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.

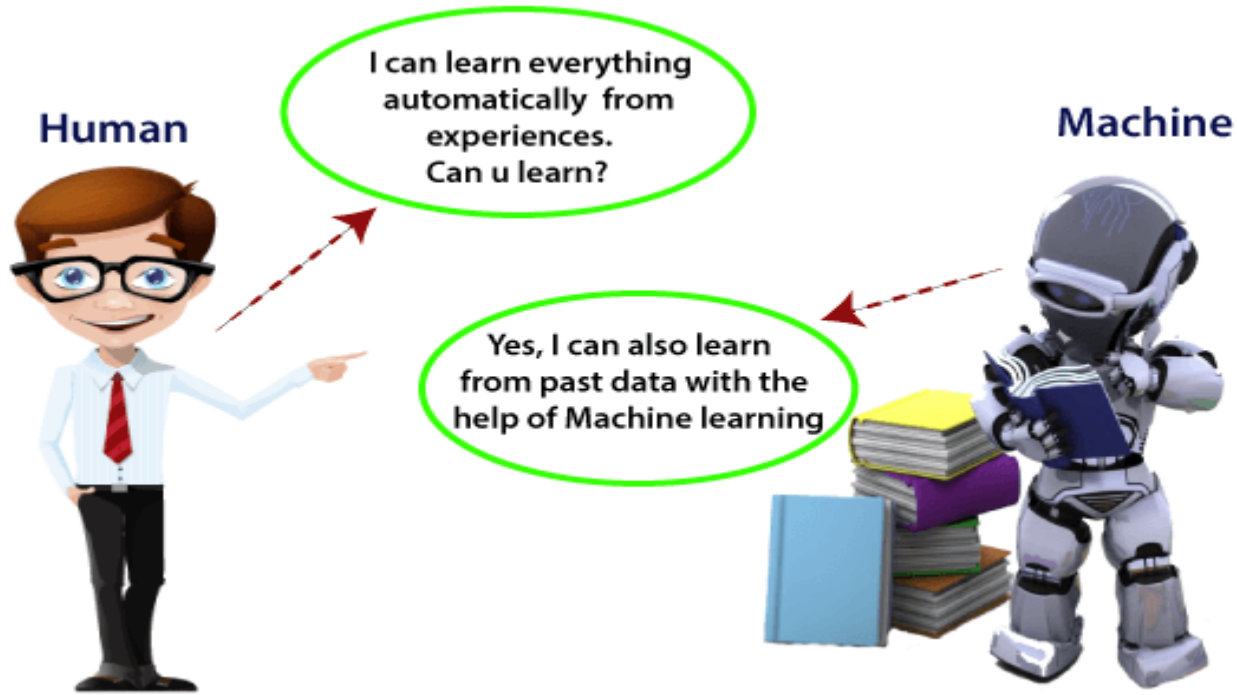
UNIT V

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

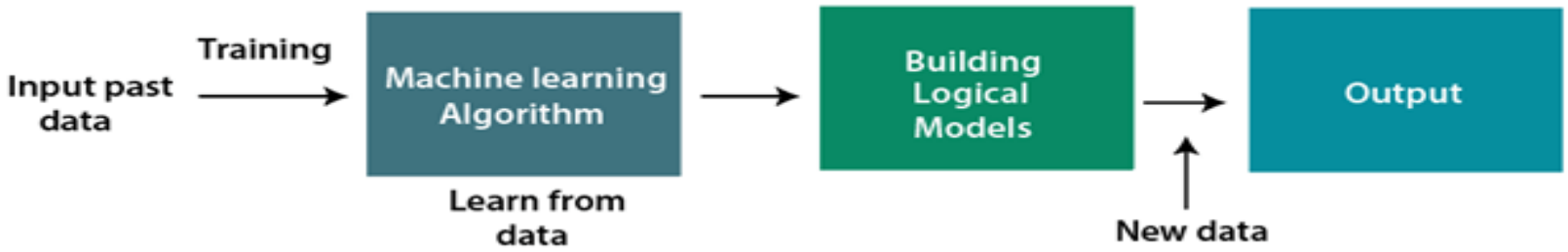
Machine learning is a growing technology which enables computers to learn automatically from past data. Machine learning uses various algorithms for building mathematical models and making predictions using historical data or information. Currently, it is being used for various tasks such as image recognition, 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:

