Machine Learning Techniques & CADD

UNIT I

Markov Chains & Hidden Markov Models: Introduction to Markov chains and HMM using Markov chains for discrimination of biological sequences. Forward and backward algorithms; Parameters estimation for HMMs. HMMs for pairwise and multiple sequence alignments. Profile HMMs.

UNIT II

Machine Learning and Bioinformatics: Introduction to various Machine Learning techniques and their applications in Bioinformatics. Support Vector Machine, Artificial Neural Network; Neural Networks and their practical applications towards the development of new models, methods and tools for Bioinformatics.

UNIT III

Machine Learning Algorithms

a) Dynamic Programming
b) Gradient Descent
c) EM/GEM Algorithms
d) Markov-Chain Monte-Carlo Methods
e) Simulated Annealing
f) Evolutionary and Genetic Algorithms

UNIT IV

CADD and Molecular Docking: Introduction, Basic Procedure; Constant and Flexible Docking.

Drug design: Drug discovery process; Target identification and validation; lead optimization and validation; Methods and Tools in Computer-aided molecular Design,

Analog Based drug design:-Pharmacophores (3D database searching, conformation searches, deriving and using 3D Pharmacophore, constrained systematic search, Genetic Algorithm, clique detection techniques, maximum likelihood method).

Structure based drug design:- Docking, De Novo Drug Design (Fragment Placements, Connection Methods, Sequential Grow), Virtual screening.

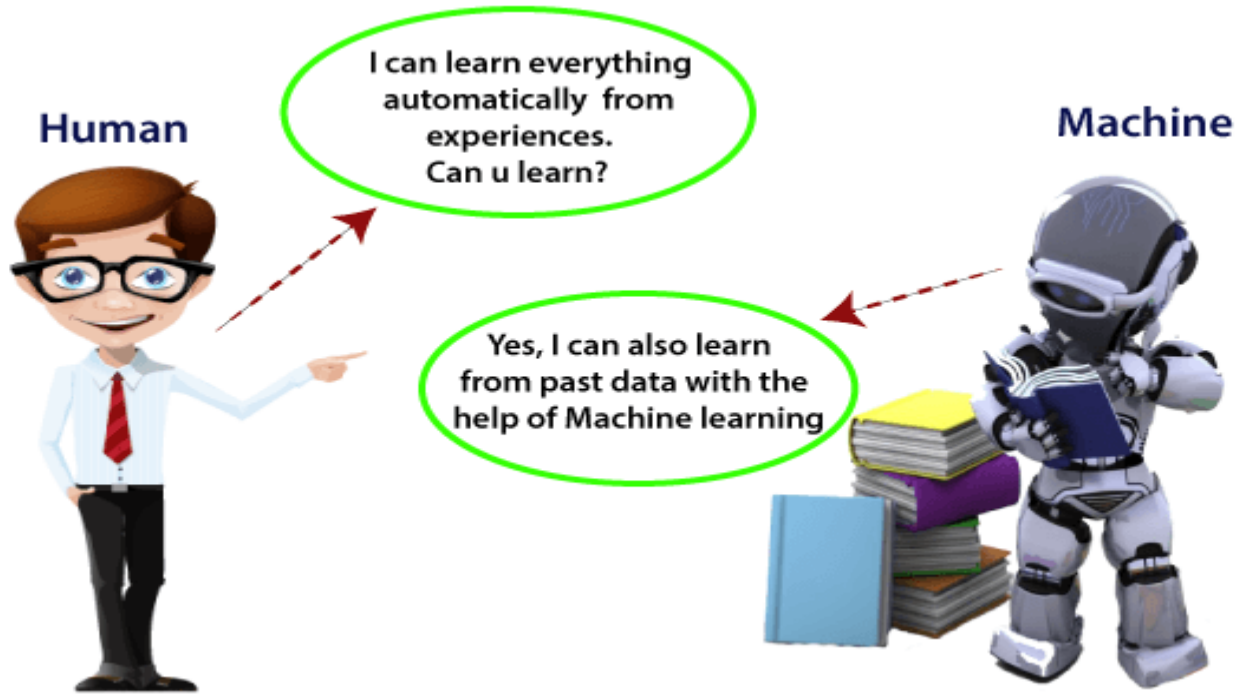
UNIT V

Structure Activity Relationship: Introduction to QSAR, QSPR, Various Descriptors used in QSARs: Electronics; Topology; Quantum Chemical based Descriptors. Regression Analysis, The Significance and Validity of QSAR Regression Equations, Partial Least Squares (PLS) Analysis, Multi Linear Regression Analysis. Use of Genetic Algorithms, Neural Networks and Principle Components Analysis in the QSAR equations.

Text References:

1. Chemoinformatics: A Textbook by Johann Gasteiger.
2. Bioinformatics second edition by David M Mount
3. Bioinformatics: Methods and Applications Genomics, Proteomics And Drug Discovery
4. Bioinformatics: the Machine Learning Approach by Pierre Baldi and Soren Brunak; Second Edition; The MIT Press

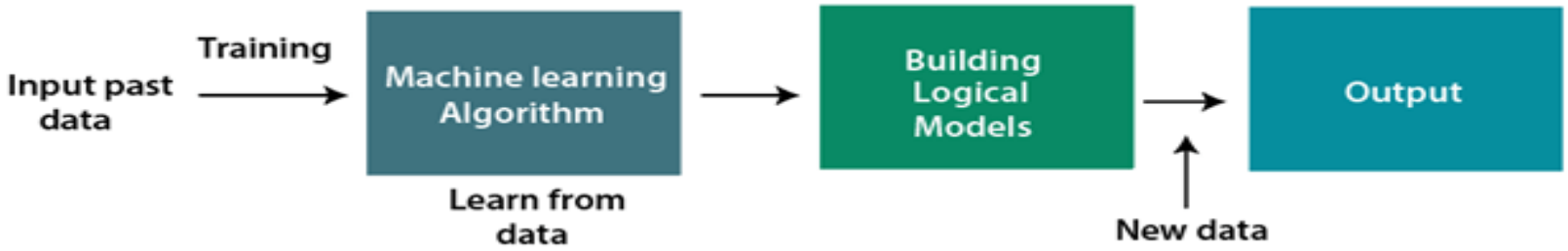
Machine learning is a growing technology which enables computers to learn automatically from past data. Machine learning uses various algorithms for building mathematical models and making predictions using historical data or information. Currently, it is being used for various tasks such as image recognition, speech recognition, email filtering, Facebook auto-tagging, recommender system, and many more.



