UNIT - 1 Introduction to machine learning

Evolution of machine learning, Paradigms for ML, learning by Rote, learning by Induction, Reinforcement learning, Types of Data, matching, stages in machine learning, Data Acquisition, feature Engineering, Data Representation, model selection, model learning, model Evaluation, model Prediction, search and learning, data sets.

UNIT - 2 Nearest Neighbour - Based Models

Introduction to Proximity measures, Distance measures, Non-metric similarity functions, Proximity blw binary Patterns, Different classification Algorithms based on the distance measures, K-Nearest Neighbor classifier, Radius Distance Nearest Neighbor Algorithm, KNN Regression, Performance of classifiers, Performance of Regression Algorithms

UNIT - 3 Models Based on Decision Trees

Decision Trees for classification, Impurity measures, Properties, Regression Based on Decision Trees, Bias - variance Trade off, Random forests for classification and Regression:

The Bayes classifier: Introduction to the Bayes classifier Bayes' Rule and Inference, The Bayes classifier and its optimality, multi-class classification / class conditional independence and Naive Bayes classifier (NBC)

UNIT - 4 Linear discriminants for machine learning

Introduction to linear Discriminants, linear

Discriminants for classification, Perception classifier, Perceptron learning Algorithm, support vector machines, linearly Non-separable case, Non-linear SVM, Kernel Trick, Logistic Regression, linear Regression, multi-layer perceptrons (MLPs), Back Propagation for Training an MLP

UNIT-5 clustering

Introduction to clustering; Partitioning of data, matrix factorization / clustering of patterns, Divisive clustering, Agglomerative clustering, Partitional clustering, K-means clustering, soft Partitioning, soft clustering, fuzzy c-means clustering Rough clustering, Rough K-means clustering Algorithm, Expectation maximization - Based clustering, Spectral clustering.

On sucessful completion of the course, students will be able to		POs
CO1	Identify machine learning techniques suitable for a given problem	PO6, PO7, PO8, PO12
CO2	Solve real-world problems using various machine learning techniques	PO6, PO7, PO8, PO9, PO12
CO3	Apply Dimensionality reduction techniques for data Preprocessing.	PO6, PO7, PO8, PO9, PO12
CO4	Explain what is learning and why it is essential in the design of intelligent machines	PO6, PO7, PO8, PO9, PO12
CO5	Evaluate Advanced learning models for language, vision, speech, decision making etc.	PO6, PO7, PO8, PO9, PO12

UNIT-1

Introduction to Machine Learning

Evolution of machine learning :-

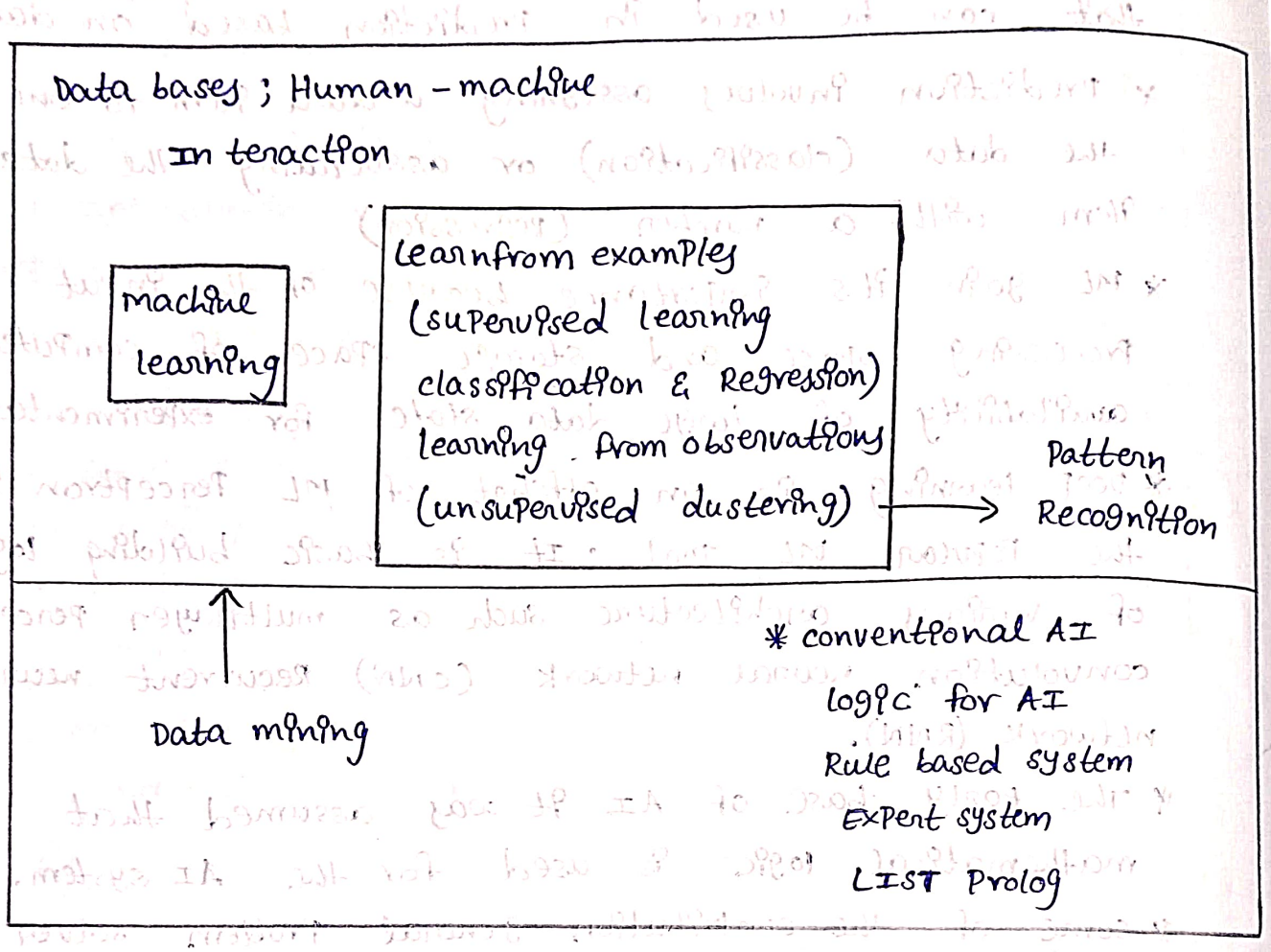
Machine learning is a process of a learning model that can be used in prediction based on data

* Prediction involves assigning a data item to one of the data (classification) or associating the data item with a number (Regression)
* ML gain its importance because of the input processing space and storage space of computer availability of large data state for experimentation.
* Deep learning is an offshot of ML Perceptron is the popular ML tool. It is basic building log of various architecture such as multilayer perceptron. convolution neural network (CNN) Recurrent neural network (RNN).
* The early base of AI it way assumed that mathematical logic is used for the AI system.
* some of the contribution general Problem solver (GPS) automatic Trainer Pruner, Rule based system, Programming language like Prolog & LISP were outcomes of this view during 20th century.

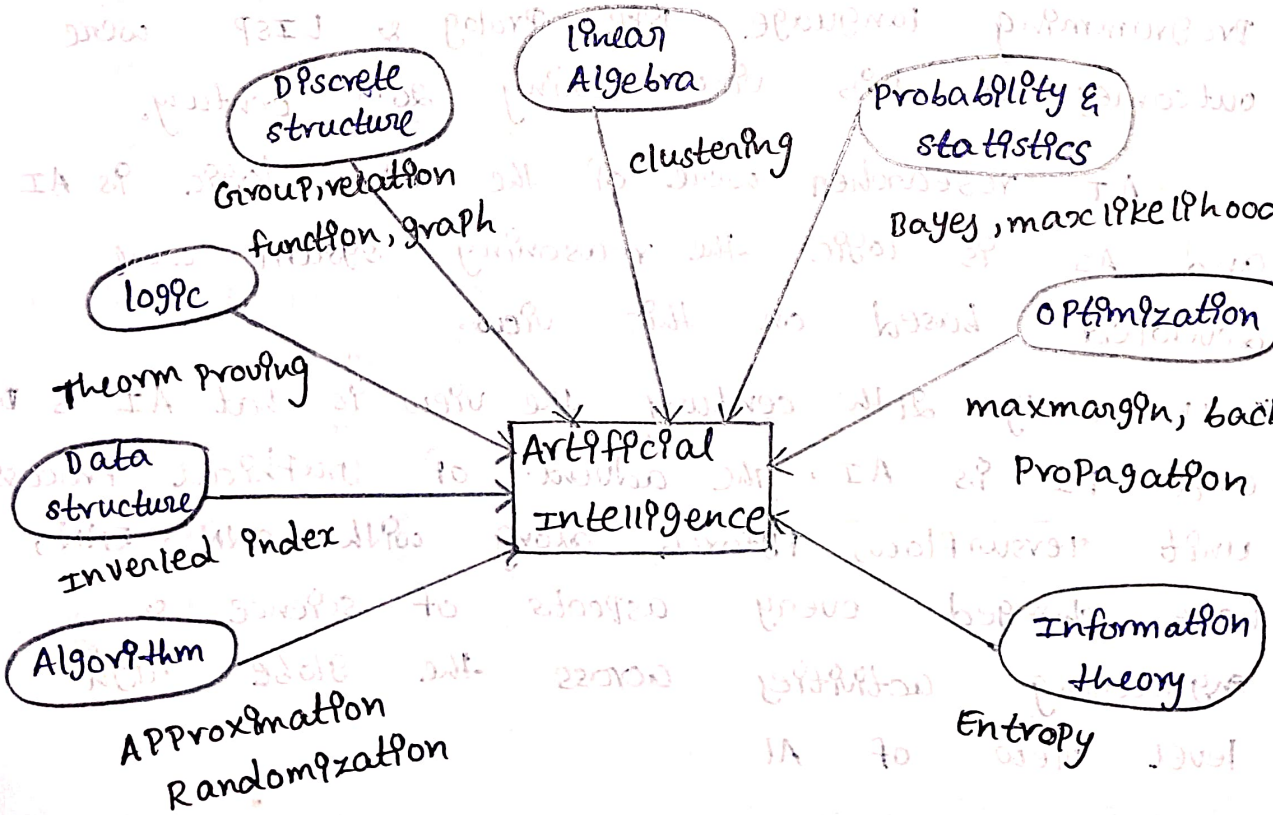
AI researcher were of the view logic is AI and AI is logic the reasoning system were developed based on this view.

* In early 21th century the view is that AI is DL and DL is AI. The advent of graphical Processing unit Tersurflow, Pytorch along with CNN, RNN, have changed every aspects of science & engineering activities across the globe. High level view of AI

* Machine learning (ML) - deals with more than pattern recognition task classification clustering are tasks of pattern recognition. where as ML deals with recognition problem also data mining is efficient organization of a database.