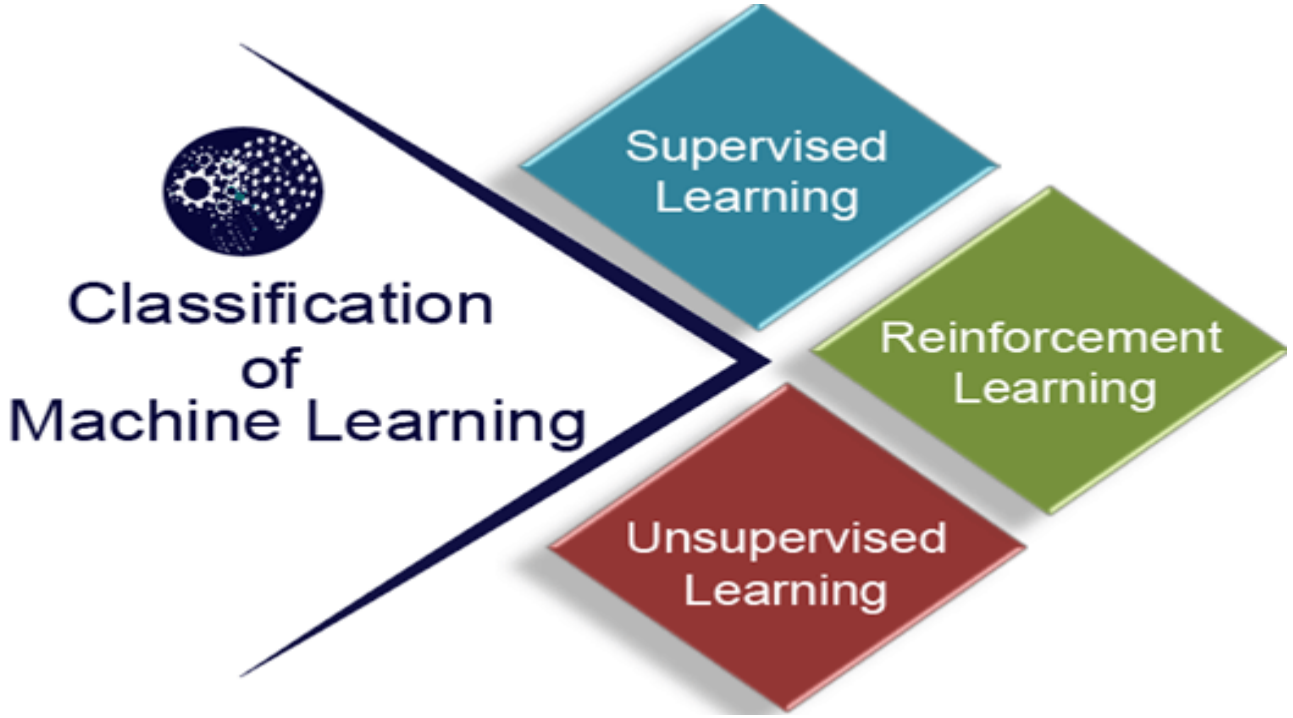
Machine learning uses data to detect various patterns in a given dataset.
It can learn from past data and improve automatically.
It is a data-driven technology.
Machine learning is much similar to data mining as it also deals with the huge amount of the data.

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



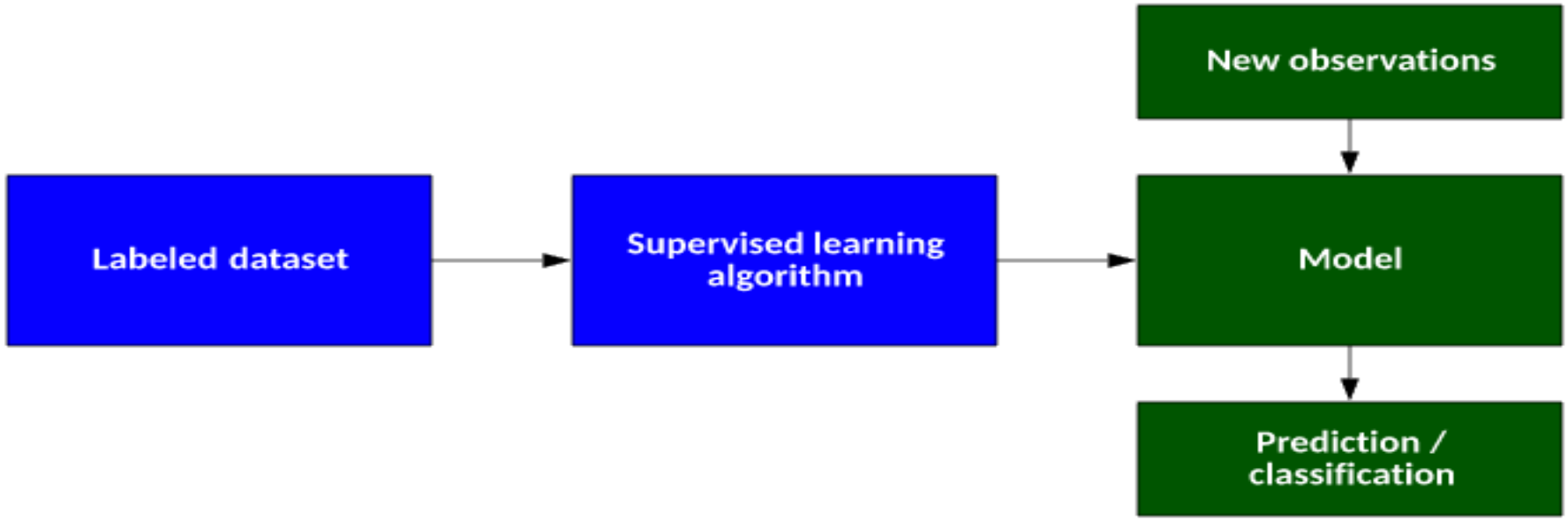
1) Supervised Learning

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**

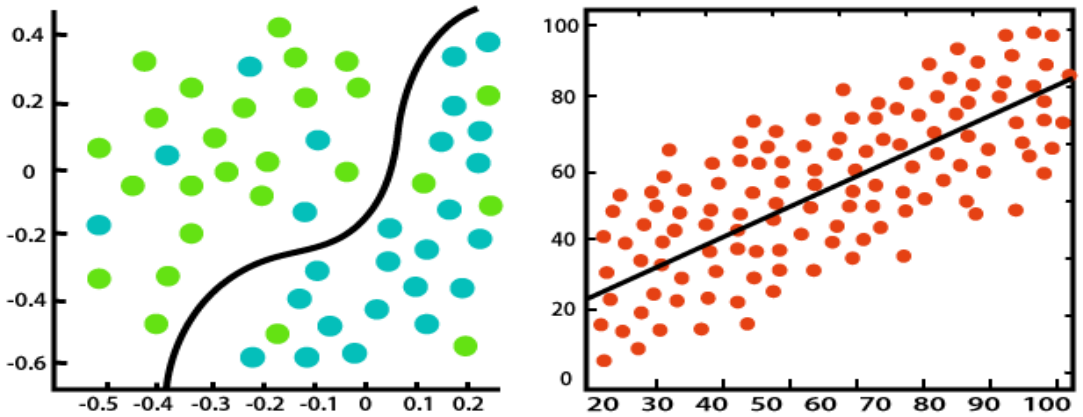


Supervised Machine Learning Categorisation

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.