How does Machine Learning work

A Machine Learning system learns from historical data, builds the prediction models, and whenever it receives new data, predicts the output for it. The accuracy of predicted output depends upon the amount of data, as the huge amount of data helps to build a better model which predicts the output more accurately.

Suppose we have a complex problem, where we need to perform some predictions, so instead of writing a code for it, we just need to feed the data to generic algorithms, and with the help of these algorithms, machine builds the logic as per the data and predict the output. Machine learning has changed our way of thinking about the problem. The below block diagram explains the working of Machine Learning algorithm:



Features of Machine Learning:

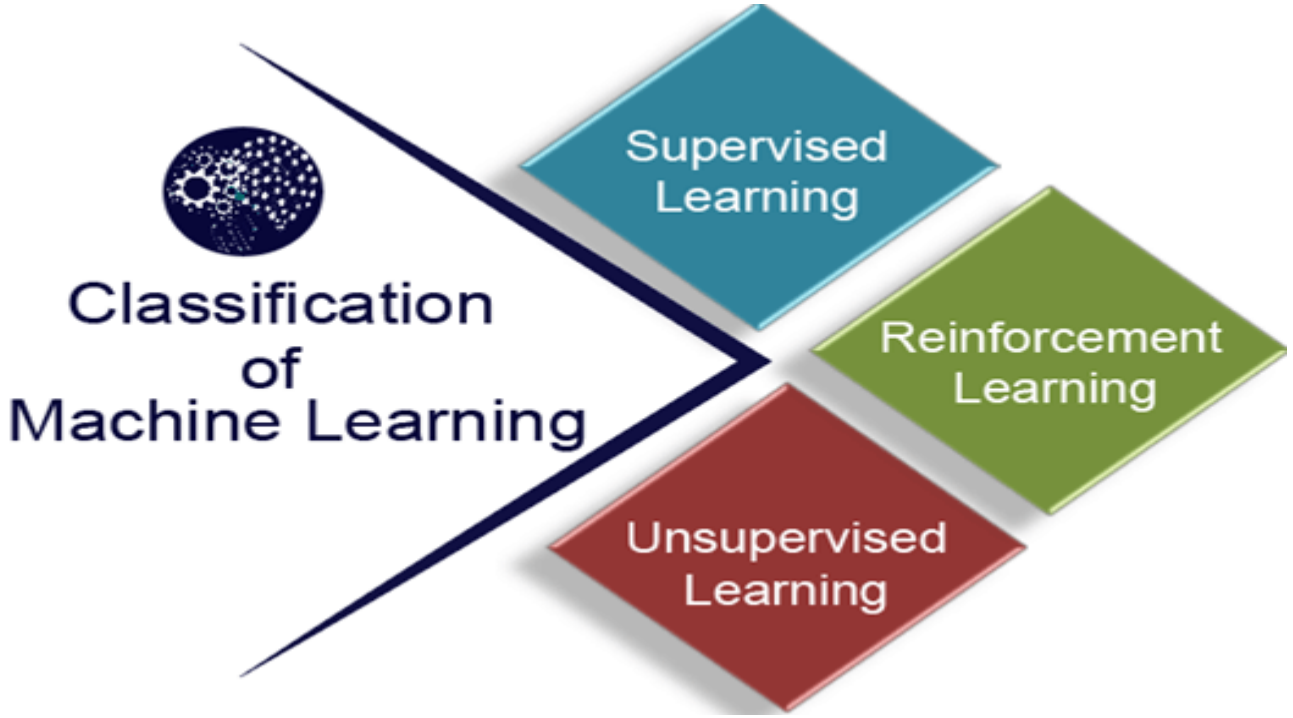
Machine learning uses data to detect various patterns in a given dataset.
It can learn from past data and improve automatically.
It is a data-driven technology.
Machine learning is much similar to data mining as it also deals with the huge amount of the data.

Following are some key points which show the importance of Machine Learning:

Rapid increment in the production of data
Solving complex problems, which are difficult for a human
Decision making in various sector including finance
Finding hidden patterns and extracting useful information from data.
Classification of Machine Learning

Machine learning can be classified into three types:

Supervised learning
Unsupervised learning
Reinforcement learning



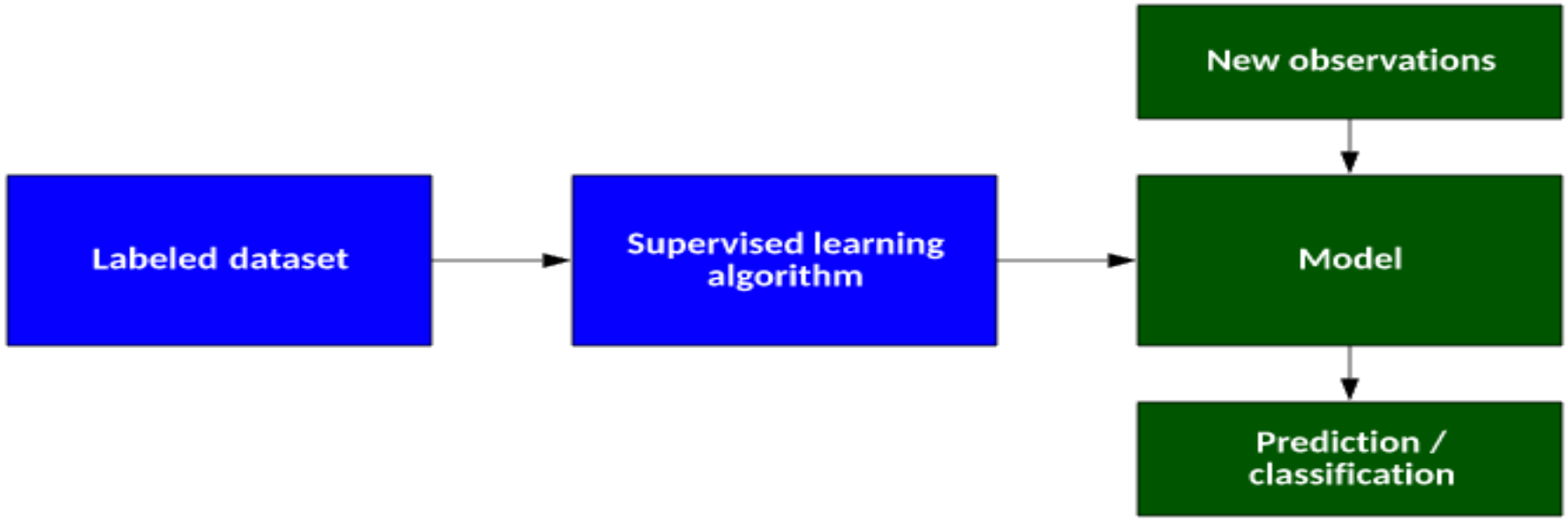
1) Supervised Learning

Supervised learning is a type of machine learning method in which we provide sample labeled data to the machine learning system in order to train it, and on that basis, it predicts the output.

The supervised learning is based on supervision, and it is the same as when a student learns things in the supervision of the teacher. The example of supervised learning is spam filtering.

Supervised learning can be grouped further in two categories of algorithms:

- **Classification**
- **Regression**

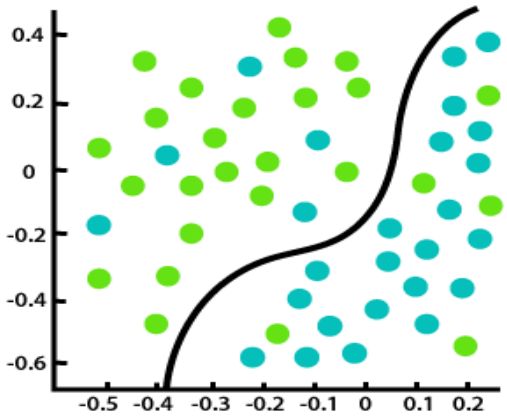


Supervised Machine Learning Categorisation

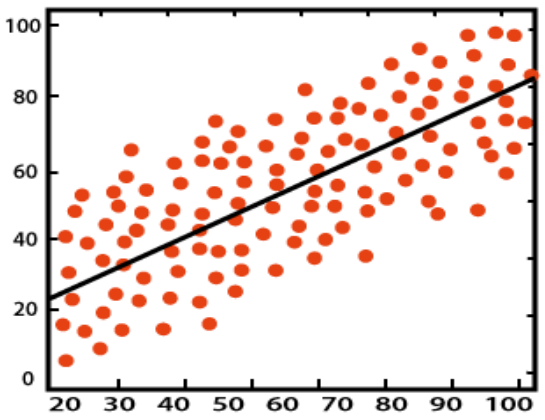
It is important to remember that all supervised learning algorithms are essentially complex algorithms, categorized as either classification or regression models.

1) Classification Models – Classification models are used for problems where the output variable can be categorized, such as “Yes” or “No”, or “Pass” or “Fail.” Classification Models are used to predict the category of the data. Real-life examples include spam detection, sentiment analysis, scorecard prediction of exams, etc.

2) Regression Models – Regression models are used for problems where the output variable is a real value such as a unique number, dollars, salary, weight or pressure, for example. It is most often used to predict numerical values based on previous data observations. Some of the more familiar regression algorithms include linear regression, logistic regression, polynomial regression, and ridge regression.



Classification



Regression

There are some very practical applications of supervised learning algorithms in real life, including:

- Text categorization
- Face Detection
- Signature recognition
- Customer discovery
- Spam detection
- Weather forecasting
- Predicting housing prices based on the prevailing market price
- Stock price predictions

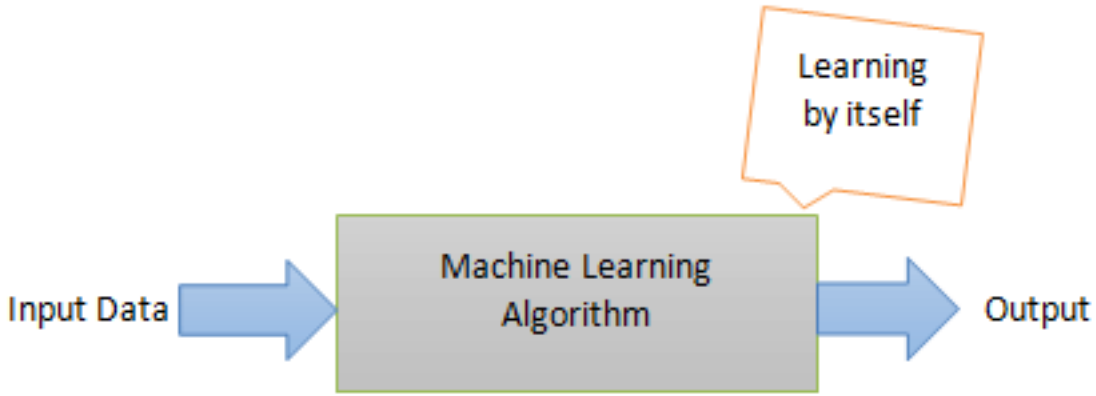
2) Unsupervised Learning

Unsupervised learning is a learning method in which a machine learns without any supervision.

The training is provided to the machine with the set of data that has not been labeled, classified, or categorized, and the algorithm needs to act on that data without any supervision. The goal of unsupervised learning is to restructure the input data into new features or a group of objects with similar patterns.