The trippical background topics of AI



Data structures & Algorithm of basics to AI systems:-

The logic and descrite system. Play an imp role in the analysis synthasis of AI system. In ML the data input may be viewed as a matrix, called a data matrix. In there are n data items each represented as L dimensional vector then the corresponding data matrix is of size nxl linear algebra is useful in analysing the weights accossiated with the edges in a neural network. the clustring may be viewed as data manuculates data factorization. the Probabity & staplists helped in estimation the distribution analying the data. optimization is essential in traping nural networks information theory concepts like entropy mutual information divergence. are important to understand. topics such as desition tree classifier futur selection, deep nural net works.

Paradigms for ML:- There are baspcally 5 types.

There are different ways Paradigms for ML such as

1) Learning by rote
2) Learning by deduction
3) Learning by abduction
4) learning by induction
5) rein forcement learning

Learning by rote:-

This involves memorization in an a effective manner. memorizing alphabets and numbers, multiplication tables Examples of rote learning. In the case of data checkend the computer values are stored so that the values no need to recompute later. the chacking is implemented by such engines and it is also

- example of rote learning the when compptatpon ps more expensive then recall this rote learning cancel the a significant large amount of time

chess master :- spend a lot of time memorizing the great gams of the past that ps called rote learning the teaches them how to think pn the chess.

learning by deduction :-

Deductive learning deals with exploitation of deductpay may earlaren this type of learning ps based on reasoning that ps truth preversiving.

* given A & pf A then B (A→B), B can deduce B we can use B along with pf B then C (B→c) to deduce C. Note that A & A→B are true, then B ps true. that ps the truth prescriving nature of learning by deduction.

consider the following statement :-

1. It ps raining
2. It ps rains, the roads are wet
3. It a road ps wet, pt ps slippary ing

from 1 & 2 we can infore using deduction that 4

4. the roads are wet.

this deduction can than be with 3 to deduce the 5

5. the roads are slippary ing

* The statements 1, 2 & 3 are true then statement 4 & 5 are automatically true
* deductive learning ps apply pn game playing.

Learning by abduction :-

Here we can infer A from B & (A→B) this ps not

truth preserving as both B & (A → B) can be true and A can be false.

→ for example

1) An Aeroplane is a flying object

Aeroplane → flying object

2) A is a flying object

A is an Aeroplane

A can be a bird or a kite

from 1 & 2 we infer using abduction that A is an Aeroplane this kind of reasoning may lead to incorrect conclusion.

"A can be a bird or a kite"

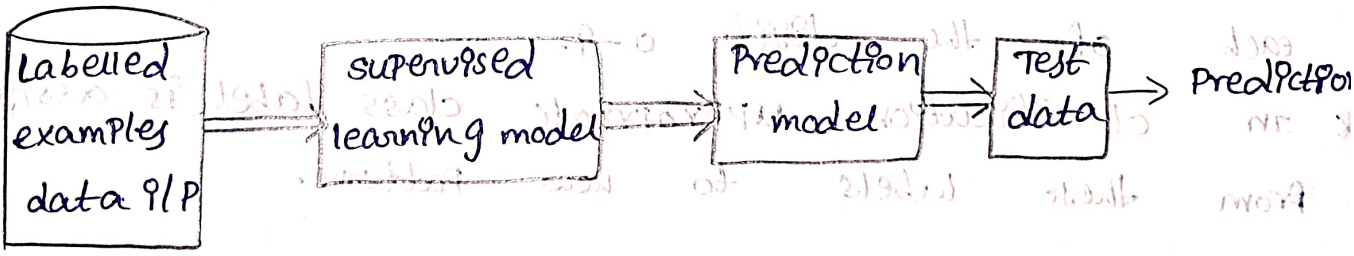
Learning by Induction:-

* Here learning by achieved with the help of examples are observations it can be may be categorized as.

i) Learning from examples (supervised learning)

ii) Learning from observations (unsupervised learning)

Learning from examples:-



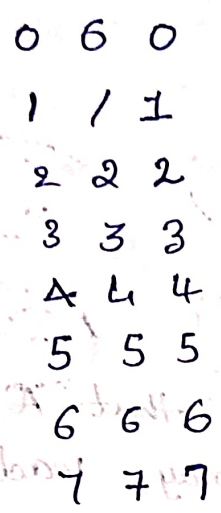
* Here a collection of label examples are provided and machine learning system uses these examples to make a prediction on a new data pattern.

* In supervised classification are learning for examples there are two machine learning models that is

classification and Regression.

classification:-

consider hand written digits shown in figure



* Here each row has three examples of each of the digit the problem is to learn on mechine learning model using this data to classify new data pattern this is called supervised learning As the model learning with help of exampler data.
* These exampler data is provided by an expert for example a medical doctor provide examples of normal patient and patients infected by covid based on test results.
* In the case of Hand written digits we have ten class labels. one class label correspond to each of the digit 0-9.
* In classification appropriate class label is assigned from these labels to new pattern.

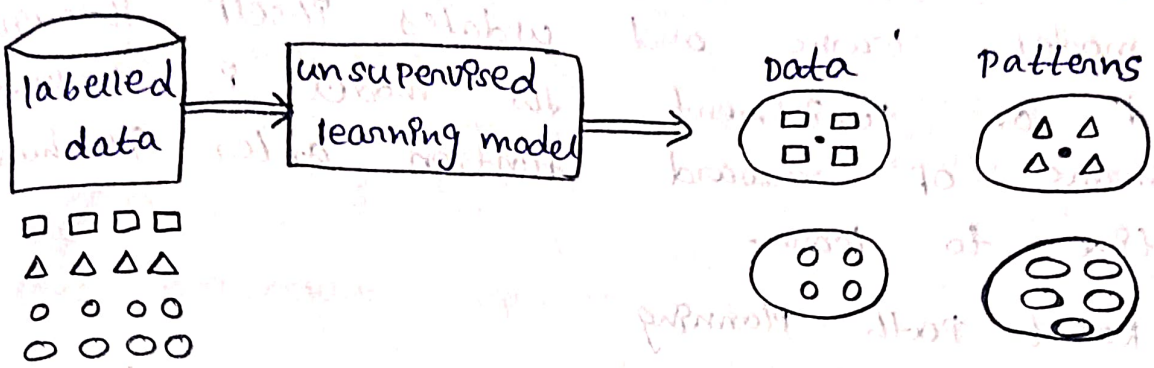
Regression:-

* There are several prediction applications where the labels come from an infinite set.

EX:- share value of a stack could be positive real number.

* the stack may have different values at a particular time and each of these values is a real number

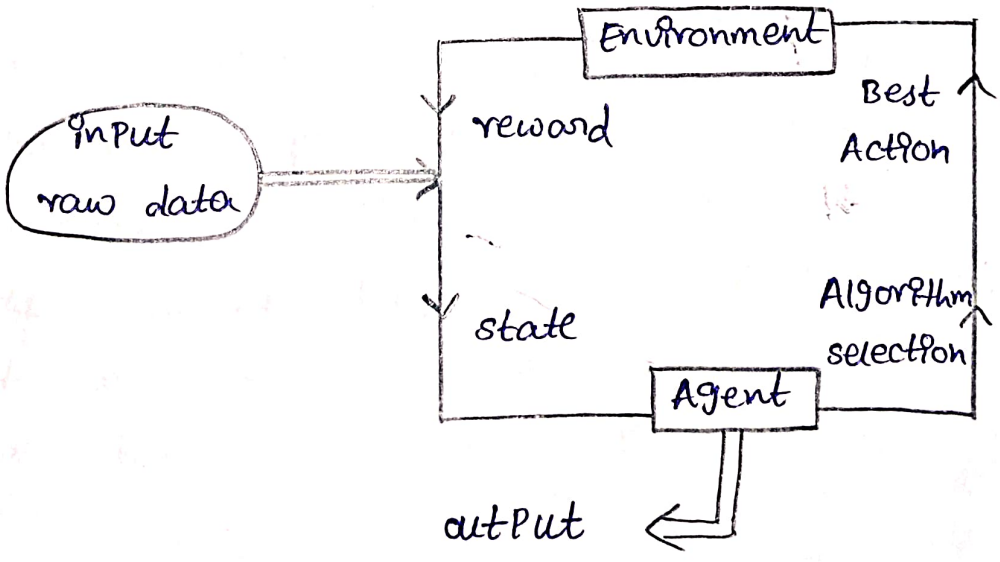
Learning from observations :-



* observation are the instance that need not be labelled in this case i be cluster (or) group the observations into a smaller number of groups.
* such grouping is performed with the help of clustering algorithms that assigns similar patterns to the same group (or) cluster.
* each cluster is represented by centroid or mean for example $x_1, x_2 --- x_p$ are p elements are cluster then the centroid is defined by

$$\frac{1}{p} \sum_{i=1}^{p} x_i$$

Reinforcement Learning :-



* In reinforcement learning the agent learns optimal policy to optimize some reward function the learned policy helps the agent in taking an action based on current state of the problem

* the model learns and updates itself through reward or punishment the model is evaluated by means of reward funtion after it had some time to learn.

Ex:- Robot Path Planning

write the difference between supervised and unsupervised Reinforcement learning.

Supervised	unsupervised	Reinforcement
* This type of learning is used when you know how to classify a given data, or in other words classes or labels are available.	* This type of learning is used when there is no idea about the class or label of a particular data. The model has to find pattern in the data.	* This type of learning is used when there is no idea about the class or label of a particular data. The model has to do the classification it will get rewarded if the classification is correct, else get punished.
* labelled training data is needed. model is built based on training data.	* Any unknown and unlabelled data set is given to the model as input and records are grouped.	* The model learns and updates itself through reward punishment.
* The model performance can be evaluated based on how many misclassifications have been done based on a comparison blw. Predicted and actual values.	* Difficult to measure whether the model did something useful or interesting Homogeneity of records grouped together is the only measure.	* Model is evaluated by means of the reward function after it had some time to learn
* There are two types of supervised learning Problems classification and regression.	* there are two types of unsupervised learning Problems clustering and association	* No such types
* simplest one to understand	* more difficult to understand and implement than supervised learning.	* Most complex to understand and apply.

standard algorithms include * Naive Bayes * K-nearest neighbour (KNN) * Decision tree * Linear regression * Logistic regression * support vector machine SVM etc	standard algorithms are * K-means * Principal component Analysis (PCA) * self-organising map (SOM) * Apriori algorithm * DBSCAN etc.	standard algorithms are * Q-learning * sarsa
practical applications include * Handwriting recognition * stock market Prediction * Disease Prediction * fraud detection etc	Practical applications include * market basket analysis * Recommender systems * customer segmentation, etc	Practical applications include * self-driving cars * Intelligent robots * AlphaGo zero (the latest version of DeepMind's AI system playing Go)

Types of data sets:-

The data can be divided into 2 types.

