Machine Learning

Urvash Shende
20051627

Machine Learning

Module - 1 : Introduction to Machine Learning

→ What is Machine learning
→ What is the need, motivation for ML
→ Well defined ML Problem
→ What are the 3 main features
→ What are the Key concepts of Machine learning

•) Machine Learning → The capability of a machine to imitate intelligent human behaviour

•) Need and motivation →

The combination of computing and communication has given rise to a society that feeds on information.

→ Info is categorized on the basis of natual of expeience

Experimental Data	Structued Human Knowledge
→ examples	→ Experience
→ Sample	→ expertise
→ measurements	→ main expressed in linguistic form (follows IF-THEN RULE)
→ patterns	↳ used in SHL.

•) Definition of Machine learning & 3 main features

→ A computer program is said to learn from experience with respect to some class of task and performance measure, if the performance and task as measured by the performance measure, improved with the experience.

→ 3 feature

1) The learning task. → 'E' → Experience.
2) The measure of performance. → 'P' → Performance.
3) The task experience. → 'T' → Task.

12/1/2023

• Feature / attributes → Inputs are the feature of ML.

→ The key concept is "learning from the experience"
→ Important aspects of "learning from the experience" are

1) Remembering and Adapting

→ Recognising that last time in a similar situation a certain action was attempted which resulted in the output so it should be tried again or the same action failed in a similar situation so something different should be tried

2) Generalizing

→ This aspect is recognizing similarity between different situation so a decision can be taken for a new situation.

• Application of ML

1) Social media (Sentiment analysis, filtering spam)
2) Transport (Safety Monitoring, Air traffic control etc)
3) Finance (Fraud detection, Portfolio management)
4) Health care centre (Drug Discovery, Disease Diagnosis)
5) eCommerce (Coustomer support, product recommendation etc)
6) Virtual assistants (Intelligent agent, NLP)

Name → Urvash Shende
Roll no → 20051627
Section → CS_11

Class Activity - 1

•) Write in about 200 - 300 words about one application of machine learning in any field.

→ SPACE EXPLORATION:

⇒ One application of machine learning in the field of space exploration is the use of autonomous systems for spacecraft navigation and control. These systems uses machine learning algorithms to analyze data from onboard sensors and make decisions about how to navigate and control the spacecraft. For example, NASA's Jet Propulsion Laboratory has developed an autonomous navigation system for the Mars rovers that uses machine learning to analyse data from cameras and other sensors to identify landmarks and obstacles in the rovers path. The system then uses this information to plan a safe and efficient path to the rover's next destination.

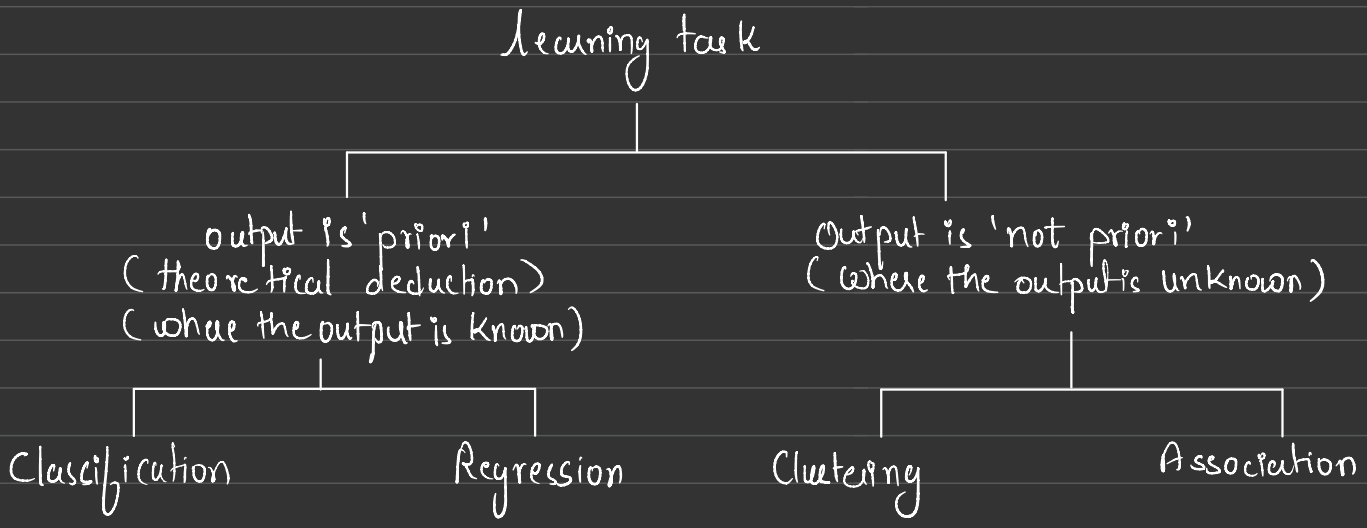
Machine learning can also be used to improve the efficiency of communication between space craft and ground control. For instance the European space agency's Hera mission, a proposed asteroid deflection test mission plans to use machine learning to optimize the data compression and transmission from the spacecraft to the ground, allowing for more data to be transmitted in a shorter amount of time.