In unsupervised learning, we don't have a predetermined result. The machine tries to find useful insights from the huge amount of data. **It can be further classifieds into two categories of algorithms:**

- **Clustering**
- **Association**



Unsupervised Machine Learning Categorization

1) Clustering is one of the most common unsupervised learning methods. The method of clustering involves organizing unlabelled data into similar groups called clusters. Thus, a cluster is a collection of similar data items. The primary goal here is to find similarities in the data points and group similar data points into a cluster.

2) Anomaly detection is the method of identifying rare items, events or observations which differ significantly from the majority of the data. We generally look for anomalies or outliers in data because they are suspicious. Anomaly detection is often utilized in bank fraud and medical error detection.

Applications of Unsupervised Learning Algorithms

Some practical applications of unsupervised learning algorithms include:

- Fraud detection
- Malware detection
- Identification of human errors during data entry
- Conducting accurate basket analysis, etc.

When Should you Choose Supervised Learning vs. Unsupervised Learning?

In manufacturing, a large number of factors affect which machine learning approach is best for any given task. And, since every machine learning problem is different, deciding on which technique to use is a complex process.

In general, a good strategy for getting right machine learning approach is to:

Evaluate the data. Is it labeled/unlabelled? Is there available expert knowledge to support additional labeling? This will help to determine whether a supervised, unsupervised, semi-supervised or reinforced learning approach should be used

Define the goal. Is the problem recurring, defined one? Or, will the algorithm be expected to predict new problems?

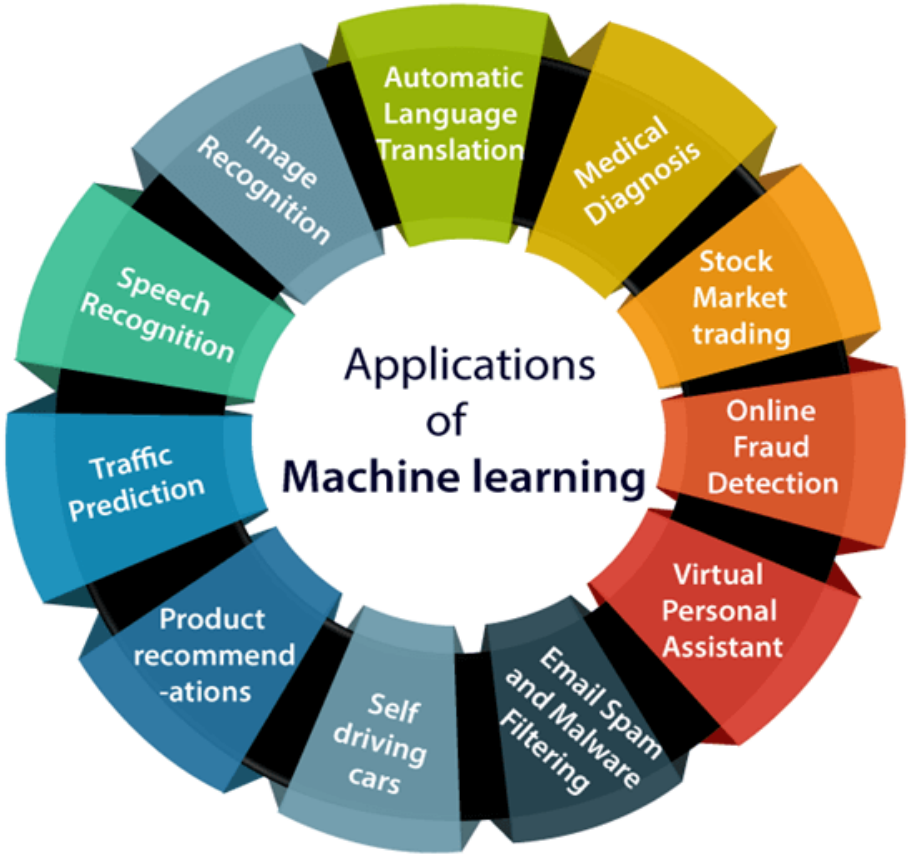
Review available algorithms that may suit the problem with regards to dimensionality (number of features, attributes or characteristics).

3) Reinforcement Learning

Reinforcement learning is a feedback-based learning method, in which a learning agent gets a reward for each right action and gets a penalty for each wrong action. The agent learns automatically with these feedbacks and improves its performance. In reinforcement learning, the agent interacts with the environment and explores it. The goal of an agent is to get the most reward points, and hence, it improves its performance.

The robotic dog, which automatically learns the movement of his arms, is an example of Reinforcement learning.

