

Module-02

Informed Search Strategies

Topic: -01: Greedy best-first search

- Greedy best-first search algorithm always selects the path which appears best at that moment.
- It is the combination of depth-first search and breadth-first search algorithms. It uses the heuristic function and search.
- Best-first search allows us to take the advantages of both algorithms.
- With the help of best-first search, at each step, we can choose the most promising node.
- In the best first search algorithm, we expand the node which is closest to the goal node and the closest cost is estimated by heuristic function,

i.e. $f(n) = h(n)$.

Where, $h(n)$ = estimated cost from node n to the goal.

The greedy best first algorithm is implemented by the priority queue.

Best first search algorithm:

Step 1: Place the starting node into the OPEN list.

Step 2: If the OPEN list is empty, Stop and return failure.

Step 3: Remove the node n , from the OPEN list which has the lowest value of $h(n)$, and places it in the CLOSED list.

Step 4: Expand the node n , and generate the successors of node n .

Step 5: Check each successor of node n , and find whether any node is a goal node or not. If any successor node is goal node, then return success and terminate the search, else proceed to Step 6.

Step 6: For each successor node, algorithm checks for evaluation function $f(n)$, and then check if the node has been in either OPEN or CLOSED list. If the node has not been in both lists, then add it to the OPEN list.

Step 7: Return to Step 2.

Advantages:

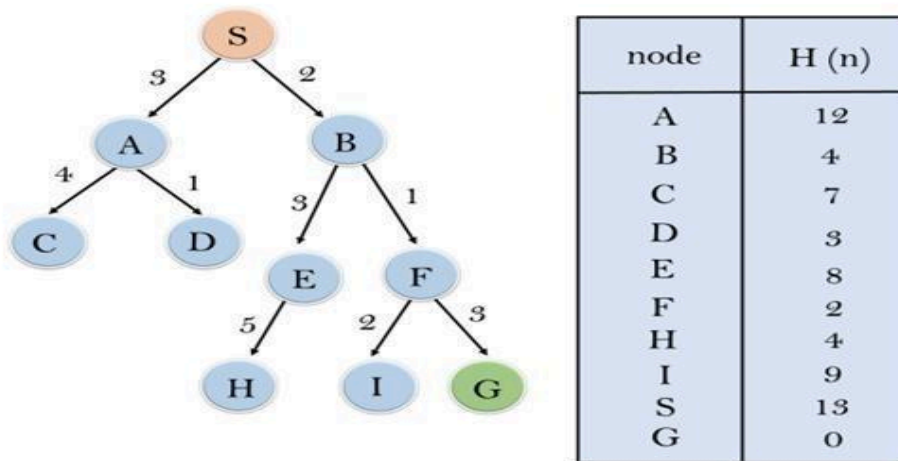
- Best first search can switch between BFS and DFS by gaining the advantages of both the algorithms.
- This algorithm is more efficient than BFS and DFS algorithms.

Disadvantages:

- It can behave as an unguided depth-first search in the worst-case scenario.
- It can get stuck in a loop as DFS.
- This algorithm is not optimal.

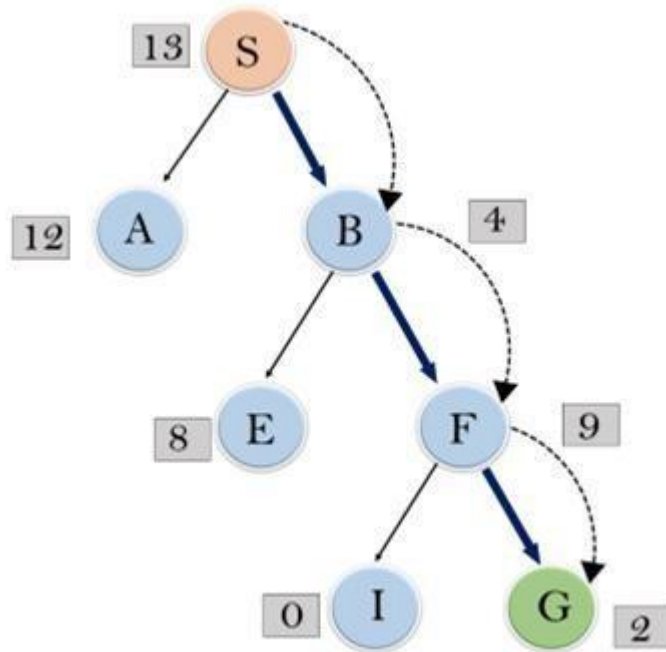
Example:

Consider the below search problem, and we will traverse it using greedy best-first search. At each iteration, each node is expanded using evaluation function $f(n)=h(n)$, which is given in the below table.



In this search example, we are using two lists which are OPEN and CLOSED Lists.

Following are the iteration for traversing the above example.



Expand the nodes of S and put in the CLOSED list

Initialization: Open [A, B], Closed [S]

Iteration 1: Open [A], Closed [S, B]

Iteration 2: Open [E, F, A], Closed [S, B]

: Open [E, A], Closed [S, B, F]

Iteration 3: Open [I, G, E, A], Closed [S, B, F]

: Open [I, E, A], Closed [S, B, F, G]

Hence the final solution path will be: **S**----> **B**---->**F** > **G**

Time Complexity: The worst-case time complexity of Greedy best first search is $O(b^m)$.

Space Complexity: The worst-case space complexity of Greedy best first search is $O(bm)$. Where, m is the maximum depth of the search space.

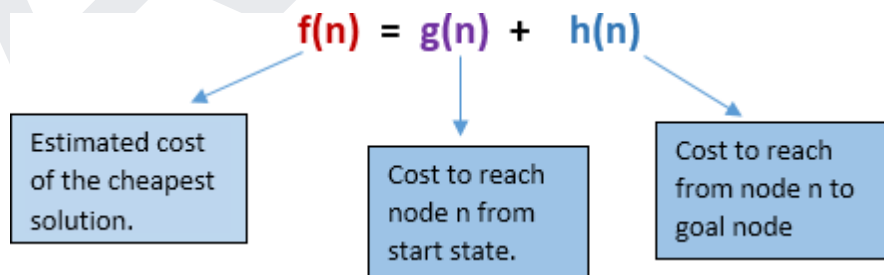
Complete: Greedy best-first search is also incomplete, even if the given state space is finite.

Optimal: Greedy best first search algorithm is not optimal.

Topic: -02: A*search

- A* search is the most commonly known form of best-first search.
- It uses heuristic function $h(n)$, and cost to reach the node n from the start state $g(n)$.
- It has combined features of UCS and greedy best-first search, by which it solves the problem efficiently.
- A* search algorithm finds the shortest path through the search space using the heuristic function. This search algorithm expands less search tree and provides optimal result faster.
- A* algorithm is similar to UCS except that it uses $g(n)+h(n)$ instead of $g(n)$.

In A* search algorithm, we use search heuristic as well as the cost to reach the node. Hence, we can combine both costs as following, and this sum is called as a **fitness number**.



Algorithm of A* search:

Step 1: Place the starting node in the OPEN list.

Step 2: Check if the OPEN list is empty or not, if the list is empty then return failure and stop.

Step 3: Select the node from the OPEN list which has the smallest value of evaluation function ($g+h$), if node n is goal node, then return success and stop, otherwise

Step 4: Expand node n and generate all of its successors, and put n into the closed list. For each successor n' , check whether n' is already in the OPEN or CLOSED list, if not then compute evaluation function for n' and place into Open list.

Step 5: Else if node n' is already in OPEN and CLOSED, then it should be attached to the back pointer which reflects the lowest $g(n')$ value.

Step 6: Return to Step 2.

Advantages:

- A* search algorithm is the best algorithm than other search algorithms.
- A* search algorithm is optimal and complete.
- This algorithm can solve very complex problems.