Another example is the use of machine learning for the analysis of data from scientific instruments on space crafts. for instance NASA's Kepler project used machine learning to analyse data from the Kepler spectrometer on Cassini spacecraft, which orbited Saturn for over a decade. The machine learning algorithms helped to identify and classify different types of particles in the Saturnain system, enabling scientist to better understand the composition and structure of the planet and its moons.

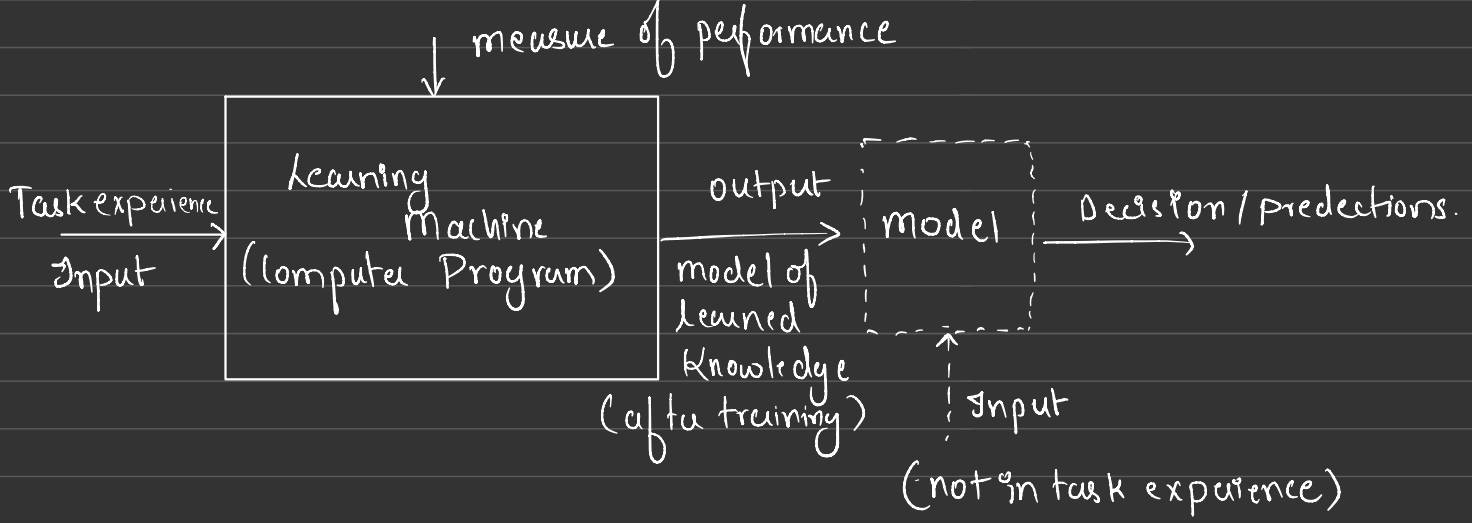
Overall, machine learning plays a crucial role in space exploration by enabling spacecraft to navigate and control themselves more efficiently analyze large amounts of scientific data, and improve communication between spacecraft and ground control, with increasing amount of data being generated by spacecraft, machine learning will continue to be an important tool for making new discoveries in the study of our solar system and beyond

→ learning task

⇒ learning task is divided into 2 types



•) Block diagram Representation of Machine learning



Input → Training Data (experimental data / Human knowledge data)
Testing Data

Step used in Machine learning

1) Collecting Data (raw data from excel, access, text files etc)

