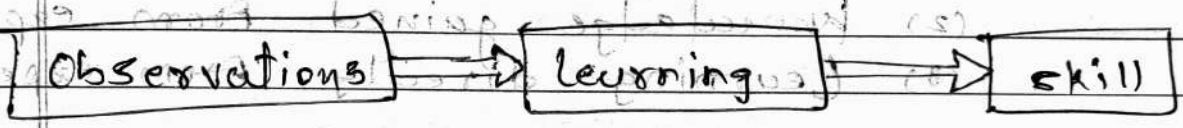
Page No.
Date

⊛ Chapter : ① : Introduction To M.Learning ⊛

Q.① What is Human Learning ?

- Human learning is observing something, identifying a pattern, building a theory to explain this pattern and testing this theory to check if its fits in most or all observations.

②



③ Human learning ⊛

- Human learning process varies from person to person to person. Once a learning process is set into the minds of people, it is difficult to change it.

- Both Human learning & machine learning generate knowledge, one residing in the set into the minds of people, it is difficult to change it.

④ Human : Learning :: Machine Learning

Intelligence ⟺ Models

Learning Materials ⟹ Data

Learning skills ⟹ skill learn { learning by heart / learning by ignoring

Q.① ~~What is Machine Learning Briefly explain Types of Machine learning.~~ ⊗

Q.② Types of Human Learning & Difference Between ML & HL.

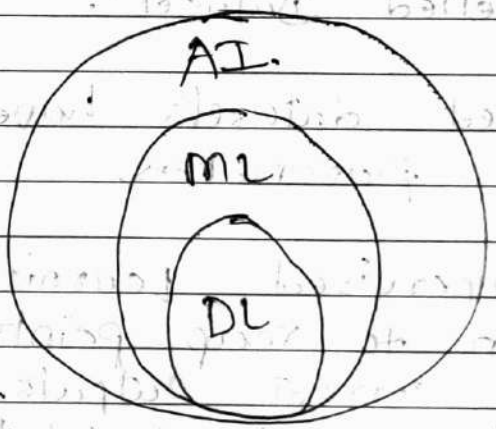
① There are three types of Human Learning:

(1) self-learning
(2) knowledge gained from expert
(3) Learning directly from expert

Human Learning	Machine Learning
(1) Humans acquire knowledge through experience either directly or shared by others.	Machines acquir knowledge through experience shared in the form of past data.
(2) Model-free & model-based mechanism can be found in HL.	knowledge based learning in ML.
process: observations → Learning → skill	process: Data → Machine learning → skill

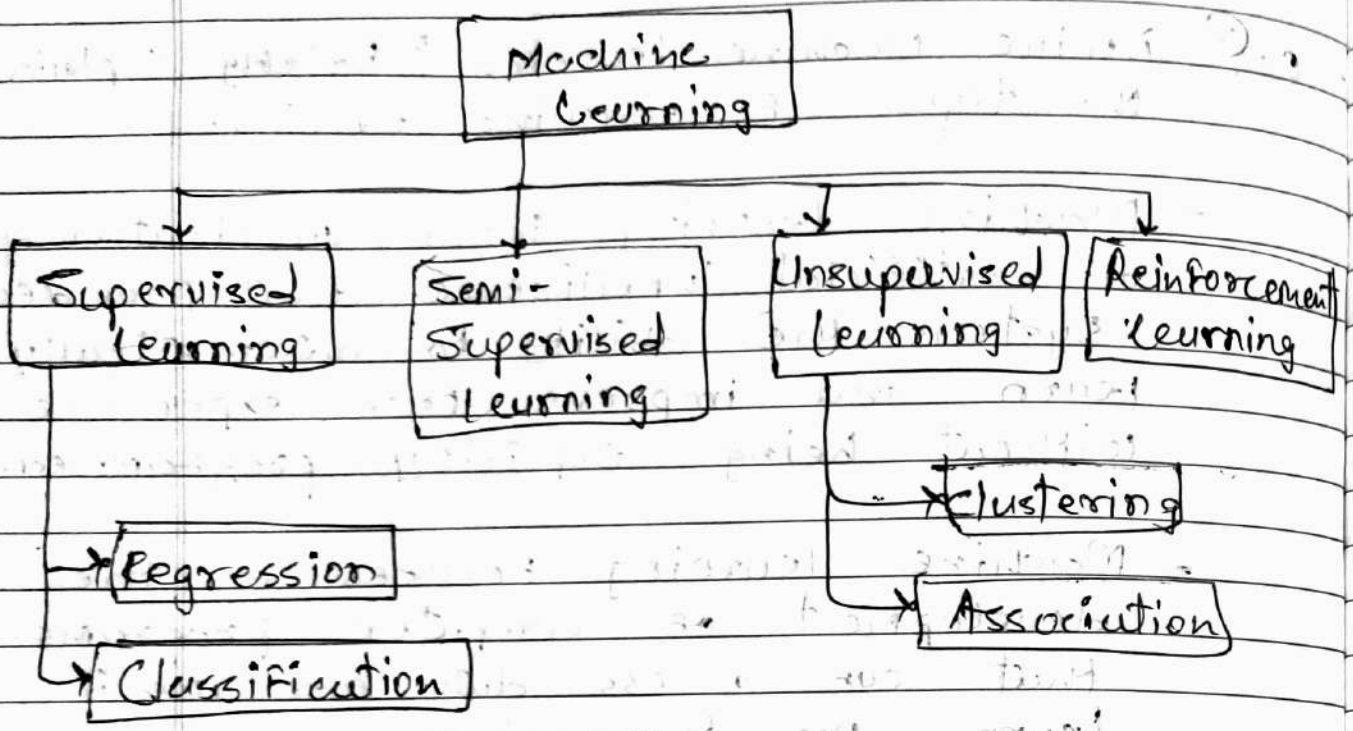
Q.③ Define Machine learning ? Briefly Explain the types of machine learning.

- Machine learning is an application of artificial intelligence that provides systems the ability to automatically learn and improve from experience without being explicitly programmed.
- Machine learning focuses on the development of computer programs that can access data and use it learn for themselves.
- The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly.



⊛ Types of Machine Learning :-

(1) Supervised Machine Learning
(2) Unsupervised Machine Learning
(3) Reinforcement Learning
(4) Evolutionary Semi-Supervised learning



[1] Supervised learning :-

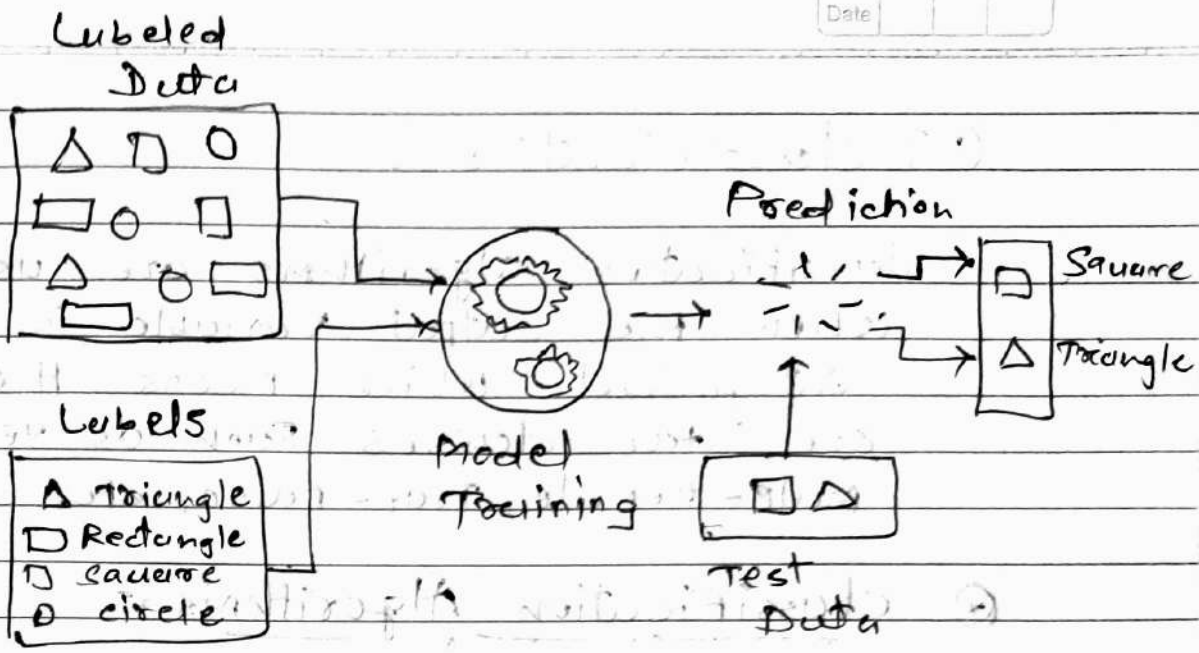
- Supervised learning is defined as when a model gets trained on a "labelled Dataset".

- Labelled datasets have both input & output parameters.

- In Supervised learning algorithms learn to map points between inputs and correct outputs. It has both training and validation datasets : labelled.

* There are two types of SL :

(1) Regression
(2) classification



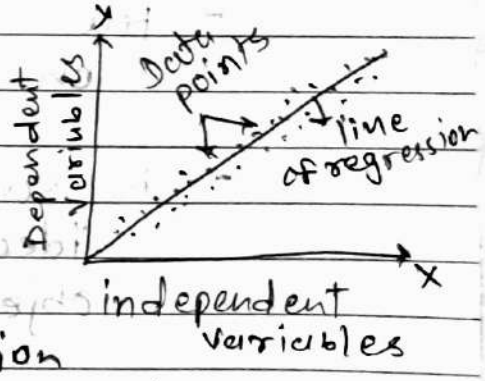
④ Supervised Machine learning ③

[1] Regression :-

- Regression algorithms are used if there is a relationship between the input & output variable.

S.L
Regression
classification

- it is used for prediction of continuous variables, such as weather forecasting, Market Trends, etc.



⑧ Regression Algorithms :-

- Lineur Regression
- Regression Trees
- Non-Lineur Regression
- Bayesian Lineur Regression
- polynomial Regression

(2] Classification :-

- classification algorithms are used when the output variable is categorical, which means there are two classes such as 'Yes-No, male-female, True-False, etc.

④ Classification Algorithms :-

- Spam Filtering
- Random Forest
- Decision Trees
- Logistic Regression
- Support Vector Machines

⑤ [3.] Unsupervised learning :-

⑥ Advantages of Supervised learning :-

- the model can predict the output on the basis of prior experiences.
- in supervised learning, we can exact idea about the classes of objects.
- SL models helps us to solve various real-world problems such as fraud detection, spam filtering, etc.

Page No.
Date

⊛ Disadvantages of supervised learning :-

- SL models are not suitable for handling the complex tasks.
- SL cannot predict the correct if the test data is different from the training dataset.
- Training require lots of computation times.
- in SL, we need enough knowledge about the classes of object.

[2] Unsupervised Learning :-

- In contrast, unsupervised machine learning algorithms are used when the information used to train is neither classified nor labeled.
- Unsupervised learning studies how systems can infer a function to describe a hidden structure from unlabeled data.
- The system doesn't figure out the right output, but it explores the data and can draw inferences from datasets to describe hidden structures from unlabeled data.

in easy language

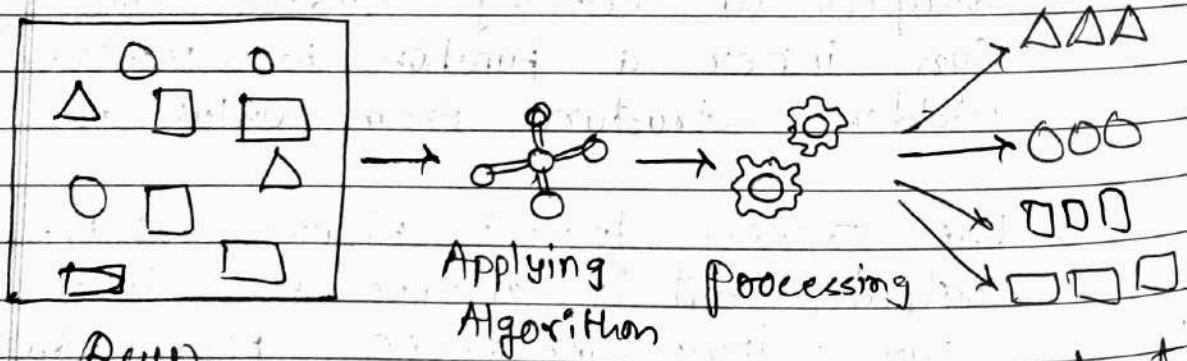
- Unsupervised learning is a type of machine learning technique in which an algorithm discovers patterns and relationships using unlabeled data.

- Unlike SL, UL doesn't involve providing the algorithm with labeled ~~data~~ target outputs.

- The primary goal of UL is often to discover hidden patterns, similarities or clusters within the data, which can then be used for various purpose, such as data exploration, visualization, dimensionality reduction & more.

④ There are two types of UL :-

(1) Clustering
(2) Association



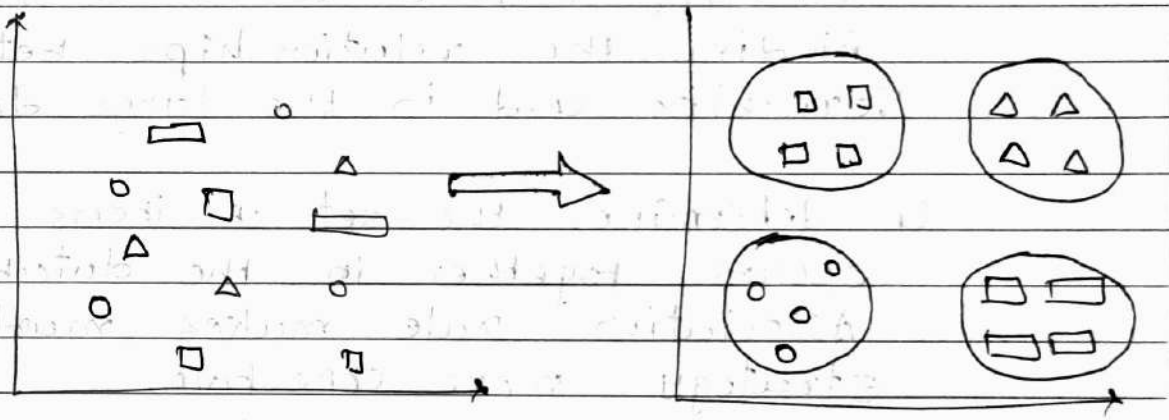
⑧ Unsupervised machine Learning ⑧

unsupervised learning

clustering | Association

[1] clustering :-

- clustering is a method of grouping the objects into clusters such that objects with most similarities remains into group and has less or no similarities with the objects of another group.
- cluster analysis finds the commonalities between the data objects & categorizes them per the presence and absense of those commonalities.



Raw Data → clustered data.

⊛ clustering ⊛

⊛ clustering Algorithms :-

- K-means clustering
- K-Nearest Neighbors (KNN) clustering
- Hierarchical clustering
- Density-Based spatial clustering of Applications with Noise (DBSCAN).

(*) clustering Methods :-

↓

(1) Density-Based Methods :

↓

(Depends on Density)

(2) Hierarchical Based Methods :

↓

(Depends on hierarchy)

(2) Association :-

- An Association rule is an unsupervised learning method which is used for finding the relationships between variables and in the large databases.

- It determines the set of items that occurs together in the databases sets Association rule makes marketing strategy more effective.

(suppose Bread)

- Such as people who buy X item are also tend to purchase Y (Butter/Jam) item.

- A typical ex of Association ## rule is market Basket Analysis.

(*) Advantages of UL :-

- UL learning is used for more complex tasks as compared to supervised learning

because, in UL, we don't have labeled input data.

- UL is preferable as it is easy to get Unlabeled data comparision to labeled data.

⊛ Disadvantages of UL :-

- UL is intrinsically more difficult than SL as it does not have corresponding output.
- The result of the UL algorithm might be less accurate as input data is not labeled, and algorithms do not know the exact output in advance.

⊛ Difference Between SL & UL :-

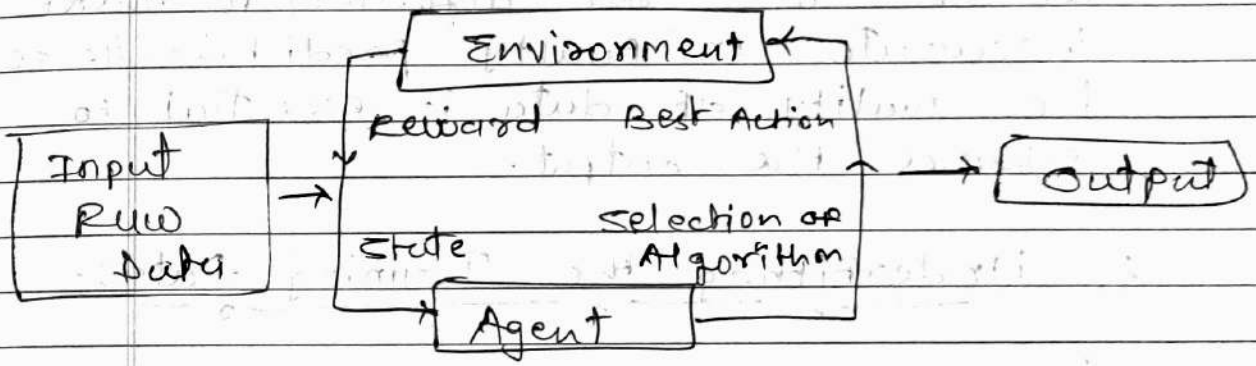
	Supervised Learning	Unsupervised Learning
(1)	SL uses known & labeled data as input.	Uses Unknown data as input.
(2)	Less computational Complexity.	More computational Complexity.
(3)	Number of classes are known.	Number of classes are unknown.

(4)	Uses off-line analysis.	uses Real-Time Analysis of Data.
(5)	Accurate and Reliable Results.	moderate Accurate & Reliable Results.
(6)	Desired output is given.	Desired output is not given.
(7)	in SL, training is used to infer model.	in UL, training Data is not used.
(8)	SL also called classification.	UL also called clustering.
(9)	We can test our model.	We can't test our model.
(10)	Ex: Spam detection Image recognition.	Ex: PCA Dimentionality reduction, etc.

Q. (A) Explain concept of penalty & reward in reinforcement learning.

- In reinforcement learning (RL), a decision-making agent takes actions in an environment and receives or rewards or penalties for its actions. The agent learns without human intervention by maximizing its reward & minimizing its penalty.

- the agent is rewarded for correct moves & punished for the wrong ones. For each good actions, the agents get possitive feedback, and for each bad action, the agent gets negative feedback or penality.



Q.⑤ Issues of Machine Learning :-

(1) poor Quality of Data
(2) Underfitting of Training Data
(3) Overfitting of Training Data
(4) Machine learning is a complex process
(5) Lack of Training Data
(6) Slow Implementation
(7) Imperfections in the Algorithm when Data grows.

(1) Poor Quality of Data :-

- Data plays a significant role in the machine learning process.

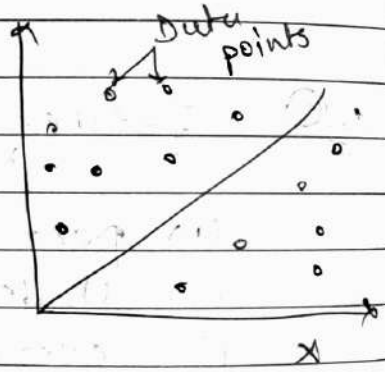
- One of the significant issues that ML proffesionals is the absense of good Quality of Data.

- Unclean & Noisy data can make the whole process extremely exhausting.

- we don't want our algorithm to make inaccurate or faulty predictions. Hense the quality of data is essential to enhance the output.

[2] Underfitting the Training Data:-

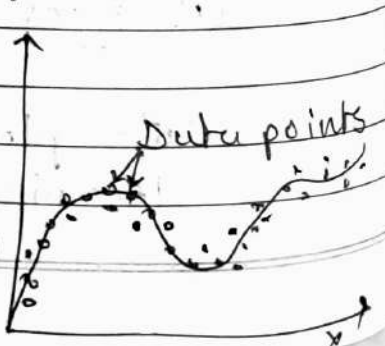
- This process occurs when data is unable to establish an accurate relationship between input & output variables.



- to overcome this issue:
  - maximize the training time
  - Add more features to data
  - Enhance the complexity of the model

[3] Overfitting of Training Data :-

- Overfitting refers to a machine learning model trained with a massive amount of data that negatively affects its performance.



- This means that the algorithm is trained with noisy and biased data, which will affects its overall performance.

[3] - to overcome this issue:

- remove outliers in training set
- select a model with lesser features

[4] Machine Learning is a complex Process :-

- the ML industry is young & is continously changing.
- Rapid hit and trail experiments are being carried on. the process is transforming, and hence there are high chances of error which makes the learning complex.

0.6 Applications of Machine Learning :-

(1) Image Recognition
(2) Speech Recognition
(3) Traffic Prediction
(4) Product Recommendations
(5) Email Spam & Malware Filtering
(6) Self Driving cars
(7) Virtual personal Assistant
(8) Online Fraud Detection
(9) Stock Market Trading
(10) Medical Diagnosis

[1] Image classif Recognition :-

- Image recognition is one of the most common applications of ML. it is used to identify objects, persons, places, digital images, etc.

- the popular use of image recognition and face detection is Automatic friend tagging suggestion.

- it is based on the project name "Deep face", which is responsible for face recognition and person identification.

[2] Speech Recognition :-

- while using Google, we get an option of "search by voice", it comes under speech recognition, and it's a popular application of machine learning.

- Speech Recognition is a process of converting voice instructions into text, and it is also known as "speech to text".

- Google Assistant, Cortana, Siri & Alexa are using speech Recognition technology to follow the voice instructions.

[3] Traffic Prediction :-

- if we want to visit a new place, we take help of Google Maps, which shows us the correct path with the shortest route and predicts the traffic conditions.

- it predicts the traffic conditions such as whether traffic is cleared, slow-moving, or heavily congested with the help of two ways:
  - Real time location
  - Average time

[4] Product Recommendation :-

- ML is widely used by various e-commerce and entertainment companies such as Amazon, Netflix, etc., for product recommendation to the user.

- Google: understand user interest using various machine learning algorithms and suggest the product as per customer interest.

- As similar: when we use Netflix, we find some recommendations for entertainment series, movies, etc.

[5] Email Spam & Malware Filtering:

- Whenever we receive a new email it is ~~fir~~ filtered automatically as important, normal, and spam. This filteration is done by machine learning.

- Below are some spam filters used by Gmail:

  - Content filter
  - Header filter
  - Rule-Based filters
  - Permission filters

- Some ml algorithms such as multi-layer perceptron, Decision Tree, and Naive Bayes classifier are used for email spam filtering & malware detection.

Q. ① What do you mean by well-posed learning problem? Explain important features that are required to well-define a learning problem with an example.

- A Computer program is said to learn from experience E with respect to some class of task T and performance measure P, if its performance at task in T, as measured by P, improves with

experience E. To have a well-defined learning problem.

= There are three important features need to be identified:

(1) the class of tasks: (T)
(2) the measure of performance to be improved (P)
(3) the source of experience (E)

Examples:-

(1) A checkers learning problem:

- Task T: playing checkers.
- Performance measure P: percent of games won against opponents.
- Training experience E: playing practice games against it-self.

(2) A self driving learning problem:

- Task T: driving on public routes using vision sensors.
- Performance measure P: average distance travelled before an error
- Training experience E: a sequence of images & steering commands recorded while observing a human drivers.

Page No.
Date

④ well-define learning problem role in ML:-

- A ML problem is well-posed if a solution to it exists. if that solution Technique is unique, and if that Solution depends on the data / experience but it is not sensitive to changes in the data / expereince.

- Learning to recognize spoken words.

- Learning to drive an autonomous vehicle.

- Learning to classify new astronomical structures.

- Learning to play word class chess all by itself.

8. ⑧ Machine Learning

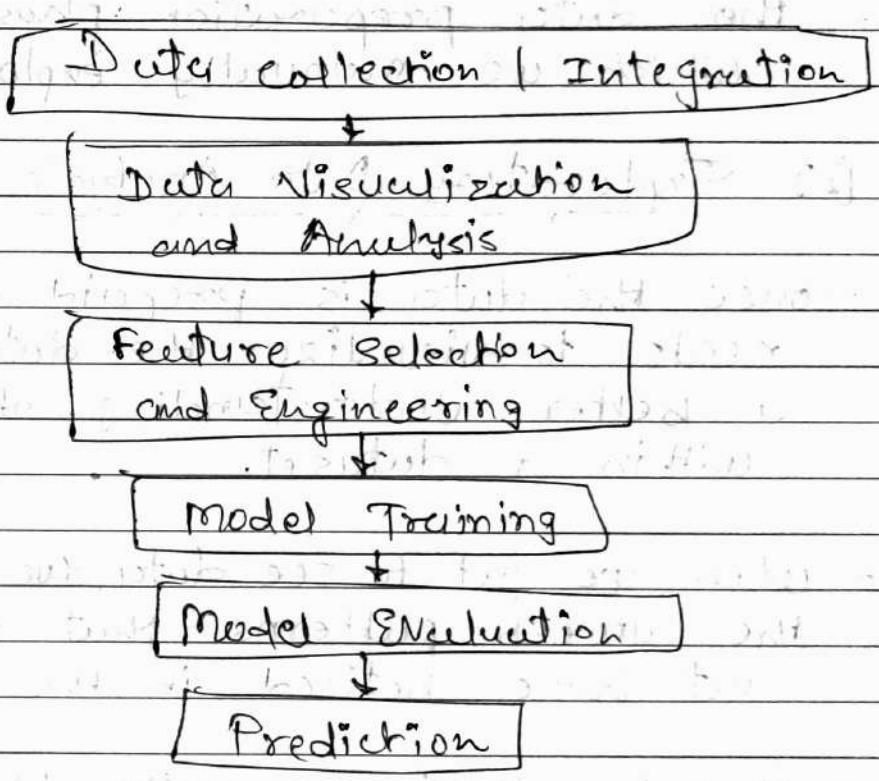
Chapter: 2: Preparing to Model

Q.1 Machine Learning Activities.