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

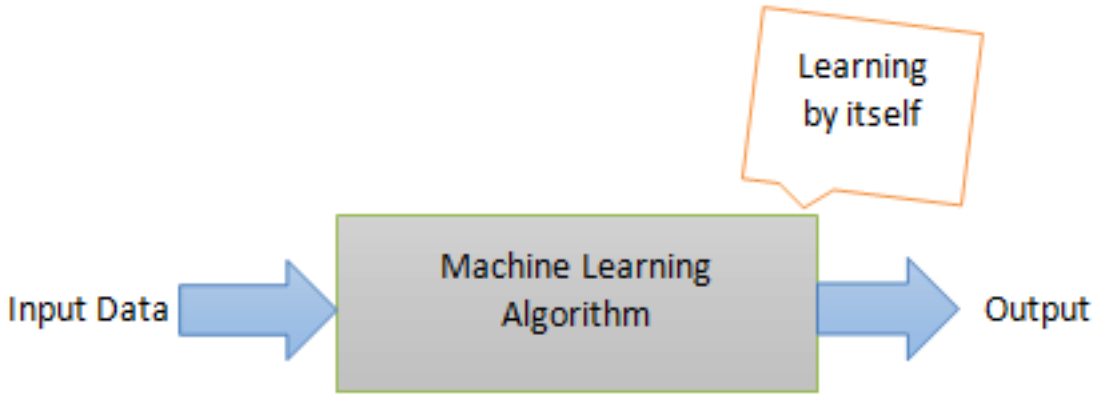
2) Unsupervised Learning

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?

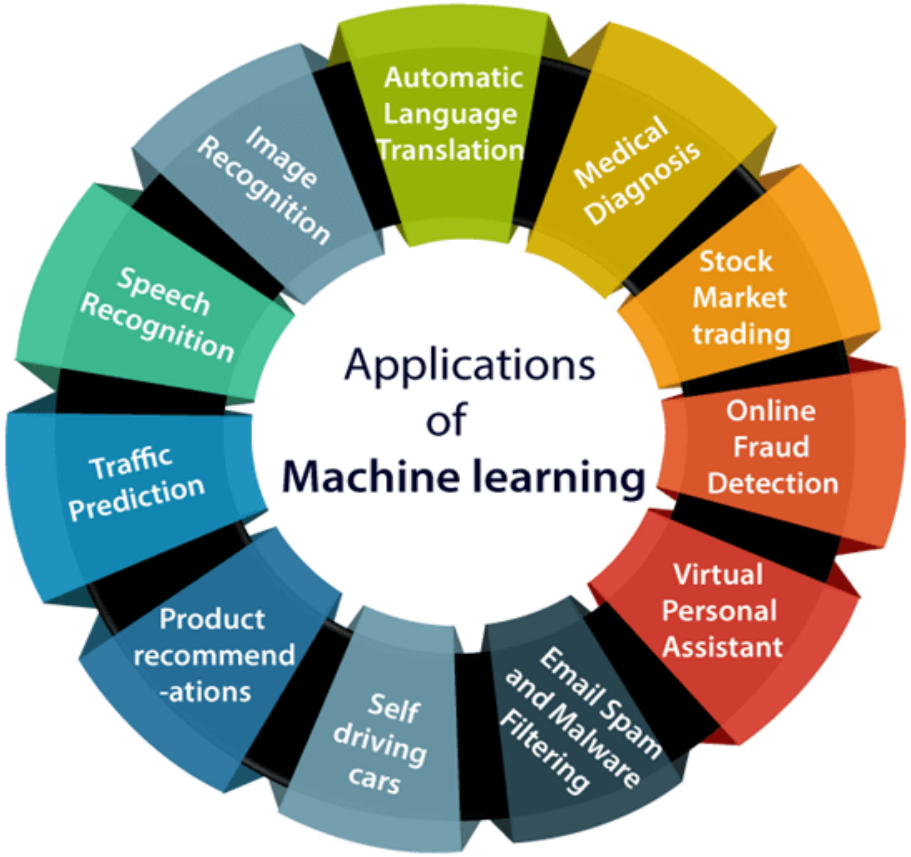
Review available algorithms that may suit the problem with regards to dimensionality (number of features, attributes or characteristics).

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

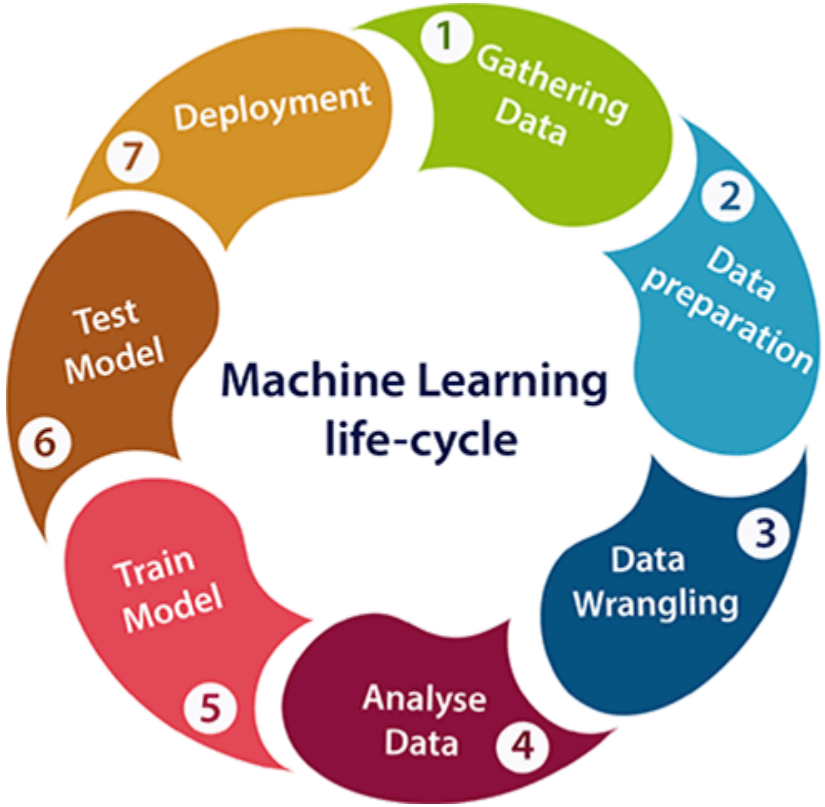


Machine Learning Application in Bioinformatics

- 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).