2) Preparing the data : Pre processing of data

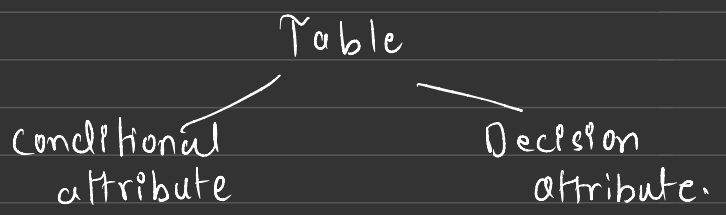
3) Training the model : There are two types of inputs 1st is training then it goes to learning machine then it goes to model where testing input is

4) Evaluating the model : To test the accuracy, the second part of the data (Test data) is used.

5) Improving the performance : This step might involve choosing a different model altogether or introducing more variable.

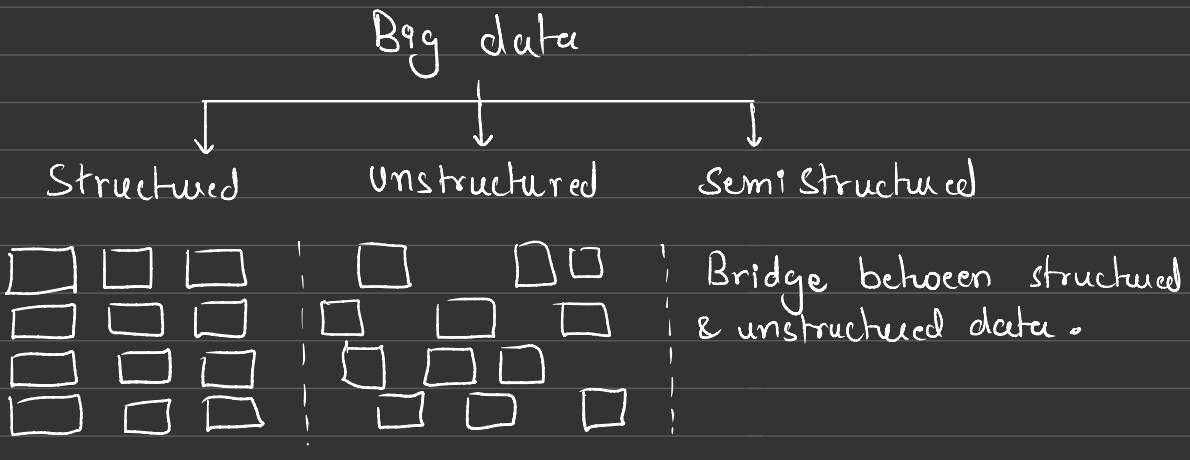
•) Data Representation

→ Raw data requires some preprocessing wrt class of tasks
→ This leads to an information system may be stored in the data warehouse used for decision making. (knowledge in the raw data.)
→ Information is divided into table in two ways.



•) Diversity of Data.

→ Data in petabytes (10^{15} byte) is called Big data.



⇒ The data which is lacks any predefined model/pattern. to the point & highly organized is referred to as structured data.

Activity (Tabulate the difference between structured and unstructured data using some parameter.)

	Parameter	Structured Data	Unstructured Data
1)	Scalability	It can be easily scaled up or down as the volume of data grows	It can be more difficult to scale as the volume & variety of data increases
2)	Data Quality	High, as data is well organised & follows pre-defined rules	Low, as data may be inconsistent or contain errors
3)	Data Accessibility	Easily accessible using standard database management systems	Can be more difficult to access and may require specialized software.
4)	Storage	Can be stored in variety of formats, such as spread sheets or databases which are optimized for structured data.	Can be more challenging to store as unstructured data can come in many different formats & may require specialized storage soln
5)	Searchability	Can be easily searched and filtered using specific criteria	Can be more difficult to search & filter, often requiring NLP or other advanced Techniques
6)	Format	It has predefined format for example consider a XL sheet which is predefined	It has variety of formats ie it comes in a variety of shapes & sizes.

19/1/23

Krish nayak for ML topics

•) The process of teaching machines can be broken down into 3 parts.

Data as Input	Abstracting the data	Generalization
Text files, spreadsheet SQL Databases	Representation of the data in a structured format	The practical application happens here where the learning from the previous step is used to develop an insight.