1. Qualitative Data (categorical data)
2. Quantitative Data (Numerical data)

Qualitative Data:-

* Qualitative data Provides information of the quality of object which cannot be measured.

Ex:- Name, Roll No etc

* It can be further divided into two types

1. Nominal data
2. ordinal data

Nominal data:-

* It has no numeric values, but a named value.
* nominal values cannot be measured.

Ex:-1. blood group: 'A', 'B', 'AB', 'O'

2. Gender : 'male', 'female'.

3. Nationality : 'Indian', 'American', 'British'...

* mathematical operations +, - etc cannot be performed on nominal data that is why mean variance can not be applied. where as a basic count is possible.
* so mode most frequently occuring value can be identified for nominal data.

ordinal data:-

* These are named object and can be arranged a sequence of increasing or decreasing value, so that we can say whether the value is better than.

another value.

Ex:- 1. customer satisfaction : 'very Good', 'Good', 'Average', 'Poor'.

2. Grades of student : 'S', 'A', 'B', 'C'...

3. Hardness of metal : 'very hard', 'hard', 'soft'

* mode can be identified.
* since ordering is possible median can be identified.

Quantitative Data:-

* Quantitative data relates to information about the quality of object hence it can be measured.

Ex:- Marks of subjects.

* Quantitative data can be divided into

1. Interval data
2. Ratio data

Interval data:-

* The Interval data is numaric data for which the order is known and the exact differences between values is also known.

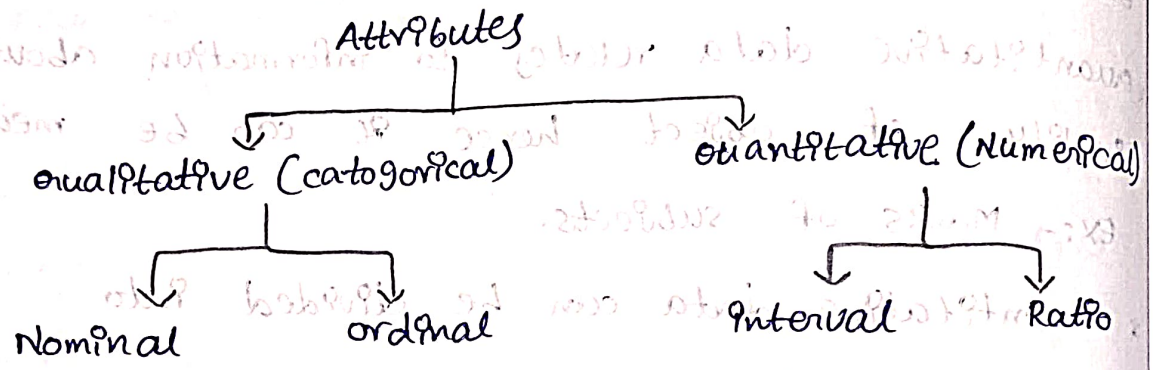
Ex:- celsius, temperature.

* The difference between each value remains same in celsius temperature.
* The Interval data do not have a true zero value there is very know zero temperature. Hence, only addition and subtraction can be applied for interval data.
* The stastical function such as mean, median, mode, standard deviation, can be calculated.

Ratio data:-

* The Ratio data represents numeric data for which exact value can be measured.
* absolute zero is available for ratio data,
* all mathematical calculation applied and all stastical calculations are applied.

Ex:- height, weight, age, salary etc



* Attributes can classified as Discrete or continuous based on the number of values that can be assigned.

Descrete Attributes:-

* It can take ~~infit~~ finite or countable number of values.

Ex:- Pincode, Rank etc.

* descrete Attribute which can assumed two values only is called binary attribute.

Ex:- Yes / No, True / false etc.

* continuous attribute it can assume any possible values which is a real number.

Ex:- length, height etc.

Matching :-

* Matching in machine learning involves techniques to find similar items (or) records across datasets (or) with in a single data set.
* It is used in both supervised and unsupervised learning.
* Matching is found by using proximity measured such as distance / dissimilarity measure (or) similarity measure.
* To data items u & v represented as l dimensional vectors and these are matched when the distance between them is smaller (or) when the similarity between them is larger.
* The distance measure is Euclidean distance and a popular similarity measure is the cosign of the angle between vectors.

$$d(u,v) = \sqrt{\sum_{i=1}^{l} (u(i) - v(i))^2}$$

cosine similarity

$$\cos(u,v) = \frac{u^t v}{\|u\| \|v\|}$$

$\|u\|$

where, $u^t v$ is the dot product between the vectors u & v.

$\|u\|$ is Euclidean distance between u and origin it is also called Euclidean Norm.

Applications of matching in machine learning :-

finding the Nearest Neighbor of a Pattern

let x be l dimensional vector and

$x = \{x_1, x_2 --- x_n\}$ is a collection of

n data vectors. the nearest neighbour of x from X denoted by $NN_x(x)$ is x_j. If

$$d(x, x_j) \leq d(x, x_i), \forall x_i \in X$$

* Assigning a set with the nearest representative centroid :-

*let $x_1, x_2, x_3 \ldots x_k$ are k sets with $x_1, x^2, \ldots x^k$ are there centroids (representative) the vector n is assign to cj if

$$d(x, x^j) \leq d(x, x^j) \text{ for } j = \{1, 2, \ldots k\}$$

where cj is jth group (or) cluster

x^j is centroid of cj this will be useful in clustering are learning from observation:

stages in machine learning :-

The figure shows number of steps involved in machine learning system.

Application domain → Data Acquisition

↓

feature Engineering = Preprocessing + Representation

↓

Model selection → choose a model ← -- Domain knocoledge

Model learning → Train the model ← --- Training data

Model Evaluation → validation the model ← --- Evaluation data (validation data)

↓

Model Prediction → learn the model ← -- Test data

Model Explanation → Explain the model ← -- Expert feedback.

* the available data is split into training validation and test data.
* Training data is used in model learning are Training.
* validation data is used to tune the machine learning model.
* Test data is used to examine how the learnt model is Performing.

Data Acquisition :-

* It is the process of gathering are collecting relevant data from various sources and storing the data.
* It is important for obtaining high quality data for model training and optimizing the performance.
* this data make machine learning organisms to learn and Perform method.

Ex:- * To distinguish between adults and children the measurement of height and weight are sufficient.

* To distinguish between covid and non covid Patient the body temperature and chest congestion are important.

Feature engineering (methods)

This step involves the combination of data PreProcessing and data representation.

1) Data PreProcessing

the raw data available needs to be updated before it can be used by an ML model. the common Problems that occur with raw data or missing values, different ranges for different variables, and

the presence of outliers.

a) missing values

occurs has a consequence of the inability to measure a feature value or due to unavailability are erroneay data entry. some ML algorithms can work whenever there are less number of missing data values in this case there is no need for preprocessing. where as in most of the cases ML models can not work with missing values therefore there are different methods for predicating the missing values.

use the nearest neighbour

x is vector and its qth component x(q) is missing

* let $X = x_1, x_2 \dots x_n$ are the set of n training vectors.
* let $x_g \in X$ is the nearest neighbour of x based on the remaining (d-1) components i.e, the predicated values x_i is $x_g(q)$ $x_i = x_g(q)$

use a larger neighbour

use the k-nearest neighbour of x to predict the missing value of x(q) let the KNN of x using the remaining d-1 components from X or $x_1, x_2 \dots x_k$ Now the predicated value x(q) is the average of qth component of this KNN.

$$x(q) = \frac{\sum_{g=1}^{k} x_g(q)}{k}$$

Ex:- the data vector is

(1,1,1) (1,1,2) (1,1,3) (1,—,2) (1,1,—) (6,6,1).

k=3

missing data vector is (1, —, 2)

KNNs are (1,1,1) (1,1,2) (1,1,3)

$\frac{1+1+1}{3} = "1"$

the data vector = (1,1,2)

cluster the data & locate the nearest cluster

this approach is based on clustering the training data and locating the cluster to which $x \in$ based on the remaining $l-1$ components. let x with its p^{th} values it's missing belongs to cluster c_q if u^q is the centroid of c_q then the predicted value of $x(p)$ is $u^q(p)$, $x(p) = u^q(p)$.

EX:-

cluster 1 : { (1,1,1) (1,2,2) (1,3,2) } centroid 1 = (1,2,2)

cluster 2 : { (3,4,3) (3,5,3) (3,3,3) } centroid 2 = (3,4,3)

cluster 3 : { (6,6,6) (6,8,6) (6,7,6) } centroid 3 = (6,7,6)

the missing data vector is

(1, −, 2)

is near to centroid 1 = (1,2,2)

the predicated value is "2"

the data vector is (1,2,2)

* the mean square error is used to find deviation of predicted value from original value. let the true values are $y_1, y_2 - - - y_n$ and predicted values are $y_1\hat{}, y_2\hat{} - - - - y_n\hat{}$ and mean square value

$$\text{MSE} = \frac{\sum_{i=1}^{n} (y_i - - - - y_i\hat{})^2}{n}$$

b) different ranges for different variables :-

The scales of value of different features could be very different this will effect the Process to depend more upon the features

that has larger values. that will reduce the contribution of the features with smaller values.

EX:- classification of adult (or) child Here height is measured in meters and weight in grams. the adult is represented by (1.6, 75000) the child is represented by (0.6, 5000) the range of height. is [0.5, 2.5] range of weight [2000, 200000]

the Euclidean distance between adult & child is

$$\text{Euclidean distance} = \sqrt{(1.6-0.6)^2 + (75000-500)^2} \approx 7000$$

$$\text{The cosine similarity} = \frac{\begin{bmatrix}1.6\\75000\end{bmatrix}\ [0.6 \cdot 5000]}{\sqrt{(1.6)^2+(75000)^2}\sqrt{(0.6)^2+(5000)^2}}$$

The Euclidean distance are the cosine of angle between the two vectors depend more upon the weight feature while the contribution of height is negiable. this can handle by scaling different compoundy differently and this process are scaling is called normalization. there are two important normalization methods.

scaling using the range:-

for any categorical features the values of two features either match are miss match and the contribution to the distance is either zero (match or 1 (mis match) in the case of numerical features the contribution should be in the range 0,1 this can be obtain by scaling the differents by the range of values of the feature.

If pth component contribution to the distance between two objects x^i & x^j is

$$\frac{|x^i(p) - x^j(p)|}{\text{Range } p}$$

where Range p is Range of pth feature the value of this term is in the range [0,1] the value of 1 is obtain when $|(x^i(p) - x^j(p)| = \text{Range } p$ and zero is obtain $x^i(p) = x^j(p)$.

standardization :-

The data is normalized so that it will have 0 mean and unit variance this is the Property of standard normal distribution

Ex:- Consider the 0th component of the data vector is 60, 80, 20, 100 & 40

$$\text{mean} = \frac{60+80+20+100+40}{5} = 60$$

zero mean data = 0, 20, -40, 40 & -20

$$\text{variance } \sigma^2 = \frac{0^2+20^2+(-40)^2+40^2+(-20)^2}{5} = 800$$

$$\text{standard deviation } \sigma = \sqrt{\sigma^2} = \sqrt{800} = 28.284$$

$$\frac{0}{28.28} \quad \frac{20}{28.28} \quad \frac{-40}{28.28} \quad \frac{40}{28.28} \quad \frac{-20}{28.28}$$

zero mean & unit variance data $\Rightarrow 0, 0.707, -1.414, 1.414, -0.707$

This data is standard Normal distribution data.