- A **Markov chain** is a stochastic model describing a sequence of possible events in which the probability of each event depends only on the state attained in the previous event

A Markov chain consists of three important components:

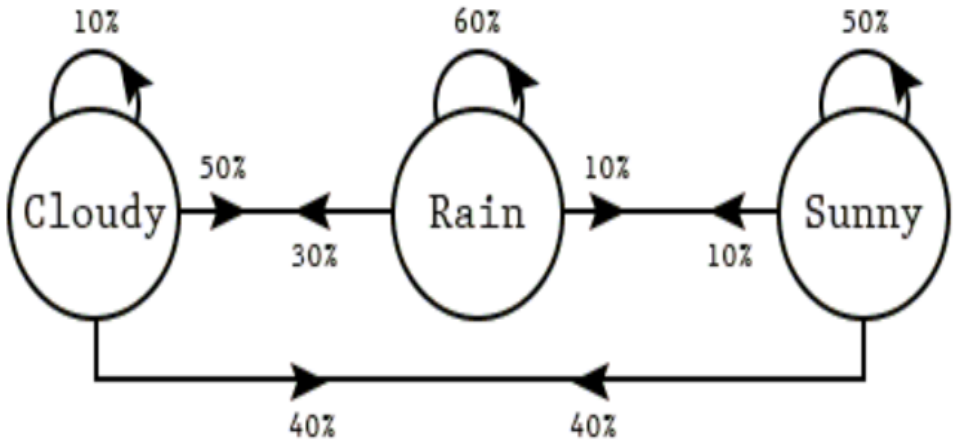
Initial probability distribution: An initial probability distribution over states, πi is the probability that the Markov chain will start in a certain state i. Some states j may have πj = 0, meaning that they cannot be initial states

One or more states

Transition probability distribution: A transition probability matrix A where each aij represents the probability of moving from state I to state j

PROPERTIES OF MARKOV CHAIN

The diagram below represents a Markov chain where there are three states representing the weather of the day (cloudy, rainy, and sunny). And, there are transition probabilities representing the weather of the next day given the weather of the current day.



There are three different states such as cloudy, rain, and sunny. The following represent the transition probabilities based on the above diagram:

If sunny today, then tomorrow:
50% probability for sunny
10% probability for rainy
40% probability for cloudy

If rainy today, then tomorrow:
10% probability for sunny
60% probability for rainy
30% probability for cloudy

If cloudy today, then tomorrow:
40% probability for sunny
50% probability for rainy
10% probability for cloudy

Using this Markov chain, what is the probability that the Wednesday will be cloudy if today is sunny. The following are different transitions that can result in a cloudy Wednesday given today (Monday) is sunny.

Sunny – Sunny (Tuesday) – Cloudy (Wednesday): The probability to a cloudy Wednesday can be calculated as 0.5 x 0.4 = 0.2
Sunny – Rainy (Tuesday) – Cloudy (Wednesday): The probability of a cloudy Wednesday can be calculated as 0.1 x 0.3 = 0.03
Sunny – Cloudy (Tuesday) – Cloudy (Wednesday): The probability of a cloudy Wednesday can be calculated as 0.4 x 0.1 = 0.04
The total probability of a cloudy Wednesday = 0.2 + 0.03 + 0.04 = 0.27.

As shown above, the Markov chain is a process with a known finite number of states in which the probability of being in a particular state is determined only by the previous state.

What are Hidden Markov models (HMM)?

The hidden Markov model (HMM) is another type of Markov model where there are few states which are hidden. This is where HMM differs from a Markov chain. HMM is a statistical model in which the system being modeled are Markov processes with unobserved or hidden states.

It is a hidden variable model which can give an observation of another hidden state with the help of the Markov assumption.

The hidden state is the term given to the next possible variable which cannot be directly observed but can be inferred by observing one or more states according to Markov's assumption.

Markov assumption is the assumption that a hidden variable is dependent only on the previous hidden state. Mathematically, the probability of being in a state at a time t depends only on the state at the time (t-1). It is termed a limited horizon assumption.

Another Markov assumption states that the conditional distribution over the next state, given the current state, doesn't change over time. This is also termed a stationary process assumption.

A Markov model is made up of two components: the state transition and hidden random variables that are conditioned on each other. A hidden Markov model consists of five important components:

Initial probability distribution: An initial probability distribution over states, πi is the probability that the Markov chain will start in state i. Some states j may have $\pi j = 0$, meaning that they cannot be initial states. The initialization distribution defines each hidden variable in its initial condition at time t=0 (the initial hidden state).

One or more hidden states

Transition probability distribution: A transition probability matrix where each aij represents the probability of moving from state i to state j. The transition matrix is used to show the hidden state to hidden state transition probabilities.

A sequence of observations

Emission probabilities: A sequence of observation likelihoods, also called emission probabilities, each expressing the probability of an observation oi being generated from a state I. The emission probability is used to define the hidden variable in terms of its next hidden state. It represents the conditional distribution over an observable output for each hidden state at time t=0.