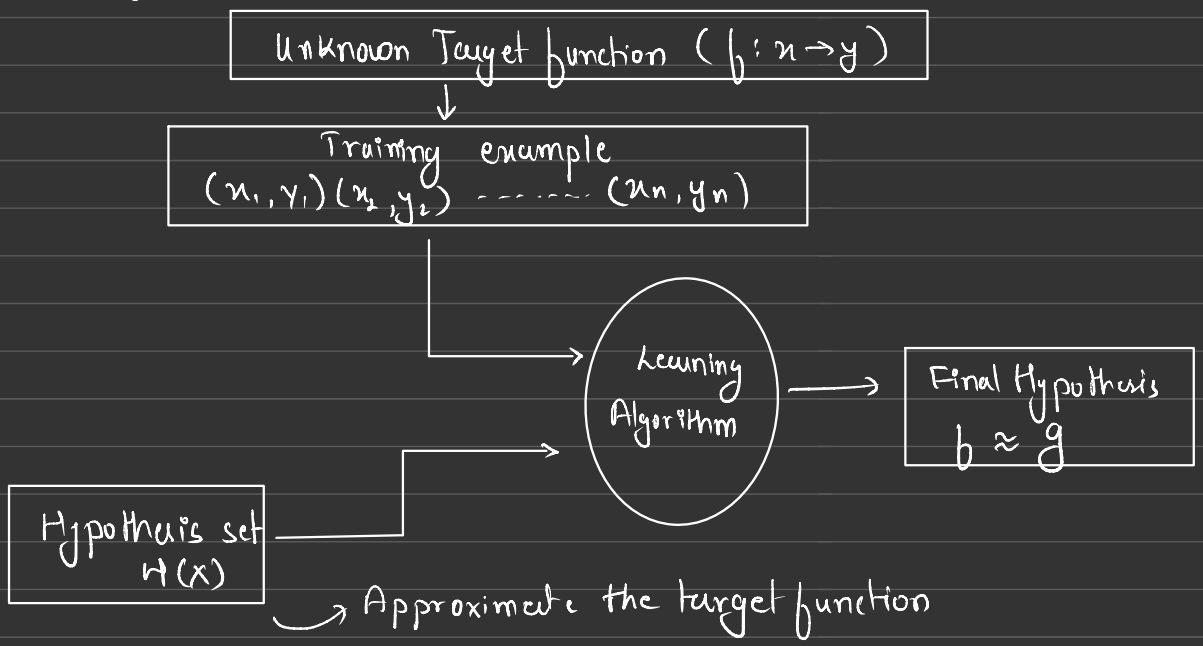
20/1/23 (Lect 4)

•) Simple learning process.
•) Design of a learning process.
→ checkers game

Well defined ML problem (T, E, P)

*) Simple learning process:

Target function (Mapping function) → maps relation between input & output.



•) Design of a learning process

1) Choosing a training experience
   → Direct training experience
   → Indirect training experience

2) Choosing the target function (maping : f(x) → y)

3) Representation of Target function :

   output : $w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3 + \cdots\cdots$

4) Error Calculation / Approximation Algorithm.

   E = Actual value - predicted value.

•) Types of Machine learning algorithms.

1) Supervised learning
   → classification
   → regression

2) Unsupervised learning
   → Clustering → Targated marketing, Recommender System, customer segmentation
   → Demensionality Reduction → meaningfull

3) Reinforcement learning
   → Game AI
   → skill application
   → learning task
   → Robot navigation

• Supervised learning.

→ The machine is designed by making use of the prior known information in the form of direct training examples consisting of observe values of system state. and the response of each state ie output vector.

$x \rightarrow$ (input vector)
$Y \rightarrow$ (output vector)

• Regression → Prediction Analysis. (Output is continous)
→
→ Futuristic Data analysis (Forcasting)
→ Deals with huge amount of data

Application : population growth analysis.
: estimating life expectancy.
: Market forcasting
: Advertisment popularity prediction.

• Classification: Application : Diagonistic.
(output is discrete) : Costoumer Retension
: Spam mail Detection.
: Image classification.

• Unsupervised : It is not associated with any kind of guidance
: No concrete data.
: Random data is allocated.
:

1) Clustering: Not at all structured data is available.
→ Creation of sets by segregating data by some certain a attributes is clustering.

2) Dimentionality Reduction: Data reduction. (on application of data)

Application → Structure Discovery.

• Reinforment learning.

→ Instant decision making comes under reinforcement learning

• UNIT 2 (Fundamentals of Machine learning)

→ Three Basic method of Machine learning.