(c) outliers in the data

The outliers in data item that is either noise are erroneous noise measuring instrument are

error data recording are responsible for Presence of outliers. A data item is called an outlier if it's values are away from average of data items or it is not similar to other items in terms of it's characteristics. outliers occur because of different reasons.

1) Data Entry Errors
2) measurement error

due to faulty devices or experiment set up Problem.

3) Experimental error:-

due to faults in experimental design

4) Intensional outliers
5) Data Performing error
6) Natural variances

It is possible to use clustering to locate search outliers.

Data representation:-

each data item is represented hay a vector in the partical applications the dimensionality of data or the size of the vector can be very large the high dimensionality data is common bio informatics information retrieval.

satellite Imaging

the diffculties in using high dimensionality data vectors are

* computation time increases with the dimensionality.

* storage space are also increases
* performance of the model high dimensionality data performance of the model
* there is a problem called peaking phenomenon that is in dimensionality increases the accuracy of the ML model increases until some values and beyond that value the accuracy starts decreasing. the model will also called overfitting that is the model will perform well for training data and fails to perform well of validation data.

the dimensionality reduction is used in ML model so that the model does not over fit for available data the dimensionality reduction approaches.

* feature selection :-

let $f = \{f_1, f_2 --- f_L\}$ are L features in the feature selection approach will select only P no. of feature $P<L$ such that these P no. of features will maximize the performance of ML model.

Feature Extraction :-

let $f = \{f_1, f_2, --- f_L\}$ are L features a set $H = \{h_1, h_2 --- h_L\}$ of $l<L$ features is extracted so this is possible with following methods.

1. linear method :-

In this $h_j = \sum_{i=1}^{L} \alpha_{ij} f_i$, $j = 1, 2 --- l$ that is each element of "H" is a linear combination of original feature. some of the methods in this category are

a. Principal component

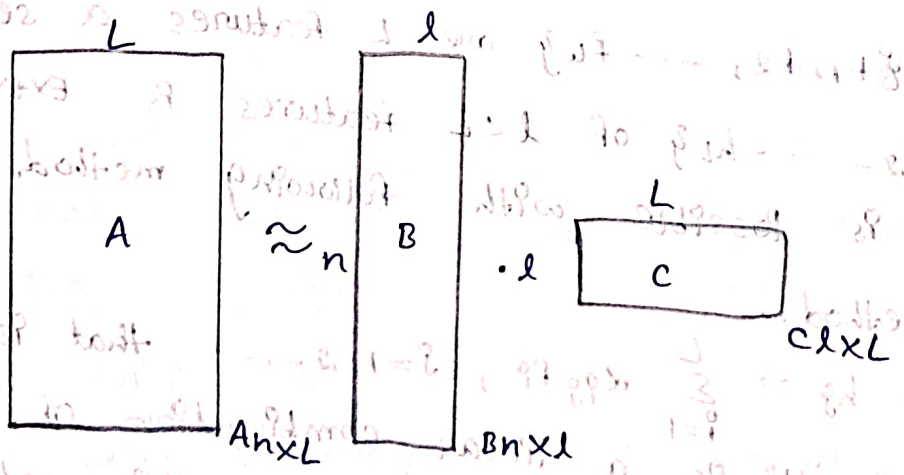
*consider data of n vectors in L-dimensional space that is AnxL the covariance matrix of "A" is of size LxL that is E_LxL and the Eigen vectors of "E" form the principal components.

* the Eigen vector corresponding to the largest Eigen value is the first principal component similarly the eigen vector corresponding to second largest eigen value is the second principal component. both the original features and principal component are use to represent any data vector.

b. Non negative matrix factorization (NMF)

NMF :- Is a method to break down larger dataset into smaller meaningful parts and ensuring that all the values are non-negative this is useful in exacting important features from the data and these features are easy to analysis and process it.

NMF is factorization of matrix AnxL into a product of Bnxl & ClxL



Non-linear feature extraction:-

H = {h1, h2---hl) is represented using $h_i = t(f_1, f_2 - f_l)$

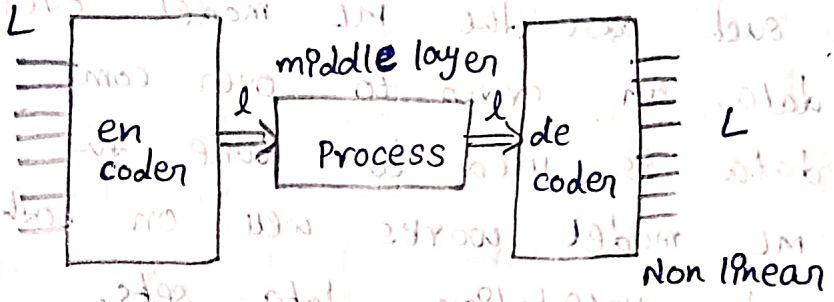
with f is a non linear function.

Ex:- $f = \underbrace{\{f_1, f_2\}}_{\text{linear}}$

$h_1 = \underbrace{a_1 f_1 + a_2 f_2 + a_3 f_1 f_2}_{\text{Non-linear}}$

This is a non linear because of the term $f_1 f_2$ in h_1

Autoencoder is a Non-linear feature extraction tool or method



Model selection

selection of the model depends upon the nature of the data and knowledge of application domain If some features or numerical and others are categorical that classifiers based on Perceptron (or) support vector machine are not suitable because the dot product between Non-numerical data is not possible.

therefore Bayesian models and decision tree models are suitable for categorical data (Non-numerical)

Model learning

this step depends on the size and type of training data. a subset of labelled data is used as training data for learning the model. another sub set is used for model validation or model evaluation. some of the ML models are highly transparent while others are opaque or black box models.

Ex:- A decision tree models or transparents.

Neural networks are opaque as the output of

Intermediate layers may not offer transparency.

Model Evaluation (model validation)

This step requires validation data If the ML model works well on the training data then we say that the model is well training sometimes it may not work well on validation data In such case the ML model overfits the training data. In order to over come over fitting validation data is used to tune the ML model so that ML model works well on both the training and validation data sets.

Model Prediction

This step deals with testing the model this step uses test data set for the for the Prediction is real world the model is used for Prediction as new input keep coming in.

Ex:- An ML model built for medical diagnosis should Predict like doctor an mains that diagnosis when a new Patient uses it.

Model explanation :-

This step is important to explain an expert are a manager. why a decision was taken by the ML model. explanation as become very important in deep learning because deals system used neural networks that are opaque or black box. such opaque behaviour has created the need for explanable AI.

search and learning:-

search is the fundamental operation in both ML & AI search was used in problem solving theorem proving planning and knowledge based systems. search is used in several computer science applications.

Ex:-

search is used in databases for Answering any queries

* In ML search is used in learning a classification model, a proximate measure for clustering and classification and learning model for regression.
* Inference is search in logic and probability.
* In algebra matrix factorization is search in optimization regularizer is used to simplify the search in finding a solution. In information theory the search for low entropy.
* optimization, inference, matrix factorization are important are ML and all are based on search.
* learning itself is search.

Explanation offered by the model:-

conventional AI systems are logic based are rule based so the reasoning process has transparency and explanability both forward and backward is used in conventional AI systems. some knowledge based systems are used in diagonesis and in teaching because of this flexibility.

Ex:-

The knowledge base used by the "MYCIN" expert system used in teaching medical students using another expert system "GUIDON". the problems

with conventional AI systems are.

1. there was no general framework for building AI systems. the experience in building one AI system did not simplify the building of another AI system.
2. Acquiring knowledge is a great challenge because different experts differed in their conclusions meeting to inconsistency.

current AI systems:-

current AI systems based on DEEP learning and "DL" is large data dependent they learn representations automatically there used multilayer neural networks and back propagation algorithms in training the models. The difficulties with "DL" systems are.

1. They are data dependent were Performance improves as the size of the data set increases.
2. learning in "DL" systems involves a simple change of weights in the neural networks this is done with the help of back propagation.
3. "DL" systems are black box systemand does not have explainable capability therefore the current "AI" researches working on explainable "AI".

Data sets used:-

Data sets for char classification:-

1. MNIST Hand writtend digit dataset

There are 10 classes corresponding to digits 0, 1 _ _ _ _ 9 and each digit is an image are size 28x28 pixels and each pixels having

values in the range 0 to 255 (0 - black, 255 white).
* there are around "6000" digits hay training patterns and around 1000 text sample in each class the class label is also provided for each of the digit.

2. FashPon MNIST Data set :-
* It is a data set of zalando's article images and consisting of a training set of 60,000 examples and text set of 10,000 examples. each example is 20x20 grace scale image and associated with a label from 10 classes.
* It is used as replacement for original MNIST data set for ML models.
* This data set is taken from in built data set of kaggle.

3. olivetti face data set
* It consist of 10 different images for each of 40 distinct objects.
* for some objects the images was taken at different times, varying the lighting.
* facial expressions (open eye, close eye, smiling, not smiling) and facial details that is glasses and not glasses.
* all images was taken against dark homogeneous background.
* each image is of the size 64x64.
* It is available scikit-learn flatform.

4. wisconsin Breast cancer data set
* we consist of 569 Examples and each is a

30 dimensional vector there are two classes benign and malignant the number of samples from benign is 357 and the number of malignant patterns is 212.

* It is available on scikit-learn.

Data sets for Regression

1. Boston Housing datasets :-
   It has 506 examples each example is 30 dimensional vector. It is available on scikit-learn flatform.

2. Airline Passengers data sets
   This data set provides monthly totals of US Airline Passengers from 1949 to 1960. This data set is taken from in built data set of "kaggle"

3. Australian weather data set
   It provides various weather record details for cities in Australian. The features include location minimum and maximum temperature. This data set is taken from an in built data set of "kaggle".

UNIT-2

Nearest Neighbour - Based Models

Introduction to Proximity measures:-

Proximity measures are used to measure the degree of similarity are dis similarity between two or more data vectors. this data vectors can represent documents images are audio or video files. Proximity measures are used by mechine learning algorithms to compare and classify are group or make predictions using these vectors. the important Proximity measure are distance measure.

1. Distance measure:-

distance measures are mathematical functions used to find how similary are dis similar two objects based on there futures the distance measure used for clustering classification and information retrival.