Applications of Machine learning

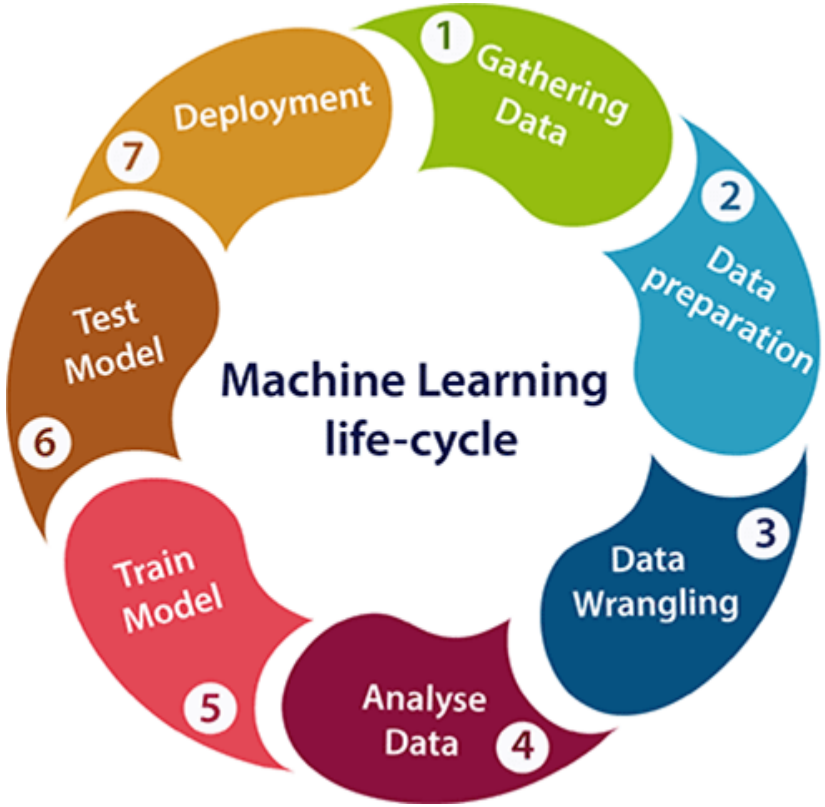


Machine Learning Application in Bioinformatics

- Comparing and aligning RNA, protein, and DNA sequences.
- Identification of promoters and finding genes from sequences related to DNA.
- Interpreting the expression-gene and micro-array data.
- Identifying the network (regulatory) of genes.
- Learning evolutionary relationships by constructing phylogenetic trees.
- Classifying and predicting protein structure.
- Molecular design and docking.

Machine learning Life cycle

Machine learning has given the computer systems the abilities to automatically learn without being explicitly programmed. But how does a machine learning system work? So, it can be described using the life cycle of machine learning. Machine learning life cycle is a cyclic process to build an efficient machine learning project. The main purpose of the life cycle is to find a solution to the problem or project.



1. Gathering Data:

Data Gathering is the first step of the machine learning life cycle. The goal of this step is to identify and obtain all data-related problems.

In this step, we need to identify the different data sources, as data can be collected from various sources such as files, database, internet, or mobile devices. It is one of the most important steps of the life cycle. The quantity and quality of the collected data will determine the efficiency of the output. The more will be the data, the more accurate will be the prediction.

This step includes the below tasks:

- Identify various data sources
- Collect data
- Integrate the data obtained from different sources

By performing the above task, we get a coherent set of data, also called as a dataset. It will be used in further steps.

2. Data preparation

After collecting the data, we need to prepare it for further steps. Data preparation is a step where we put our data into a suitable place and prepare it to use in our machine learning training.

In this step, first, we put all data together, and then randomize the ordering of data.

This step can be further divided into two processes:

- **Data exploration:**

It is used to understand the nature of data that we have to work with. We need to understand the characteristics, format, and quality of data.

A better understanding of data leads to an effective outcome. In this, we find Correlations, general trends, and outliers.

- **Data pre-processing:**

Now the next step is preprocessing of data for its analysis.

3. Data Wrangling

Data wrangling is the process of cleaning and converting raw data into a useable format. It is the process of cleaning the data, selecting the variable to use, and transforming the data in a proper format to make it more suitable for analysis in the next step. It is one of the most important steps of the complete process. Cleaning of data is required to address the quality issues.

It is not necessary that data we have collected is always of our use as some of the data may not be useful. In real-world applications, collected data may have various issues, including:

Missing Values
Duplicate data
Invalid data
Noise

So, we use various filtering techniques to clean the data.
It is mandatory to detect and remove the above issues because it can negatively affect the quality of the outcome.

4. Data Analysis

Now the cleaned and prepared data is passed on to the analysis step. This step involves:

Selection of analytical techniques
Building models
Review the result

The aim of this step is to build a machine learning model to analyze the data using various analytical techniques and review the outcome. It starts with the determination of the type of the problems, where we select the machine learning techniques such as Classification, Regression, Cluster analysis, Association, etc. then build the model using prepared data, and evaluate the model.

Hence, in this step, we take the data and use machine learning algorithms to build the model.

5. Train Model

Now the next step is to train the model, in this step we train our model to improve its performance for better outcome of the problem.
We use datasets to train the model using various machine learning algorithms. Training a model is required so that it car understand the various patterns, rules, and, features.

6. Test Model

Once our machine learning model has been trained on a given dataset, then we test the model. In this step, we check for the accuracy of our model by providing a test dataset to it.

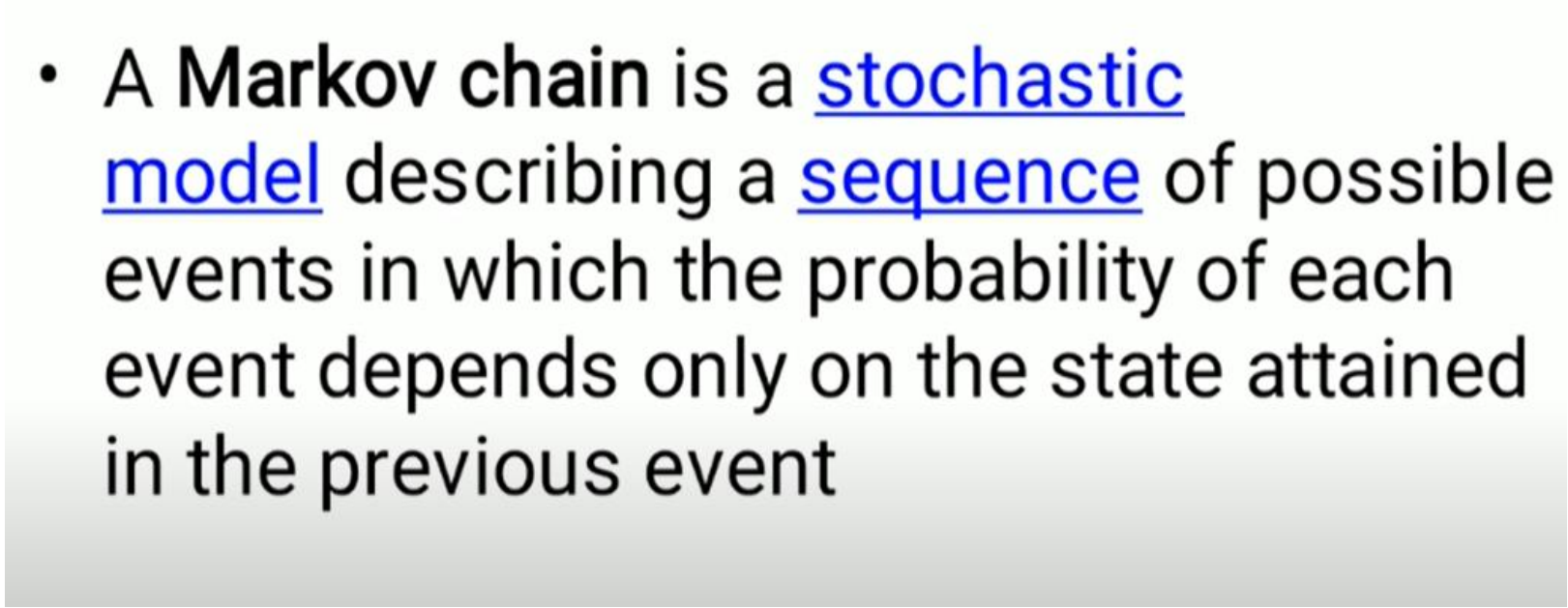
Testing the model determines the percentage accuracy of the model as per the requirement of project or problem.