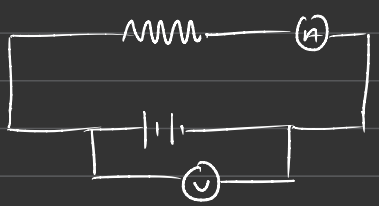
1) Least Square method
2) KNN method ( )
3) Distance based method

1) Least Squared Method

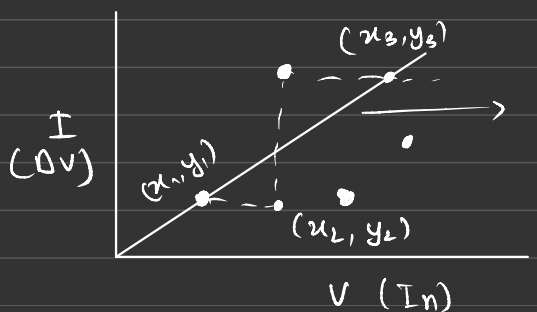
→ Used to find best fit line

Example:

Ohm's law



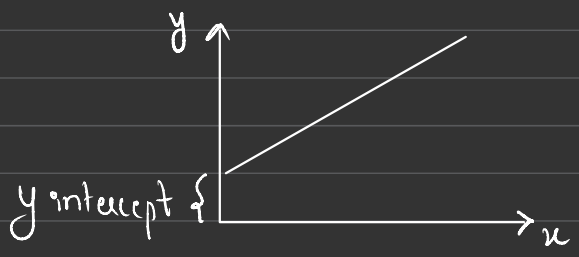
Voltage – Independent variable.
Current – Dependent variable.



→ line of best fit / regression line

• Best fit line ⟶ $y = mx + c$ (c → y intercept)

- dependent variable → target variable
- Slope (m)
- Independent variable → Predicator variable



⇒ Definition: It is based on the idea that the square of the errors obtained must be minimized to the most possible extent hence the name is given least square method.

formula:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [y_i - f(x_i)]^2 \quad \text{should be least.}$$

Simple linear Regression Model.

The least square line $y = a + bx$ for the given data

We can determine the values of a and b in the simple linear regression equation using the give data by minimising the error function. which is

$$E(a,b) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$E(a,b) = \sum_{i=1}^{n} (y_i - a + bx_i)^2$$

Differentiating the error function with respect to a and b and equating them to zero constitutes a set of two equation

$$\Rightarrow \frac{\partial E(a,b)}{\partial a} = -2\sum_{i=1}^{n}(y_i - a - bx_i) = 0$$

$$= \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} a - \sum_{i=1}^{n} bx_i = 0$$

$$= \Sigma y - na - b\Sigma x = 0$$

$$= na + b\Sigma x = \Sigma y \rightarrow (1)$$

$$\Rightarrow \frac{\partial E(a,b)}{\partial b} = -2\sum_{i=1}^{n} x_i(y_i - a - bx_i) = 0$$

$$= \sum_{i=1}^{n} x_i y_i - a\sum_{i=1}^{n} x_i - \Sigma bx_i^2 = 0$$

$$= \Sigma xy - \Sigma ax - \Sigma bx^2 = 0$$

$$\Rightarrow \Sigma xy = \Sigma ax + \Sigma bx^2 \rightarrow (2)$$

(1) & (2) are called normal equation.

Q) Find the least square line $y = a + bx$ for the given data

x	-2	-1	0	1	2
y	1	2	3	3	4

x	y	x^2	xy
-2	1	4	-2
-1	2	1	-2
0	3	0	0
1	3	1	3
2	4	4	8

$\Sigma x = 0 \quad \Sigma y = 13 \quad \Sigma x^2 = 10 \quad \Sigma xy = 7$

$5a = 13 \quad = \quad a = 13/5 = 2.6$

$0.0 + 10b = 7 \quad b = 10/7 = 0.7$

$y = 2.6 + (0.7)x$

Q)

Year	1958	1959	1960	1961	1962
Sales	65	95	80	115	105

deviation (n = odd)