(i) Euclidean distance

It represents the shortest distance between two vectors it is used in clustering analysis.

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

It is used for contineous numerical data

2) Manhattan distance

It determines the obsalute difference, the pair of quardinates in a plane P quaridinates are (x_1, y_1) & quardinates (x_2, y_2) the manhatten distance between P and Q is

$$|x_1 - x_2| + |y_1 - y_2|$$

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

* It is used for high dimensional data. It is also called city block distance.

3) Min kowskI distance

* It is generalized distance measure that include both Euclidean and manhatten distance has special cases and controlled by a parameter P.

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

* It is used for flexible distance calculations for P=1 it becomes manhatten for P=2 it becomes Euclidean.

4) Hamming distance

* Number of positions were two vectors differ and it is used for error detection and sequence comparision.

$$d(x,y) = \sum_{i=1}^{n} [x_i \neq y_i]$$

where, $[x_i \neq y_i] = 1$ if symbols differ $= 0$ else

* It is used for binary string DNA sequences error correction.

EX:- 1. karolin

2. kathrin

$d(x,y) = 3$

similarity measures.

It is used to calculate how similar the objects are the higher value of similarity for more similar object.

1. cosine similarity

* measure cosine distance of angle between two vectors
* it focuses on orientation rather than magnitude.

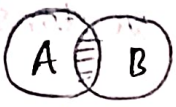
$$\text{cosine}(x,y) = \frac{xty}{\|x\|\,\|y\|} \rightarrow \text{similarity formula}$$

$$\text{distance formula} = 1 - \frac{x^{ty}}{\|x\|\,\|y\|}$$

* it is used for text mining "NLP" it also used in recommandation systems. Natural language Process

Jaccard Index

Intersection union



A∩B

A∪B

$$* \quad J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

* the Jaccard distance is set based distance that compliment this similarity by the ratio of intersection two union.

$$d(A,B) = 1 - \frac{|A \cap B|}{|A \cup B|}$$

* It is used for binary are categorical data specially for sets.

Pearson correlation

It measures linear relation between variables

---

distances measure

* distances measure is used to find the dis similarity between patterns represented as vectors.
* Patterns which are more similar could be closer the distance function (d) is a matric are non-matric

the most popular distance metric is minkowski metric. a metric is type of a measure that satisfies three attributes:

1. Positive reflexivity

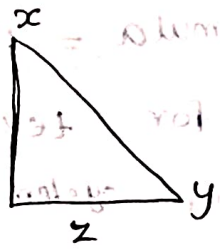
$$d(x,x)=0$$

2. symmentry

$$d(x,y)=d(y,x)$$

3. Triangular inequality

$$d(x,y) \leq d(x,z)+d(z,y)$$



where xyz are three patterns.

* the different types of dissimilarity/similarity measures between pattern (data, I/P) vectors.

1. Minkowski distance

The minkowski metric can be defined has $d^r(P,Q)=$

$$\left(\sum_{k=1}^{L}(|P_k-Q_k|^P)\right)^{1/P}$$

where,

* R is a parameter that determines the type of metric
* P and Q or L-dimensional vectors based on the value of R.

* L_∞ norm :- $r=\infty$, $d(P,Q)=$ max distance $k(|P(k)-Q(k)|)$

Here $r=\infty$ and $d(P,Q)=(|P(k)-Q(k)|)\ k\in\{1,---L\}$

* L2 norm :- Here $r=2$ and $d(P,Q)=\left(\sum_{k=1}^{L}(|P(k)-Q(k)^2|)\right)^{1/2}$

is euclidean distance

* L_1 norm :- Here $r=1$ and $d(P,Q)=\left(\sum_{k=1}^{L}(|P(k)-Q(k)|)\right)$

is the city-block distance.

fractional norm :- It is possible that R is a fraction the resulting distance is called fractional norm. It is not a metric as a does not satisfy the triangular inequality. normalization of feature values is used to keep all the features in the same range the mahalanohis distance is popular distance measure that is used in classification and it is computed using the covarience metrics. the squared mahalanohis distance is given by

$d^2(x,y) = (x-y)^t \Sigma^{-1} (x-y)$ where Σ is covarience matrix.

EX:- If $x = (5, 2, 4)$ $\quad x_1\ y_1\ z_1$

$y = (3, 4, 2)$ $\quad x_2\ y_2\ z_2$

Euclidean distance $= \sqrt{(x_2-x_1)^2+(y_2-y_1)^2+(z_2-z_1)^2}$

$= \sqrt{(3-5)^2+(4-2)^2+(2-4)^2}$

$= \sqrt{(-2)^2+(2)^2+(-2)^2}$

$= \sqrt{4+4+4} \quad = \sqrt{12}$

$\Rightarrow 3.56$

Manhatten distance

$d(x,y) = \sum_{i=1}^{n} |2| + |2| + |2| = 6$

2. Non - metric similarity function

* this category includes the similarity functions that do not opaque. the triangular inequality of similarity.
* this are commonly used for image or string data and they are does not include outliers of noisy data.

EX:- A non metric similarity function k-median

A non metric similarity function is k-median distance between 2 vectors.

$$x = \{x(1), x(2) \cdots x(l)\}$$

$$y = \{y(1), y(2) \cdots y(l)\}$$

the formula called k-median distance is

$$d(x,y) = k\text{-median}\{|x(1)-y(1)|, x(2)-y(2)-|x(2)-y(1)|\}$$

where

k-median operator returns the k^{th} value of the ordered difference vector.

* the another similarity measure is the cosine of the angle between vectors (x & y). It is symmetric because $\cos(\theta) = \cos(-\theta)$. the cosine similarity measure is

$$s(x,y) = \cos(x,y) = \frac{x \cdot y}{\|x\| \|y\|}$$

. the distance function corresponding to cosine similarity is $d(x,y) = 1 - s(x,y)$, $d(x,y)$ is also symmetric & it does not satisfies triangular inequality.

Ex:- let x, y, z are 3 vectors, the angle between x and z is 45° the angle between z & y is 45°. show that it does not satisfies triangular inequality.

x, y & z

x & z is 45°, z & y is 45°

$$d(x,z) = 1 - \cos(x,z)$$

$$= 1 - \frac{1}{\sqrt{2}} = 0.2928$$

$$d(z,y) = 1 - \cos(z,y) = 1 - \frac{1}{\sqrt{2}} = 0.2928$$

$$d(x,y) = 1 - \cos 90° = 1$$

$$d(x,z) + d(z,y) = 0.2928 + 0.2928 = 0.586$$

$$d(x,y) \leq d(x,z) + d(z,y)$$

$$1 \not\leq 0.586$$

3. Proximity between binary patterns :-

let P & Q are 2 'l' bits, binary vector

(1) Hamming distance (HD) :-

If $P(i) = Q(i)$ then P & Q are matched at ith bit

else if $P(i) \neq Q(i)$ then P & Q mismatched on ith bit,

the hamming distance is no. of mismatched bits of the 'l' bit locations.

Ex :- $P = [0\ 1\ 0\ 1\ 1\ 1\ 0\ 0\ 1]$

$Q = [1\ 0\ 0\ 1\ 0\ 0\ 1\ 1\ 0]$

hamming distance is "7"

because they mismatched in 7 locations

2. simple matching co-efficient (SMC)

let us define the following: m01 is No. of bits where p is 0 and q is 1, m10 is No. of bits where p is 1 and q is 0, m00 is No of bits where P=0 and q is 0, m11 is No. of bits where p is 1 and q is 1

$$SMC(p,q) = \frac{m_{11}+m_{00}}{m_{00}+m_{01}+m_{10}+m_{11}}$$

Jacord co-efficient

Jacord co-efficient is defined as

$$JC(p,q) = \frac{m_{11}}{m_{10}+m_{01}+m_{11}}$$

$$p = 1\ 0\ 0\ 1\ 0\ 0\ 1\ 1\ 0\ 1$$

$$q = 0\ 1\ 1\ 1\ 0\ 0\ 0\ 1\ 1\ 1$$

$$\therefore m_{01} = 3$$

$$\therefore m_{10} = 2$$

$$\therefore m_{00} = 2$$

$$\therefore m_{11} = 3$$

$$SMC(p,q) = \frac{m_{11}+m_{00}}{m_{00}+m_{01}+m_{10}+m_{11}}$$

$$= \frac{3+2}{2+3+2+3}$$

$$= \frac{5}{10} = 0.5$$

$$JC(p,q) = \frac{m_{11}}{m_{10}+m_{01}+m_{11}}$$

$$= \frac{3}{2+3+3}$$

$$= \frac{3}{8}$$

$$= 0.375$$

Different classification algorithm based on distance measure

1. Nearest Neighbor classifier (NNC)

* Here a test pattern T is classified based on its nearest neighbor in the training data. let
* let $x = \{(x_1, l_1)\ (x_2, l_2) \dots (x_n, l_n)\}$ is labeled training data set of n patterns. the first component is pattern vector. and second component is its class label.
* If is class labels if there are P classes with their labels coming from the said $C = \{c_1, c_2 \dots c_P\}$

  $l_i \in C \quad i = 1, 2 \dots n.$
* The nearest neighbor of test pattern T is given by