Or

Machine Learning steps / Life cycle.

(*) There are mainly 6 Steps / Activities of Machine Learning:



[1] Data collection and Integrity :-

- The first step of ML Activities involves the collection of data and integration of data.

- Data collected acts as an input to the model. inputs are called Features.

- The more the Data is, more the better our model becomes.

- Once the data is collected we need to integrate and prepare the data.

- Integration of data means placing all related data together.

- Then data preperation phase starts in which we manually explore the data

(2) Exploratory Data Analysis & Visualizatio

- once the data is prepaid developer needs to visualize the data to have a better understanding of relationship within a dataset.

- When we get to see data, we can notice the unseen patterns that we may not have noticed in the first phase.

- It helps developers easily identify missing data and outliers.

- Data visulization can be done by plotting histograms, scatter plots, etc.

- After visualization is done data is analyzed so that developer can decide what ML technique he may use.

[3] Feature selection and Engineering :-

- Feature selection means selecting what features the developer wants to use within the model.

- Features should be selected so that a minimum correlation exists between them & a maximum correlation exists between the selected features & output.

- Feature Engineering is the process of converting raw data into useful data or getting the maximum out of orginal data.

- it deals with accuracy & Precision of data.

[4] Model Training :-

- After the three phase are done completely we enter the model training phase.

- it is the first step officially when the developer gets a train the model on basis of data.

- to train the model, data is split into three parts - Training Data, Validation Data and Testing Data.

- Training Data is used to train the model. Around 70% - 80% of data goes into the training data set.

- Validation Data is used to evaluate model. Around 10% - 15% of data used in Validation data.

- Res Testing data is used to test model. Rest 20% - 25% of data used in testing data.

- Data can be randomize using skikit learn in Python.

[5] Model Evaluation :-

- After the model training, Validation or development data is used to evalute the model.

- To get the most accurate predictions to test data may be used for further model evaluation.

- A Confusion matrix is created after model evaluation to calculate accuracy and precision numerically.

- After model evaluation, our model enters the final stage that is prediction.

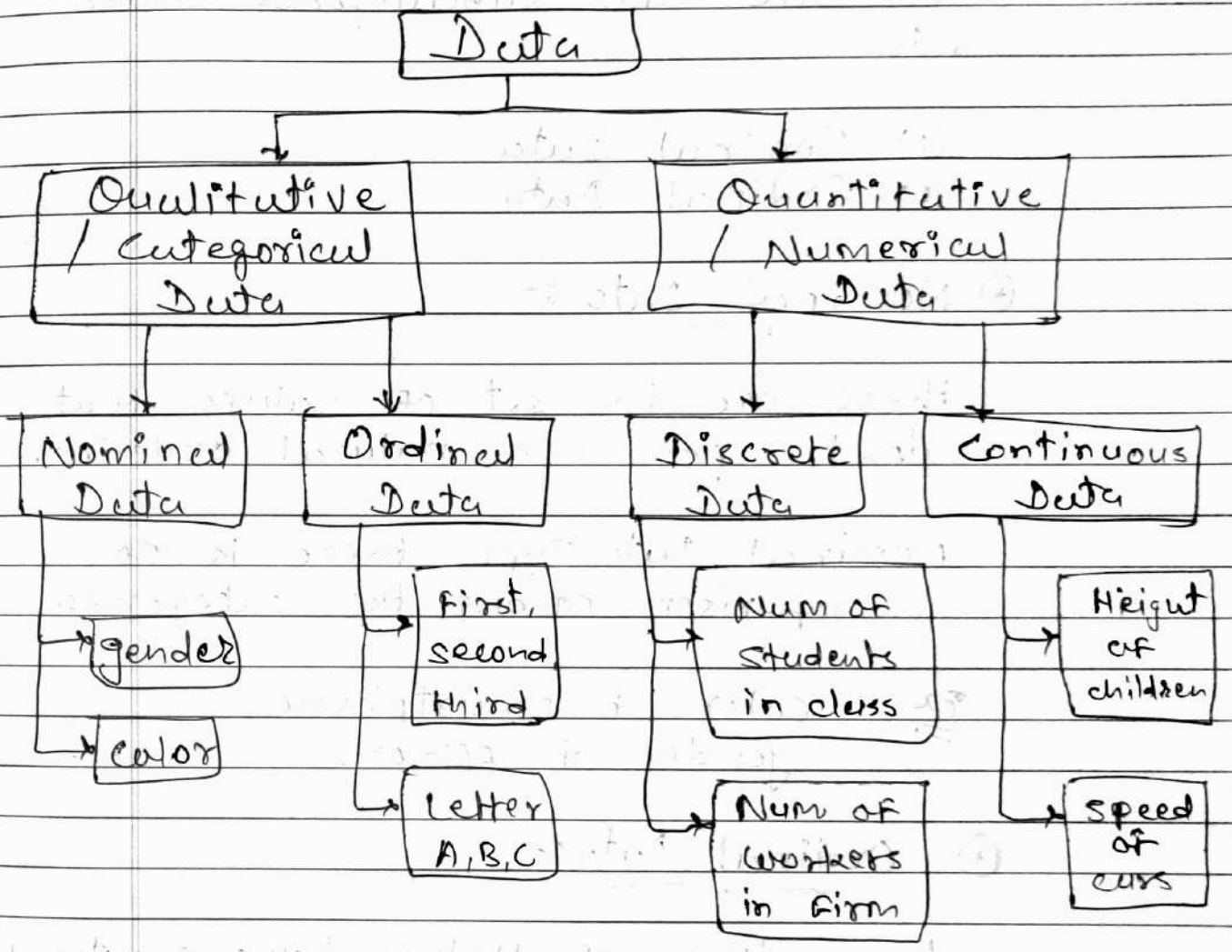
[6] Prediction :-

- In the prediction phase developer deploys the model.

- After the model deployment, it becomes ready to make predictions.

- predictions are made on training data and test data to have a better understanding of the build model.

Q.② Explain types of Data in ML.



⊛ Qualitative Data / categorical :-

- Categorical Data describes the object under consideration using a finite set of discrete class.

- it means this type of data can't be counted or measured using numbers & therefore divided [easily] into categories.

Ex. gender of person.

- there are two subcategories under this:

(1) Nominal Data
(2) Ordinal Data

⊛ Nominal Data :-

- These are the set of values that don't possess a natural ordering.

- Nominal Data Type there is no comparison among the categories.

Ex color of smartphone,
gender of person.

⊛ Ordinal Data :-

- These types of values have a natural ordering while maintaing their class or values.

- These categories help us deciding which encoding strategy can be applied to which type of data.

Ex. Small < medium < large < Extra large.
C < B < A (Grade).

⊛ Quantitative Data / Numerical :-

- This data type tries to Quantify things and it does by considering numerical values and that make it Countable in nature.

- There are two subcategories of this :

(1) Discrete Data
(2) Continuous Data

⊛ Discrete Data :-

- The numerical values which fall under are integers or whole numbers are placed under this category.

Ex. The Number of students in class.
The Numer of Employees in Firm.

⊛ Continuous Data :-

- The fractional Numbers are considered as continuous data.

Ex. The speed of cars.
Heights of children.

Q.③ Measures of Central Tendency :-

- Central tendency in ML are the numerical values that are used to represent mid-value or central value a large collection of numerical data.

- Some of the most commonly used measures of central tendency are:

(1) mean
(2) Median
(3) mode

Ex

3	3	5	6	8

① mean = $\frac{\sum x_i}{n} = \frac{3+3+5+6+8}{5} = \boxed{\frac{26}{5}}$

② Median = $\frac{n+1}{2} = \frac{6}{2} = 3$ = 5

↓

Value of observation at $\left[\frac{(n+1)}{2}\right]^{th}$ position

if N is ~~odd~~ even

↓

else $\left(\frac{n}{2}\right)$.

③ mode = Value which has maximum frequency.

= 3

⊛ Standard Deviation & Variance ⊛

⊛ SD :-

- How far our given set of data varies along with the mean of the data is measured in standard Deviation.
- we can also define the SD as the square root of the variance.

⊛ Formula :-

$$S = \sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

↓ For sample S.D

Where,

σ = Standard Deviation
N = Number of observations
x_i = i^{th} value (observation) ~~in the populat~~
$\bar{x}$ = mean

⊛ Variance :-

- Variance is defined as "the measure of how far the set of data is dispread from their mean value".
- Variance is represented as σ^2.

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$ → For sample Variance

⊛ relationship between σ & σ^2 :

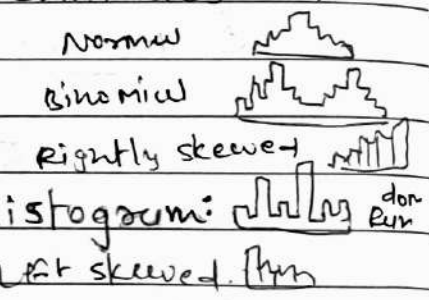
Variance = (Standar Deviation)^2

$\sqrt{\text{(Variance)}}$ = Standard Deviation.

Q. (4) Box plots & Histogram for take cur of outliers in data.

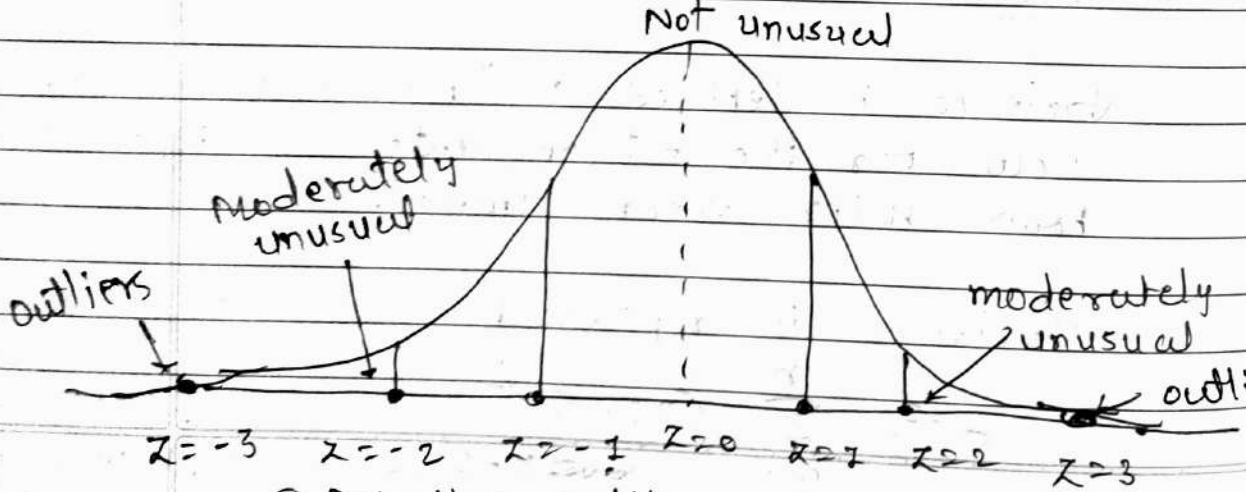
- Below are the some techniques of detecting outliers :-

(1) Boxplots
(2) Z-score
(3) Histogram:



⊛ Z-score :-

- Z-score is also called standard score.
- This scale helps to understand if a data value is greater or smaller than mean and how far away it is from the mean.

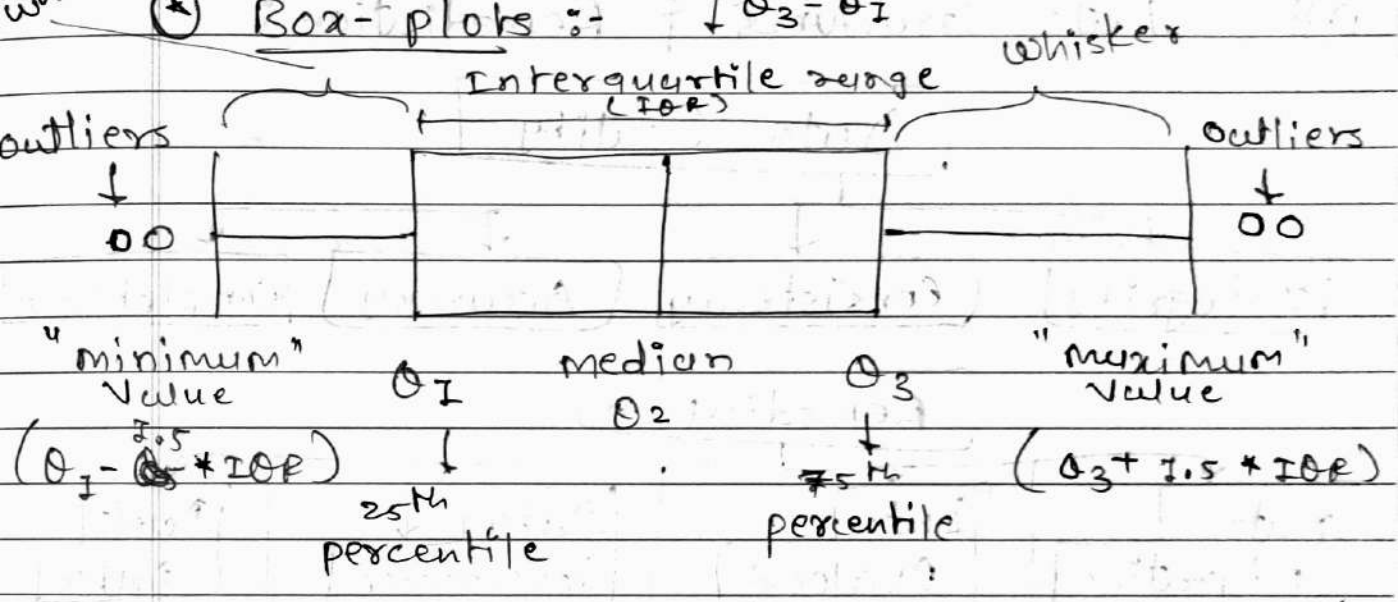


⊛ Detecting outliers with z-score ⊛

- any data point whose z-score falls out of 3rd standard deviation is an outlier.

$$\text{Z-Score} = \frac{(X - \text{mean})}{\text{std. deviation}}$$

⑦ Box-plots :-



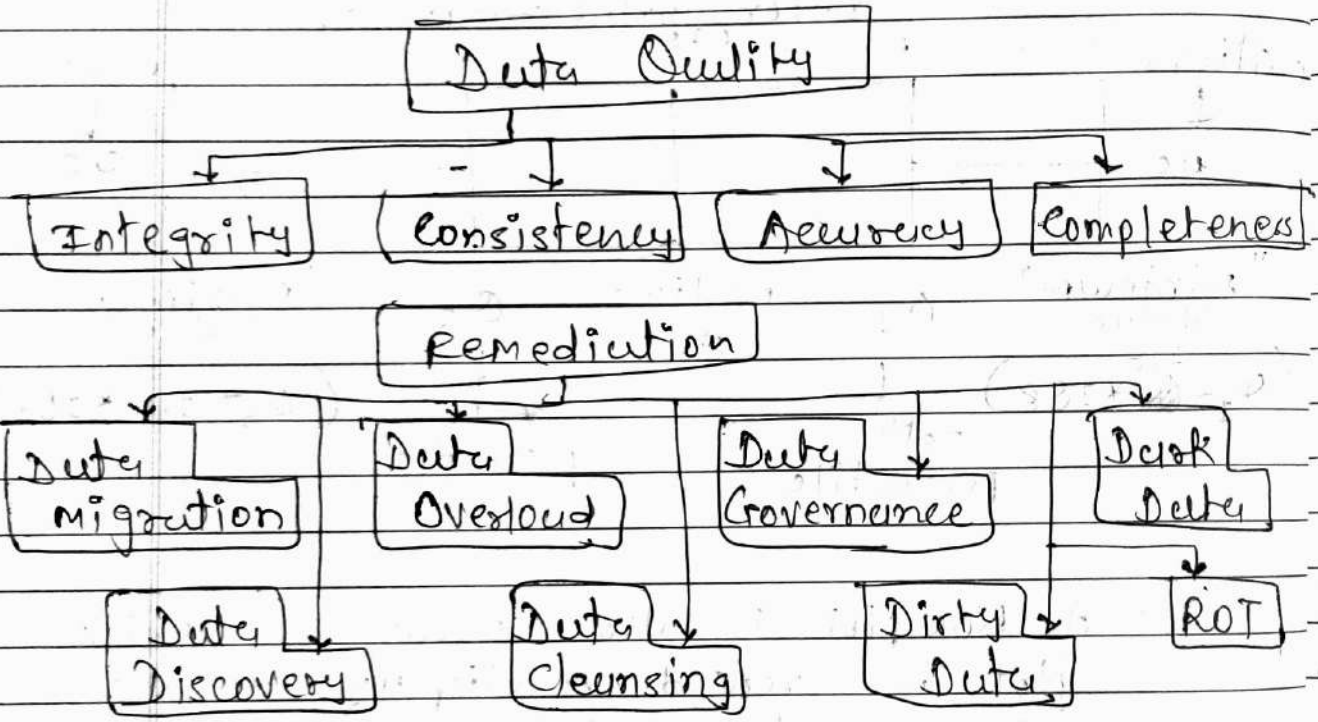
- A Box plot is also known as whisker plot is created to display the summary of the set of Data values having properties like minimum, first quartile, median, third quartile and maximum.

- in the boxplot, a box is created from the first quartile to the third quartile, a vertical line is also there which goes through the box at the median.

- Data points that lie 1.5 times of IQR above Q3 and below Q1 are outliers.

Lower Bound = $Q_1 - 1.5 * IQR$
Higher Bound = $Q_3 + 1.5 * IQR$
$IQR = Q_3 - Q_1$.

Q.⑤ Data Quality & Remediation :



⑤ Stages of Remediation :-

(1) Assessment
(2) Organizing & segmentation
(3) Indexation & classification
(4) Migrating
(5) Data cleansing

(*) Data PreProcessing Techniques :-

Q.(6) Dimentionality Reduction :-

(1) Dimentionality Reduction
(2) Feature subset selection.

(*) Data preprocessing :-

- Data preprocessing is an important step in data mining process IML.

- It refers to the cleaning, transforming, and integrating of data in order to make it ready for analysis.

- The goal of Data preprocessing is to improve the quality of the data & to make it more suitable for the specific data mining task.

(*) some of common D.P Techniques :

(1) Dimensionality Reduction
(2) Feature subset selection.

(*) Dimensionality Reduction :-

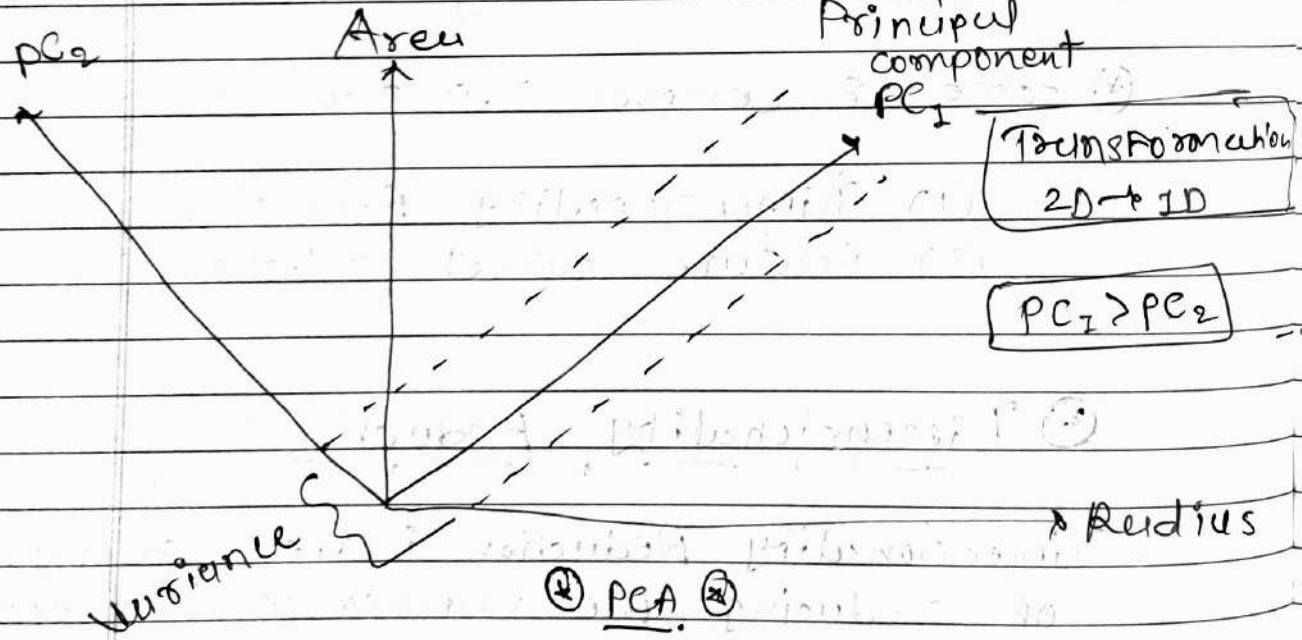
- Dimensionality Reduction is the process of reducing the number of random variables under consideration by obtaining a set of principal variables.

- Most ML Techniques are not be effective for high-dimensional data. Query accuracy and efficiency degrade rapidly as the dimension increases.

- The "dimensionality" simply refers to number of features.

⊛ Principal Component Analysis :-

- principal component Analysis (PCA) is used to reduce the dimensionality of a data set by finding a new set of variables, smaller than the original set of variables, retaining most of the samples information, and useful for the regression & classification of data.



⊛ PCA ⊛

- PCA is a technique for dimensionality reduction that identifies a set of orthogonal axes, called principal components,

that captures the maximum variance in the data.

- PCA can be used for a variety of purposes, including data visualization, feature selection, & data compression.

Steps of PCA :-

(1) Standardization
(2) Covariance Matrix computation
(3) Compute Eigenvalues & Eigenvectors of covariance Matrix to identify Principal components.

Advantages of PCA :-

(1) Dimensionality Reduction
(2) Feature selection
(3) Data Visualization
(4) Multicollinearity
(5) Noise Reduction
(6) Data compression

Feature Subset Selection :-

- Feature selection is a process that chooses a subset of feature from the original features so that the feature space is optimally reduced according to a certain criterion.

- Feature selection is a critical step in the feature construction process.

- There are three general classes of feature selection algorithms:

  - Filter Methods
  - Wrapper Methods
  - Embedded Methods.

- The role of feature selection in ML is,

  - to reduce the dimensionality of feature
  - to speed up a learning algorithm.
  - to improve the predictive accuracy of a classification algorithm.
  - to improve the comprehensibility of the learning results.

⊛ Feature Selection Algorithms are as follows

(1) Instance Based Approaches
(2) Nondeterministic Approaches
(3) Exhaustive Complete Approaches

---

(1) Instance Based Approaches :-

- There is no explicit procedure for feature subset generation.

- Many small data samples are sampled from the data.

- Features are weighted according to their roles in differentiating instances of different classes for a data sample.
- Features with higher weights can be selected.

[2] Nondeterministic Approaches :-

- Genetic Algorithms and simulated annealing are also used in feature selection.

[3] Exhaustive, Complete Approaches :-

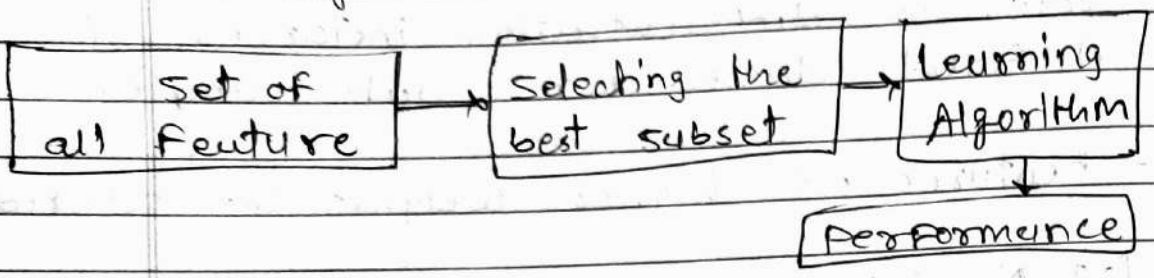
- Branch and Bound evaluates estimated accuracy and ABB checks an inconsistency measure that is monotonic.
- Both start with a full feature set until the preset bound cannot be maintained.