⇒

Year	Sales	x	x^2	xy	
1958	65	−2	4	−130	225
1959	95	−1	1	−95	
1960	80	0	0	0	
1961	115	1	1	115	325
1962	105	2	4	210	
Total	$\Sigma y = 460$	$\Sigma x = 0$	$\Sigma x^2 = 10$	$\Sigma xy = 100$	

$$Na + b\Sigma x = \Sigma y$$

$$5a = 460$$

$$a = 92$$

$$\Sigma xy = \Sigma ax + \Sigma bx^2$$

$$100 = b\,10$$

$$a = 92,\ b = 10$$

$$y = 92 + 10x$$

• Equation for Best fit line ($y = a + bx$)

Normalize equation

$$na + b\Sigma x = \Sigma y$$

$$a\Sigma x + b\Sigma x^2 = \Sigma xy$$

Example 3

Find the trend values

1953.5

Year :	1950	1951	1952	1953	1954	1955	1956	1957
Value :	346	411	392	512	626	640	611	796

Year	New value ($\tilde{x} = x - 1953.5$) x	y	x^2	xy
1950	-3.5	346	12.25	-1211
1951	-2.5	411	6.25	-1027.5
1952	-1.5	392	2.25	-588.0
1953	-0.5	512	0.25	-256.0
1954	0.5	626	0.25	313.0
1955	1.5	640	2.25	960.0
1956	2.5	611	6.25	1527.5
1957	3.5	796	12.25	2786.0
	$\Sigma x = 0$	$\Sigma y = 4334$	$\Sigma x^2 = 42$	2504.0

$$y = 541.75 + (59.6)x$$

Trend values

333.15
392.75
452.35

$$na + b\Sigma x = \Sigma y$$
$$a\Sigma x + b\Sigma x^2 = \Sigma xy$$

•> Nearest Neighbour Method. / Non Parametric Method.
Instance based / Memory based learning method. / Distance based.

The steps in KNN classification

step 1 : Look at the data

Step 2 : Calculate the distance.

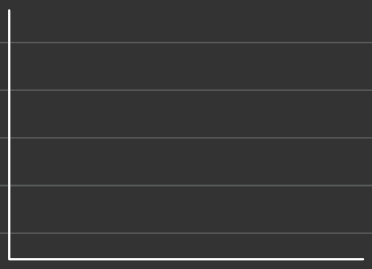
Step 3 : Find neighbour

Step 4 : Vote on labels

Example

•> Predict the weight of the person with ID 11 based on his height and age.

Height	Age	Weight
5	45	77
5.11	26	47
5.6	30	55
5.9	34	59
4.8	40	72
5.8	36	60
5.3	19	40
5.8	28	60
5.5	23	45
5.6	32	58
5.5	38	69.6 ?



Id 1 5 6

$\frac{77 + 72 + 60}{3} = 69.6$ kg

K=3 ID 11 weight = 69.6 kg
K=5 ID 11 weight = 65.2 kg

KNN works best when data is not too large.

$K = 5$

$I_{D11} = \underset{(x_2)}{5.5} \quad \underset{(y_2)}{38} \quad$ Wunknown

$I_{D1} = \sqrt{(x_2 - x_1)^2 - (y_2 - y_1)^2} = \sqrt{(5.5-5)^2 + (38-45)^2} = 7.017$

$I_{D2} = \sqrt{(5.5 - 5.11)^2 - (38-26)^2} = 12.04$

$I_{D3} = 8.0006$

$I_{D4} = 4.019$

$I_{D5} = 2.11$

$I_{D6} = 2.022$

$I_{D7} = 19.001$

$I_{D8} = 10.0044$

$I_{D9} = 15$

$I_{D10} = 6.0008$

• Distance Based learning Method.

→ Distance can be defined in multiple ways

Non-negativity : $d_{il} \geq 0$

Self-proximity : $d_{ii} = 0$

Symmetry : $d_{il} = d_{li}$

Triangle Inequality : $d_{il} \leq d_{ik} + d_{kl}$

• Properties of Distance

1) If the two objects are not same then distance between them should be greater than zero. (Positivity)

2) Symmetry

3) Self proximity.

4) Distance between two object is less than the summation of these distance where k is common to both i & L is called triangle inequality.

5) If a distance satisfies all its properties then it is called as metric

Types of Distances

1) Euclidean Distance

$$d_{il} = \sqrt{(x_1^{(i)} - x_1^{(l)})^2 + (x_2^{(i)} - x_2^{(l)})^2 + \cdots + (x_n^{(i)} - x_n^{(l)})^2}$$

• Statistical Distance

$$d_{ij} = \sqrt{(x^i - x^j)^T \Sigma^{-1} (x^i - x^j)}$$

x^i & x^j are n-dimensional vectors of measurement value of pattern iff. Σ is co variance matrix of these vectors.