Real-world examples of Hidden Markov Models (HMM)

Retail scenario: Now if you go to the grocery store once per week, it is relatively easy for a computer program to predict exactly when your shopping trip will take more time. The hidden Markov model calculates which day of visiting takes longer compared with other days and then uses that information in order to determine why some visits are taking long while others do not seem too problematic for shoppers like yourself. Another example from e-commerce where hidden Markov models are used is the recommendation engine. The hidden Markov models try to predict the next item that you would like to buy.

Travel scenario: By using hidden Markov models, airlines can predict how long it will take a person to finish checking out from an airport. This allows them to know when they should start boarding passengers!

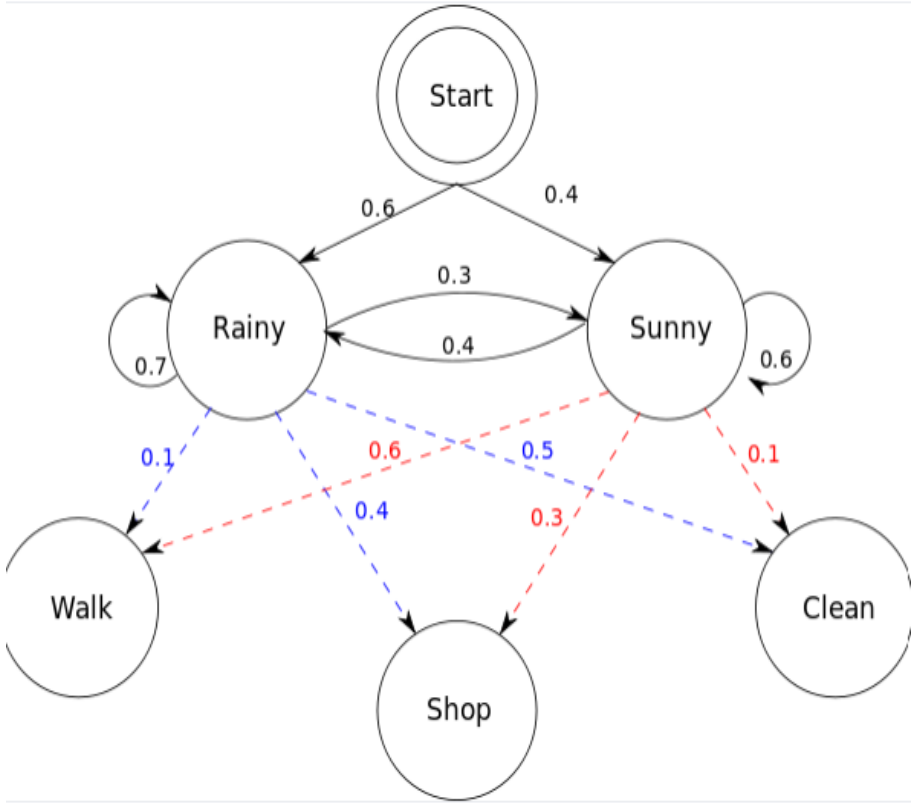
Medical Scenario: The hidden Markov models are used in various medical applications, where it tries to find out the hidden states of a human body system or organ. For example, cancer detection can be done by analyzing certain sequences and determining how dangerous they might pose for the patient. Another example where hidden Markov models get used is for evaluating biological data such as RNA-Seq, ChIP-Seq, etc., that help researchers understand gene regulation. Using the hidden Markov model, doctors can predict the life expectancy of people based on their age, weight, height, and body type.

Marketing scenario: As marketers utilize a hidden Markov model, they can understand at what stage of their marketing funnel users are dropping off and how to improve user conversion rates.

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!!

Let’s understand the above using the hidden Markov model representation shown below:



The hidden Markov model in the above diagram represents the process of predicting whether someone will be found to be walking, shopping, or cleaning on a particular day depending upon whether the day is rainy or sunny. The following represents five components of the hidden Markov model in the above diagram:

```
states = ('Rainy', 'Sunny')

observations = ('walk', 'shop', 'clean')

start_probability = {'Rainy': 0.6, 'Sunny': 0.4}

transition_probability = {
   'Rainy' : {'Rainy': 0.7, 'Sunny': 0.3},
   'Sunny' : {'Rainy': 0.4, 'Sunny': 0.6},
   }

emission_probability = {
   'Rainy' : {'walk': 0.1, 'shop': 0.4, 'clean': 0.5},
   'Sunny' : {'walk': 0.6, 'shop': 0.3, 'clean': 0.1},
   }
```

Let’s notice some of the following in the above picture:

There are two hidden states such as rainy and sunny. These states are hidden because what is observed as the process output is whether the person is shopping, walking, or cleaning.
The sequence of observations is shop, walk, and clean.

An initial probability distribution is represented by start probability

Transition probability represents the transition of one state (rainy or sunny) to another state given the current state

Emission probability represents the probability of observing the output, shop, clean and walk given the states, rainy or sunny.

The Hidden Markov model is a special type of Bayesian network that has hidden variables which are discrete random variables. The first-order hidden Markov model allows hidden variables to have only one state and the second-order hidden Markov models allow hidden states to be having two or more two hidden states.