⊛ Feature Selection Algorithms :-

⊛ Filter methods :-

- These methods are generally used while doing the pre-processing step.
- These methods select features from the dataset irrespective of the use of any

ML Algorithm.



- Some techniques used are:

(1) Information Gain
(2) chi-square test → $x^2 = \sum \frac{(\text{observed value} - \text{Expected value})^2}{\text{Expected value}}$
(3) Fisher's score
(4) Correlation Coefficient
(5) Variance Threshold
(6) Dispresion Ratio
(7) Mutul Dependence
(8) Relief

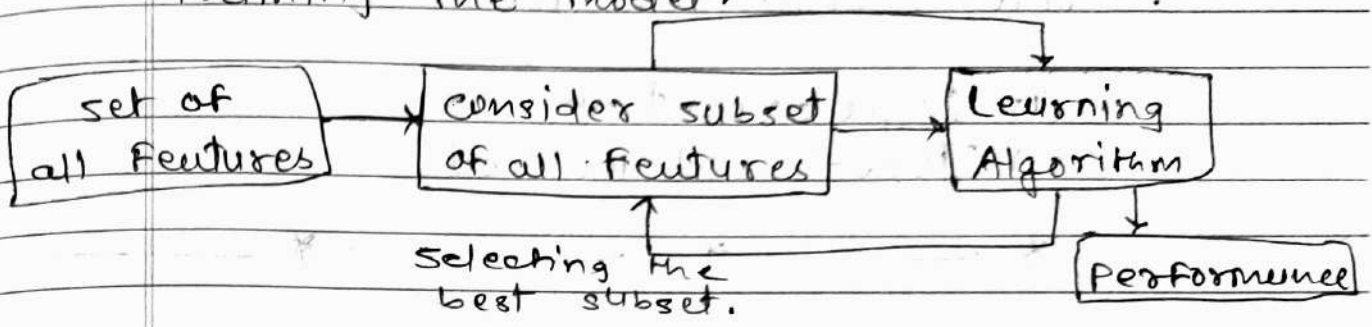
⊛ Wrapper Methods :-

- Wrapper methods, also reffered to as greedy algorithms train the algorithm by using a subset of features in an iterative manner.

- Based on the conclusions made from training in prior to the model, addition and removal of features takes place.

- The main advantage of wrapper methods over the filter methods ~~has~~ is that they provide an optimal set of features for

training the model.



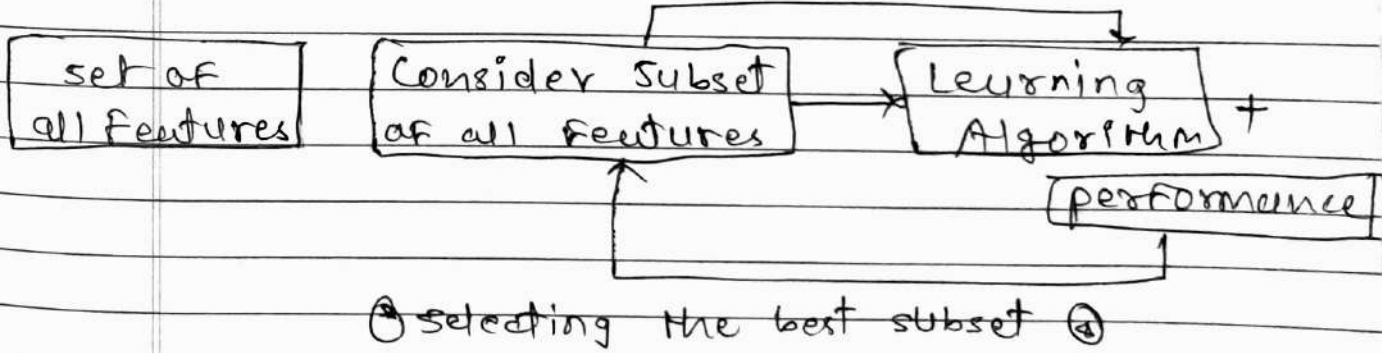
- some techniques used are:

(1) Forward selection
(2) Backward selection
(3) Bi-directional elimination
(4) Exhaustive selection
(5) Recursive Elimination

③ Embedded Methods :-

- In Embedded Methods, the Feature selection algorithm is blended as part of the learning algorithm, thus having its own built-in Feature selection methods.

- Embedded Methods encounter the drawbacks of Filter and wrapper methods and more accurate advantages.



- Some techniques used are:

(1) Regularization
(2) Tree-based Methods

* * *

⊛ Chapter : ③ : Modeling & Evaluation ⊛

Q.① Explain model selection & its techniques.

- Model selection is the process of selecting one final machine learning model from among a collection of candidate ml models for a training dataset.

- Model selection is a process that can be applied both across different types of models. (eg. Logistic regression, SVM, KNN).

- Model selection is the process of choosing one of the models as the final model that address the problem.

- The best approach to model selection requires "sufficient" data, which may be nearly infinitive depending on the complexity of the problem.

- There are two main classes of techniques to approximate this ideal case of model selection :

(1) Probabilistic Measures :

- choose a model via in-sample error and complexity.

(2) Resampling Methods :

- cho... via estimated out-of-sample error.

⑧ Probabilistic Measures :-

- probabilistic measures involve analytically scoring a candidate model using both performance on the training dataset and the complexity of the model.

- It is known that training error is optimistically biased, and therefore is not a good basis for choosing a model.

- A model with fewer parameters is less complex, and because it is preffered because it is likely (of this) to generalize better on average.

- Four commonly used probabilistic model selection measures include:

(1) Akaike Information Criterion (AIC)
(2) Bayesian Information Criterion (BIC).
(3) Minimum Description Length (MDL).
(4) Structural Risk Minimization (SRM).

- probability measures are appropriate when using simple linear models like linear regression or logistic regression where the calculating of model complexity penalty is known and tractable.

(*) Resampling methods :-

- Resampling methods seek to estimate the performance of a model on out-of-sample data.

- This is achieved by splitting the training dataset into subtrain and test sets, fitting a model on the sub train set, and evaluating it on the test sets.

- This process may then be repeated, multiple times and then mean performance across each trail is reported.

- It is a type of Monte Carlo estimate of model performance on out-of-sample data, although each trail is not strictly independent as depending on the resampling method chosen the same data may appear multiple times in different training datasets or test datasets.

- Three common resampling model selection methods include:

  (1) Random train/test split.
  (2) Cross-Validation
  (3) Bootstrap.

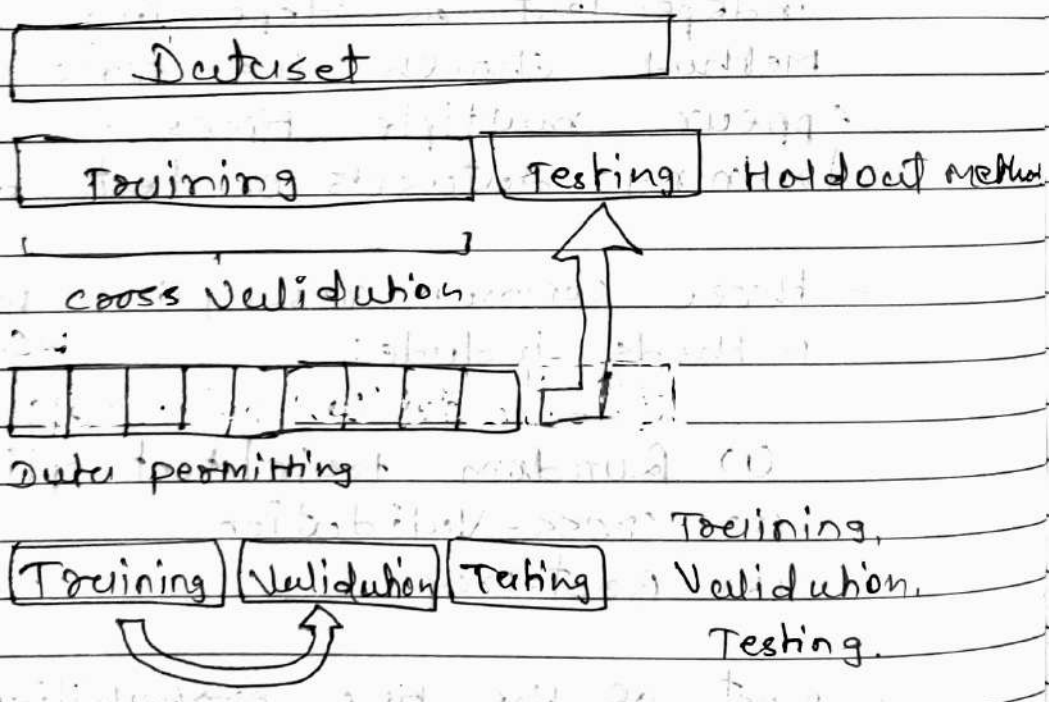
- Most of the time probabilistic measures are not available, therefore resampling methods are used.

Q.② Explain Cross-Validation.

- To overcome over-fitting problems, we use a technique called cross-validation.

- Cross validation is a technique for evaluating performance by training several machine learning models on subsets of the available input data and evaluating them on the complementary subset of the data.

- Use cross-validation to detecting overfitting, i.e, failing to generalize a pattern.



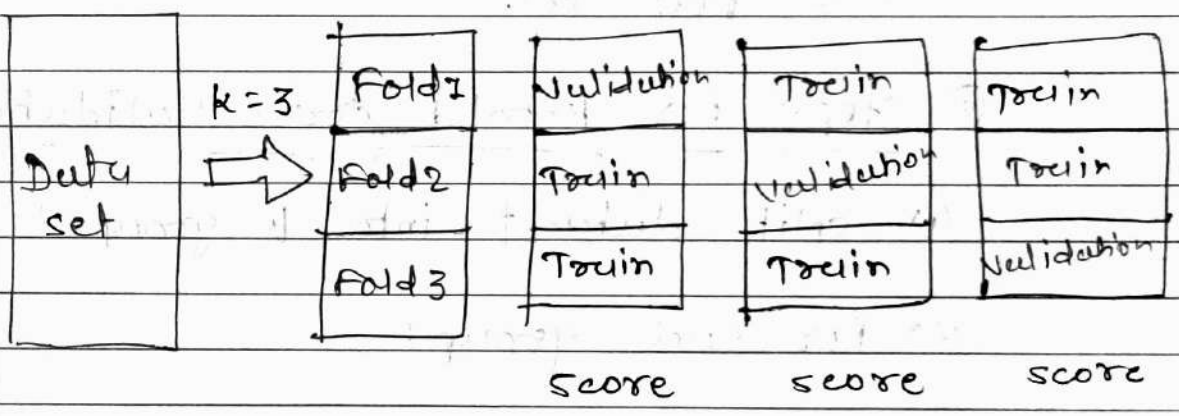
⊛ Cross Validation ⊛

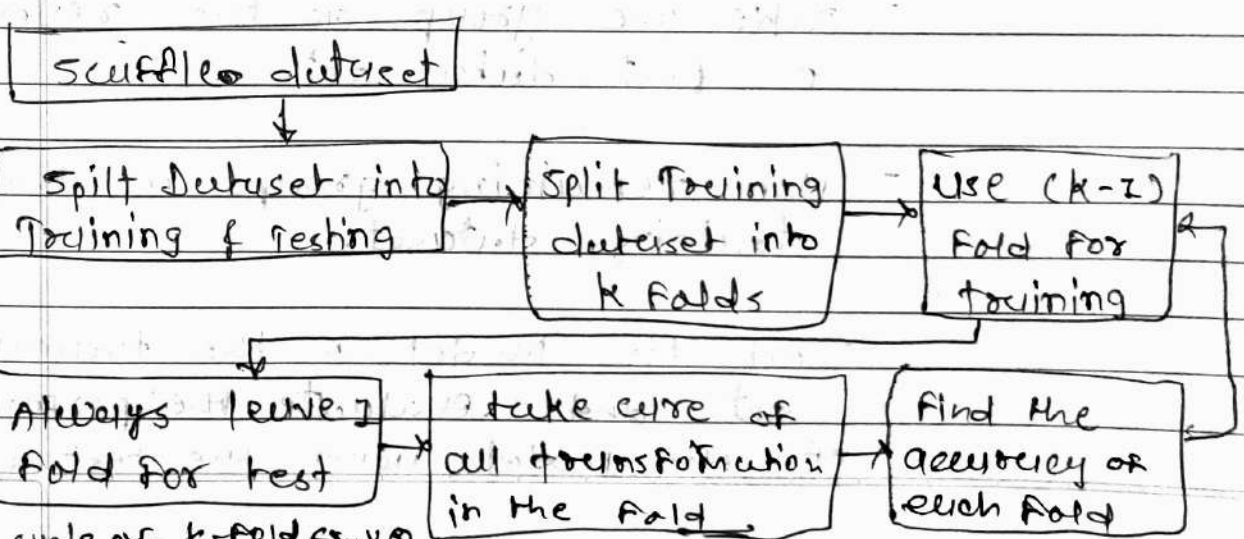
- Cross-Validation is a resampling technique with the fundamental idea of splitting the dataset into 2 parts training data and test data.

⊛ Methods For Cross-Validation :-

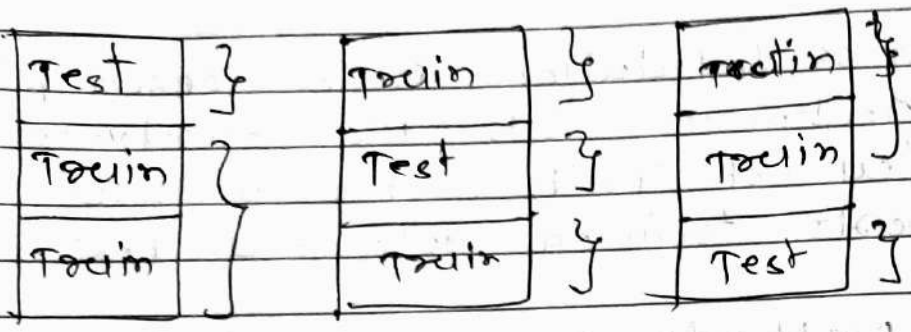
(1) Validation set approach
(2) Leave p-out cross Validation
(3) Leave one out cross validation
(4) k-Fold cross Validation
(5) Stratified k-Fold cross Validation

⊛ k-Fold Cross-Validation :-





⊛ life cycle of k-Fold cr...v ⊛



- K-Fold cross-Validation approaches divides the input into k-groups of samples of equal of samples of equal sizes. 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.

⊛ Steps for k-fold cross Validation :-

(1) split dataset into k groups:

(2) For each group:

- Take one group as the reverse or test data set.

- Use remaining groups as the training dataset.

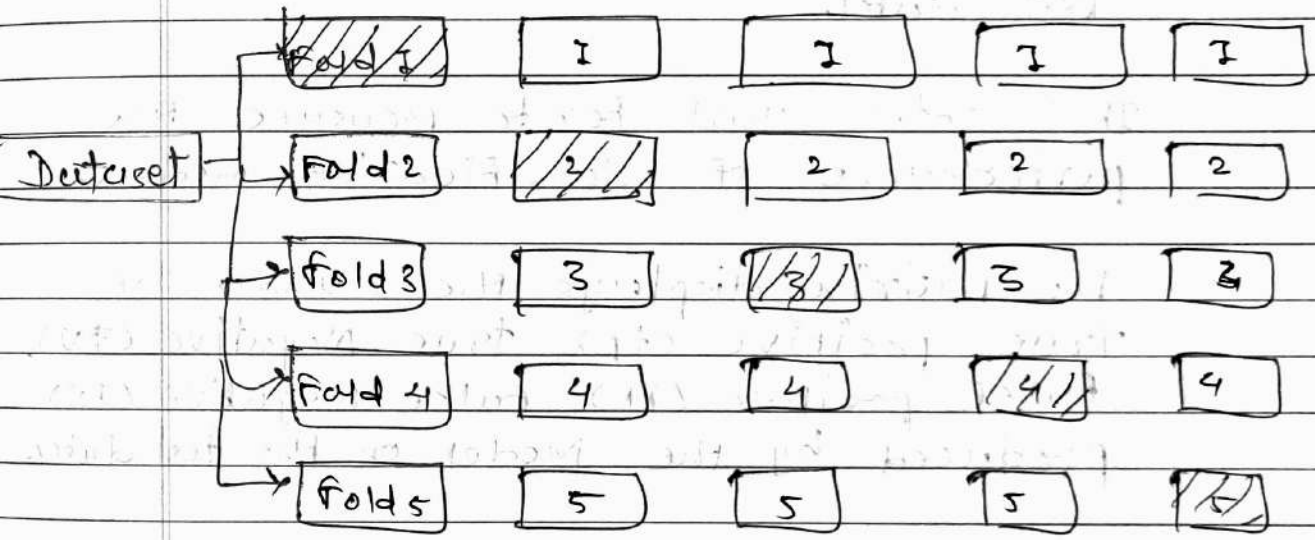
- Fit the Model on the training set and evaluate the performance of the Model using the test set.

Ex. 5-Fold cross Validation.

so, the dataset is grouped into 5 folds.

- on 1st iteration, the first fold is reserved for the test model, & rest are used to train the model.

- on 2nd iteration, the second fold is used to test the model & rest are used to train the model. this process continue until each fold is not used for the test fold.



▨ - Testing

▯ - Training.

- Cross-Validation is used to overcome the disadvantage of train/test split.

Q.③ Explain Model Evaluation Techniques.

(1) Cross-Validation
(2) Confusion Matrix
(3) Chi-square Test
(4) Root-mean-squared Error (RMSE)
(5) ~~etc~~. Accuracy, ~~etc~~.
(6) Hotbot, etc.

⊛ Confusion Matrix :-

- A confusion matrix is a matrix that summarizes the performance of a machine learning model on a set of test data.

- It is often used for to measure the performance of classification models.

- The matrix displays the number of true positive (TP), true Negative (TN), false positive (FP), false Negative (FN) produced by the model on the test data.

		Actual Values	
		Positive	Negative
predicted values	Positive	True (TP) positive	False (FP) positive
	Negative	False (FN) Negative	True (TN) Negative

⊛ Confusion matrix ⊛

(1) True Positive (TP):

- it is the total counts having both predicted & actual values

- the predicted value matches the actual value, or the

- the actual values was positive & the model predicted a positive value.

(2) True Negative (TN):

- the predicted value matches the actual value.

- the actual value was negative, & the model predicted a negative value.

(3) False Positive :- (Type-I Error)

predicted.

- the predicted value was Fulsy

- the actual value was negative, but the model predicted a positive value.

(4) False Negative :- (Type-2 Error)

predicted.

- the predicted value was Fulsy

- the actual value was positive, but the model predicted a negative value.

(*) Terms Related Confusion Matrix (*)

(1) Accuracy
(2) Precision
(3) Recall
(4) F1-Score

[1] Accuracy :-

- Accuracy is used to measure the performance of the model.
- it is the ratio of total correct instances to the total instances.

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

[2] Precision :-

- Precision is a measure of how accurate a model's positive predictions
- it is defined as the ratio of true positive predictions to the total number of positive predictions made by the model.

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

(3) Recall :-

- Recall measures the effectiveness of a classification model in identifying all relevant instances from a dataset.
- it is the ratio of the number of true positive (tp) instances to the sum of true positive and false negative instances.

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

[4] F1-Score :-

- F1-score is used to evaluate the overall performance of a classification model.
- it is the harmonic mean of precision and recall.

$$\text{F1-score} = 2 * \frac{\text{Precision} \cdot \text{Recall}}{\text{precision} + \text{Recall}}$$

Ex.

	Predicted +	Predicted -
Actual +	60	15
Actual -	6010	15

Ans:

	Predicted +	Predicted -	
Actual +	60	15	75
Actual -	10	15	25
	70	30	

$$\text{Accuracy} = \frac{60+15}{60+10+15+15} = \frac{75}{100} = 75\%$$

$$\text{Precision} = \frac{60}{60+10} = 0.8571$$

$$\text{Recall} = \frac{60}{60+15} = 0.80$$

$$\text{F1-score} = 2 * \frac{0.8571 * 0.80}{0.8571 + 0.80}$$

$$= 2 * \frac{1.6757}{1.6771} (0.7000)$$

$$= \frac{1.4000}{1.6771}$$

$$= 0.8358$$

Q.(4) Model Representation & Interpretability :-

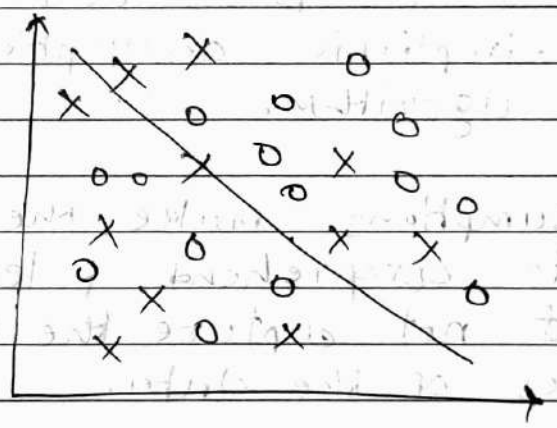
(1) Underfitting & Overfitting
(2) Bias Variance

(*) Underfitting :-

- A statistical model or a ML model is said to be have underfitting when a model is too simple to capture data complexities.

- An underfit model's are inaccurate.

- The underfitting model has High bias & low variance.



Reason for Underfitting :

- Training Data is too small
- Less Number of Features
- Noisy data

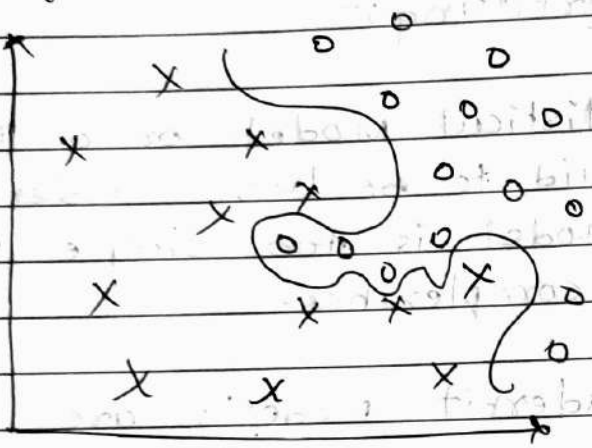
(*) Overfitting :-

- A ML model is said to have overfitting when the model does not make accurate predictions on testing data.

- when a model gets trained with so much data, it starts learning from

the noise & inaccurate data entries in our data set.

- Overfitting model has low bias & High Variance.



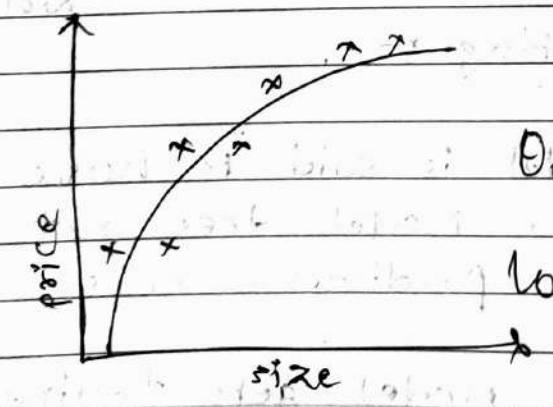
Reasons for Overfitting:

- Noisy data
- too ~~small~~ big data
- Large Amount of Features.

⊖ Bias :-

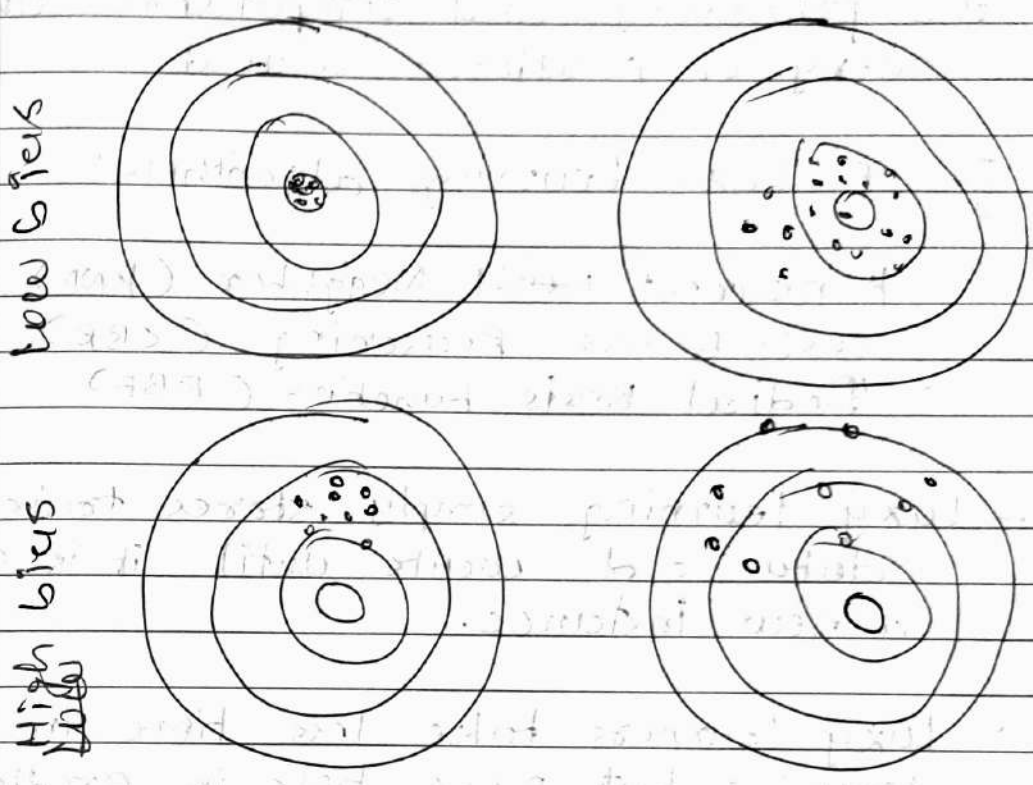
- Bias refers to the error due to overly simplistic assumptions in the learning algorithm.