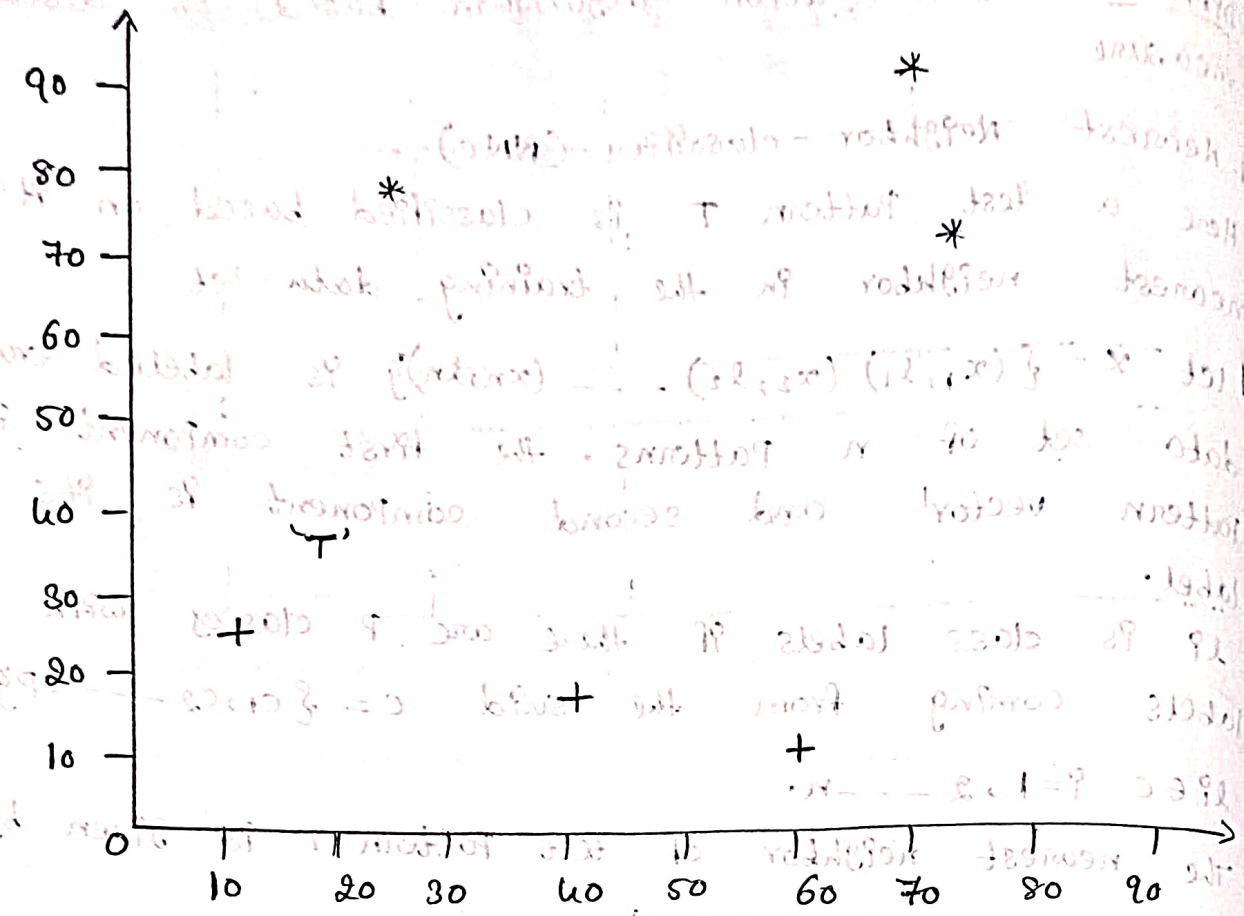
$$NN(T) = \arg\min_{x_i \in x} d(x_i, T)$$

where x_i is ith training pattern and $d(x_i, T)$ is the distance between x_i and T, NN (T) is at the minimum distance from T and it is maximum similar to T the NN rule assigns the T with the class label of NN (T).

EX:- find the class of test pattern using Nearest Neighbor classifier.

	Brightness	saturation	class
x_1	40	20	Red
x_2	50	50	Blue
x_3	60	90	Blue
x_4	10	25	Red
x_5	70	70	Blue
x_6	60	10	Red
x_7	25	80	Blue
	20	35	?

Red '+'
Blue '*'



$$d_T x_1 = \sqrt{(20-40)^2+(35-20)^2} = \sqrt{20^2+15^2} = 25$$

$$d_T x_2 = \sqrt{(20-50)^2+(35-50)^2} = \sqrt{30^2+15^2} = 33.54$$

$$d_T x_3 = \sqrt{(20-60)^2+(35-90)^2} = \sqrt{40^2+55^2} = 68.01$$

$$d_T x_4 = \sqrt{(20-10)^2+(35-25)^2} = \sqrt{10^2+10^2} = 14.14$$

$$d_T x_5 = \sqrt{(20-70)^2+(35-70)^2} = \sqrt{50^2+35^2} = 61.03$$

$$d_T x_6 = \sqrt{(20-60)^2+(35-10)^2} = \sqrt{40^2+25^2} = 47.16$$

$$d_T x_7 = \sqrt{(20-25)^2+(35-80)^2} = \sqrt{5^2+45^2} = 45.27$$

NN (T) is x_4

The class of x_4 is assign to T

class of T is "Red"

Brightness	saturation	class
20	35	Red

let the training

	Brightness	saturation	class	distance	
	40	20	Red	25	
	50	50	Blue	33.54	
	60	90	Blue	68.01	
x_4	10	25	red	14.14	smallest
	70	70	blue	61.03	
	60	10	Red	47.16	
	25	80	blue	45.27	

* let the training set consist of the following two dimensional patterns with associated labels.

$x_1 = (0.7, 0.7)$, $l_1 = 1$;

$x_2 = (0.8, 0.8)$, $l_2 = 1$;

$x_3 = (1.1, 0.7)$, $l_3 = 1$

$x_4 = (1.1, 1.1)$, $l_4 = 1$

$x_5 = (3.7, 2.7)$, $l_5 = 2$

$x_6 = (4.1, 2.7)$, $l_6 = 2$

$x_7 = (3.7, 3.1)$, $l_7 = 2$

$x_8 = (3.4, 0.6)$, $l_8 = 3$

$x_9 = (3.1, 0.8)$, $l_9 = 3$

$x_{10} = (3.1, 0.6)$, $l_{10} = 3$

1 = "Δ"

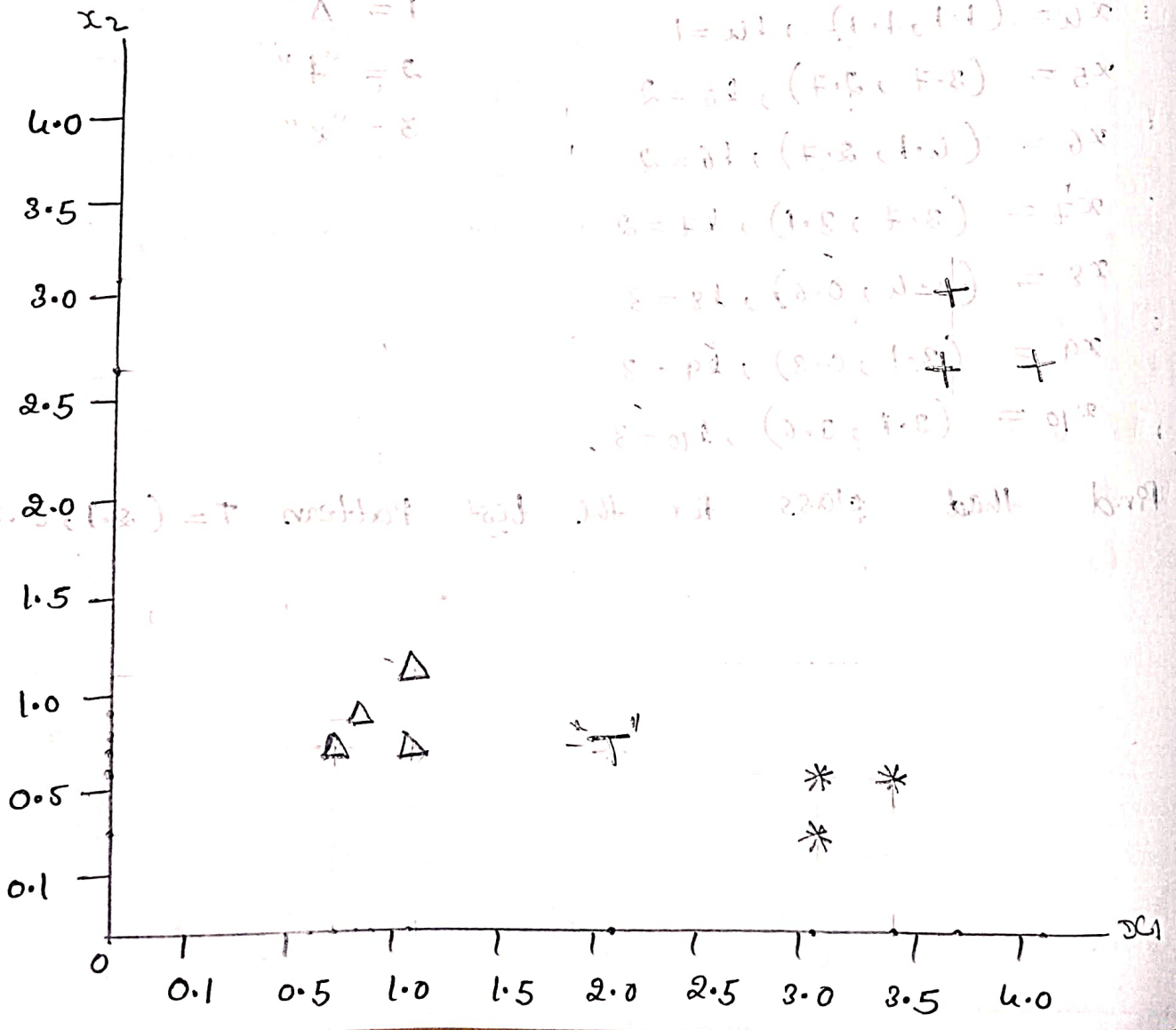
2 = "+"

3 = "*"

find that class for the test pattern $T = (2.1, 0.7)$

	Brightness	saturation	class	
x_1	0.7	0.7	1 △	l_1
x_2	0.8	0.8	1	l_2
x_3	1.1	0.7	1	l_3
x_4	1.1	1.1	1	l_4
x_5	3.7	2.7	2 +	l_5
x_6	4.1	2.7	2	l_6
x_7	3.7	3.1	2	l_7
x_8	3.4	0.6	3 *	l_8
x_9	3.1	0.3	3	l_9
x_{10}	3.1	0.6	3	l_{10}

Brightness	saturation	class
2.1	0.7	?



$$dTx_1 = \sqrt{(2.1-0.7)^2+(0.7-0.7)^2} = \sqrt{(1.4)^2+(0)^2} = 1.4$$

$$dTx_2 = \sqrt{(2.1-0.8)^2+(0.7-0.8)^2} = \sqrt{(1.3)^2+(0.1)^2} = 1.303$$

$$dTx_3 = \sqrt{(2.1-1.1)^2+(0.7-0.7)^2} = \sqrt{(1)^2+(0)^2} = 1$$

$$dTx_4 = \sqrt{(2.1-1.1)^2+(0.7-1.1)^2} = \sqrt{(1)^2+(0.4)^2} = 1.0077$$

$$dTx_5 = \sqrt{(2.1-3.7)^2+(0.7-2.7)^2} = \sqrt{(1.6)^2+(2)^2} = 2.561$$

$$dTx_6 = \sqrt{(2.1-4.1)^2+(0.7-2.7)^2} = \sqrt{(2)^2+(2)^2} = 2.828$$

$$dTx_7 = \sqrt{(2.1-3.7)^2+(0.7-3.1)^2} = \sqrt{(1.6)^2+(2.4)^2} = 2.884$$

$$dTx_8 = \sqrt{(2.1-3.4)^2+(0.7-0.6)^2} = \sqrt{(1.3)^2+(0.1)^2} = 1.303$$

$$dTx_9 = \sqrt{(2.1-3.1)^2+(0.7-0.3)^2} = \sqrt{(1)^2+(0.4)^2} = 1.0077$$

$$dTx_{10} = \sqrt{(2.1-3.1)^2+(0.7-0.6)^2} = \sqrt{(1)^2+(0.1)^2} = 1.004$$

Brightness	saturation	class	distance	
0.7	0.7	1	1.4	
0.8	0.8	1	1.303	
1.1	0.7	1	1	→ smallest
1.1	1.1	1	1.077	
3.7	2.7	2	2.561	
4.1	2.7	2	2.828	
3.7	3.1	2	2.884	
3.4	0.6	3	1.303	
3.1	0.3	3	1.0077	
3.1	0.6	3	1.004	

NN (T) x_3

the class of x_3 is assign to T

class of T is "1"

Brightness	saturation	class
0.7	0.7	1

2. K-Nearest Neighbor classifier

* In K-Nearest Neighbor (KNN) algorithm we find the K-nearest neighbor for the test pattern "T" from the training data "x" then assign majority class label among K-Neighbors to the "T".
* To determinate the K-Nearest neighbor of "T" it is necessary to calculate the distance b/w "T" and each of the "N" Training patterns.
* The distance matric used depends on specific Problem the distance matric can be, Euclidean distance, Manhatten distance are cosine distance
* The class label of "T" is determinate based on the majority class label among its K-nearest neighbor.
* By using this method the errors in classification can be reduce when ever the training patterns are noisy.
* The case of large data set "K" can be increased to reduce the error.
* The value of "K" can be selected such that the lowest error occur in classification the value of "K" usually selected of odd number (K=5).