7. Deployment

The last step of machine learning life cycle is deployment, where we deploy the model in the real-world system.

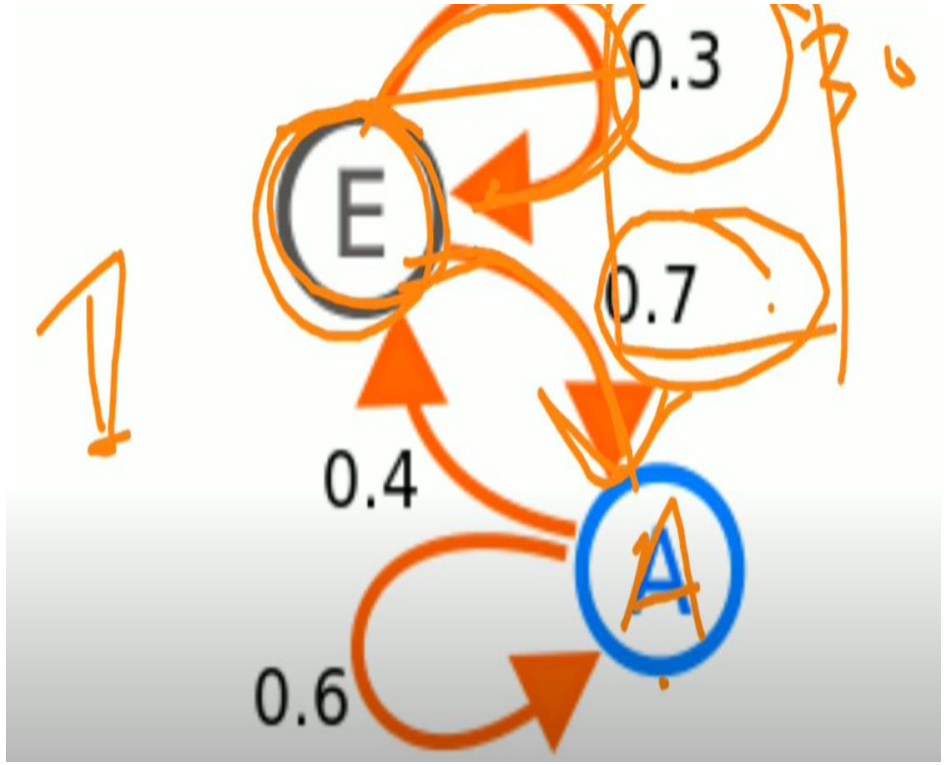
If the above-prepared model is producing an accurate result as per our requirement with acceptable speed, then we deploy the model in the real system. But before deploying the project, we will check whether it is improving its performance using available data or not. The deployment phase is similar to making the final report for a project.

Definition: A Markov Chain is a sequence of random variables, within a finite state space with values in S, for which the transitional probability P, of the state at the time t, is given by the transitional from the state and the time t−1, with probability p (Markov assumption).