- these assumptions make the model easier to comprehend & learn about but might not capture the underlying complexities of the data.



⊕ $\theta_0 + \theta_1 x + \theta_2 x \theta_2$

low - Bias



⊛ Variance :-

- Variance, on the other hand, is the error due to the model's sensitivity to fluctuations in the training data.
- it is the variability of the model's predictions for different instances of training data.

Q.5 Lazy vs Eager Learners.

⊛ Lazy Learning :-

- Lazy learning is also known as instance-based learning & memory-based learning. it postpones most of

the processing and computation until a query or prediction request.

Ex of Lazy learning algorithms:

- K-nearest Neighbour (KNN)
- Case-Based Reasoning (CBR).
- Radial Basis Function (RBF)

- Lazy learning simply stores training data and waits until it is given a new instance.

- Lazy learners take less time in training but more time in predicting.

② Eager learning :-

- Eager learning, also known as Model-Based learning, is an approach where the ML algorithm constructs a generalized model during training.

- This methods try to uncover the relations & patterns hidden in training data. Hence resulting model is a compact and abstract representation of the training dataset used.

- Eager learners adjust the model performance parameters during training to minimize the cost function. Once the model is trained. it

can make predictions about new inputs.

Ex of Super Learning algorithms.

- Decision Tree
- Support Vector Machine (SVM)
- Naive Bayes
- Artificial Neural Network (ANN)

Q. (16) (4) How to improve the performance of the model:

(1) choosing the Right Algorithms
(2) Use the Right Quantity of Data
(3) Quality of Training Data
(4) Supervised or Unsupervised ML
(5) Model Validation & Testing
(6) Tuning Model parameter

Q. (7) How to training the predictive model.

(1) First Determine the type of training dataset.

(2) Collect / Gather the labelled training data.

(3) Split the training dataset into training dataset, test dataset and Validation dataset.

(4) Determine the input data features of the training dataset, which should have enough knowledge so that the model can accurately predict the output.

(5) Determine the suitable algorithm for the model, such as support vector machine, decision tree, etc.

(6) Execute the algorithm on the training dataset. Sometimes we need validation sets as the control parameters, which are the subset of training datasets.

(7) Evaluate the accuracy of the model by providing the test set. If the model predicts the correct input. Output, which means our model is accurate.

Q.(8) Explain Predictive & Descriptive Analytics.

→ (A) Predictive Analytics :-

- Analytics that help you forecast future performance and results.

- predictive modeling is the process of taking known results and developing a model that can predict values for new occurences.

- it uses historical data to predict future events.

- these are very different types of predictive modeling techniques including linear regression, logistic regression, ridge regression, time series, decision tree, neural network.

⊛ Descriptive Analytics :-

- the descriptive analysis uses mainly unsupervised learning approaches for summerizing, classifying, extracting rules to answer what happens / was happened in the past.

- the descriptive models are different in nature from predictive models since they don't need to perform as accurately as the predictive models need to.

- Ex: Association rules or Market-Basket Analysis, clustering, Feature extraction.

⊛ Chapter : ④ : Basics of Feature Engineering ⊛

Q.① Explain Feature Engineering.

- Feature Engineering is the process of transforming raw data into features that better represent the underlying problem to the predictive models, resulting in improved model accuracy on unseen data.

⊛ Goals of Feature Engineering :-

- preparing the proper input dataset, compatible with the machine learning algorithm requirements.
- improving the performance of ML models.

⊛ → Feature Engineering is divided into 3 broad categories :

(1) Feature Selection
(2) Feature Transformation
(3) Feature Extraction.

[1] Feature Selection :-

- it is all about selecting a small subset of features from large pool of features.

- we select those attributes which best explain the relationship of an independent variable with the target variable.

- there are certain features which are more important than other features to the accuracy of the model.

- it is different from dimensionality reduction because the dimensionality reduction does so by combining existing attributes, whereas the feature selection method include or excludes those features.

Ex. chi-square test,
correlation coefficient,
LASSO,
Ridge Regression.

[2] Feature Transformation :-

- it means transforming our original feature to the functions original features.

Ex. Scaling, binning, discretization, and filling missing data values are the most common forms of data transformation.

- To reduce right skewness of the data, we use log.

[3] Feature Extraction :-

- when the data to be processed through an algorithm is too large, it's generally considered redundant.

- Analysis with a large number of variables uses a lot of computation power and memory, therefore we should reduce the dimensionality of these types of variables.

- it is a term for constructing combinations of the variables.

- For tabular data, we use PCA to reduce features.

- For image, we can use line or edge detection.

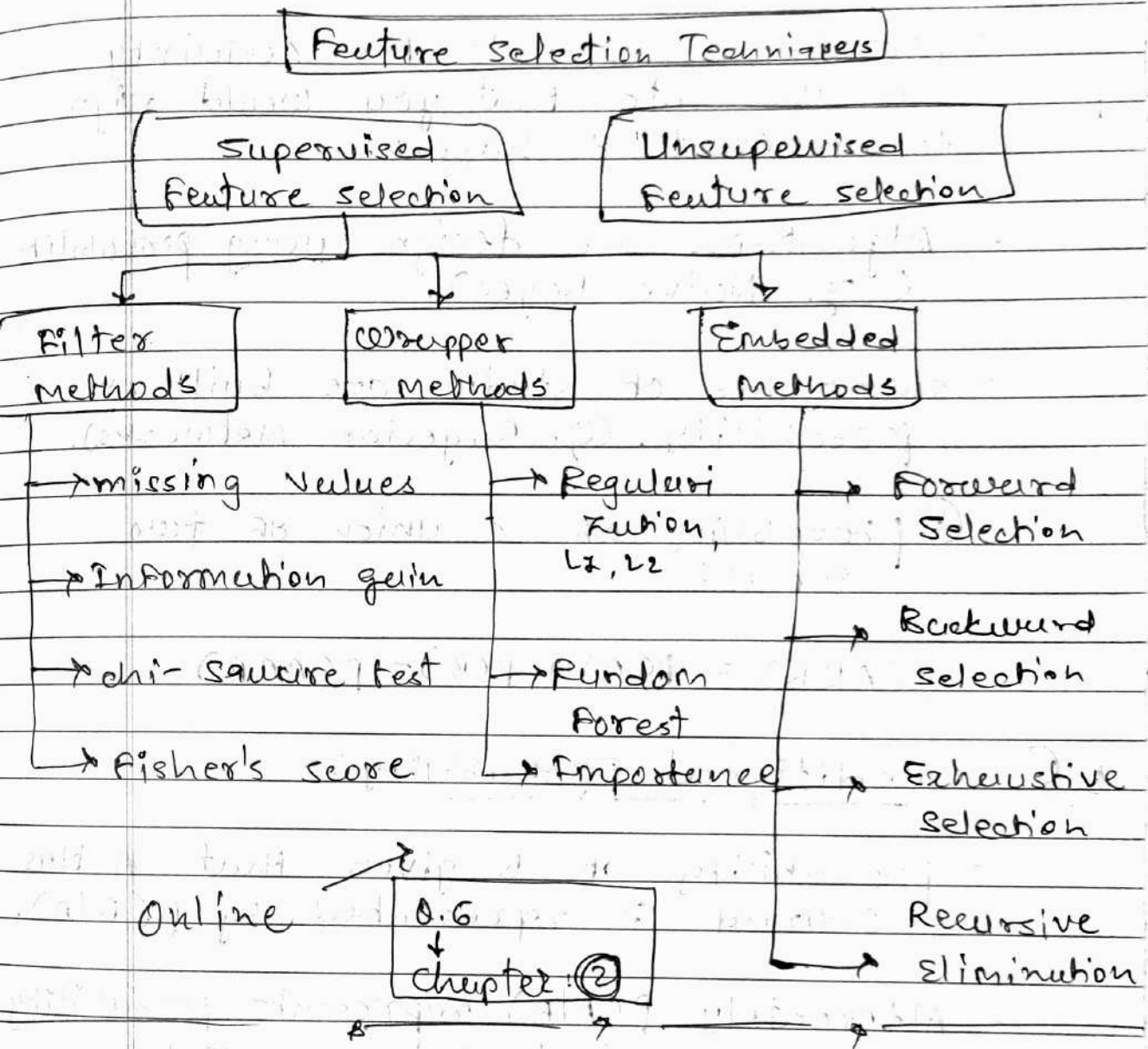
---

⑧ Feature Engineering Techniques :-

(1) Imputation
(2) Handling outliers
(3) Binning
(4) Log Transform
(5) One-Hot Encoding
(6) Grouping Operations
(7) Feature split
(8) Scaling
(9) Extracting Date

} Online

Q.② Feature selection Techniques:-



② Read measures of Feature relevance and Redundancy from Technical publication pg no. (4-6).

⭐ Chapter: ⑤ : Overview of Probability ⭐

Q. ① Concept of Probability.

- probability represents the certainty is the rate that you would assign to an event to happen.
- Algorithms are design using probabilities (Eg. Naive Bayes).
- Sub-fields of study are built on probability. (Eg. Bayesian Networks).

⑧ probability of a union of two events:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Q. ② Conditional probability :-

- probability of B given that A has occurred is represented by $P(B/A)$
- Alternately $P(A/B)$ represents probability of A given that B has occurred.
- $P(B/A)$ and $P(A/B)$ are called Conditional Probabilities.

$$P\left(\frac{B}{A}\right) = \frac{P(B \cap A)}{P(A)} \qquad P\left(\frac{A}{B}\right) = \frac{P(A \cap B)}{P(B)}$$

Q.② Binomial Distribution :-

- Binomial means 'two numbers'

- The outcomes of health research are often measured by (whether they have occured or not. eg

ex. admitted to hospital, recovered from disease, died, etc.

- The binomial distribution occurs in game of chance, quality inspection, opinion polls, medicine & so, on.

- it may be modelled by assuming that the number of events 'n' has a binomial distribution with a fixed probability of event p.

- Binomial distribution written as B(n,p), where n is total number of events and p = probability of an event.

⊛ properties of Binomial Distribution :-

1. Experiment consist of n identical trials.
2. Each trial has only two outcomes.
3. The probability of one outcome is p and the other is $q = 1 - p$.
4. The trails are independent.
5. We are interested in X, the number of success observed during the n trials.

- Trails satisfying the above properties are called Bernoulli trails.

- The probability function x,

$$f(x) = \binom{n}{x} p^x q^{n-x}$$

$f(x) = 0$ otherwise.

- The distribution on of X with probability function is called the binomial distribution.

- The mean (M) of B.d is,

$$\mu = np$$

- The variance is,

$$\sigma^2 = npq$$

③ Poisson Distribution :-

- Poisson Distribution named after its inventor Simeon poisson.

- He found that if we have a rare event (i.e p is small) and we know the expected or mean (or M) number of occurances, the probabilities of 0, 1, 2 events are given by:

$$p(R) = \frac{e^{-\mu}\mu^{R}}{R!}$$

- p.d is a distribution the number of rare events that occur in a unit of time, distance, & space & so on.

Ex. 1. Number of insurances claims in a unit of time.
2. Number of Airplane crash in triangle area.

4. Bernoulli Distribution :-

- The most basic of all discrete random variables is the Bernoulli.

- x is said to have a Bernoulli Distribution if $X = 1$ occurs with probability p & $X = 0$ occurs with probability $p-1$.

$$f(x) = \begin{cases} p & X = 1 \text{ (success)} \\ p-1 & X = 0 \text{ (Failure)} \\ 0 & \text{otherwise (failure)} \end{cases}$$

$p(\text{success}) = p$

$p(\text{Failure}) = q = 1-p$

- Suppose an experiment has only two possible outcomes, "success" & "Failure" and let π be the probability of success.

- if we let X denote the number of successes, the X will be Bernoulli distribution.

$$P(X=x) = p^{x}(1-p)^{1-x}$$

$$P(x) = \begin{cases} p & \text{if } X = 1 \\ q & \text{if } X = 0 \end{cases}$$

Q.(5) Multinomial Distribution :-

- The multinomial distribution is a generalization of the binomial distribution to k categories instead of just binary (success / Fail).

- For n independent trials each of which leads to a success for exactly one of k categories, the multinomial distribution gives the probability of any particular combination of numbers of successes for the various categories.

- The multinomial distribution can be used to compute the probabilities in situations in which there are more than two possible outcomes.

eg Rolling a die N times.

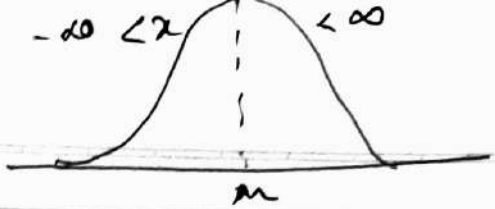
Q.(6) Normal / Gaussian Distribution :-

- Gaussian distribution is also called Normal Distribution.

- It is defined for continuous random variables. The PDF for a Gaussian random variable is given as,

Gaussian PDF : $f_X = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-m)^2/2\sigma^2}$ Here, M is Mean.

diagram of Gaussian Distribution:



Continuous Nature

Q. ⑦ Likelihood Probability :-

- In machine learning, likelihood probability is a function that describes how likely a specific data point is to fit an existing data distribution.

- likelihood is used when given some results, one wants to know how likely is that those results fit a specific distribution.

- Eg. in a coin toss, the likelihood of obtaining a "heads" outcome in a single toss, assuming the coin is fair, is 0.5 since there are two equally likely possibilities (heads or tails).

- likelihood can be determined by applying the equation,

$$L(\text{conditions}) = P(\text{event}).$$

Q. ⑧ posterior Probability :-

- in machine learning, posterior probability is a revised probability that takes into account new information.

- It is a type of conditional probability that results from updating the prior probability with information summarized by the likelihood via an application of Bayes's rule.

- posterior probability represents the revised belif or confidence in a hypothesis or event after considering observed data.

- It combines the prior knowledge or belifs with the new evidence provided by the data to obtain an updated probability estimate.

- For example, let there be two urns, urn A having 5 black balls and 10 red balls and urn B having 10 black balls and 5 red balls.

- Now if an urn is selected at random, the probability that urn A is choosen is 0.5

- The formula for calculating posterior probability is

$$P\left(\frac{A}{B}\right) = \frac{P(A \cap B)}{P(B)} = \frac{P(A) * P(B/A)}{P(B)}$$

- A & B are events
- P(B/A) is the probability of B occuring given that A is true
- P(A) & P(B) are the probabilities of A & occuring & B occuring independently of each other.

Q. (9) Monte Carlo Approximation :-

- Monte Carlo method is used for drawing a sample at random from the empirical distribution.

- Using the Monte Carlo technique, we can approximate the expected value of any function of a random variable by simply drawing samples from the population of the random variables, and then computing the arithmetic mean of the function applied to the samples.

- These methods are used in cases where analytical or numerical solutions don't exists or are too difficult to implement

- Monte Carlo methods generally follow the following steps :

(1) Determine the statistical properties of possible inputs

(2) Generate many sets of people possible inputs which follows the above properties

(3) perform a deterministic calculation with these sets

(4) Analyze the statistically the results.

- Monte carlo integration uses random sampling of a function to numerically compute an estimate of its integral.

- Suppose that we want to integrate the one-dimensional function f(x) from a to b:

$$F = \int_a^b f(x)\,dx$$

- we can appropiate this integral by averaging samples of the func^n F at uniform random points within the interval.

(*) Some example of Monte carlo Sampling Methods include:

(1) Direct Sampling
(2) Importance Sampling
(3) Rejection sampling

(1) Direct Sampling :-

- sampling the distribution directly without prior information.

(2) Importance Sampling :-

- Sampling from a simpler distribution approximation of the target distribution.

(3) Rejection Sampling :-

- sampling from a broader distribution and only considering samples within a region of the sampled distribution.

* —— * —— *

Chapter : 6 : Bayesian Concept Learning

posterior

① Baye's Theorem :-

- Baye's Theorem is a method to revise the probability of an event given additional information.

- Baye's Theorem calculates a conditional probability called a posterior or revised probability.

- Baye's Theorem is a result in probability theory that relates conditional probabilities.

  - If A & B denotes two events, P(A/B) denotes the conditional probability of A occuring, given that B occurs.

  - The two conditional probabilities P(A/B) and P(B/A) are generally different.

- Baye's Theorem gives relation Between P(A/B) and P(B/A). An important application of Baye's Theorem is that :

  - it gives a rule how to update or revise the strengths of evidence-based beliefs in light of new evidence a posteriori.

②

$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$ → Prior pro.

likelihood pro. → Marginal pro.

pro.

P prior probability is an initial probability value originally obtained before any additional information is obtained.

A posterior probability is a probability value that has been revised by using additional information that is later obtained.

Suppose that $B_1, B_2, \ldots B_n$ partition the outcomes of an experiment and that A is another event.

- For Any number k, with $1 \leq k \leq n$, we have the Formula:

$$P\left(\frac{B_k}{A}\right) = \frac{P(A|B_k) \cdot P(B_k)}{\sum_{i=1}^{n} P(A|B_i) \cdot P(B_i)}$$

* * *

Consistent learners & Baye's optimal classifier.

(*) consistent learners :-

- The group of learners who commit zero error over the training data and output the hypothesis are called "consistent learners".

- if the training data and is noise free and deterministic and if there is uniform prior probability distribution over H. then every consistent learner outputs the MAP Hypothesis.

(*) Bayes Optimal Classifier :-

- Baye's classifier is a classifier that minimizes the error in a probabilistic manner.

- if it is Baye's optimal, then the errors are weighted using the join probability distribution between the input and output sets.

- the Baye's error is then the error of the Baye's classifier.

Q.(3) Naive Baye's Classifier :-

- Naïve Bayes classifier are a family of simple probabilistic classifiers based on applying Baye's theorem with strong independence assumptions between the features.

- it is highly scalable, requiring a number of parameters linear in the number of variables in a learning problem.

A Naive Bayes classifier is a program which predicts a class value given a set of attributes.

For each known class value,

1. Calculate probabilities for each attribute, conditioned on the class value.

2. Use the product rule to obtain a joint conditioned probability for the attributes.

3. Use Bayes rule to obtain a joint conditioned probability for the attributes.

- Once this has been done for all class values, output the class with the highest probability.

- A key benifit of Naive Bayes classifier is that it requires only a little bit of training information to guage the parameters essential for the classification.

(*) Advantages :-

1. simple to implement.
2. Calculation is fast & produce effective result.

3. suitable for noisy & missing data
4. works well for small number of data.

Q.(4) Bayesian Belief Network :-

- "A Bayesian network is a probabilistic graphical model which represents a set of variables and their conditional dependencies using a directed acyclic graph."

- it is also called Bayes Network, decision Network, belief Network or Bayesian Model.

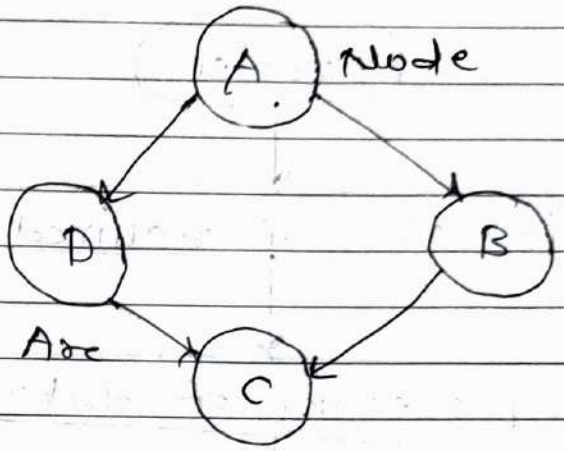
- Bayesian Belief Network are probabilistic because these networks are built from a probability distribution, and

- Bayesian Network can be used for building models from data & experts opinions, and it consists of two parts:

  (1) Direct Acyclic Graph
  (2) Table of conditional Probabilities

- The generalized form of Bayesian Belief Network that represents & solve decision problems under certain knowledge is known as an Influence Diagram.

- A Bayesian Network graph is made up of nodes & Arcs (directed links), where:



- Each node corresponds to the random variables, and a variable can be continuous or discrete.

- Arc or directed arrows represent the causal relationship or conditional probabilities between random variables.

  - In the above diagram A, B, C & D are random variables represented by the nodes, the network graph.

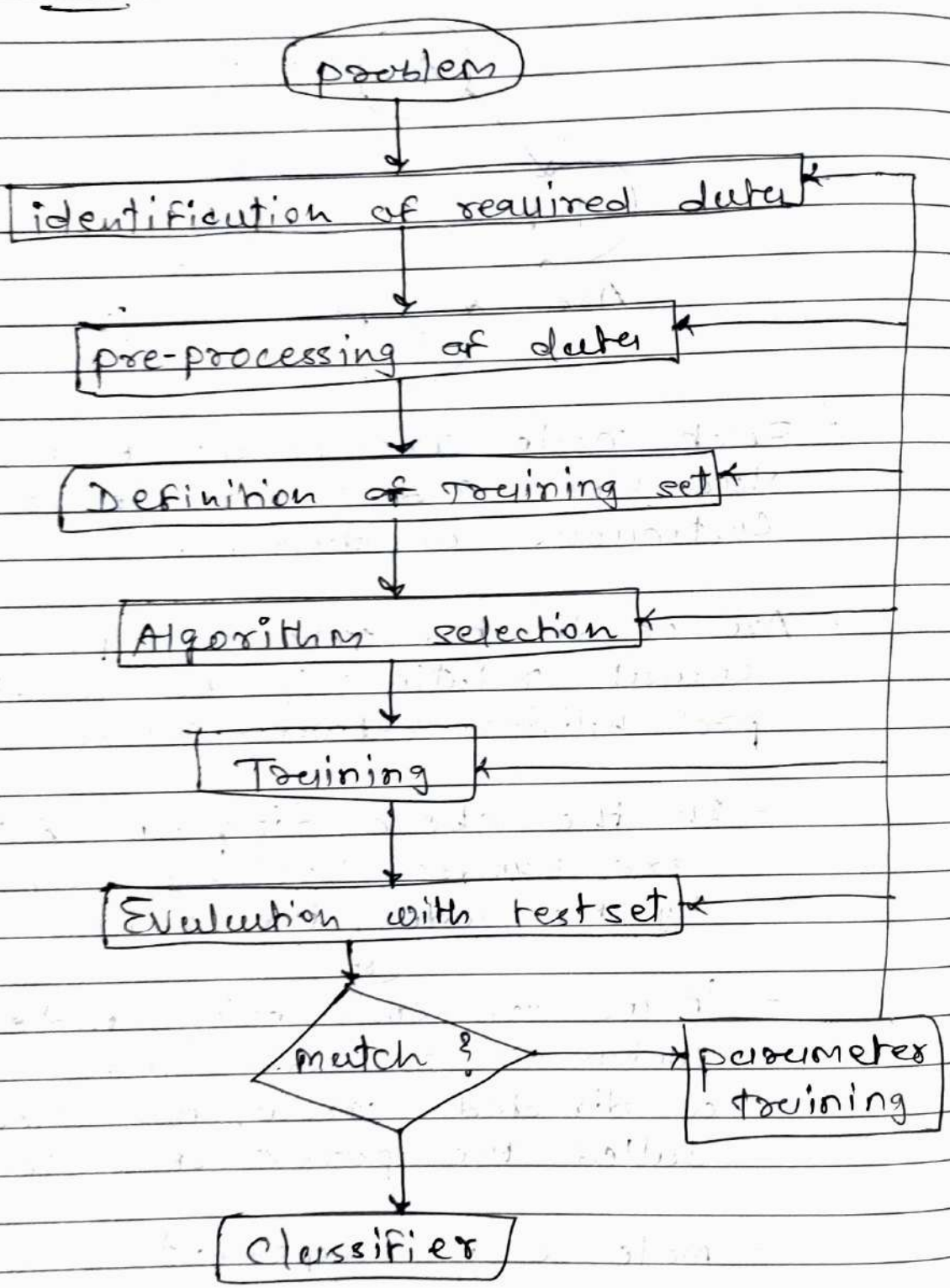
  - if we are considering node B, which is connected with node A by a directed arrow, then node A is called the parent of Node B.

  - Node C is independent of node A.