• Manhattan Distance

→ $$d_{ij} = \sum_{i=1}^{n} |x_i - y_i|$$

• Minkowoski

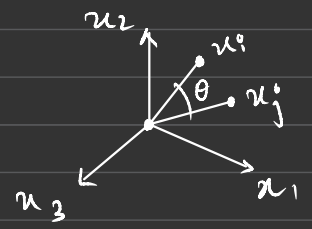
→ $$L_p(x^i, x^j) = \left(\sum_{j=1}^{n} |x_j^{(i)} - x_j^{j}|^p \right)^{1/p}$$

• Hamming Distance.

Using XOR operation

Calculate distance between two binary vectors.

• Cosine Similarity :-



The Cosine Similarity between two vectors x_i & x_j is defined as cosine of the angle θ between them.

$$\cos\theta = \frac{x_1^i x_1^j + x_2^i x_2^j + \cdots + x_N^i x_N^j}{\sqrt{\sum_{j=1}^{N} (x_j^i)^2} \times \sqrt{\sum_{j=1}^{N} (x_j^i)^2}}$$

→ for a given new data point $B = (1.4, 1.6)$ find the cosine similarity wrt the data given.

Cosine similarity

Sol

Sl No	A_1	A_2
1	1.5	1.7
2	2	1.9
3	1.6	1.8
4	1.2	1.5

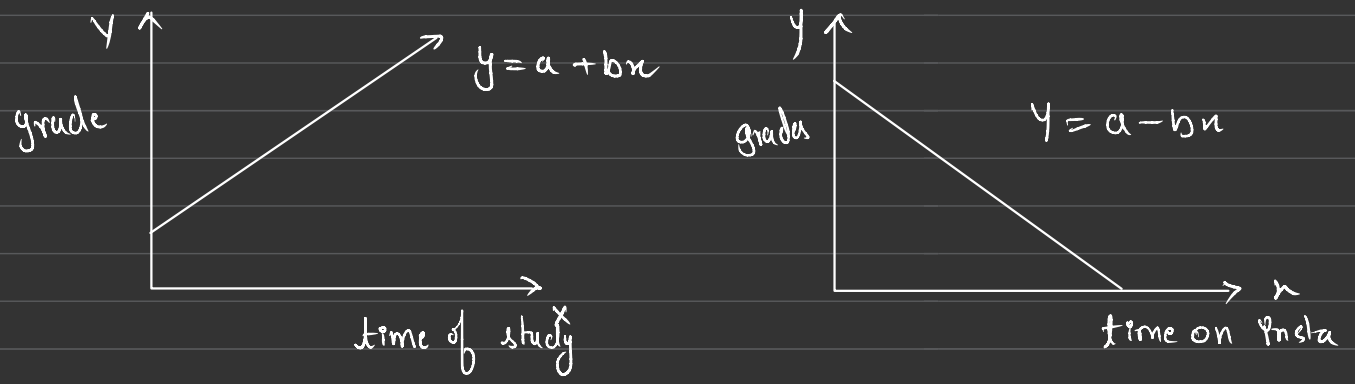
$$\frac{1.5 \times 1.4 + 1.7 \times 1.6}{\sqrt{1.5^2+1.7^2} \times \sqrt{1.4^2+1.6^2}} = \frac{4.82}{4.820041493 6} = 0.999999$$

$$\frac{2 \times 1.4 + 1.9 \times 1.6}{\sqrt{2^2+1.9^2} \times \sqrt{1.4^2+1.6^2}} = \frac{5.84}{5.864912616 6} = 0.99575$$

$= 0.999997$

$= 0.999903$

Unit-3 (Linear Regression)



We will study this.

• Regression – predicts values of responsible variable from attribute variables.

→ Variable – Continous numeric values

Linear Regression

Simple linear Regression	Multiple linear Regression.
$y = w_0 + w_1 x$	$y = w_0 + w_1 x_1 + w_2 x_2$
w_0 → Intercept	
w_1 → weight function / regression coefficient	
one - one	many - one.

The linear regression analysis is used to predict the value of response variable based on the value of attribute variable. where the output or the response variable is continous in nature. here the relation between dependent & independent variable is linear.