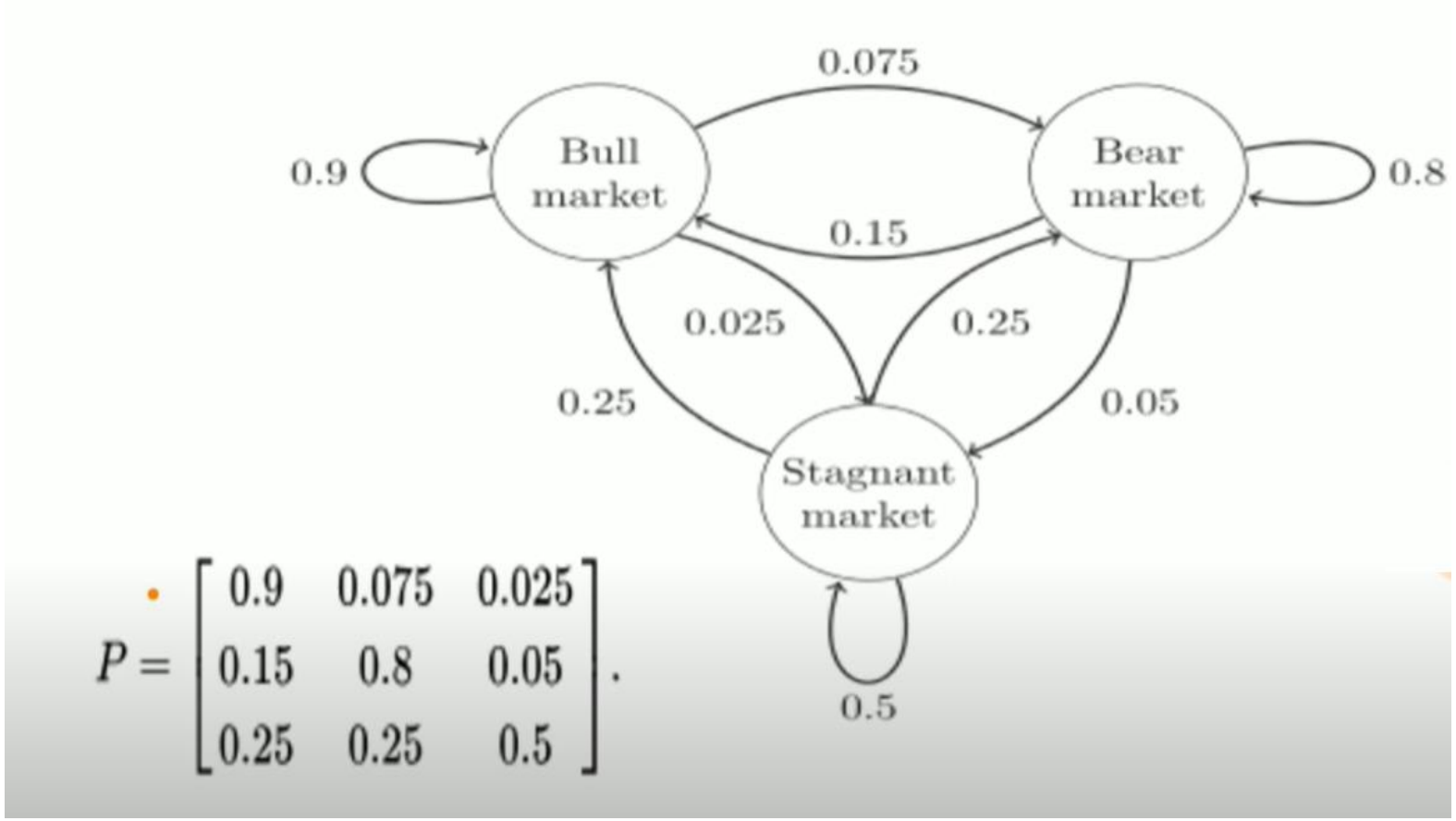


PROPERTIES OF MARKOV CHAIN

- In probability theory and related fields, a **Markov process** Markov property " memorylessness"
- Markov chain as a Markov process in either discrete or continuous time with a countable state space



PROPERTIES OF MARKOV CHAIN TRANSITION DIAGRAM AND TABLE



- $$P = \begin{bmatrix} 0.9 & 0.075 & 0.025 \\ 0.15 & 0.8 & 0.05 \\ 0.25 & 0.25 & 0.5 \end{bmatrix}.$$

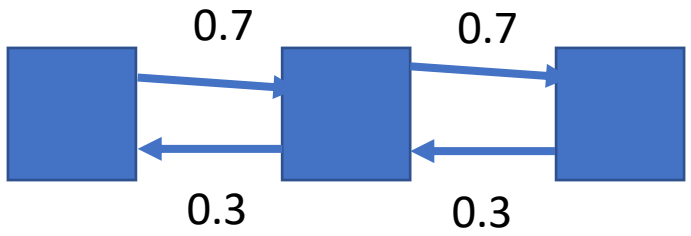
The Markov property "The future depends on the past through the present."

$$\mathbb{P}(X_{n+1} = s_j | X_0 = i_0, X_1 = i_1, \ldots, X_{n-1} = i_{n-1}, X_n = i) = \mathbb{P}(X_{n+1} = s_j | X_n = i)$$

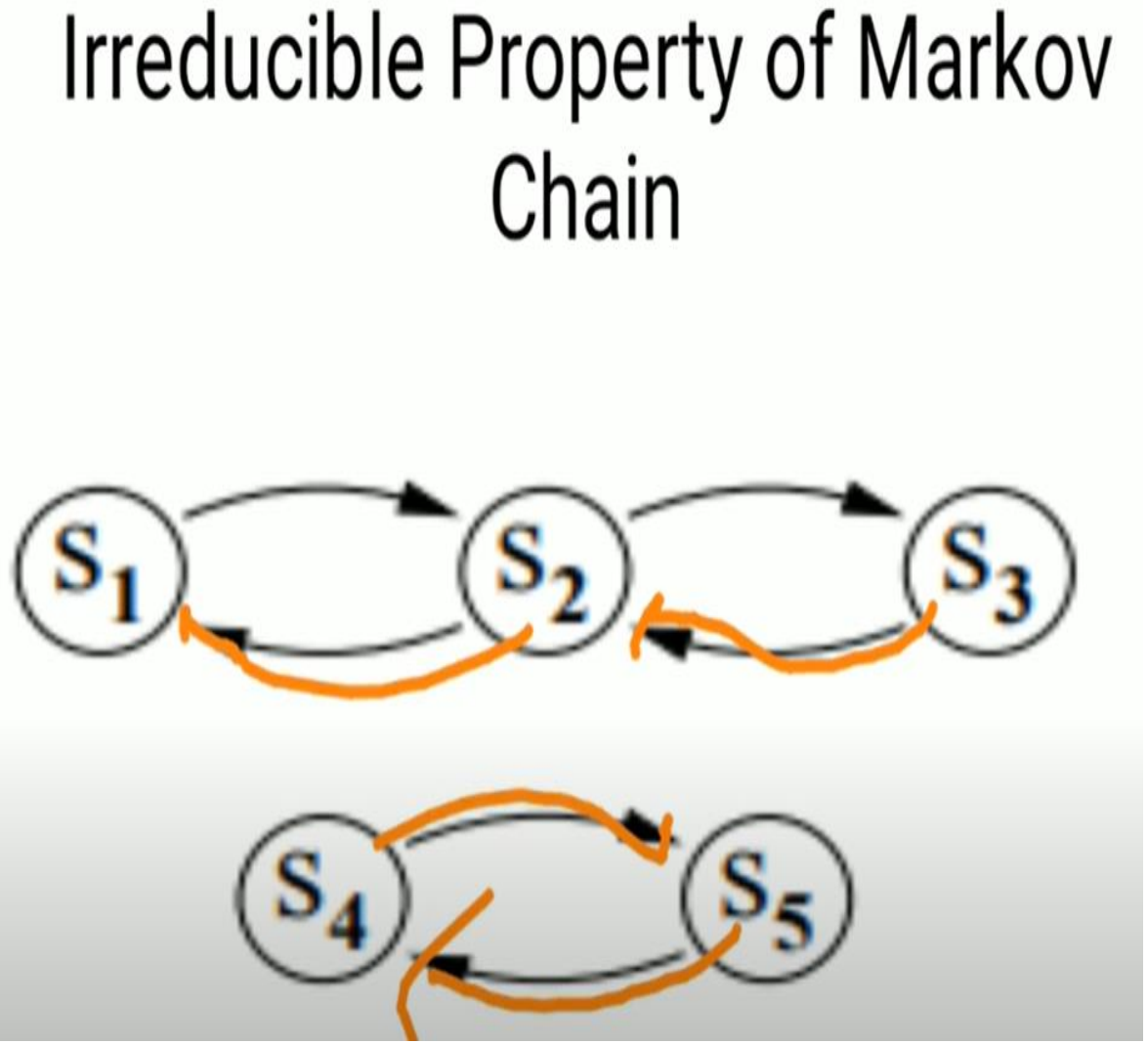
Representation I: transition matrix We can represent the Markov chain by a matrix containing the transition probabilities, or ...