EX:-

* let training set consist of following data points with the label

$x_1 = (0.7, 0.7)$, $l_1 = 1$

$x_2 = (0.5, 0.5)$, $l_2 = 1$

$x_3 = (1.1, 0.7)$, $l_3 = 1$

$x_4 = (1.1, 1.1)$, $l_4 = 1$

$x_5 = (3.7, 2.5)$, $l_5 = 2$

$x_6 = (4.1, 2.7)$, $l_6 = 2$

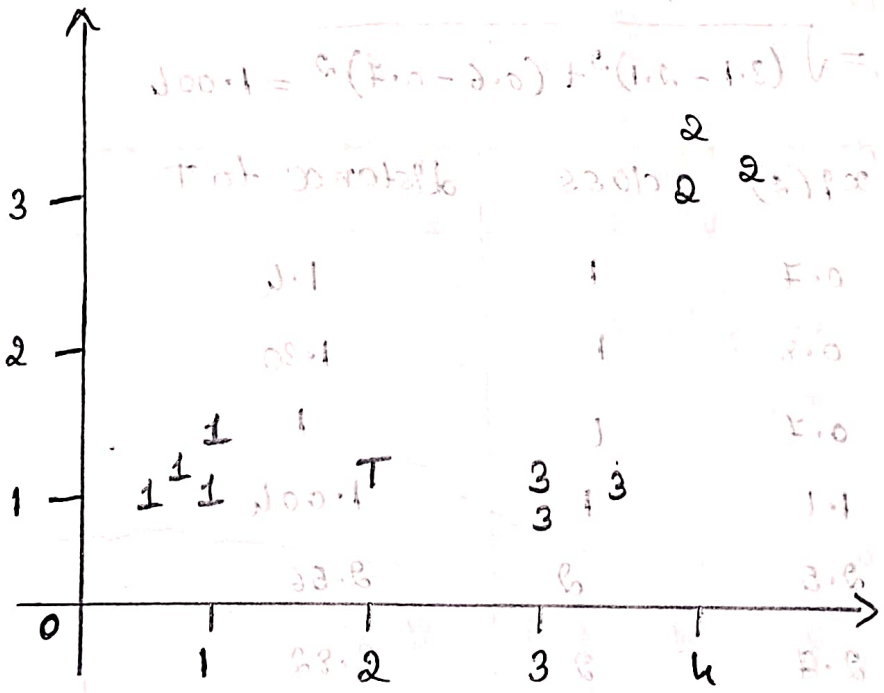
$x_7 = (3.7, 3.1)$, $l_7 = 2$

$x_8 = (3.4, 0.6)$, $l_8 = 3$

$x_9 = (3.1, 0.3)$, $l_9 = 3$

$x_{10} = (3.1, 0.6)$, $l_{10} = 3$

find the label of test pattern $T = (2.1, 0.7)$, $l = ?$ using KNN for $k = 5$



To find the class of test pattern we have to the calculate the distance from test pattern to the each of the training data using eualidean distance.

$$d(x_9,T) = \sqrt{(x_9(1)-T(1))^2+(x_9(2)-T(2))^2}$$

$$d(x_1,T) = \sqrt{(0.7-2.1)^2+(0.7-0.7)^2} = 1.4$$

$$d(x_2,T) = \sqrt{(0.8-2.1)^2+(0.8+0.7)^2} = 1.30$$

$$d(x_3,T) = \sqrt{(1.1-2.1)^2+(0.7-0.7)^2} = 1$$

$$d(x_4,T) = \sqrt{(1.1-2.1)^2+(1.1-0.7)^2} = 1.07$$

$$d(x_5,T) = \sqrt{(3.7-2.1)^2+(2.5-0.7)^2} = 2.56$$

$$d(x_6,T) = \sqrt{(4.1-2.1)^2+(2.7-0.7)^2} = 2.82$$

$$d(x_7,T) = \sqrt{(3.7-2.1)^2+(3.1-0.7)^2} = 2.88$$

$$d(x_8,T) = \sqrt{(3.4-2.1)^2+(0.6-0.7)^2} = 1.30$$

$$d(x_9,T) = \sqrt{(3.1-2.1)^2+(0.3-0.7)^2} = 1.07$$

$$d(x_{10},T) = \sqrt{(3.1-2.1)^2+(0.6-0.7)^2} = 1.004$$

	$x_9(1)$	$x_9(2)$	class	distance to T
x_1	0.7	0.7	1	1.4
x_2	0.8	0.8	1	1.30
x_3	1.1	0.7	1	1
x_4	1.1	1.1	1	1.004
x_5	3.7	2.5	2	2.56
x_6	4.1	2.7	2	2.82
x_7	3.7	3.1	2	2.88
x_8	3.4	0.6	3	1.30
x_9	3.1	0.3	3	1.077
x_{10}	3.1	0.6	3	1.004

	x1(1)	x1(2)	class	distance to T
x_3	1.1	0.7	①	1
x_{10}	3.1	0.6	3	1.004
x_4	1.1	1.1	①	1.04
x_9	3.1	0.3	3	1.077
x_2	0.8	0.8	①	1.30

Since the value of $k=5$ we will consider first 5 rows the majority class with in the first 5 nearest neighbors to the test pattern is 1 therefore the label of test pattern is 1.

Test Pattern

$T = (2.1, 0.7), l = 1$

find the class of test pattern using Nearest Neighbor classifier.

Brightness	saturation	class	distance
40	20	Red	25
50	50	blue	33.54
60	90	blue	68.01
10	25	Red	14.14
70	70	blue	61.03
60	10	Red	47.16
25	80	Blue	45.27

Brightness	saturation	class	distance
10	25	(Red)	14.14
40	20	(Red)	25
50	50	blue	33.54
25	80	blue	45.27
60	10	(Red)	47.16

T = (20, 35) class = Red

consider a scenario where a realstate agency wants to Predict the Prize of a house based on features such as area (square field) and age of the house (Yes) explain how KNN recursion can be used. In this case given a small data set of 5 houses demonstrate the steps to Predict the Prize of a new house using k=3 Include the distance calculation, selection of neighbors and competition of Predicted Prize discuss how different values of k might effect the Prediction outcome.

House Io	Area (sq.ft)	Bedrooms	Age (year).	Price in (1000s)
H_1	2100	3	10	500
H_2	1600	2	5	330
H_3	2400	4	20	550
H_4	1410	3	15	299
H_5	1800	2	8	410

New house to Predict

* Area = 1900 sq.ft
* Bedroom = 3
* Age = 10 years

3. Radius distance Nearest Neighbour :-

* this algorithm consider all the neighbours with in a specified distance "r" of the test Pattern this algorithm can be described as.

1. Given a Point T identify the data Points that fall within the radius centered at "T" denoted by

$$Br(T) = \{xp \in x \text{ s.t } \|T - xp\| \leq r\}$$

2. If BR(T) is empty the output is the majority class of entire dataset

3. If BR(T) is notempty the output is the majority class of class of the data points with in BR(T)

4. this algorithm is useful for identifying outlayers the choice of value of radius (or) is important because it can effect performance of the algorithm.

5. this algorithm is more efficient than KNN algorithm.

* for the training data set given below

$x_1 = (0.7, 0.7), l_1 = 1;$

$x_2 = (0.8, 0.8), l_2 = 1;$

$x_3 = (1.1, 0.7), l_3 = 1;$

$x_4 = (1.1, 1.1), l_4 = 1;$

$x_5 = (3.7, 2.7), l_5 = 2;$

$x_6 = (4.1, 2.7), l_6 = 2;$

$x_7 = (3.7, 3.1), l_7 = 2;$

$x_8 = (3.4, 0.6), l_8 = 3;$

$x_9 = (3.1, 0.3), l_9 = 3;$

$x_{10} = (3.1, 0.6), l_{10} = 3;$

find the class of test pattern $T = (1.7, 2)$ using radius distance Nearest neighbor algorithm with the radius of 1.45

To find the data points with in the radius of 1.45 we need to find the Euclidean distance between each of the data point is the training set and test pattern. the euclidean distance

$$d(x_i, T) = \sqrt{((x_i(1) - T(1))^2 + (x_i(2) - T(2))^2}$$

Now find out distance

$$d(x_1,T) = \sqrt{(0.7-1.7)^2+(0.7-2)^2} = 1.64$$

$$d(x_2,T) = \sqrt{(0.8-1.7)^2+(0.8-2)^2} = 1.58$$

$$d(x_3,T) = \sqrt{(1.1-1.7)^2+(0.7-2)^2} = 1.43$$

$$d(x_4,T) = \sqrt{(1.1-1.7)^2+(1.1-2)^2} = 1.08$$

$$d(x_5,T) = \sqrt{(3.7-1.7)^2+(2.7-2)^2} = 2.11$$

$$d(x_6,T) = \sqrt{(4.1-1.7)^2+(2.7-2)^2} = 2.29$$

$$d(x_7,T) = \sqrt{(3.7-1.7)^2+(3.1-2)^2} = 2.28$$

$$d(x_8,T) = \sqrt{(3.4-1.7)^2+(0.6-2)^2} = 2.20$$

$$d(x_9,T) = \sqrt{(3.1-1.7)^2+(0.3-2)^2} = 2.20$$

$$d(x_{10},T) = \sqrt{(3.1-1.7)^2+(0.6-2)^2} = 1.97$$

	xp(1)	xp(2)	label	distance to T
x_1	0.7	0.7	1	1.64
x_2	0.8	0.8	1	1.58
x_3	1.1	0.7	(1)	(1.43)
x_4	1.1	1.1	(1)	(1.08)
x_5	3.7	2.7	2	2.11
x_6	4.1	2.7	2	2.29
x_7	3.7	3.1	2	2.28
x_8	3.4	0.6	3	2.20
x_9	3.1	0.3	3	2.20
x_{10}	3.1	0.6	3	1.97

From the table the patterns located with in a radius of 1.45 or x_3 and x_4 the class of this patterns are 1. therefore the text pattern assigned with

$$\boxed{T = (1.7, 2),\ \text{class} = 1}$$

***

4. KNN Regression

* let x is set of n label data

$x = \{(x_1, y_1)\ (x_2, y_2)\ (x_3, y_3) \dots (x_n, y_n)\}$

where,

x = is a data vector

Y = scalar values

Here x_i, $i = 1, 2 \dots n$ is a data vector

y_i, is a scaler

the regression model needs to use x to find the value of Y for a text vector T. In the case of regression based on KNN the following steps are followed:

* find k- Nearest neighbor of "T" from n data vectors let KNN are $x_1', x_1 2 \dots x^k$
* Consider Y values associated with this k nearest neighbors. let them r $y_1', y 2 \dots y k$
* take the average value of this y values and assign this average value as Predicted value of y to "T"

the Predicted value of Y is $y\lambda = \frac{1}{k} [y' + y^2 \dots + y^k]$

Ex:- consider the data shown in table find the Predicted target value for T = (0.3, 0.4) using KNN regression.