- The B.N has mainly two components:
  (1) Casual component (2) Actual numbers

⊛ chapter : ⑦ : supervised learning ⊛

Q.① Process of supervised Learning model :-



⊛ supervised learning model process ⊛

⑥ Steps of S.learning model :-

(1) Problem identification
(2) Identification of Required data
(3) Data pre-processing
(4) Definition of training data set
(5) Algorithm selection
(6) Training
(7) Evaluation with the test data set

---

[1] Problem identification :-

- First step of supervised learning is problem identification.
- problem statement must be well defined it contains goals and benefits.

[2] Identification of Required data :-

- the required data set that precisely represents the identified problem needs to be identified / evaluated.

[3] Data Pre-Processing :-

- This is related to be cleaning/transforming the data set.
- This step ensures that all the unnecessary / irrelevant data elements are removed.

[4] Definition of training data set :-

- Before starting the analysis, the user should decide what kind of data set is to be used as a training set.

[5] Algorithm selection :-

- This involves determining the structure of the learning function & the corresponding learning algorithm.

[6] Training :-

- The learning algorithm identified is run on the gathered training set for further fine tuning.

[7] Evaluation with the test data set :-

- training data is run on the algorithm, and its performance is measured here.

Q.② Classification Algorithms :-

(1) K-Nearest Neighbour (KNN)
(2) Decision Tree
(3) Support Vector Machine

Page No.
Date

⊛ K-Nearest Neighbor (KNN) ⊛

- K-nearest neighbour is one of the simplest ML algorithms based on supervised learning technique.

- KNN algorithm assumes the similarity between the new data & available cases & put the new data into the category that is most similar to the available categories.

- KNN algorithm stores all available data and classifies a new data point based on the similarity.

- KNN is a Non-parametric algorithm, which means it does not make any assumption on underlying data.

- KNN is also called lazy learner algorithm because it doesn't learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.

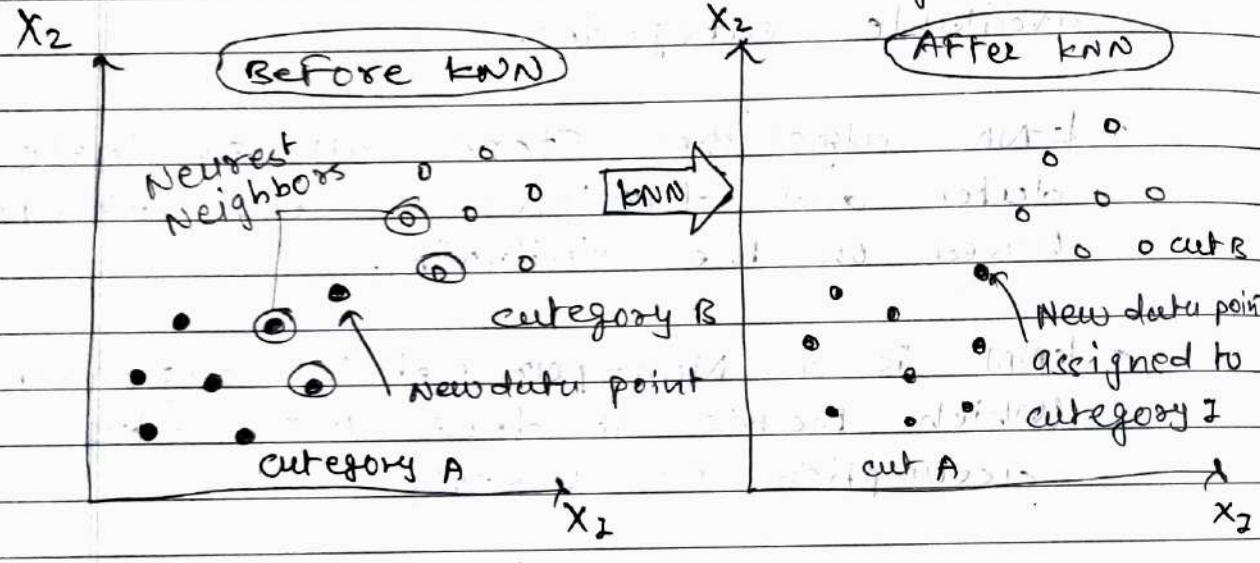
Example :

- Suppose there are two categories, i.e, category A and category B, and we have a new data point

X_1, so this data point will lie in which of these categories.

- To solve this type of problem, we need a KNN algorithm. With the help of KNN, we can easily identify the category or class of a particular dataset.

- Consider the below diagram:



(*) Working steps of KNN :-

(1) Select the number ~~of~~ k of the Neighbors

(2) Calculate the Euclidean distance of k number of neighbors

(3) Take the k-Nearest neighbors as per the calculated Euclidean Distance.

(4) Among these k Neighbors, Count the number of data points in each category.

(5) Assign the new data point to that category for which the number of the nig neighbor is maximum.

(6) Our model is ready.

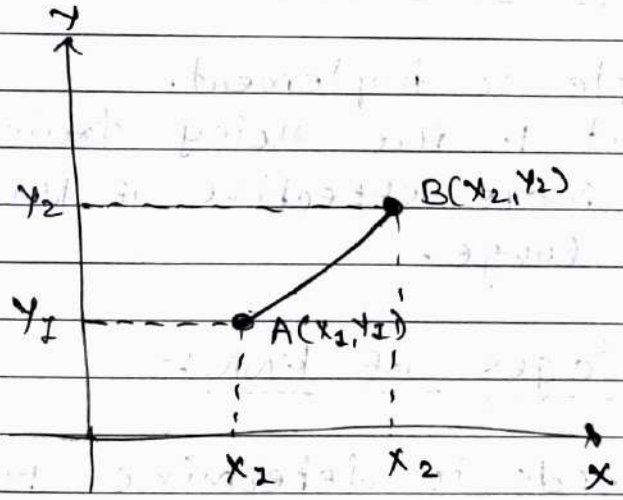
⊛ Distance Metrics used in KNN :-

(1) Euclidean Distance
(2) Manhattan Distance
(3) Minkowski Distance

⊛ Euclidean Distance :-

- This is nothing but the cartesian distance between the two points which are in the hyperplane.

- $$E.D = \sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2}$$



↓ Euclidean Distance Between A2 and B1.

④ Manhattan Distance :-

- Manhattan Distance metric is generally used when we are interested in the total distance traveled by the object instead of the displacement.

$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

⑤ Minkowski Distance :-

- Minkowski Distance is a special case of Manhattan Distance.

$$d(x, y) = \left(\sum_{i=1}^{n} (x_i - y_i)^p \right)^{1/p}$$

⑥ Advantages of KNN :-

- it is simple to implement.
- it is robust to the noisy training data.
- it can be more effective if the training data is large.

⑦ Disadvantages of KNN :-

- Always needs to determine the value of k which may be complex.
- Computation cost is very High.

Decision Tree

- Decision tree learning is a method for approximating discrete-valued target functions, in which the learned function is represented by a decision tree.

- A decision tree is a tree where each node represents a feature, each link (branch) represents a decision (rule) and each leaf represents an outcome (categorical or continues values).

- A decision tree or a classification tree is is a tree in which each internal-node is labeled with an input feature.

- The arcs coming from a node labeled with a feature are labeled with each of the possible value of the feature.

- A decision tree has two kinds of nodes,

  1. Each leaf node has a class lable, determined by majority vote of training examples reaching that leaf.

  2. Each internal node is a question on features, it ~~becomes~~ branches out according to the answers.

- A decision tree is a tree where -

a. Each non-leaf node has associated with it an feature (attribute).

b. Each leaf node has associated with it a classification (+ or -)

c. Each arc has associated with it one of the possible values of the attribute at the node from which the arc is directed.

- A decision tree is a flow-chart like tree structure, where each node denotes a test on an attribute value, each branch represents an outcomes of the test & tree leaves represents classes or class distribution.

* There are several steps involved in Building of decision tree :-

(1) Begin the tree with the root node, says S, which contains the complete dataset.

(2) Find the best attribute in the dataset using Attribute Selection Measure (ASM).

(3) Divide the S into subsets that contains possible values for the best attributes

(4) Generate the decision tree node, which contains the best attribute.

(5) Recursively make new decision tree using the subsets of the dataset ~~of~~ the dataset created in step-3.

- Continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf node.

Example :-

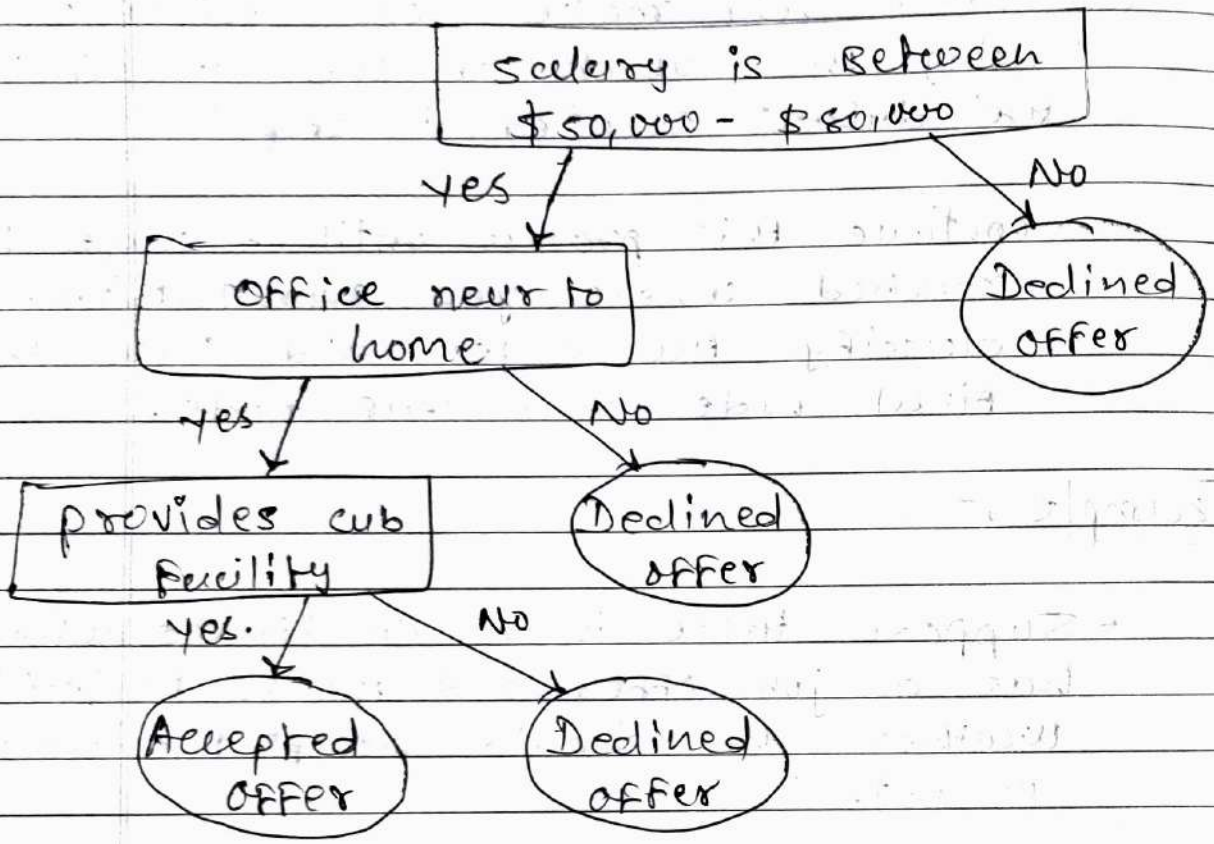
- Suppose there is a candidate who has a job offer and wants to decide weather he should accept the offer or Not.

- To solve this problem, the decision tree starts with the root node (salary attributes by ASM).

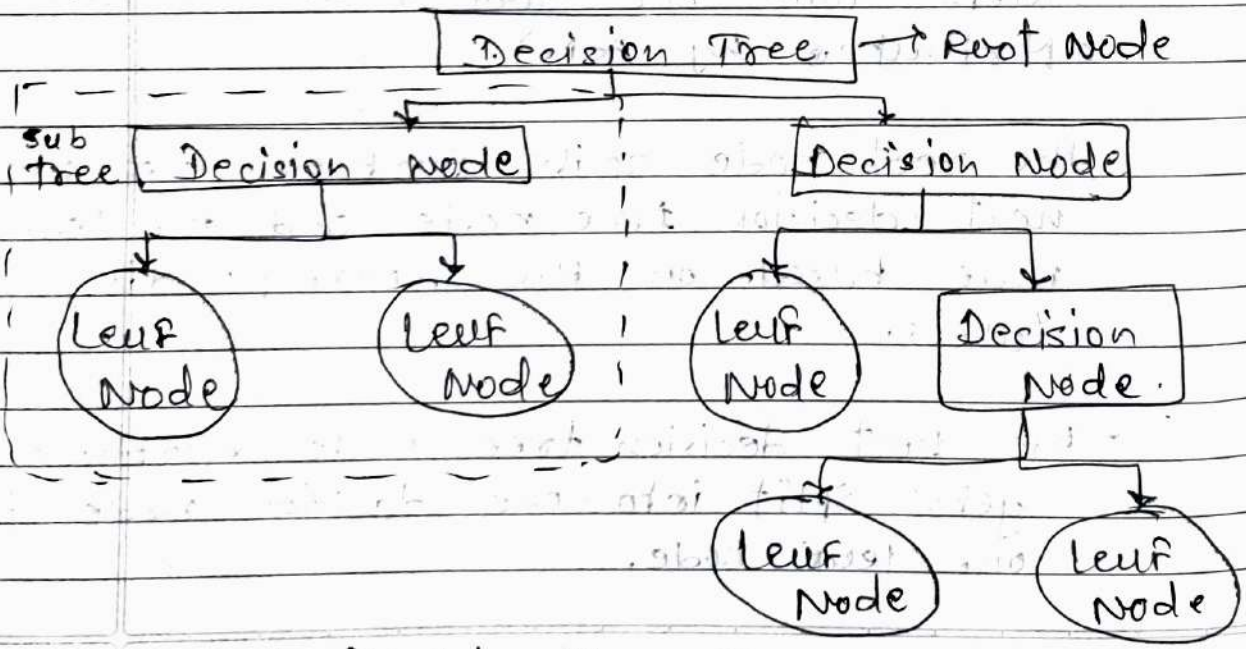
- The root node splits further into the next decision ~~tree~~ node and one leaf node based on the corresponding labels.

- The next decision tree node further gets split into one decision node & one leaf node.

- Finally, the decision tree node splits into two leaf nodes (Accepted offers & Declined offers).



★ Decision Tree ★



★ Basic Decision Tree ★

⊛ Attribute Selection Measures ⊛

- While implementing a Decision Tree, the main issue arises that how to select the best attribute for the root node and for sub-nodes.

- So, to solve such problems there is a technique which is called as Attribute selection measure or ASM.

- There are two popular techniques for ASM, which are:

  (1) Information Gain
  (2) Gini Index

⊛ Information Gain :-

- Information Gain is the measurement of changes in entropy after the segmentation of a dataset based on an attribute.

- It calculates how much information a feature provides us about a class.

Information Gain = Entropy (S) − [(Weighted Avg) * Entropy (Each Feature)]

⊛ Entropy :-

- Entropy is a metric to measure the impurity in a given attribute.
- It specifies randomness in data. Entropy can be calculated as :

$$\text{Entropy}(S) = -p(\text{yes}) \log_2 p(\text{yes}) - p(\text{no}) \log_2 p(\text{no})$$

where,

S = Total number of samples
P(yes) = probability of yes
P(no) = probability of no

⊛ GINI Index :-

- Gini index is a measure of impurity or purity used while creating a decision tree in the CART (classification & Regression Tree) algorithm.
- An attributes with the low Gini index should be preferred as compared to the high Gini index.
- Gini index can be calculated using the below formula :

$$\text{Gini Index} = 1 - \Sigma_j P_j^2$$

Tree

(*) Pruning :- Getting an optimal Decision

- Pruning is a process of deleting the unnecessary nodes from a tree in order to get the optimal decision tree.

- A too large tree increase the risk of overfitting, and a small tree may not capture all the important features of the dataset.

- there are mainly two types of tree Pruning technology used:

  (1) Cost Complexity Pruning
  (2) Reduced Error Pruning.

(*) Advantages of the Decision Tree :-

- it is simple to understand as it follows the same process which a human follow while making any decision in real life.

- it can be very useful for solving decision - related problems.

(*) Disadvantages of the Decision Tree :-

- the decision tree contains lots of layers, which makes it complex.

- it may have an overfitting issue, which can be resolved using the Random Forest Algorithm.

⊛ Support Vector Machine ⊛

- Support Vector Machine or SVM is one of the most popular Supervised Learning Algorithms, which is used for classification as well as Regression problems.

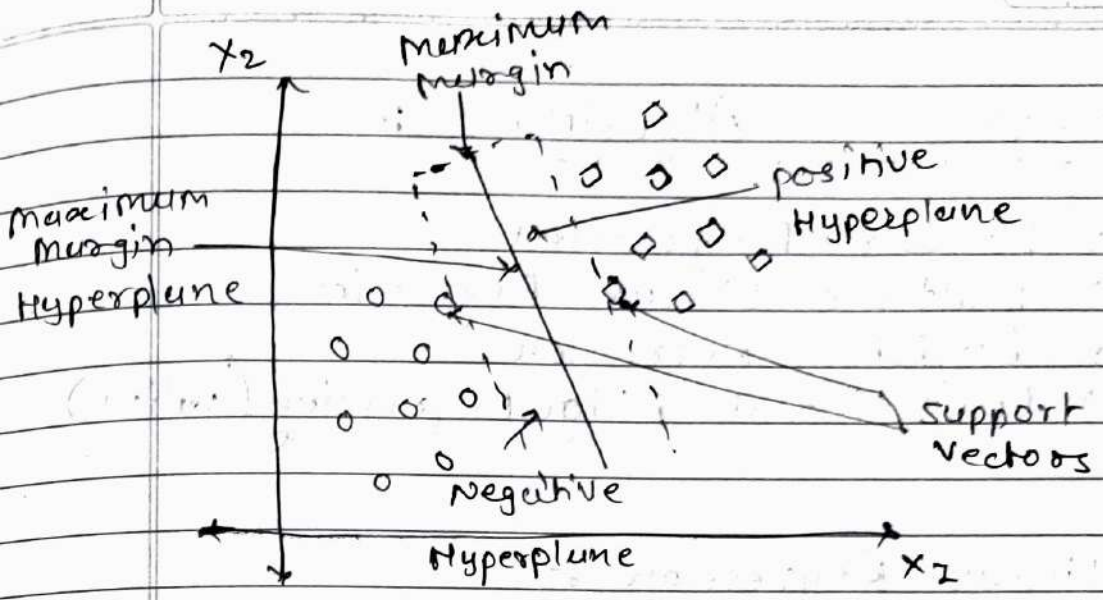
- The goal of the SVM algorithm is to create the best line or decision boundry that can segregate n-dimensional space into classes so that we easily put the new data point in the correct category in the future. This is the Best Decision Boundry is called a "Hyperplane".

- SVM chooses the extreme vectors that help in creating the hyperplane. These extremes cases are called as support vectors, and hence algorithm is termed as Support Vector Machine.

- Consider the below diagram in which there are two different categories that are classified using a decision boundry or hyperplane:

- SVM algorithm can be used for

  - Face detection
  - Image classification
  - Text categorization.



② SVM can be two types :-

(1) Linear SVM :-

- linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier.

(2) Non-linear SVM :-

- Non-linear SVM is used for non-linearly separated data, which means if a dataset cannot be classified by using a straight line, then such data is termed as non-linear data and classifier used is called as Non-linear SVM classifier.

Q.3 Regression Algorithms :-

(1) Linear Regression
(2) Multiple Linear Regression
(3) Logistic Regression
(4) Lasso and Ridge Regression (online)

① Linear Regression :-

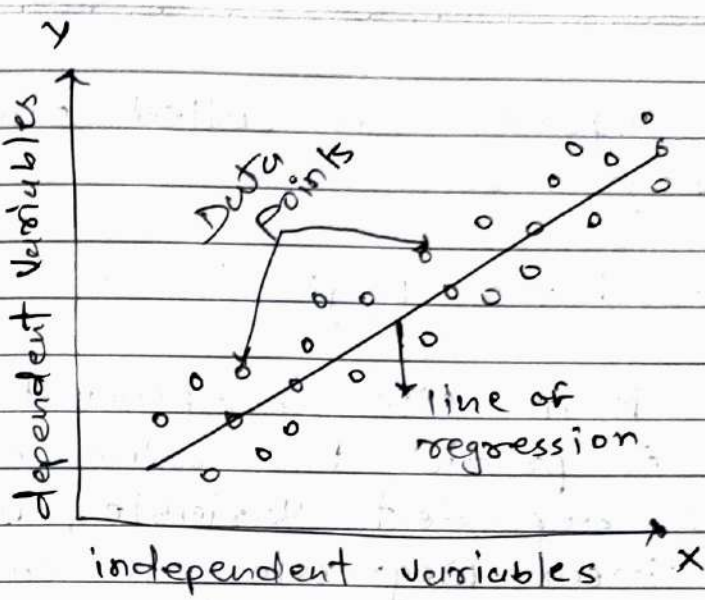
- linear Regression is one of the most popular & easiest ML algorithm.

- it is a statistical method for predictive analysis. linear regression makes predictions for continuous/real or numeric variables such as sales, salary, age, product, price, etc.

- linear Regression algorithm shows a linear relationship between a dependent (y) and one or more independent (x) variables, hence called as linear Regression.

- linear Regression shows the linear Relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

- the linear regression model provides a sloped straight line representing the relationship between the variables.



- Mathematically, we can represent a linear regression as:

$$y = a_0 + a_1 x + \varepsilon$$

where,

y = Dependent Variable (Target Variable)
x = independent Variable (Predictor Variable)
a_0 = intercept of the line
a_1 = linear regression coefficient
ε = random error

- The value for x & y are training datasets For Linear Regression (variables) Model representation.

⊛ Types of Linear Regression :-

(1) Simple Linear Regression :-

- If a single independent variable is used to predict the value of a numerical dependent variable, then such a linear

regression algorithm is called simple Linear Regression.

(2) Multiple linear Regression :-

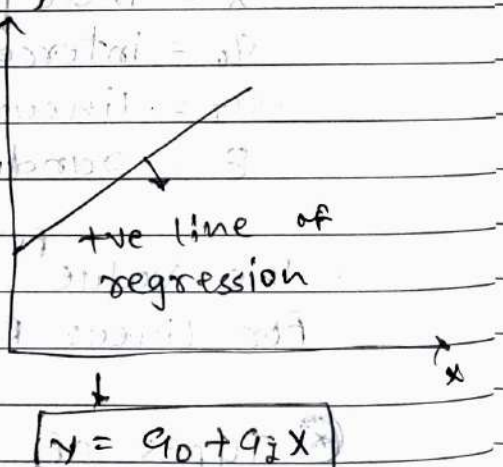
- if more than one independent variable is used to predict the value of a numerical dependent variable, then such a linear Regression is called Multiple Linear Regression.

⊛ Linear Regression Line :-

- A linear line showing the relationship between the dependent & independent variables is called a regression line.

⊛ Positive Linear Relationship :-

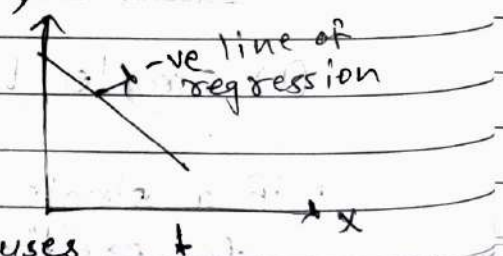
- if the dependent variable increase on the Y-axis and independent variable increases on X-axis, then such a relationship is termed as a positive linear relationship.



$Y = a_0 + a_1 X$

⊛ Negative linear Relationship :-

- if the dependent variables decreases on the Y-axis & independent variables increases on X-axis, then such ...negative...ship.

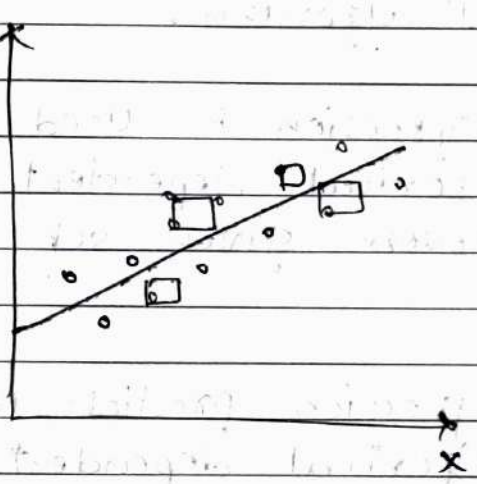


$Y = -a_0 + a_1 X$

⑧ Sum of Square :-

- Sum of squares (SS) is a statistical tool that is used to identify the dispersion of data as well as how well the data can fit the model in regression analysis.

- The sum of squares got its name because it calculated by finding the sum of squared differences.



- The regression sum of squares describes how well a regression model represents the modeled data.

- A Higher regression sum of squares indicates that the model does not fit the data well.