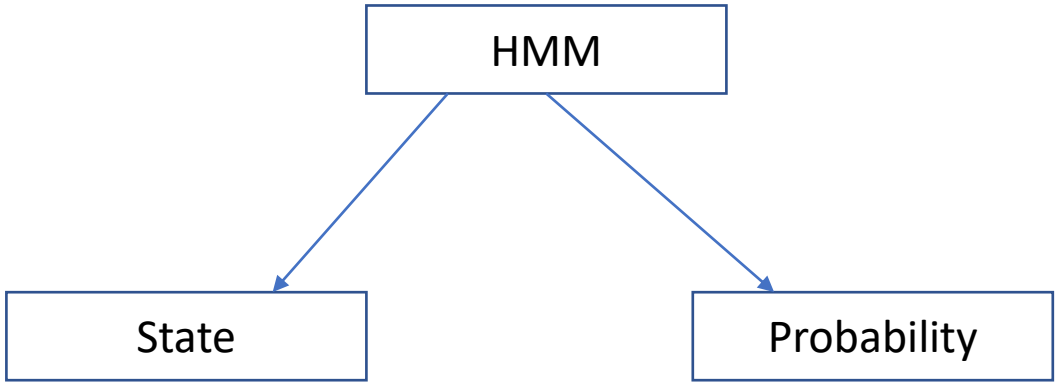
The hidden Markov model represents two different states of variables: Hidden state and observable state. A hidden state is one that cannot be directly observed or seen. An observable state is one that can be observed or seen. One hidden state can be associated with many observable states and one observable state may have more than hidden states. The hidden Markov model uses the concept of probability to identify whether there will be an emission from the hidden state to another hidden state or from hidden states to observable states.

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

We need Some Important point to developed Hidden Markov model;



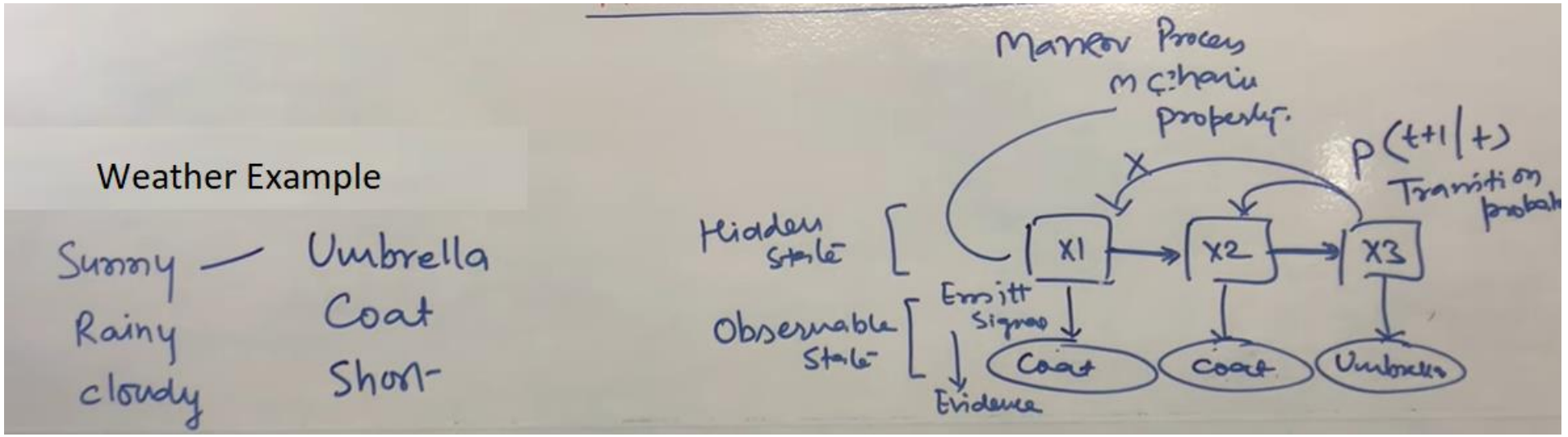
1. Hidden State
2. Observable

1. Transition
2. Emission

Scenario 1:
Suppose you are seating inside the basement, and we can not be able to see outside the weather.
So how can we fine outside weather is totally dependent on observation .

1. If the person coming with umbrella, then we will make sense might be outside sunny or cloudy
2. if the person coming with raincoat so we can assume outside will be rainy.
3. If the person coming with short dress, then might me Sunny .

Question: If the coat is given then we will predict Weather ?