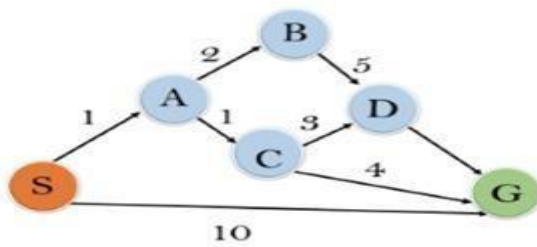
Disadvantages:

- It does not always produce the shortest path as it mostly based on heuristics and approximation.
- A* search algorithm has some complexity issues.
- The main drawback of A* is memory requirement as it keeps all generated nodes in the memory, so it is not practical for various large-scale problems.

Example:

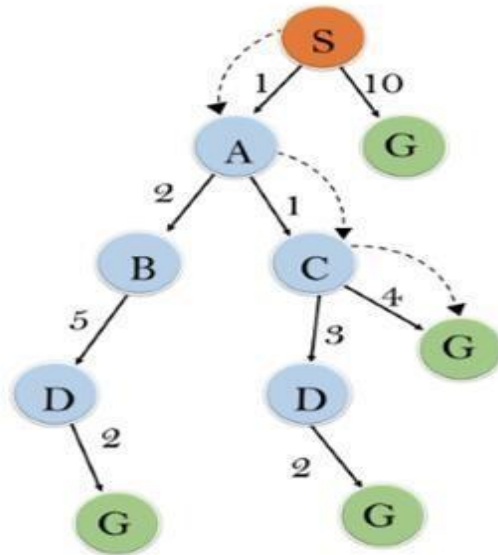
In this example, we will traverse the given graph using the A* algorithm. The heuristic value of all states is given in the below table so we will calculate the $f(n)$ of each state using the formula $f(n) = g(n) + h(n)$, where $g(n)$ is the cost to reach any node from start state.

Here we will use OPEN and CLOSED list.



State	$h(n)$
S	5
A	3
B	4
C	2
D	6
G	0

Solution:



Initialization: $\{(S, 5)\}$

Iteration1: $\{(S \rightarrow A, 4), (S \rightarrow G, 10)\}$

Iteration2: $\{(S \rightarrow A \rightarrow C, 4), (S \rightarrow A \rightarrow B, 7), (S \rightarrow G, 10)\}$

Iteration3: $\{(S \rightarrow A \rightarrow C \rightarrow G, 6), (S \rightarrow A \rightarrow C \rightarrow D, 11), (S \rightarrow A \rightarrow B, 7), (S \rightarrow G, 10)\}$

Iteration 4 will give the final result, as $S \rightarrow A \rightarrow C \rightarrow G$ it provides the optimal path with cost 6.

Points to remember:

- A* algorithm returns the path which occurred first, and it does not search for all remaining paths.
- The efficiency of A* algorithm depends on the quality of heuristic.
- A* algorithm expands all nodes which satisfy the condition $f(n)$

Complete: A* algorithm is complete as long as:

Branching factor is finite.

Cost at every action is fixed.

Optimal: A* search algorithm is optimal if it follows below two conditions:

Admissible: the first condition requires for optimality is that $h(n)$ should be an admissible heuristic for A* tree search. An admissible heuristic is optimistic in nature.

Consistency: Second required condition is consistency for only A* graph-search.

If the heuristic function is admissible, then A* tree search will always find the least cost path.

Time Complexity: The time complexity of A* search algorithm depends on heuristic function, and the number of nodes expanded is exponential to the depth of solution d . So, the time complexity is $O(b^d)$, where b is the branching factor.

Space Complexity: The space complexity of A* search algorithm is $O(b^d)$

Topic: -03: Heuristic functions

- Heuristic functions in order to reach the goal node in a more prominent way.
- Therefore, there are several pathways in a search tree to reach the goal node from the current node.
- The selection of a good heuristic function matters certainly.
- A good heuristic function is determined by its efficiency.
- More is the information about the problem, more is the processing time.

Example:

- Consider the following **8-puzzle problem** where we have a start state and a goal state.
- Our task is to slide the tiles of the current/start state and place it in an order followed in the goal state. There can be four moves either left, right, up, or down.
- There can be several ways to convert the current/start state to the goal state, but we can use a heuristic function $h(n)$ to solve the problem more efficiently.