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

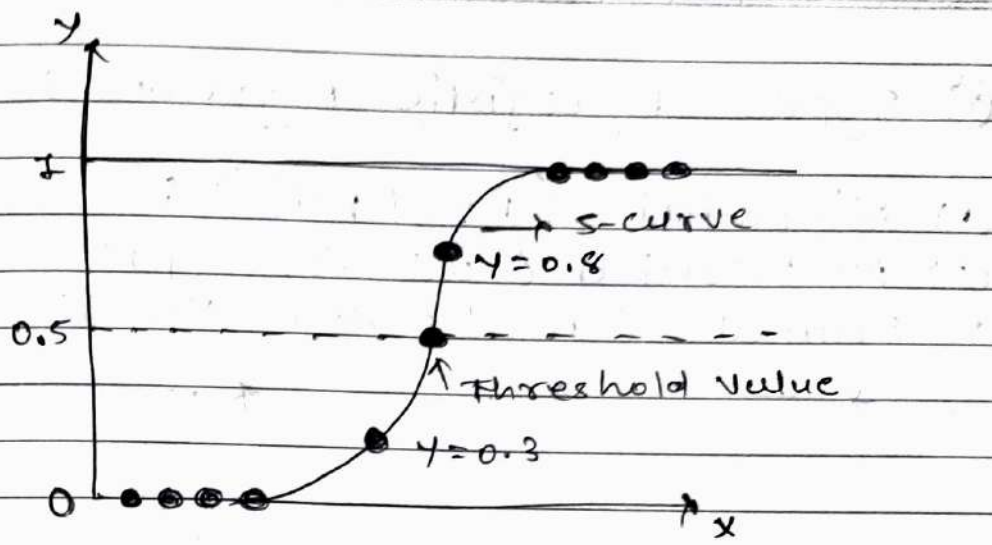
where,

$\hat{y}_i$ = the value estimated by the regression line

$\bar{y}$ = the mean of value of a sample

(*) Logistic Regression :-

- logistic Regression is one of the most popular ML algorithm.
- logistic Regression is used for predicting the categorical dependent variable using a given set of independent variables.
- logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. it can be either Yes or No, 0 or 1, true or false, etc. but instead of giving the exact value as 0 & 1, it gives the probabilistic values which lie between 0 and 1.
- In logistic Regression, we fit an 'S' shaped logistic function, which predicts two maximum values (0 or 1).



⊛ Sigmoid Function :-

- The sigmoid function is a mathematical function used to map the predicted values to probabilities.

$$y = \frac{1}{1 + e^{-x}}$$

where, y = dependent variable
x = independent variable
e = euler's constant (2.718)

- it maps any real value into another value within a range 0 to 1.

- The value of the logistic regression must be between 0 & 1, which can't go beyond this limit, so it forms a curve like the "S" form. The S-form curve is called the sigmoid function or logistic function.

⊛ Types of Logistic Regression :-

(1) Binomial (Pass / Fail) :-
(2) Multinomial (cat, dog, sheep) :-
(3) Ordinal (Low / Medium (High) :-

⊛ Chapter : 8 : Unsupervised Learning ⊛

Q.① Unsupervised learning Applications :-

(1) Customer segmentation :-

- Unsupervised learning algorithms can group customers based on their purchasing behaviour, allowing business to tailor marketing strategies.

(2) Anomaly Detection :-

- By identifying abnormal patterns or outliers, unsupervised learning can help detect fraud, network intrusions, or manufacturing defects.

(3) Image and Text clustering :-

- Unsupervised learning can automatically group similar images or texts, aiding in tasks like image organization, document clustering, or content recommendation.

(4) Genome Analysis :-

- Unsupervised learning algorithms can analyze genetic data to identify patterns and relationships, leading to insights in personalized medicine & genetic research.

(5) Social Network Analysis :-

- Unsupervised learning can be used to identify communicates or influential individuals within social networks, enabling targeted marketing or deteting online communities.

(*) Unsupervised learning Algorithms :

(1) K-means
(2) K-Mediods
(3) Association Rules → Application of Association Rules
(4) Market-Basket Analysis ↙
(5) The Apriori Algorithm ← type of Association rule

Q.(2) K-means clustering :-

- K-means clustering is an Unsupervised learning Algorithm that is used to solve the clustering problems in ML or Data science.

- K-means clustering is an Unsupervised learning algorithm, which groups the unlabeled dataset into different clusters.

- Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, & for K=3,

there will be there three clusters, & so on.

- it allows us to cluster the data into different groups & a convinient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training.

- it is a centroid-based algorithm where each cluster is associated with a centroid.

- the main aim of this algorithm is to ~~min~~ minimize the sum of distances between the data point & their corresponding clusters.

- the algorithm takes the unlabeled dataset as input, divides the dataset into k-numbers of clusters, & repeats the process until it does not find the best clusters. the value of k should be predetermined in this algorithm.

- this k-means clustering algorithm mainly performs two tasks:

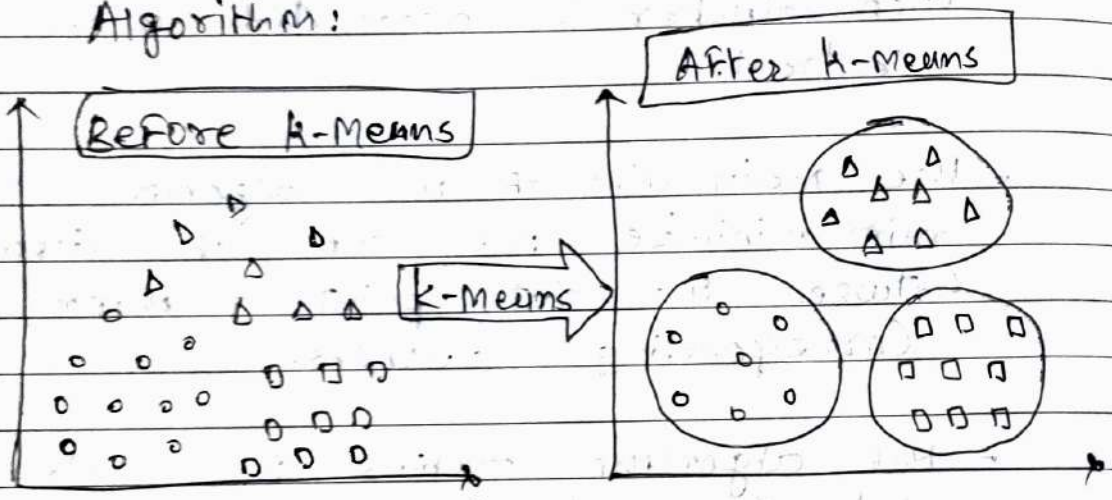
  - Determines the best value for k center points or centroids by an iterative process.

  - Assigns each data point to its closest k-center. those data points which

are near to the particular k-center, create a cluster.

- Hence each cluster has datapoints with some commonalities, & it is away from other clusters.

- The below diagram explains the working of the k-means clustering Algorithm:



② Working of k-means clustering :-

(1) Select the number k to decide the number of clusters.

(2) Select Random k points or centroids.

(3) Assign each data point to their closest centroid, which will form the predefined k clusters.

(4) Calculate the variance & place a new centroid of each cluster.

(5) Repeat the third steps, which means ~~re~~ reassign each datapoint to the new closest centroid of each ~~other~~ cluster.

(6) If any reassignment occurs, then go to step-4 else go to FINISH.

(7) The model is ready.

Q.(3) k-Medoids :-

- k-Medoids (also called partitioning Around Medoid) algorithm was proposed by Kaufman & Rousseeuw.

- A Medoid can be defined as a point in the cluster, whose dissimilarities with all the other points in the cluster are minimum.

- The dissimilarities of the Medoid (C_i) & Object (P_i) is calculated by using

$$E = |P_i - C_i|$$

- The cost in k-Medoids algorithm is given as,

$$C = \sum_{C_i} \sum_{P_i \in C_i} |P_i - C_i|$$

⊛ Algorithm :-

(1) Initialize :

- select k random points out of the n data points as the medoids.

(2) Associate each data point to the data set to point to the closest medoid by using any common distance metrics methods.

(3) while the cost decreases:

- for each medoid m, for each data o point which is not a medoid:

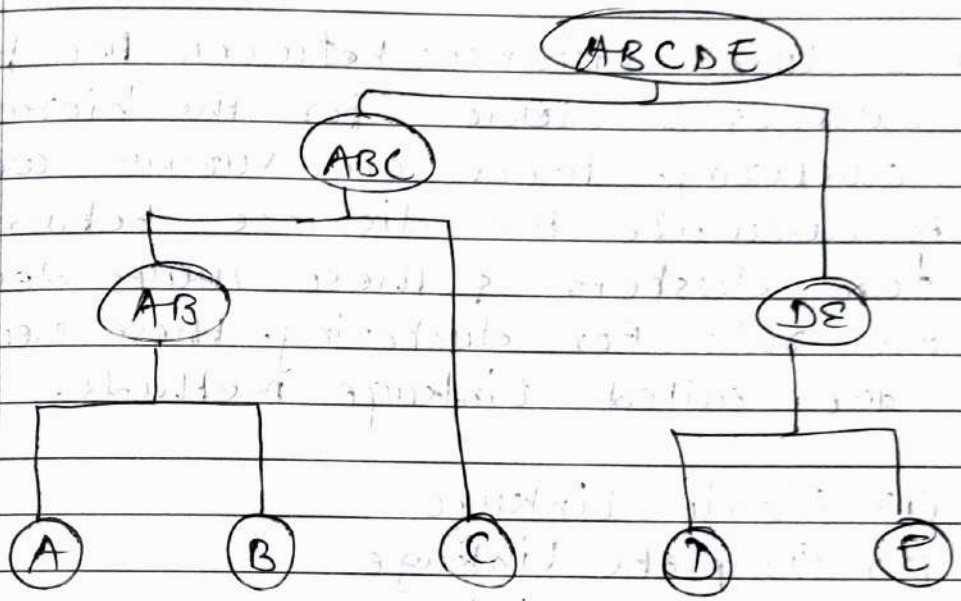
  - swap m & o, associate each data point to the closest medoid, & recompute the cost.

  - if the total cost is more than that in the previous step, undo the swap.

Q. ④ Hierarchical clustering :-

- Hierarchical clustering is used to group the unlabeled datasets into a cluster and also known as hierarchical cluster Analysis. (HCA).

In this algorithm, we develop the hierarchy of clusters in the form of tree, and this tree-shaped structure is known as the dendogram.



(*) dendogram (*)

- The Hierarchical clustering technique has two approaches:

(1) Agglomerative:

- Agglomerative is a bottom-up approach, in which the algorithm starts with taking all data points as single clusters and merging them until one cluster is left.

(2) Divisive:

- Divisive algorithm is ~~a bottom-up~~ the reverse of the agglomerative algorithm

as it is a top-down approach.

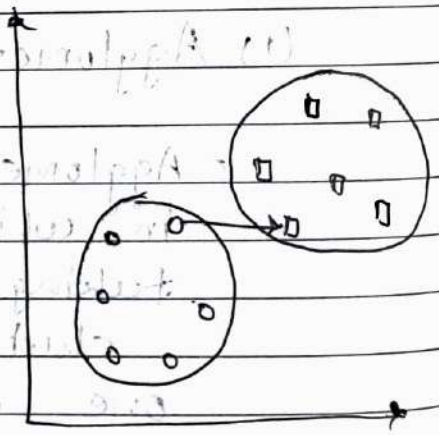
(*) Measure for the distance between two clusters :-

- The closest Distance between the two clusters is crucial for the hierarchical clustering. There are various ways to calculate the distance between two clusters, & these ways decide the rule for clustering. These measures are called Linkage methods.

(1) Single Linkage
(2) Complete Linkage
(3) Average Linkage
(4) Centroid Linkage

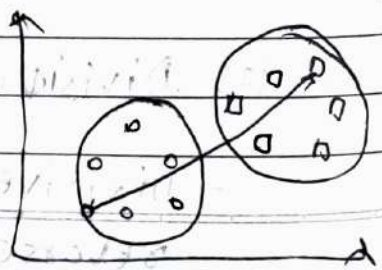
(*) Single Linkage :-

- it is the shortest Distance between the closest point of the clusters.



(*) Complete linkage :-

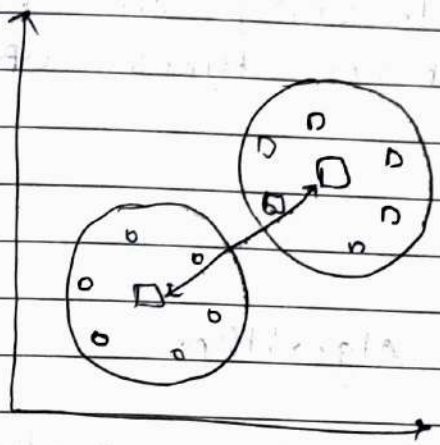
- it is the Farthest distance Between the two points of two different clusters.
it is one of the popular linkage method as it forms tighter clusters than single linkage.



Page No.
Date

(*) Avg :- Avg of distance.

(*) centroid linkage :-



Association Rule :-

Association rule learning is a type of Unsupervised learning technique that checks for the dependency of one data item on another data item & maps accordingly so that it can be more profitable.

- it tries to find some interesting relations or associations among the variable of dataset.
- it is based on different rules to discover the interesting relations between variables in the databases.

The associate rule learning is one of the very important concepts of ML, & it is

employed in Market Basket Analysis, Web usage mining, continuous production, etc. +

- Association rule learning can be divided into three types of algorithms:

(1) Apriori
(2) Eclat
(3) F-P Growth Algorithm

- Association rule learning works on the concept of If and Else statement, such if A then B.

- Here the If element is called antecedent and then statement is called as Consequent.

- These types of relationships where we can find out some association or relation between two items is known as single cardinality.

- It is all about creating rules, and if the number of items increases, then cardinality also increases accordingly so, to measure the associations between thousands of data items, there are several

metrics. these metrics are given below:

(1) Support
(2) Confidence
(3) Lift

⊛ Support :-

- Support is the frequency of A or how frequently an item appears in the dataset.

- It is defined as the fraction of the transaction T that contains the itemset X. if there are X datasets, then for transactions T, it can be written as:

$$\text{Supp}(X) = \frac{\text{freq}(X)}{T}$$

⊛ Confidence:-

- Confidence indicates how often the rule has been found to be true.

- It is the ratio of the transaction that contains X & Y to the number of records that contain X.

$$\text{Confidence} = \frac{\text{freq}(X,Y)}{\text{freq}(X)}$$

⑤ Lift :-

- it is the strength of any rule, which can be defined as below formula:

$$\text{Lift} = \frac{\text{Supp}(X,Y)}{\text{Supp}(X) \times \text{supp}(Y)}$$

⑥ Applications of Association Rules :-

- ⑥ Market Basket Analysis
- ⑥ Medical Diagnosis
- ⑥ Protein Sequence

Q.⑥ Apriori Algorithm :-

- Apriori Algorithm refers to the algorithm which is used to calculate the association rules between objects.

- It means how two or more objects are related to one another. In other words, we can say that the apriori algorithm is an association rule leaning that analyzes that people who bought product A also bought product B.

- The primary objective of the apriori algorithm is to create the association rule between different objects.

- The association rule describes how two or more objects are related to one another. Apriori algorithm is also called frequent pattern mining.

- Apriori algorithm refers to an algorithm that is used in mining frequent products sets and relevant association rules.

- Apriori algorithm helps the customers to buy their products with ease and increases the sales performance of the particular store.

(*) Components of Apriori algorithm :-

- The given three components comprise the apriori algorithm.

(1) Support
(2) Confidence
(3) Lift

Eg Suppose you have 4000 customers transactions in a Big Bazar. you have to calculate the support, confidence, and Lift for two products, and you may say Biscuits and chocolate. This is because customers frequenty buy these two items together.

out of 4000 transactions, 400 contain Biscuits, whereas 600 contain chocolate, and these 600 transactions include a 200 that includes Biscuits and chocolates. ~~This is because customers frequently buy these two items together.~~ using this data find out the support, confidence & lift.

(*) Support :-

- support refers to the default popularity of any product. you find the support as a quotient of the division of the number of transactions comprising that product by the total number of transactions. Hence, we get

Support (Biscuits) = (Transactions relating biscuits) / (Total Transactions)

$= \frac{400}{4000}$ | = 10%

(*) Confidence :-

- Confidence refers to the possibility that the customers bought biscuits and chocolates together.

- so, you need to divide the number of transactions that comprise both biscuits

and chocolates by the total number of transactions to get the confidence.

Hence,

$$\text{Confidence} = \frac{(\text{Transaction releasing both biscuits \& chocolates})}{(\text{Total Transactions involving Biscuits})}$$

$$= \frac{200}{4000} = 50\%$$

- it means that 50 % of the customers who bought biscuits bought chocolates also.

(*) Lift :-

- Lift refers to the increase in the ratio of the sales of chocolates when you sell biscuits. The mathematical equations of lift are given below.

$$\text{Lift} = \frac{(\text{Confidence}(\text{Biscuits} - \text{chocolates}))}{(\text{Support}(\text{Biscuits}))}$$

$$= \frac{50}{10} = 5$$

- it means that the probability of people

buying both biscuits and chocolates together is five times more than that of purchasing the biscuits alone.

- if the lift value is below one, it requires that the people are unlikely to buy both the items together.

- Larger the value, the better is the combination.

Q. (7) (*) Market Basket Analysis :-

- Market Basket Analysis is a data mining technique used by retailers to increase sales by understanding customer purchasing patterns.

- it involves analyzing large data sets, such as purchase history, to reveal product groupings and products that are likely to be purchased together.

- market Basket Analysis is modelled on Association rule.

i.e, the IF ?, THEN ? construct.

Ex if the customer buy Bread, THEN he is likely to buy butter as well.

- Association rules are usually represented as: {Bread} → {Butter}

- some terminologies of MBA:

(1) Antecedent:-

- items or 'itemsets' found within a data are antecedents.

(2) Consequent :-

- A consequent is an items found in combination with the antecedent.

(3) Types of Market Basket Analysis :-

(1) Descriptive MBA
(2) Predictive MBA
(3) Differential MBA

⊛ chapter: ⑨: Neural Network ⊛

Q.① Explain Neural Network & its Types.

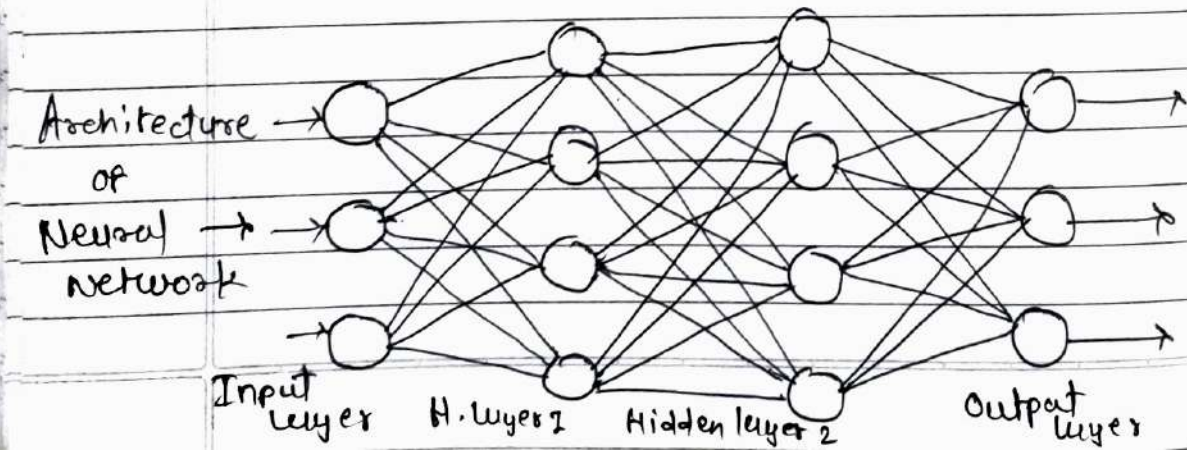
- A Neural Network is a solution that leverages ml algorithms to "mimic" the operations of human brain.

- Neural Network process data more efficiently and feature improved pattern recognition and problem-solving capabilities.

- Neural Networks are also known as Artificial Neural Networks (ANNs) or simulated Neural Networks (SNNs).

- Neural Networks are a subtype of machine learning & an essential element of deep learning algorithms.

- the architecture of Neural Network is also based on the Human brain. it is highly interlinked structure allows it to imitate the sing signaling process of biological neurons.



- the architecture of a Neural Network comprises node layers that are distributed across an input layer, single or multiple Hidden layers, & an output layer.

- Unlike Traditional Computers which process data sequentially, Neural Network can learn and multitask.

② Types of Neural Networks :-

- (1) Perceptron
(2) Feed Forward Neural Network
(3) Multilayer Perceptron
(4) Convolutional Neural Network
(5) Recurrent Neural Network
(6) Long short Term Memory (LSTM)
(7) Modular Neural Network
(8) Generative Adversarial Networks

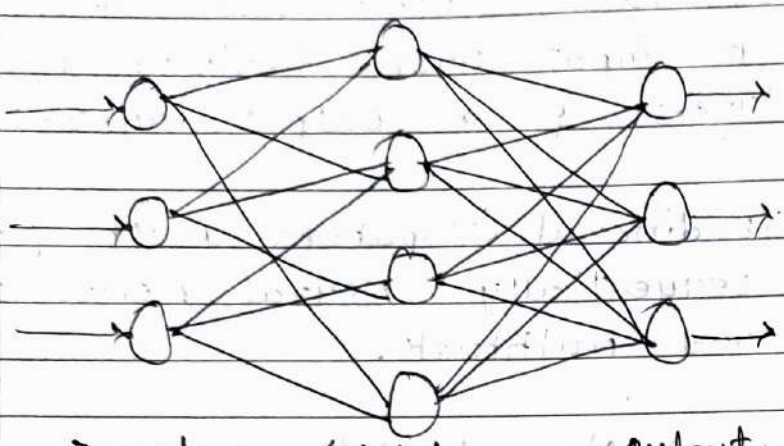
Q.② Explain Feed Forward Neural Network.

- A Feed Forward Neural Network is an artificial Neural Network in which the nodes are connected circularly.

- A Feed-Forward Neural Network, in which some routes are cycled, is the polar opposite of a Recurrent Neural Network.

- A Feed-Forward model is the basic type of neural network because the input

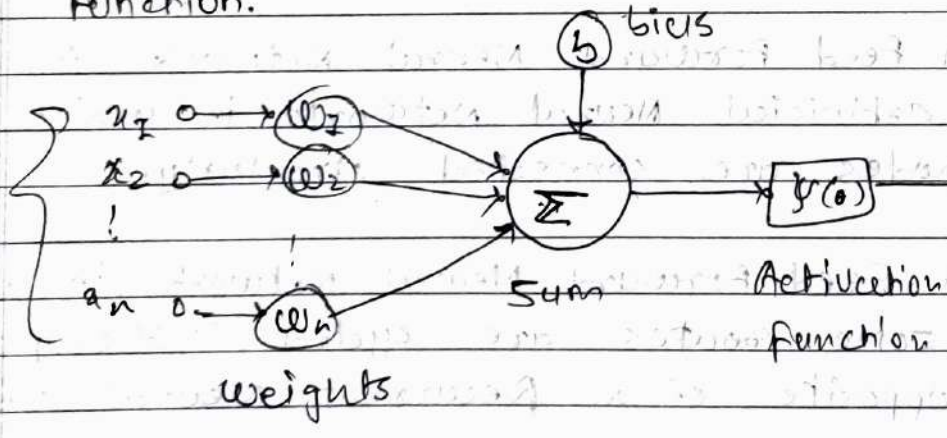
is only proceed in one direction.



- A weight is being applied to each input to an artificial neuron.

- First, inputs are multiplied by their weights, and then a bias is applied to the outcome. This is called weighted sum.

- After that, the weighted sum is proceed via an Activation function, a non-linear function.



⊛ weighted sum of all inputs ⊛

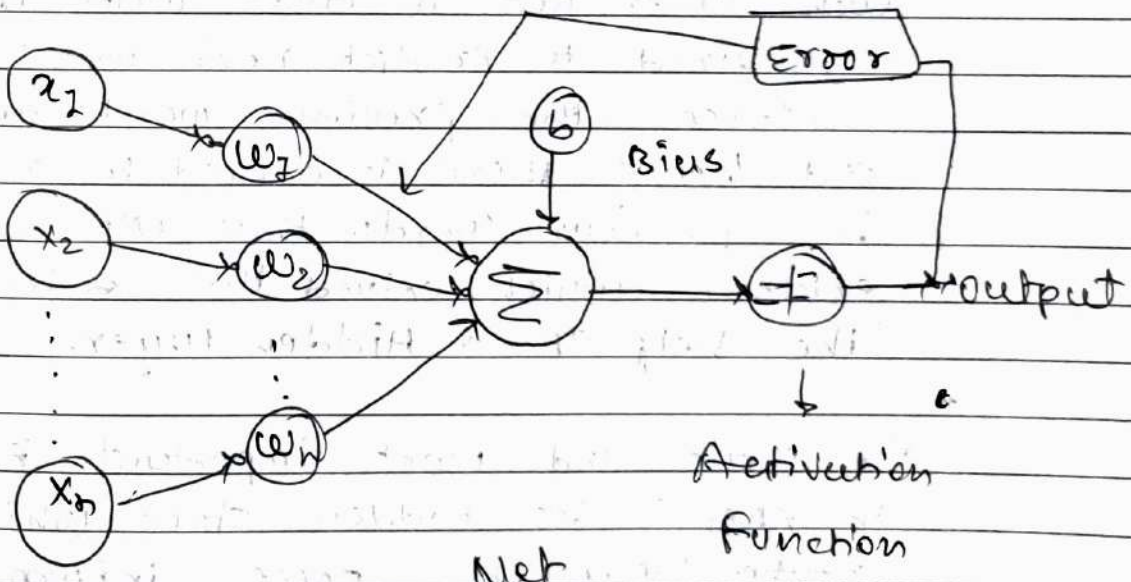
Q. (3) Explain Perceptron.