	$x_i(1)$	$x_i(2)$	target y_i
x_1	0.2	0.4	8
x_2	0.4	0.2	8
x_3	0.6	0.4	12
x_4	0.8	0.6	16
x_5	1.0	0.7	19
x_6	0.8	0.4	14
x_7	0.6	0.2	10
x_8	0.5	0.5	12
x_9	0.2	0.6	10

To find k nearest neighbors we apply Euclidean distance b/w data vectors and test pattern

$$d(x_1,T) = \sqrt{x_1(1)-T(1)^2-(x_1(2)-T(2))^2}$$

$$d(x_1,T) = \sqrt{(0.2-0.3)^2+(0.4-0.4)^2} = 0.1$$

$$d(x_2,T) = \sqrt{(0.4-0.3)^2+(0.2-0.4)^2} = 0.223$$

$$d(x_3,T) = \sqrt{(0.6-0.3)^2+(0.4-0.4)^2} = 0.3$$

$$d(x_4,T) = \sqrt{(0.8-0.3)^2+(0.6-0.4)^2} = 0.52$$

$$d(x_5,T) = \sqrt{(1.0-0.3)^2+(0.7-0.4)^2} = 0.76$$

$$d(x_6,T) = \sqrt{(0.8-0.3)^2+(0.4-0.4)^2} = 0.5$$

$$d(x_7,T) = \sqrt{(0.6-0.3)^2+(0.2-0.4)^2} = 0.36$$

$$d(x_8,T) = \sqrt{(0.5-0.3)^2+(0.5-0.4)^2} = 0.223$$

$$d(x_9,T) = \sqrt{(0.2-0.3)^2+(0.6-0.4)^2} = 0.223$$

$x_p(1)$	$x_p(2)$	Target y_p	distance
0.2	0.4	8	0.1
0.4	0.2	8	0.223
0.6	0.4	12	0.3
0.8	0.6	16	0.52
1.0	0.7	19	0.76
0.8	0.4	14	0.5
0.6	0.2	10	0.36
0.5	0.5	12	0.223
0.2	0.6	10	0.223

x_1	0.2	0.4	8	0.1
x_8	0.5	0.5	12	0.223
x_2	0.4	0.2	8	0.223
x_9	0.2	0.6	10	0.223
x_3	0.6	0.4	12	0.3

If K=1 then the Nearest neighbor is x_1

If K=5 then the 5 Nearest neighbor the text pattern x_1, x_8, x_2, x_9, x_3 and the labels of this neighbors are 8, 12, 8, 10, 12 the Predicted target value

$$y\hat{} = \frac{1}{5}(8+12+8+10+12)$$

$$= \frac{1}{5}(50) = 10$$

the target value $\boxed{T = (0.3, 0.4), y\hat{} = 10}$

Performance measure

Performance of classifier :-

i) classification Accuracy

let n the total no. of patterns that are classified by a classification algorithm. let nc is the No. of correctly classified patterns classification Accuracy

$$= \frac{nc}{n}$$

ii) confusion matrix :-

It is used to store to result simplified matrix formate this data is further simplified to analyze the result.

Ex :-

let they are three classes c_1, c_2 and c_3 there are 200, 100 and 50 patterns from classes c_1, c_2 and c_3 respectively. let a classifier classify this patterns into three classes as shown in the table.

True / predicted	c_1	c_2	c_3
c_1	(180)	15	5
c_2	5	(85)	10
c_3	3	2	(45)

* the first row of table shows that 200 Patterns from C_1 are classified as
* 180 are assigned to C_1, 15 are assigned to C_2 and 5 are assigned to C_3. similarly second row for 100 Patterns from C_2 and third row shows the assignment of 50 Patterns from C_3
* the classification Accuracy is obtained by considering the correctly classified Patterns that is 180 Patterns from C_1, 85 Patterns from C_2, and 45 Patterns from C_3 therefore No. of correctly classified Patterns.

$$= 180 + 85 + 45$$
$$= 310$$

total no. of Patterns $= C_1 = 200$
$C_2 = 100$
$C_3 = 50$
350

the classification Accuracy $= \frac{n_c}{n} = \frac{310}{350} = 0.8857$

the percentage classification Accuracy is 88.57 %

* It is possible to simplify the confusion table with respect to C_1 the compact table is of size 2×2

True / Predicted	C_1	$\overline{C_1}$ (Not C_1)
C_1	TP 180	FN 20
$\overline{C_1}$	FP 8	TN 142

* The first row of the table shows that 180 Patterns of C_1 are correctly classified therefore thes are called True Postives (or) TP. TP = 180
* The second column and first row has 20 patterns these Patterns are from C_1 that are Not classified

as c_1 they are assigned to $\bar{c_1}$ (Not c_1) they are called false negative (or) FN. FN = 20.

* the first entry in second row shows that 8 patterns from other class are assigned to c_1 so they are called false positive (or) FP. FP = 8.

* The second column in the second row shows that 142 patterns from $\bar{c_1}$ (not c_1) assigned to $\bar{c_1}$ they are called True negative (or) TN. TN = 142 the confusion matrix is given by

True / Predicted	c_1	$\bar{c_1}$
c_1	TP	FN
$\bar{c_1}$	FP	TN

this matrix is used to find following measures

* Precision :-

It is the ratio of TP to total no. of Predicted positives (TP + FP)

$$\text{Precision} = \frac{TP}{TP+FP} = \frac{180}{180+8} = \frac{180}{188}$$

$$= 0.9574$$

* Recall :-

It is the ratio of TP to total no. of positives in the training data (TP + FN)

$$\text{Recall} = \frac{TP}{TP+FN} = \frac{180}{180+20} = \frac{180}{200}$$

$$= 0.9$$

* F1 - score

It is the harmonic mean of precision and recall

$$f_1 - \text{score} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}}$$

$$= \frac{2}{\left(\frac{1}{0.9574}\right)+\left(\frac{1}{0.9}\right)}$$

$$= 0.9278$$

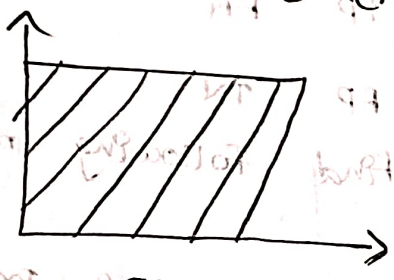
* True Positive Rate or TPR

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

* false Positive Rate or FPR

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

* Area under the curve (AUC)



AUC is obtained by plotting the graph b/w FPR on X-axis and TPR on Y-axis calculate the area under discard this is also called the receive of (receiver operator characteristics) or (AUC)

Performance of Regression algorithm:-

1. Mean squared Error (MSE)

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

where n is no. of Patterns

y_i is the target value for ith Pattern

$\hat{y}_i$ is the Predicted value of ith Pattern using regression model.

2. Mean absolute Error:- (MAE)

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|$$

It is average of difference b/w target and predicted value.

Ex:- If $TP=2$, $FN=0$, $FP=1$, $TN=1$ find Precision, Recall, f_1-score, True Positive Rate, False Positive Rate and classification accuracy.

True / Predicted	C_1	$\overline{C_1}$
C_1	$TP=2$	$FN=0$
$\overline{C_1}$	$FP=1$	$TN=1$

Precision :-

$$\frac{TP}{TP+FP} = \frac{2}{2+1} = \frac{2}{3} \Rightarrow 0.6666$$

Recall :-

$$\frac{TP}{TP+FN} = \frac{2}{2+0} = \frac{2}{2} = 1$$

f_1 - score :-

$$\frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} = \frac{2}{\frac{1}{0.6666} + \frac{1}{1}}$$

$$\Rightarrow 1.3332$$

True Positive Rate :-

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

$$= \frac{2}{2+0} = \frac{2}{2} \Rightarrow 1$$

False Positive Rate :-

$$FPR = \frac{FP}{FP+TN} = \frac{1}{1+1} = \frac{1}{2} \Rightarrow 0.5$$

classification Accuracy

$$\frac{nc}{n}$$

$$nc = TP+TN = 2+1 = 3$$

$$n = TP + TN + FP + FN$$

$$= 2 + 1 + 1 + 0$$

$$= 4$$

$$\frac{nc}{n} = \frac{3}{4} = 0.75$$

The Percentage of Accuracy is 75%.

2) TP = 2, FN = 0, FP = 0, TN = 2 calculate classification Performance measures

True/ Predicted	C1	$\overline{C1}$
C1	TP = 2	FN = 0
$\overline{C1}$	FP = 0	TN = 2

Precision :-

$$\frac{TP}{TP+FP} = \frac{2}{2+0} = \frac{2}{2} = 1$$

Recall :-

$$\frac{TP}{TP+FN} = \frac{2}{2+0} = \frac{2}{2} = 1$$

f1 score :-

$$\frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} = \frac{2}{\frac{1}{1} + \frac{1}{1}} = \frac{2}{2} = 1$$

True positive rate :-

$$TPR = \frac{TP}{TP+FN} = \frac{2}{2+0} = \frac{2}{2} = 1$$

False positive rate :-

$$FPR = \frac{FP}{FP+TN} = \frac{0}{0+2} = \frac{0}{2} = 0$$

classification Accuracy

$$\frac{nc}{n}$$

$$nc = TP + TN = 2 + 2 = 4$$

$$n = TP + TN + FP + FN$$

$$= 2 + 2 + 0 + 0$$

$$= 4$$

$$\frac{nc}{n} = \frac{4}{4} = 1$$