Representation II: transition graph ... we can represent the Markov chain with a transition graph where a positive transition probability is represented by an arrow.

Time homogeneity The property that the transition probabilities doesn't change over time.



Irreducible Property of Markov Chain



HMM Applications

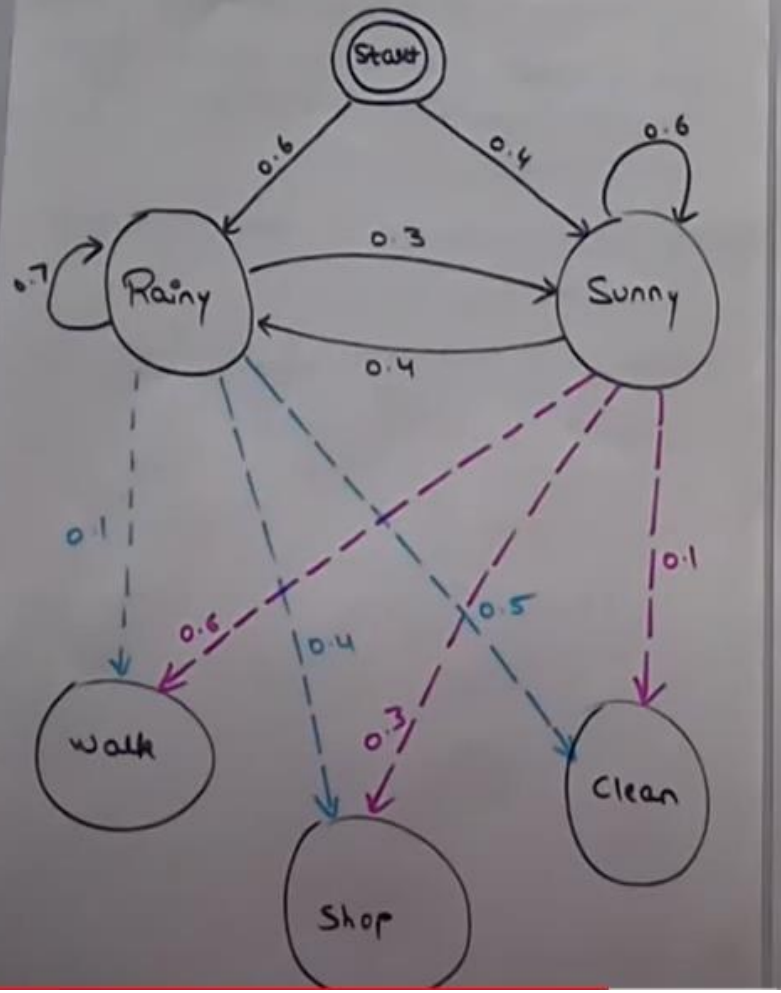
- Stock market: bull/bear market hidden Markov chain, stock daily up/down observed, depends on big market trend
- Speech recognition: sentences & words hidden Markov chain, spoken sound observed (heard), depends on the words
- Digital signal processing: source signal (0/1) hidden Markov chain, arrival signal fluctuation observed, depends on source
- Bioinformatics!!

What Is Stochastic Modeling?

Stochastic modeling is a form of financial model that is used to help make investment decisions. This type of modeling forecasts the probability of various outcomes under different conditions, using random variables.

Stochastic modeling presents data and predicts outcomes that account for certain levels of unpredictability or randomness. Companies in many industries can employ stochastic modeling to improve their business practices and increase profitability. In the financial services sector, planners, analysts, and portfolio managers use stochastic modeling to manage their assets and liabilities and optimize their portfolio

In some industries, a company's success or demise may even hinge on it.
कुछ उद्योगों में, किसी कंपनी की सफलता या मृत्यु भी उस पर निर्भर हो सकती है।



Terminologies in HMM

1. Hidden
2. Observation
3. Transition probability
4. Emissions

Property

1. Memoryless
2. Its future & past are independent i.e. prediction depends on current state.

Goal

The goal is to make a sequence of decisions where a particular decision may be influenced by earlier decision

* Hidden Markov model consists of a set of states with transition probabilities as well as an observation probability distribution.

* HMM is the extension of Markov model.

* The states in an Hidden Markov Model need not correspond to observable states.

* HMM models a process that produces a sequence of observable symbols as output.

* This model can be constructed to produce the symbols from the given sequence of symbols.