- Perceptron is a Building Block of Artificial Neural Networks.

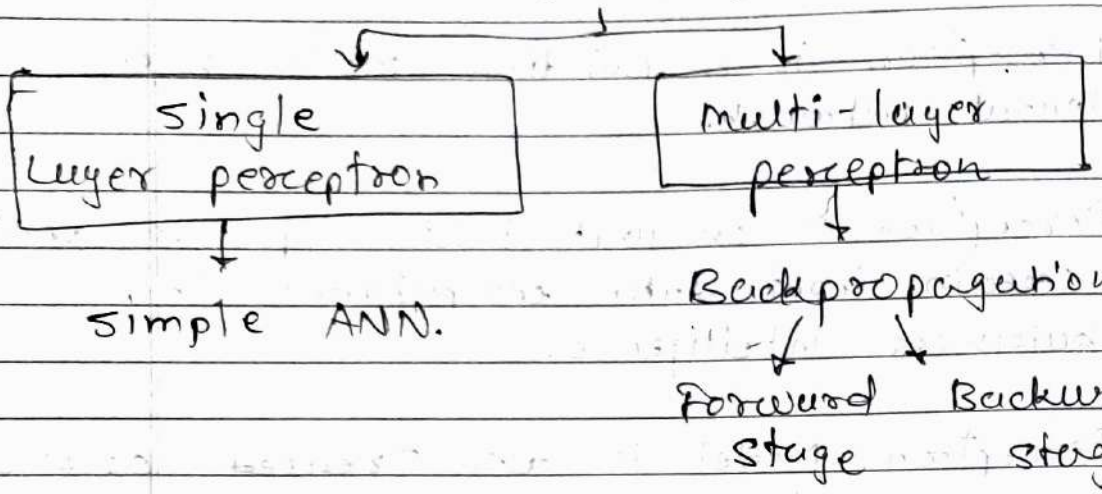
- Perceptron is an unit that helps to detect certain input data computations in buisiness intelligence.

- Perceptron model is also treated as one of the best & simplest type of ANN. However, it is a supervised learning algorithm of binary classifiers. Hence, we can consider as a single-layer neural network with four main parameters:

  - input values
  - weights & Bias
  - net sum
  - Activation Function.



⑧ Types of Perceptron Models :-

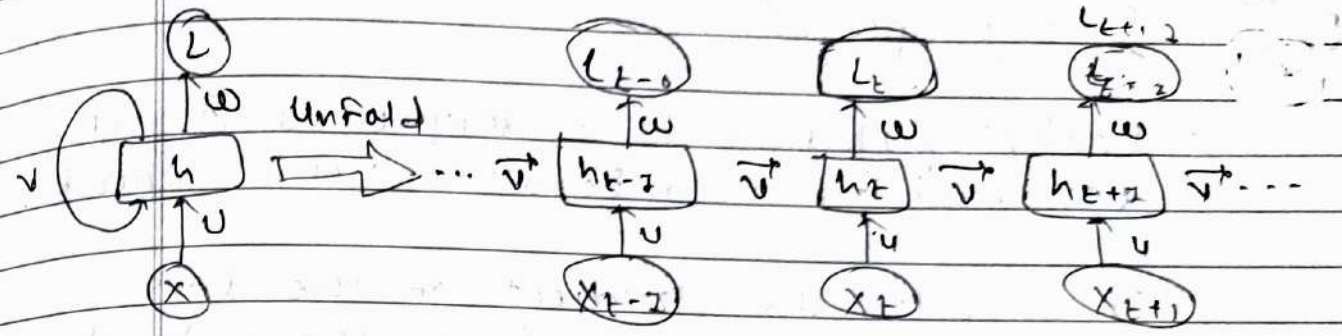


Q. ⑪ Recurrent Neural Network :-

- Recurrent Neural Network is a type of Neural Network where output from the previous step is fed as input to the current step.

- In traditional Neural Networks, all the inputs & outputs are independent of each other, But in cases when it required to predict next word of sentence, the previous words are required and hence there is a need to remember the previous words. Thus RNN came into existence which solved this issue with the help of a Hidden layer.

- The main and most important feature of RNN is its Hidden state, which remembers some information about a sequence.



⊛ RNN Architecture ⊛

⊛ Formula for Hidden state:

$$h_t = F(h_{t-1}, x_t)$$

Where, h_t → current state
h_{t-1} → previous state
x_t → input state

⊛ Formula for Applying Activation Func^n:

$$h_t = \tanh(W_{hh} h_{t-1} + W_{xh} x_t)$$

Where, W_{hh} = weight at recurrent neuron
W_{xh} = weight at input neuron

⊛ The Formula for calculating output:

$$y_t = W_{hy} h_t$$

Where, y_t → output
W_{hy} → weight at output layer

Q.⑤ Backpropagation :-

- Backpropagation is widely used Algorithm for training Feed Forward neural networks.

- It computes the gradient of the loss function with respect to the network weights.

- The backpropagation algorithm works by computing the gradient of the loss function with respect to each weight via the chain rule, computing the gradient layer by layer, and iterating backward from the last layer to avoid redundant computation of intermediate terms in the chain rule.

⊛ training is done in three stages:

(1) Feed-forward of input training pattern
(2) Calculation & backpropagation of the error
(3) Updation of the weight.

⊛ Backpropagation Algorithm :-

online.

↓

Diagram → online

Page No.
Date

Q.⑥ Activation Functions :-

- the activation function decides whether a neuron should be activated or not by calculating the weighted sum & further adding bias to it.

- the purpose of the Activation function is to introduce non-linearity into the output of neuron.

⊛ Types of Activation functions :-

(1) linear function
(2) Sigmoid function
(3) Tanh function
(4) RELU function
(5) Leaky RELU function
(6) Softmax function.

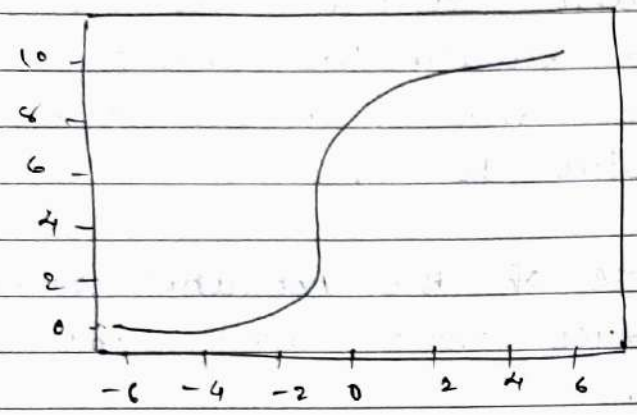
⊛ linear function :-

- linear function has the equation similar to a straight line, $Y = X$.

- it ranges between $-\infty$ to $+\infty$.

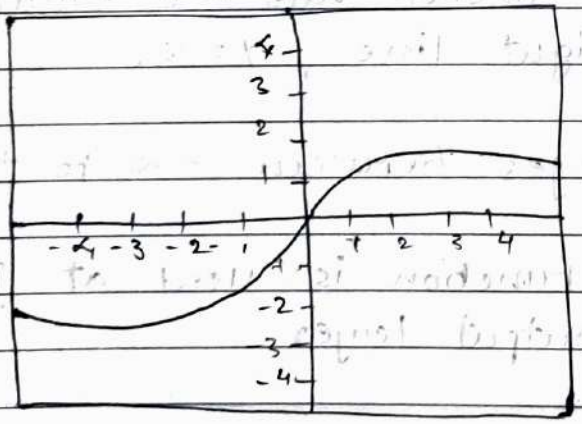
- linear function is used at just one place i.e output layer.

⊛ Sigmoid function :-



- it is a function which is plotted as 'S' shaped graph.
- Equation = $Y = \frac{1}{1+e^{-z}}$
- it is Non-linear in nature.
- usually used for Binary classification where results lies between 0 to 1.

⊛ Tanh function :-

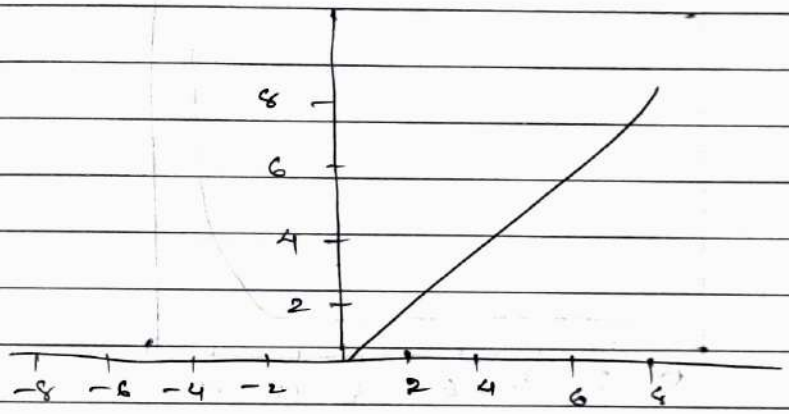


- the activation function that works almost always better than sigmoid function is Tanh function also known as Tangent Hyperbolic function.

$$f(x) = \tanh(x) = \frac{2}{1+e^{-2x}} - 1$$

- it ranges between -1 to $+1$.

(*) RELU function :-



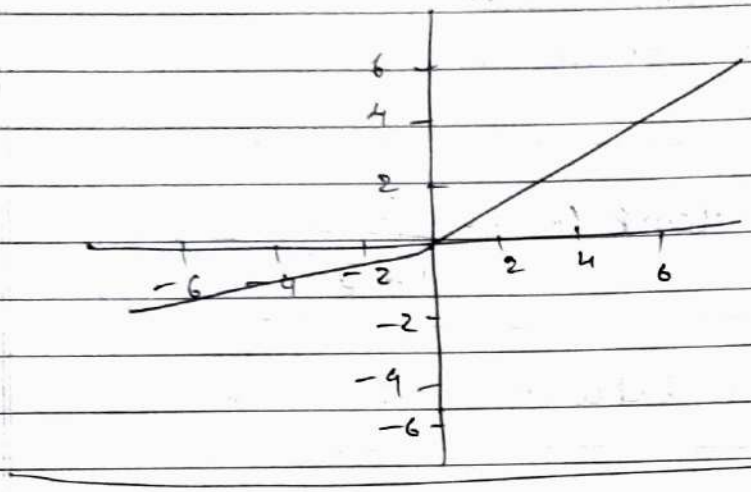
- it stands for Rectified linear unit. it is the most widely used activation function.

$$Y(x) = \max(0, x)$$

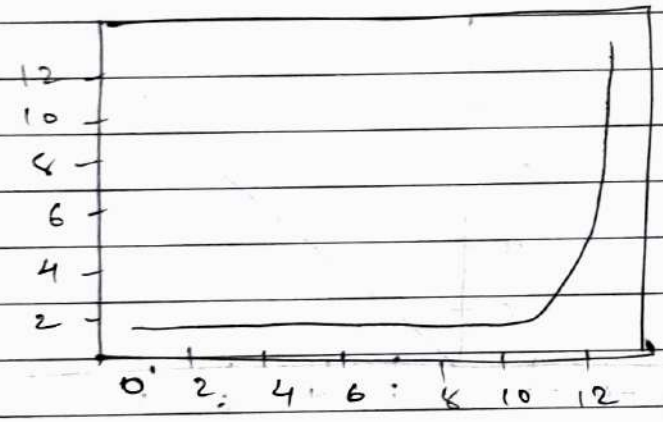
- it ranges between $[0, \infty)$.

- RELU is less computationally expensive than tanh & sigmoid.

⑧ Leaky RELU function :-



⑧ Softmax function :-



- The softmax function is also a type of sigmoid function but is handy when we are trying to handle multi-class classification problems.

- if your output is multiclass classification then, Softmax is very useful to predict the probabilities of each classes.

OTHER QUESTIONS

Singular Value Decomposition :-

- The singular value decomposition (SVD) of a matrix is a factorization of that matrix into three matrices.
- The SVD of m×n matrix A is given by the formula:

$$A = U\Sigma V^T$$

where,

U = m×n matrix of the orthogonal eigenvectors of AA^T.

Σ = diagonal matrix with r elements equal to the root of the positive eigenvalues of AA^T or A^TA.

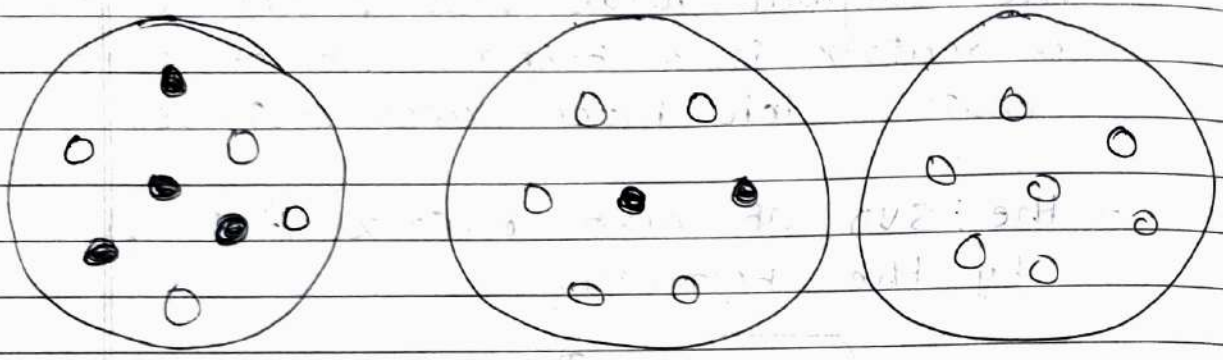
V^T = transpose of a n×m matrix containing the orthogonal eigenvectors of A^TA.

Entropy :-

- Entropy is defined as the randomness or measuring the disorder of the information being processed in ML.

- entropy is ml metric that measures the unpredictibility or impurity in the system.

Very impure Less impure Minimum impure



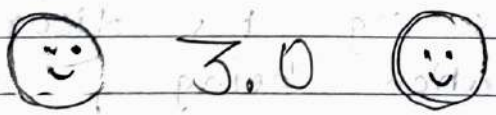
● - impure
O - pure

$$E = -\sum_{i=1}^{N} P_i \log_2 P_i$$

where, P_i = probability of randomly selecting an example in class i.

- Entropy is lies between 0 to 1.

Thank You From Ouh.



⊗ Apriori Algorithm Example ✓

- minimum Support = 50%
- Threshold Confidence = 70%

Ex:

TID	Items
100	1 3 4
200	1 3 5
300	1 2 3 5
400	2 5

(I) →

Itemset	Support	
1	2/4 → 50%	
2	3/4 → 75%	
3	3/4 → 75%	
4	1/4 → 25%	⊗
5	3/4 → 75%	

(Itemset → 1, 2, 3, 5)

(II)

Itemset	Support	
{1, 2}	1/4 → 25%	⊗
{1, 3}	2/4 → 50%	
{1, 5}	1/4 → 25%	⊗
{2, 3}	2/4 → 50%	
{2, 5}	3/4 → 75%	
{3, 5}	2/4 → 50%	

(III)

Itemset	Support	
{1,3,5}	1/4 = 25%	✗
{2,3,5}	2/4 = 50%	
{1,2,3}	1/4 = 25%	✗

(IV)

Rules	Support	Confidence	
(2 ∧ 3) → 5	2	2/2 = 100%	
(3 ∧ 5) → 2	2	2/2 = 100%	
(2 ∧ 5) → 3	2	2/3 = 66%	✗
2 → (3 ∧ 5)	2	2/3 = 66%	✗
5 → (2 ∧ 3)	2	2/3 = 66%	✗
3 → (2 ∧ 5)	2	2/3 = 66%	✗

(*) Confidence :

$= S(A \cup B) / S(A)$

Eg $(2 \wedge 3) \rightarrow 5$

$(A) \rightarrow (B)$

$\frac{S((2 \wedge 3) \cup 5)}{S(2 \wedge 3)}$

$= \frac{2}{2} = 100\%$

Final Rules of Association : $(2 \wedge 3) \rightarrow 5$, $(3 \wedge 5) \rightarrow 2$

⊛ Conditional probability :-

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

$P(A/B) \rightarrow$ probability of Hypothesis given that Evidence is true/occured.

&

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

Hypothesis Evidence

⊛ Bayes Theorem :-

Steps : $P(A/B) \cdot P(B) = P(A \cap B)$ ..(1)

$P(B/A) \cdot P(A) = P(A \cap B)$ ..(2)

Both LHS & RHS is same,

$\therefore P(A/B) \cdot P(B) = P(B/A) \cdot P(A)$

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

$$\therefore P\left(\frac{A}{B}\right) = \frac{P\left(\frac{B}{A}\right) \cdot P(A)}{P(B)}$$

- $P(A)$ → Prior probability
- $P(B)$ → marginal probability
- $P\left(\frac{B}{A}\right)$ → Likelihood probability.
- $P\left(\frac{A}{B}\right)$ → posterior probability

⊛ Naive Baye's Classifier Example :-

Fruit = { yellow, sweet, long }

Fruit	yellow	sweet	long	total
Orange	350	450	0	650
Banana	400	300	350	400
Others	50	100	50	150
Total	800	850	400	1200

N.B.C Formula $= P\left(\frac{A}{B}\right) = \frac{P(B/A) \cdot P(A)}{P(B)}$

$$P\left(\frac{\text{Yellow}}{\text{orange}}\right) = \frac{P\left(\frac{\text{orange}}{\text{Yellow}}\right) \cdot P(\text{yellow})}{P(\text{orange})}$$

$$= \frac{350/800 * 800/1200}{\frac{650}{1200}}$$

$$= 0.5$$

$P(\text{sweet}/\text{orange}) = 0.69$

$P(\text{long}/\text{orange}) = 0$

$P\left(\frac{\text{Fruit}}{\text{Orange}}\right) = 0.53 \times 0.69 \times 0 = 0$

$P\left(\frac{\text{Fruit}}{\text{Banana}}\right) = 1 \times 0.75 \times 0.87 = \boxed{0.65}$

$P\left(\frac{\text{Fruit}}{\text{others}}\right) = 0.33 \times 0.66 \times 0.33 \boxed{= 0.072}$

Highest probability of Fruit is Banana.

(*) Bernoulli N.B.C :-

or Bernoulli Distribution :-

$P(\text{success}) = p$

$P(\text{Failure}) = \bar{x} = 1-p$

$X = 1$ [success]

$X = 0$ [Failure]

'X' has Bernoulli Distribution

- $P(X=x) = p^x \cdot (1-p)^{1-x}$

- $p(x) = \begin{cases} p & \text{if } x = 1 \\ q & \text{if } x = 0 \end{cases}$

- All trails in this distribution is called "bernoulli trails".

⊛ Multinomial N.B.C or Multinomial Distribution :-

- Discrete count

$$P(X_1 = x_1, \ldots, X_k = x_k) = \frac{n!}{x_1! \ldots x_k!} p_1^{x_1} \ldots p_k^{x_k}$$

Blood group	O	A	B	AB
p	0.44	0.42	0.10	0.04

→ 6 indians,

I : O, 2 : A, 2 : B, 1 : AB

$$P(X_1 = 1, X_2 = 2, X_3 = 2, X_4 = 1)$$

$$= \frac{6!}{1!\,2!\,2!\,1!} \cdot (0.44)^1 (0.42)^2 (0.10)^2 (0.04)^1$$

$$= \frac{720}{4} \times 7.76 \times 10^{-6}$$

$$= 0.0014$$

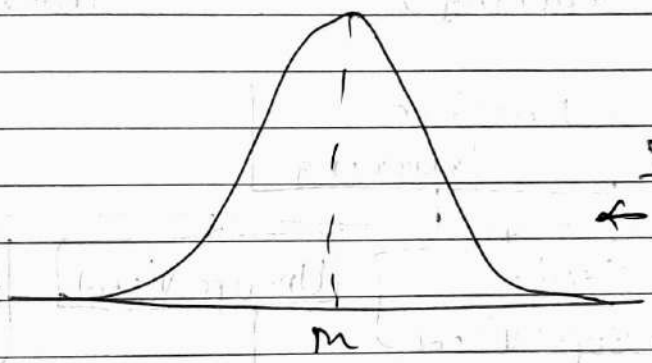
⑧ Gaussian / Normal . N.B.C :-

OR Gaussian / Normal Distribution :-

- Continuous Random Variable.
- probability Density Function (PDF) :

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} \cdot e^{-(x-m)^2/2\sigma^2}$$

where, m = mean
σ = standard deviation.



← Distribution graph

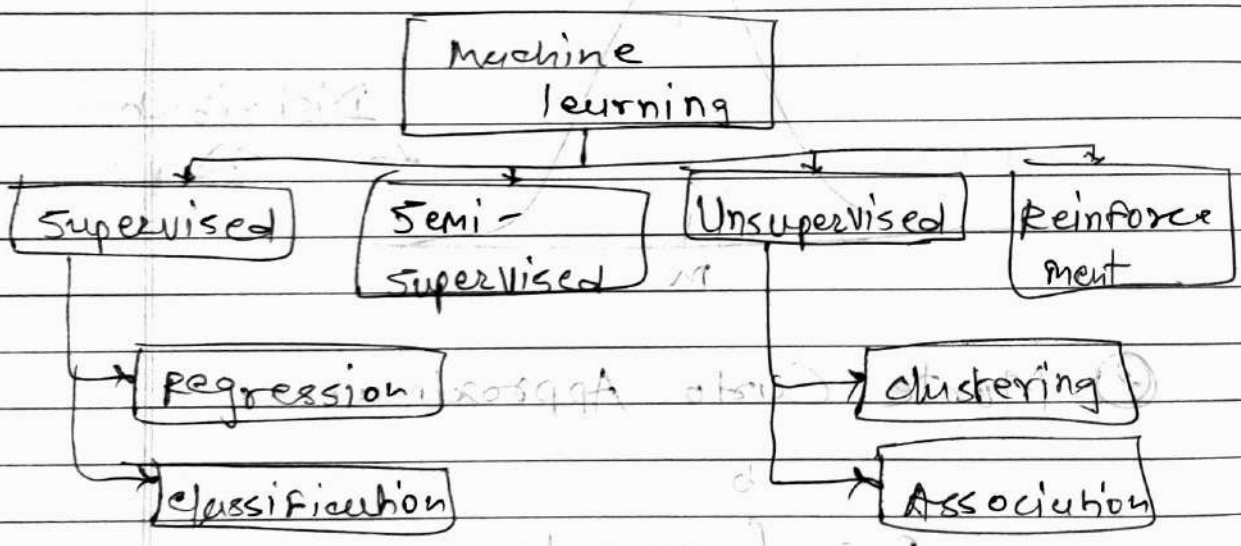
⑧ Monte Carlo Approximation :

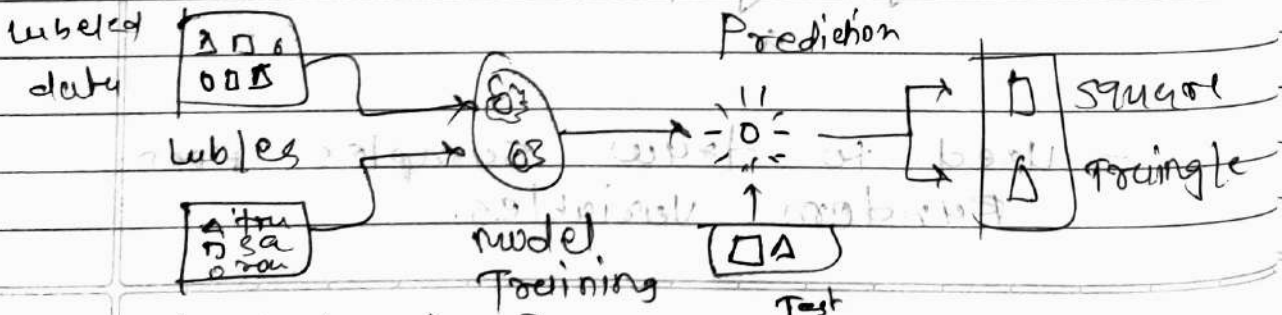
$$I = \int_a^b f(x)\, dx$$

- Used to draw samples from Random Variables.