1	2	3
8	6	
7	5	4

Start State

1	2	3
8		4
7	6	5

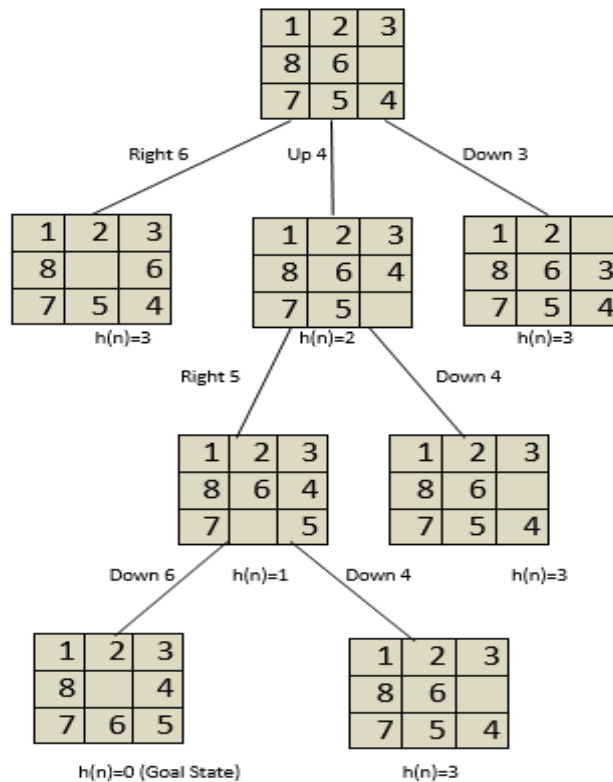
Goal State

A heuristic function for the 8-puzzle problem is defined below:

$h(n)$ =Number of tiles out of position.

So, there is total of three tiles out of position i.e., 6,5 and 4. Do not count the empty tile present in the goal state). i.e. $h(n)=3$. Now, we require to minimize the value of $h(n) = 0$.

We can construct a state-space tree to minimize the $h(n)$ value to 0, as shown below:

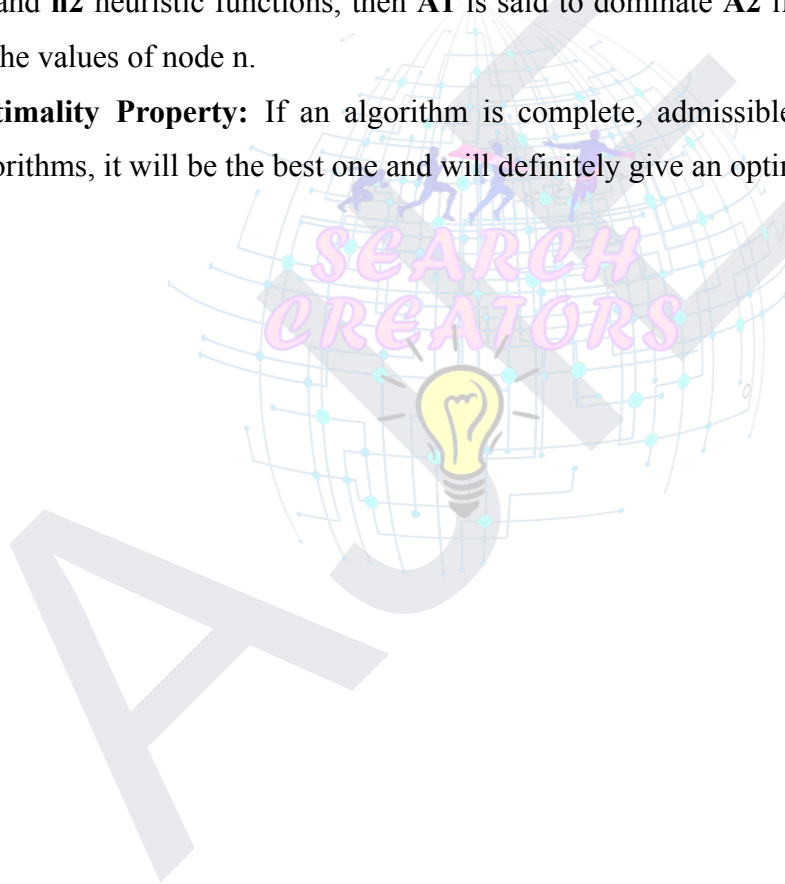


- It is seen from the above state space tree that the goal state is minimized from $h(n)=3$ to $h(n)=0$. However, we can create and use several heuristic functions as per the requirement.
- It is also clear from the above example that a heuristic function $h(n)$ can be defined as the information required to solve a given problem more efficiently.
- The information can be related to the nature of the state, cost of transforming from one state to another, goal node characteristics, etc., which is expressed as a heuristic function.

Properties of a Heuristic search Algorithm

Use of heuristic function in a heuristic search algorithm leads to following properties of a heuristic search algorithm:

- **Admissible Condition:** An algorithm is said to be admissible, if it returns an optimal solution.
- **Completeness:** An algorithm is said to be complete, if it terminates with a solution (if the solution exists).
- **Dominance Property:** If there are two admissible heuristic algorithms **A1** and **A2** having **h1** and **h2** heuristic functions, then **A1** is said to dominate **A2** if **h1** is better than **h2** for all the values of node **n**.
- **Optimality Property:** If an algorithm is complete, admissible, and dominating other algorithms, it will be the best one and will definitely give an optimal solution.



Chapter-02: Introduction to Machine Learning

What is Machine Learning

In the real world, we are surrounded by humans who can learn everything from their experiences with their learning capability, and we have computers or machines which work on our instructions.

But can a machine also learn from experiences or past data like a human does? So here comes the role of **Machine Learning**.

Introduction to Machine Learning

A subset of artificial intelligence known as machine learning focuses primarily on the creation of algorithms that enable a computer to independently learn from data and previous experiences.

Arthur Samuel first used the term "machine learning" in 1959.

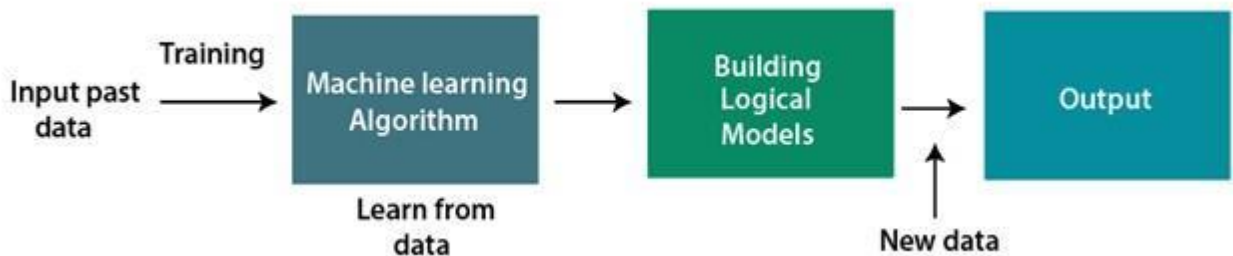
It could be summarized as follows:

- Without being explicitly programmed, machine learning enables a machine to automatically learn from data, improve performance from experiences, and predict things.
- Machine learning algorithms create a mathematical model that, without being explicitly programmed, aids in making predictions or decisions with the assistance of sample historical data, or training data.
- For the purpose of developing predictive models, machine learning brings together statistics and computer science.
- Algorithms that learn from historical data are either constructed or utilized in machine learning. The performance will rise in proportion to the quantity of information we provide.

A machine can learn if it can gain more data to improve its performance.