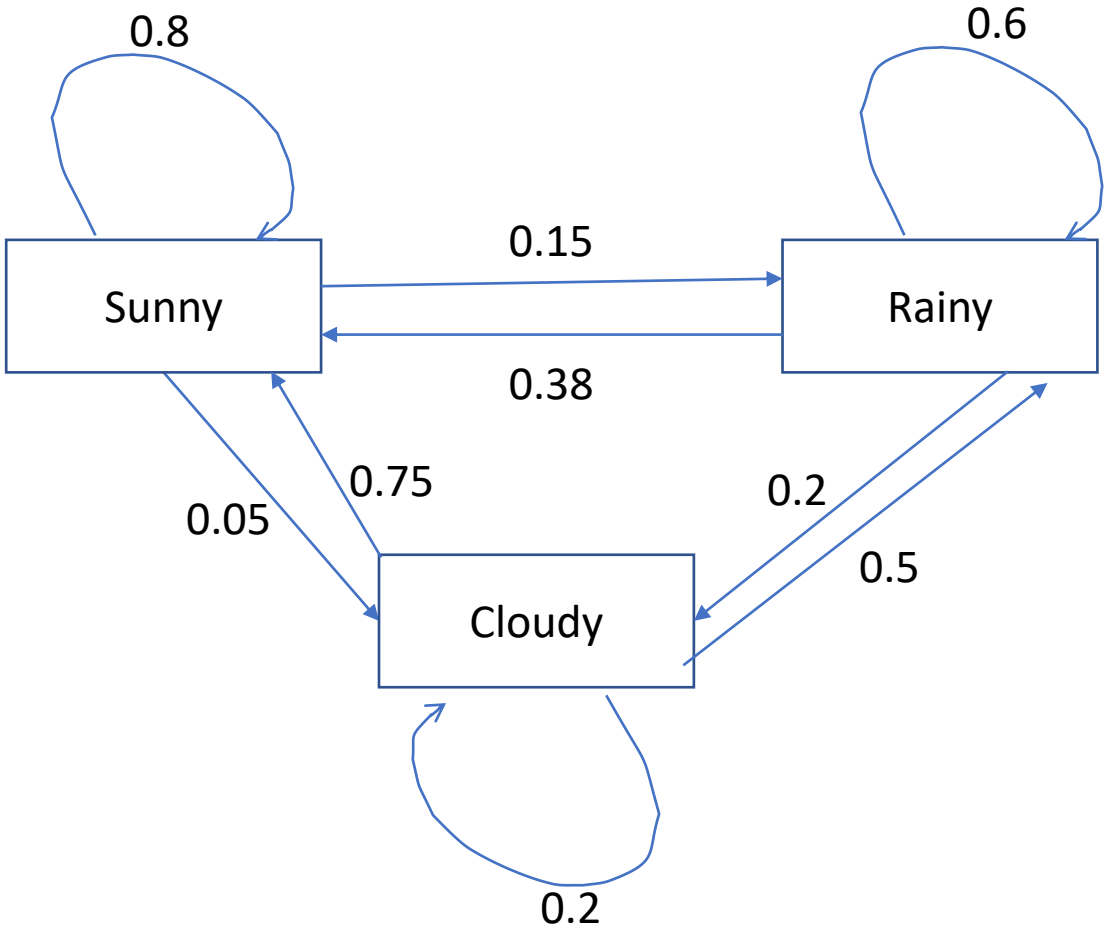


Scenario 2: Suppose you are seating in the one room and your friend is seating in the other room.
on the observation basis we have to find my friend is happy or not .

POS(Part of speech)

Scenario 3: I book my ticket

Find out the hidden state if Observation state is given.



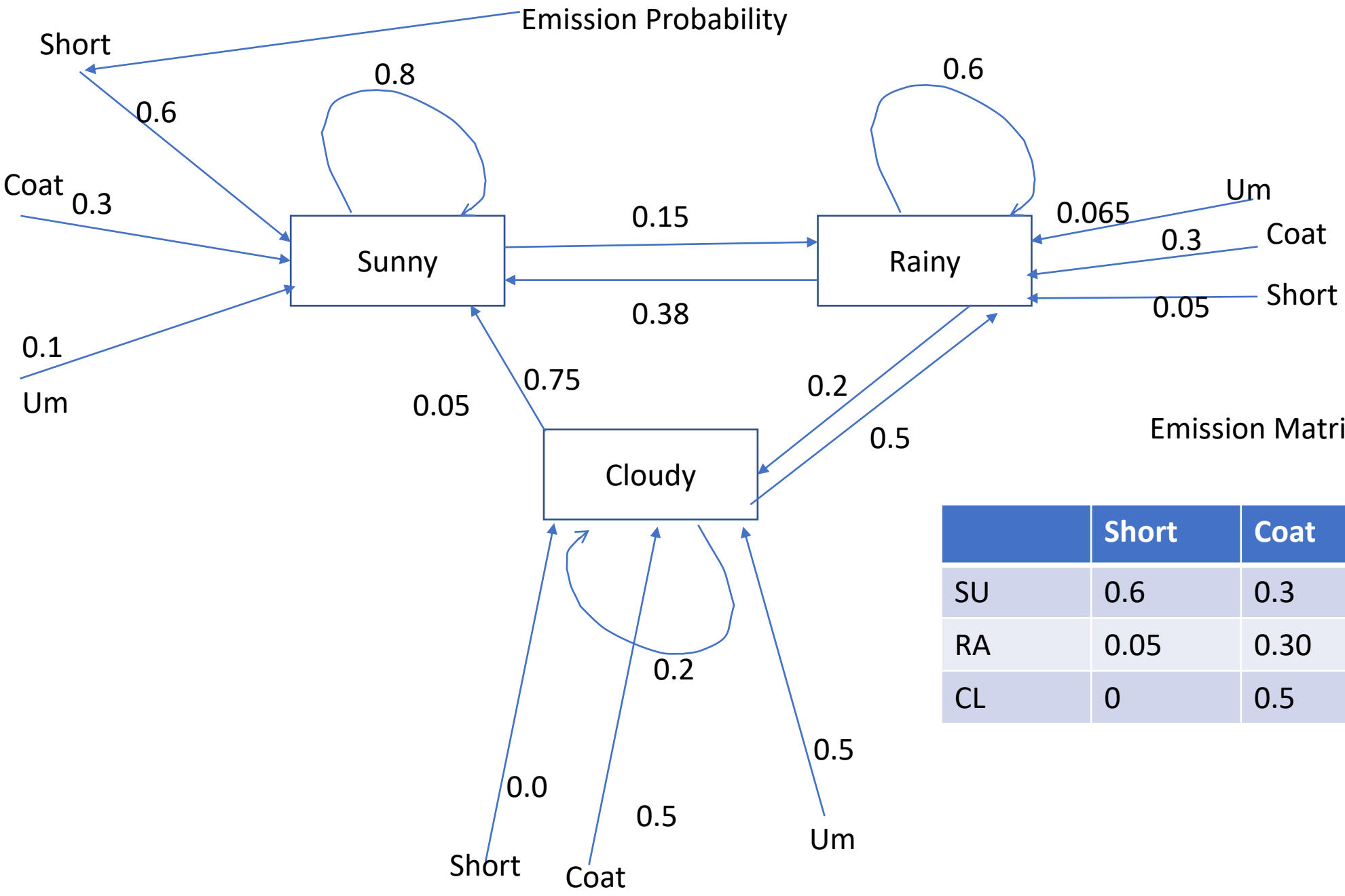
Transition State Diagram

Transition matrix(A)

	SU	RA	CL
SU	0.8	0.15	0.05
RA	0.38	0.6	0.02
CL	0.75	0.5	0.2

Initial State:

$$\pi = [\underset{Su}{0.75} \quad ,\underset{RA}{0.2} \quad , \underset{CL}{0.05}]$$