⊛ Human learning :-

Human learning	Machine learning
Intelligence	models
↓	↓
Learning Materials	Data
↓	↓
Learning skill	skillburn
(model-based & model Free learning)	(knowledge-based learning)





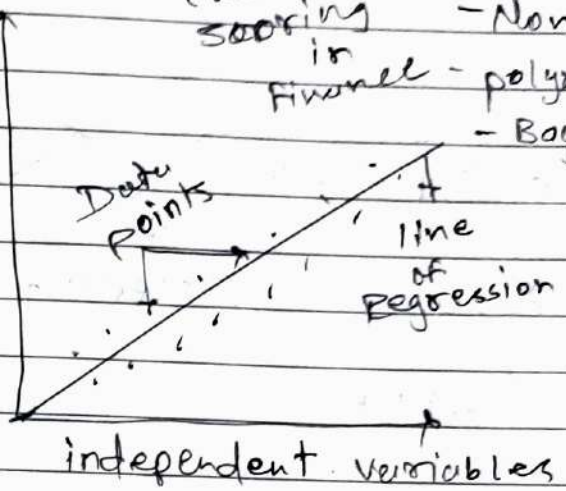
⊛ Supervised learning ⊛

- Stock price prediction
- Health outcome prediction
- Retail sales forecasting

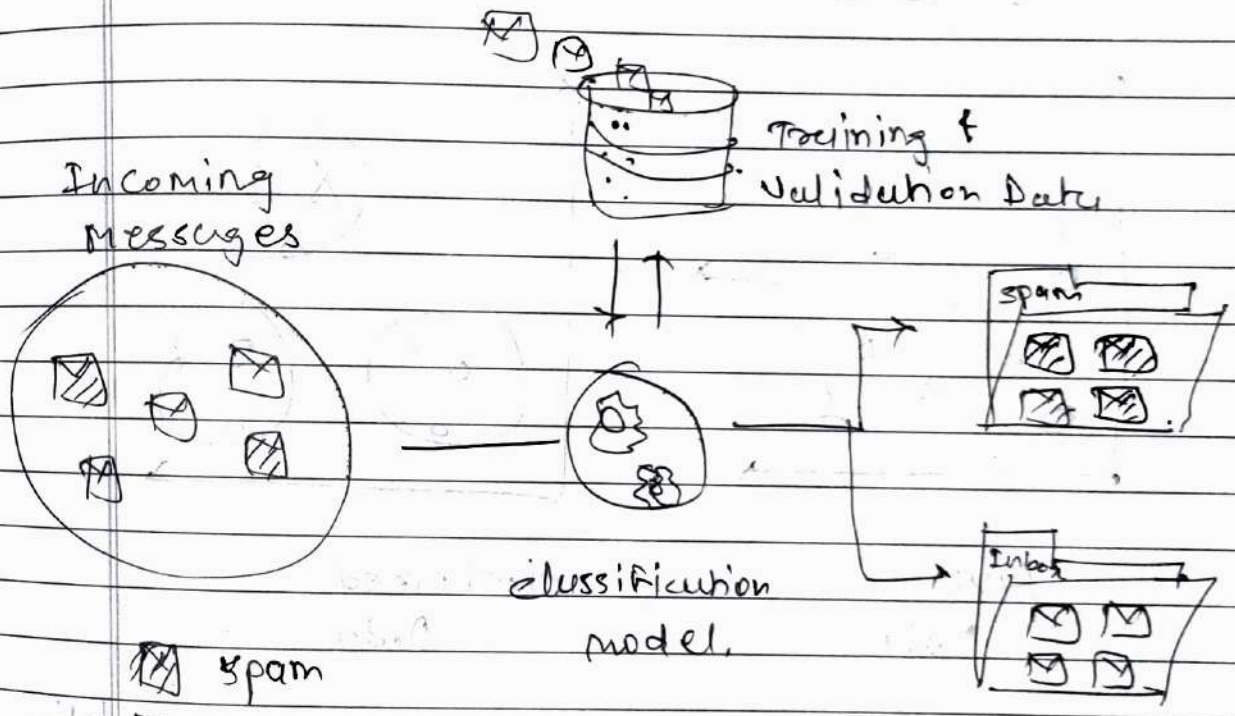
⊛ Regression :-

- Real estate price prediction
- Credit scoring in Finance

- Ridge Regression
- Lasso Regression
- linear regression
- Non-linear regression
- polynomial regression
- Bayesian learning regression.



⊛ Classification :-

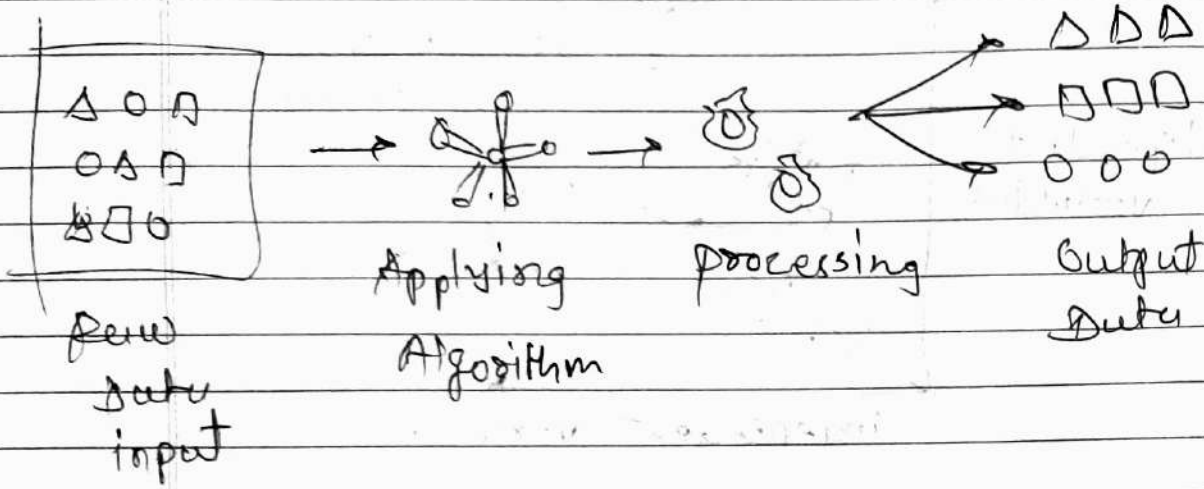


- spam
- No spam

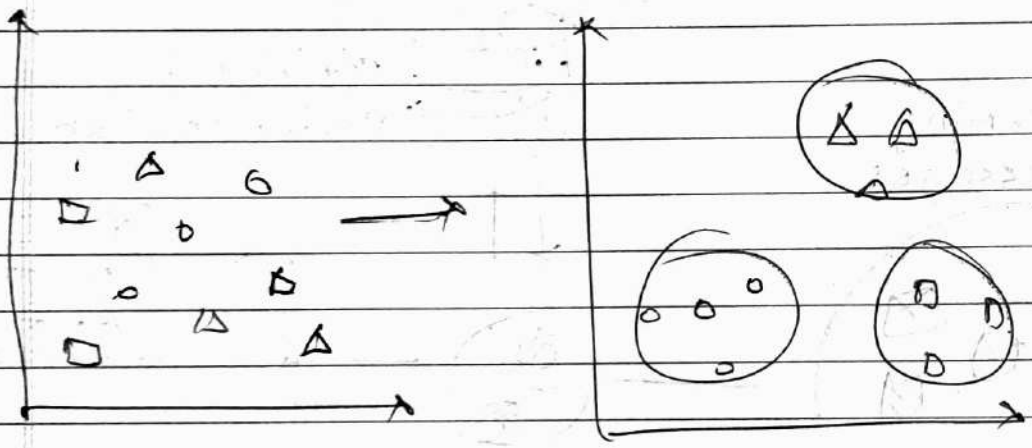
- Random Forest
- Decision Tree
- Logistic Regression
- Support Vector machine

- Spam Email filtering
- Medical diagnosis
- Image recognition
- sentiment Analysis in social media
- customer churn Prediction.

② Unsupervised learning :-



③ Clustering :-



- k-means clustering
- Hierarchical clustering
- Density Based clustering

- Customer segmentation
- Anomaly detection
- Document clustering
- Genomic Data Analysis
- Recommendation system.

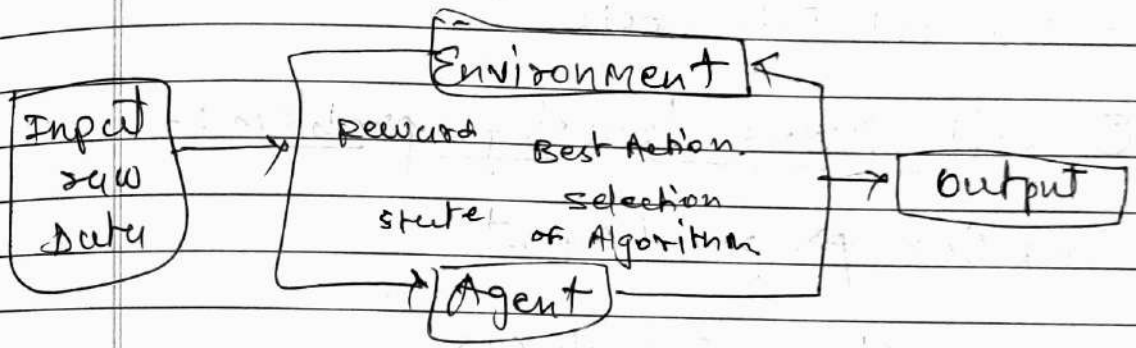
Page No.
Date

⊛ Association :-

IF A ⇒ B
Then

{Bread} → {Butter}

⊛ Reinforcement learning :-



⊛ Issue of ML :-

1. Poor Quality of Data
2. Overfitting of Training Data
3. Underfitting of Training Data
4. ML is a complex process
5. Lack of Training Data
6. Slow implementation
7. imperfection of Algorithms when Data grows
8. Bias & ~~Variance~~ Fairness

⊛ Application of ML :-

1. Image Recognition
2. Speech Recognition
3. Traffic prediction
4. product Recommendation
5. Email Spam & Malware filtering
6. self Driving cars
7. Virtual personal Assistant
8. Online fraud detection
9. Stock Market trading
10. Medical Diagnosis

⊛ well posed learning problem :-

Computer ~~program~~ / Agent

- Class of Task (T)
- Measure of performance to be improved (P)
- Source of experience (E)

- self Driving cars
- Hand written Digit Recognition
- checkers learning problem

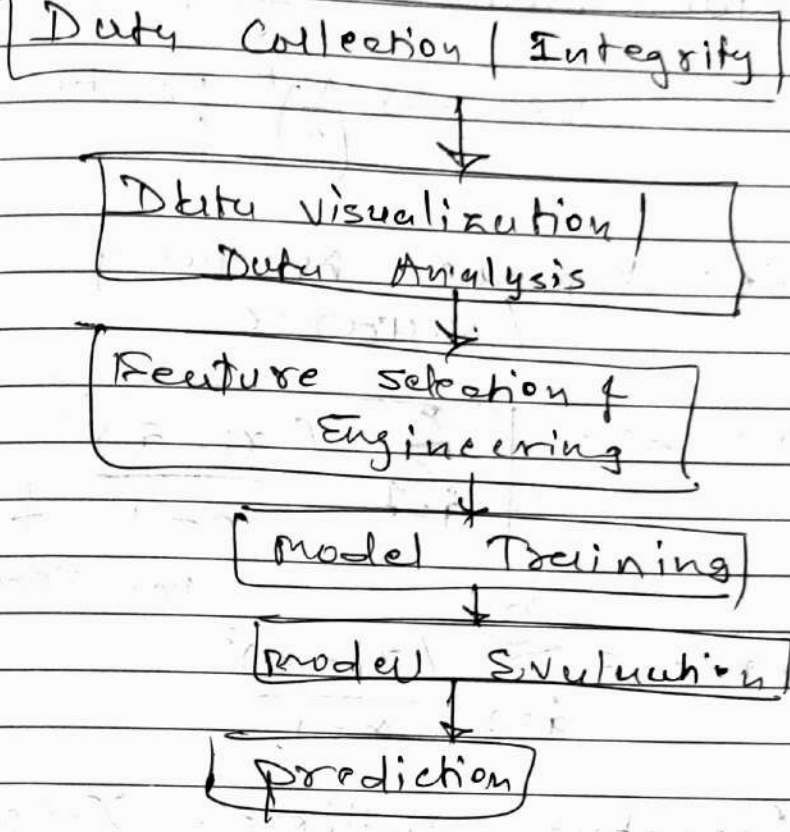
Ex. T: driving on public on public Routes using vision sensors

P: Average distance traveled before an error

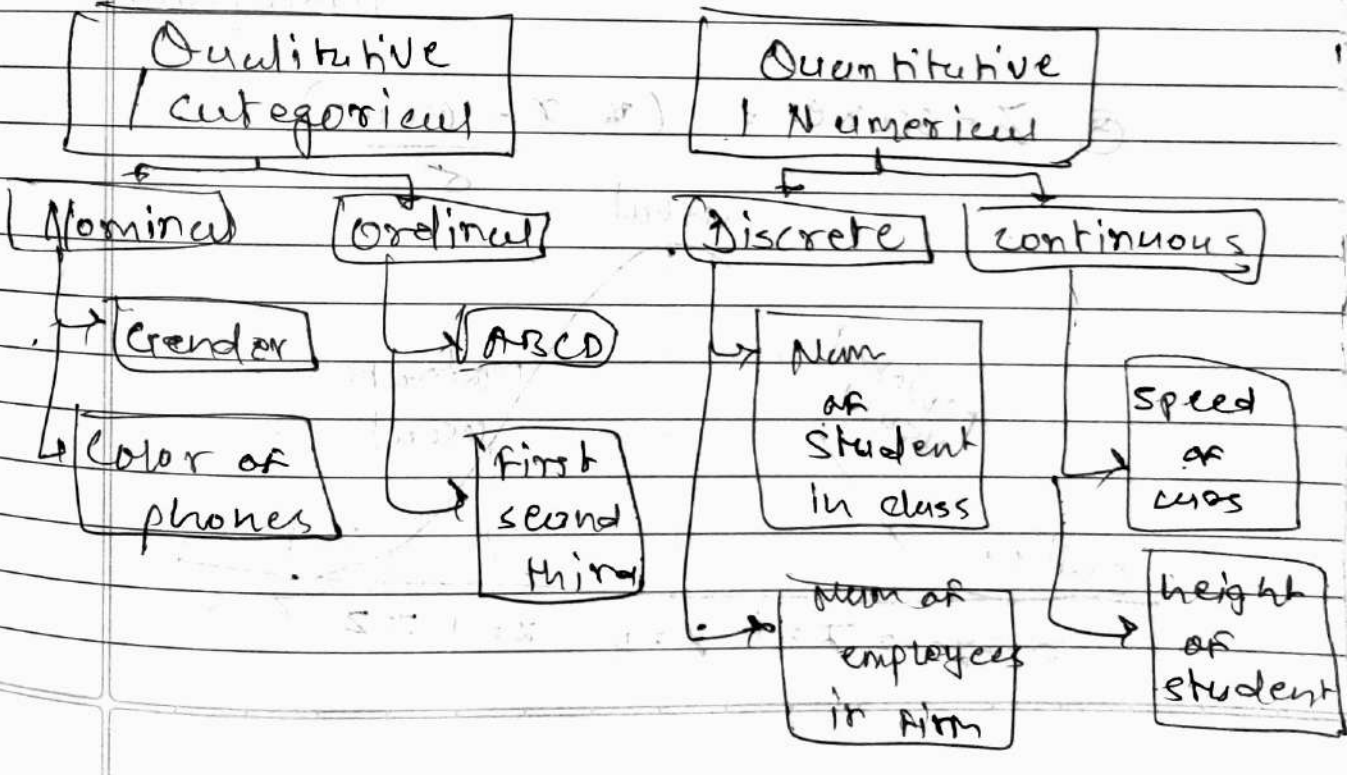
E: A sequence of images & steering Commands recommended while observing Human Driver.

Page No.
Date

⊛ M2 steps / lifecycle :-

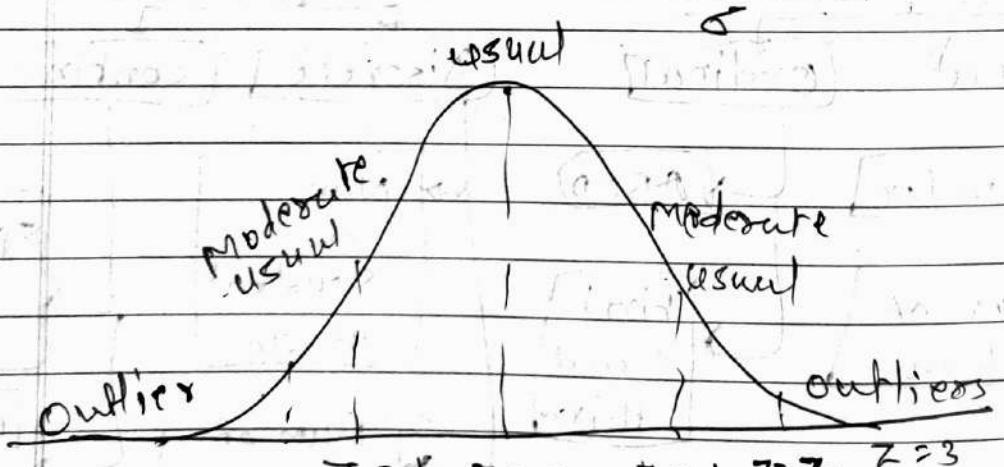


⊛ Types of Data



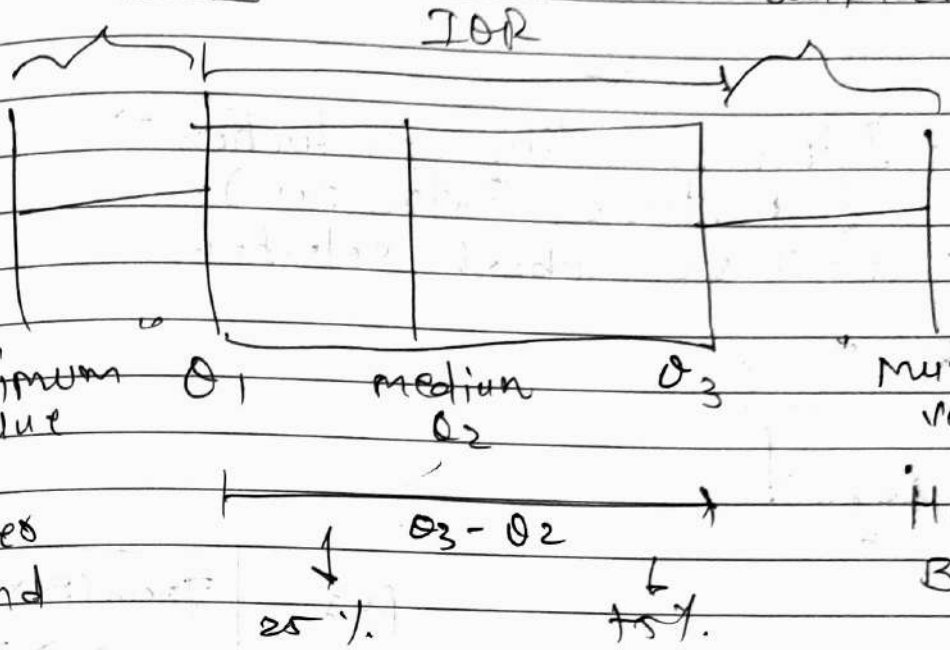
Central Tendency :-

- Mean : $\frac{\sum x_i}{n}$
- Median : odd → $\left(\frac{n+1}{2}\right)^{th}$ position
  even → $\left(\frac{n}{2}\right)^{th}$ position
- Mode : Maximum Frequency / occurence
- S.D : $\sigma = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}}$ For sample S.D
- Variance = $\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$
- Detecting outliers : Z-score
  Box-plot
  Histogram
- Z-score = $\frac{(x - \text{mean})}{\sigma}$



Page No.
Date

⑧ Box-plot :-



⑨ Data Quality

- Integrity
- Consistency
- Accuracy
- Completeness

⑩ Data Remediation

- Data Migration
- Data Overload
- Data Governance
- Dark Data
- Data Discovery
- Data cleansing
- Dirty Data
- ROI

Stages of Remediation :

Assessment
Organizing & segmentation
Indexing & classification
Migrating
Data cleansing

Page No.
Date

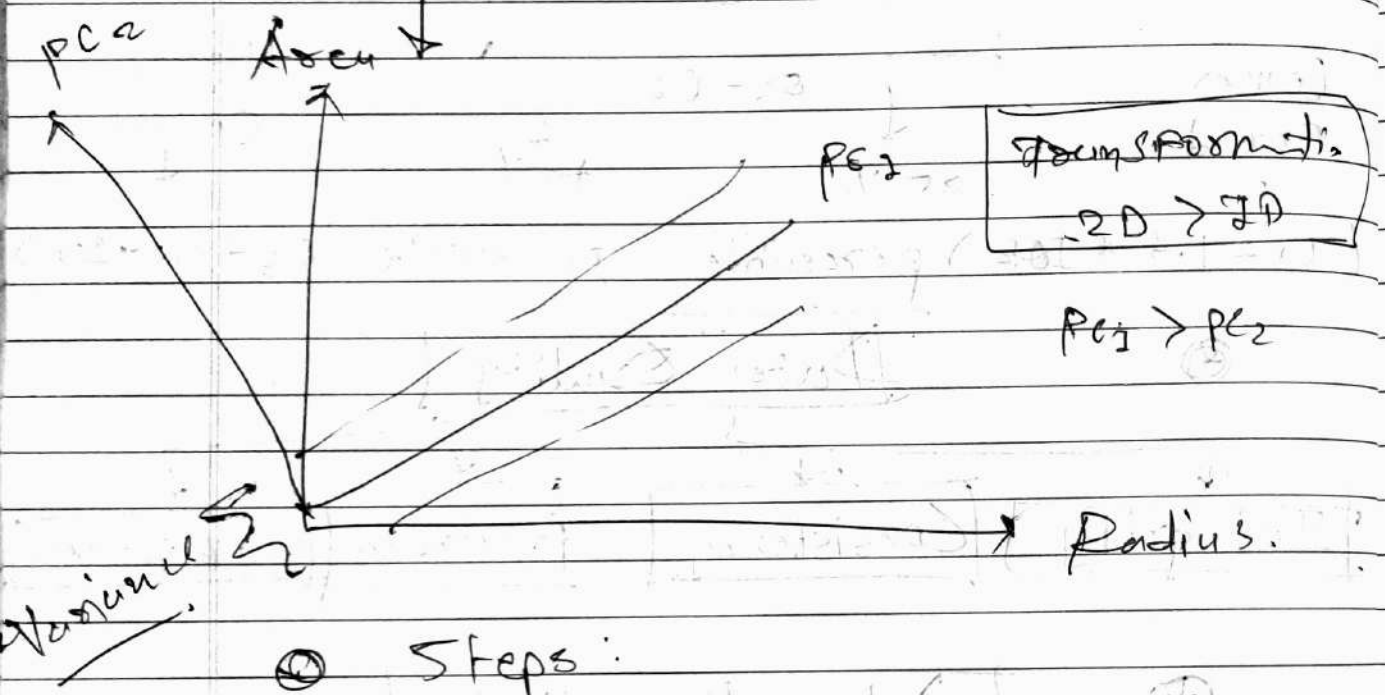
⊛ Data preprocessing

↓

- Dimensionality Reduction → (Feature Extraction)
- Feature Subset selection

⊛ PCA:

↓



⊛ Steps:

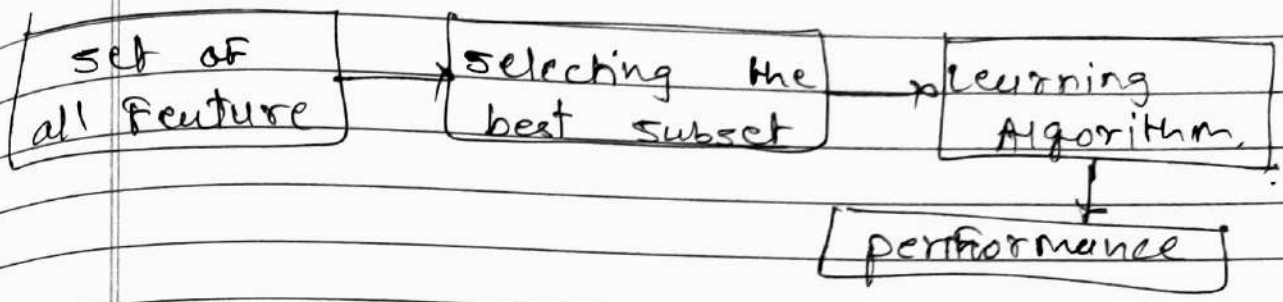
(1) Standardization
(2) Covariance Matrix Computation
(3) Eigen Vector & Eigen Values
Find From CV matrix
Then Find pc components.

⊛ Feature Subset selection:

(1) Filter
(2) Wrapper
(3) Embedded

Page No.
Date

① Filter Methods ②



(1) Information Gain
(2) chi-square test → $\chi^2 = \sum \frac{(O.V - E.V)^2}{E.V}$
(3) Fisher's Score
(4) correlation coefficient
(5) Variance Threshold
(6) Dispersion Ratio.
(7) mutual Dependence
(8) relief

② Wrapper Methods : Backward selection
Forward selection
Bidirectional elimination
Exhaustive selection
Recursive elimination

③ Embedded methods : Regularization
Tree-based methods.

④ Model Section Techniques
choose a model

(1) Probabilistic Measure : Via in-sample error & complexity

(2) Resampling Measure : out-of-sample error