this is known as train data

Test data :- Test data is used to make the prediction from the model that is already trained.

Cross-validation dataset :- It is used to overcome the disadvantage of (train) test split by splitting the data set into groups of (train) test split and average the result

overfitting & underfitting

overfitting :- Performs well on training data but underperforms on test data

underfitting :- Performing poorly over the train and test set

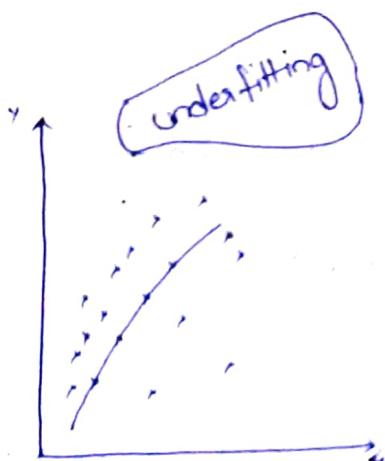
Bias and variance

Bias - Difference prediction by model and the correct value

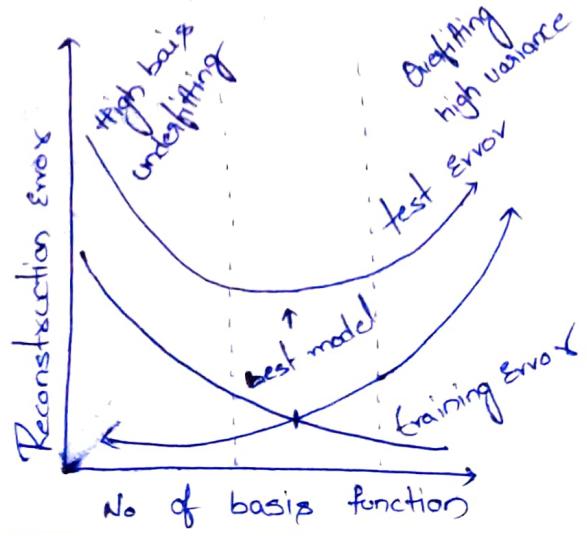
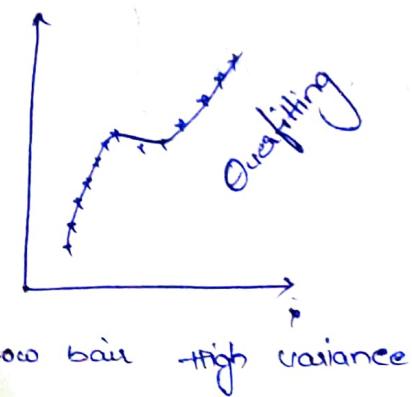
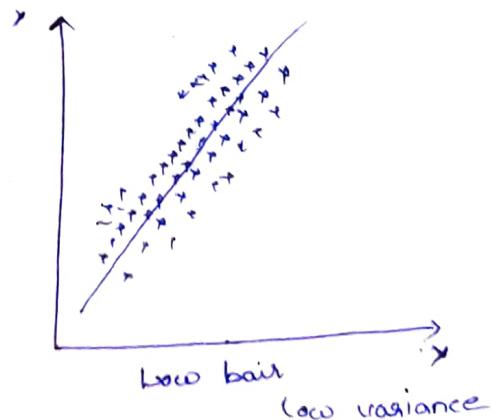
high bias :- underfitting , straight line prediction

Variance :- spread of our data

high variance :- overfitting , fitting accuracy



high bias high variance



R-square vs Adjusted R-squared

R^2 :- Assumes that every single variable explain the variations in the dependent variable

Adj R^2 :- It tells you the percentage of variation explained by only the independent variables that actually affect the dependent variable.

Accuracy vs precision vs Recall

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

$\left\{ \frac{\text{Correctly predicted}}{\text{Total observation}} \right\}$

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

$\left\{ \frac{\text{Correctly predicted } +^ve \text{ observation}}{\text{Total } +^ve \text{ observation}} \right\}$

Recall

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

$\left\{ \frac{\text{Correctly predicted } +^ve \text{ observation}}{\text{Total observation in a class}} \right\}$

Cost Function vs loss function

Loss function : Error for a single training Example

Cost function :- average Errors/ loss over the entire training data set

Covariance vs Correlation

Covariance :- indicates the direction of the linear relationship

Correlation :- explains both strength and direction or how much proportion change in one variable leads change in second variable