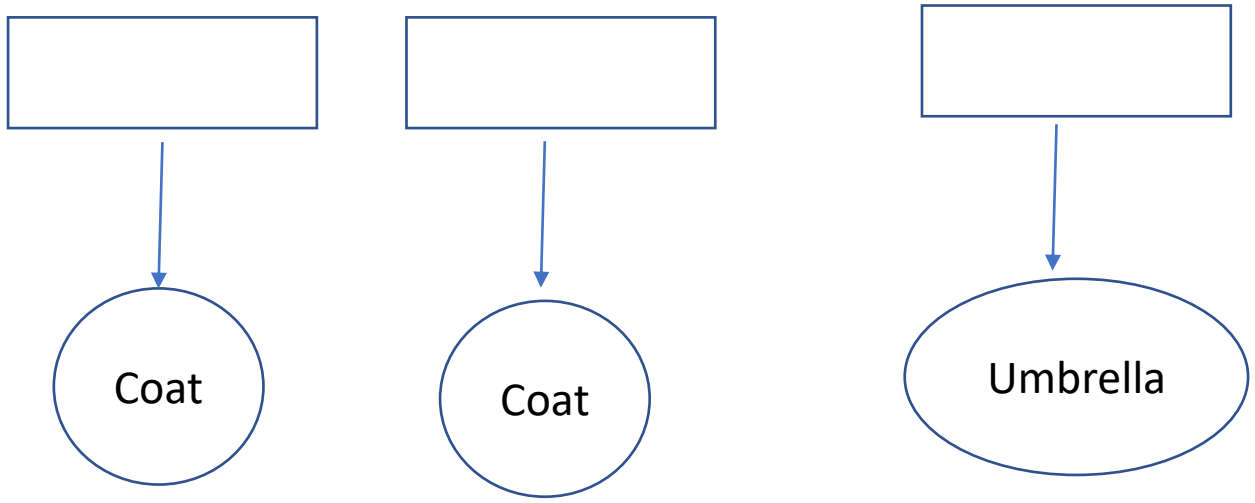


Emission Matrix

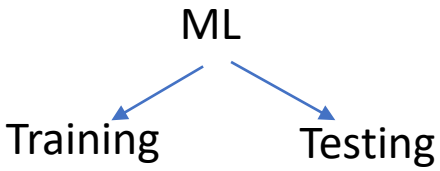
	Short	Coat	UM
SU	0.6	0.3	0.1
RA	0.05	0.30	0.65
CL	0	0.5	0.5

1. What will be the hidden state for Observable state

SU	CL	CL
SU	SU	SU
SU	RA	CL
---	----	-----



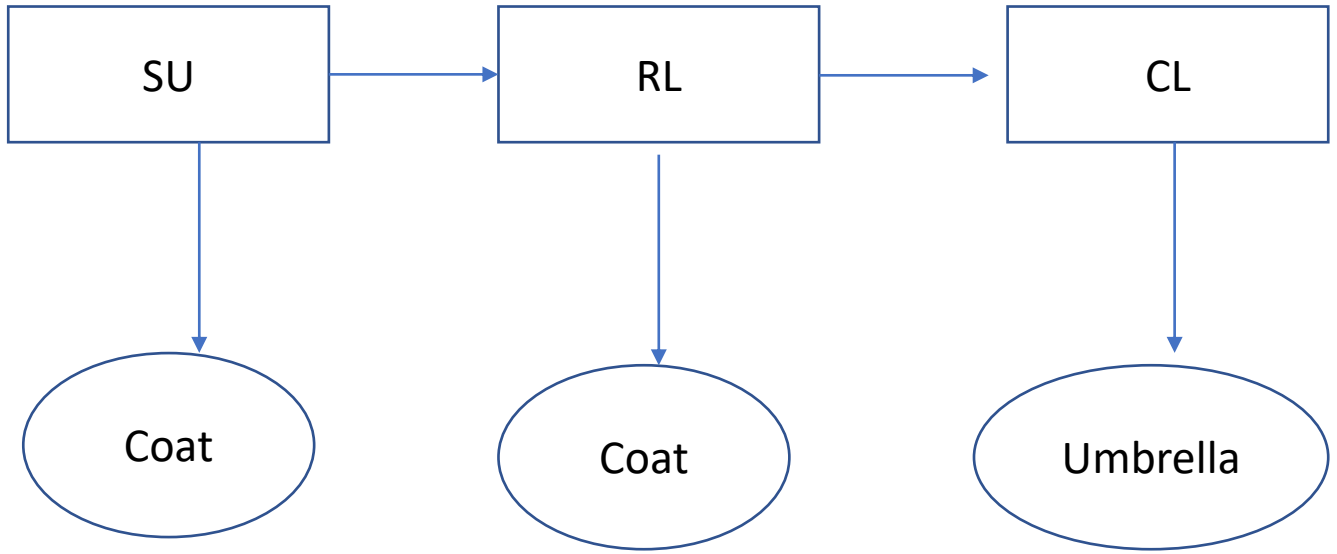
N= No. of Hidden State=3
T= No. of Observable State=3
$M = N^T = 3^3 = 27$



HMM

π
A
B

Suppose we picked up
First Combination :



What will be the hidden state when observable state is given ?
Suppose Coat is given then what will be the hidden state for Coat?

Joint probability :

$$P(a,b) = p(a/b)\ p(b)$$

What will be the hidden state when observable state is given ?
Suppose Coat is given then what will be the hidden state for Coat?

P(COAT,SU)= P(COAT/SU) P(SU)

If all the observable state is denoted by O and all Hidden state is denoted by Q so
P(O,Q)= P(O/Q) P(Q)

Now we take Complete Shape

P(C,C,U,SU,RA,CL) = P(C/SU) P(C/RA) P(U/CL) P(SU) P(RA/SU) P(CL/RA)
0.3* O.30 *0.5* 0.75*0.15* 0.02= ?