How does Machine Learning work

- A machine learning system builds prediction models, learns from previous data, and predicts the output of new data whenever it receives it.
- The amount of data helps to build a better model that accurately predicts the output, which in turn affects the accuracy of the predicted output.
- Let's say we have a complex problem in which we need to make predictions.
- Instead of writing code, we just need to feed the data to generic algorithms, which build the logic based on the data and predict the output.
- Our perspective on the issue has changed as a result of machine learning.
- The Machine Learning algorithm's operation is depicted in the following block diagram:



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.

Need for Machine Learning

- The demand for machine learning is steadily rising.
- Because it is able to perform tasks that are too complex for a person to directly implement, machine learning is required.
- Humans are constrained by our inability to manually access vast amounts of data; as a result, we require computer systems, which is where machine learning comes in to simplify our lives.
- By providing them with a large amount of data and allowing them to automatically explore the data, build models, and predict the required output, we can train machine learning algorithms.
- The cost function can be used to determine the amount of data and the machine learning algorithm's performance.
- We can save both time and money by using machine learning.

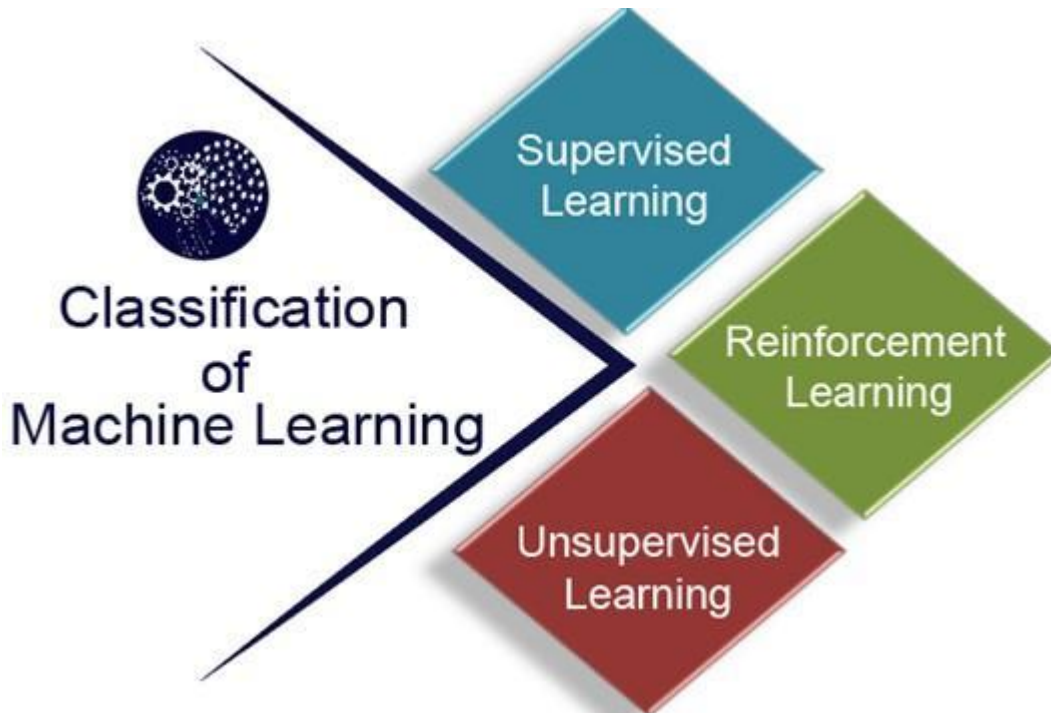
Following are some key points which show the importance of Machine Learning:

- 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

At a broad level, machine learning can be classified into three types:

1. Supervised learning
2. Unsupervised learning
3. Reinforcement learning



1) Supervised Learning

In supervised learning, sample labeled data are provided to the machine learning system for training, and the system then predicts the output based on the training data.

- The system uses labeled data to build a model that understands the datasets and learns about each one.
- After the training and processing are done, we test the model with sample data to see if it can accurately predict the output.
- The mapping of the input data to the output data is the objective of supervised learning. The managed learning depends on oversight, and it is equivalent to when an understudy learns things in the management of the educator.
- Spam filtering is an example of supervised learning.

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

1. **Classification**
2. **Regression**

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 classified into two categories of algorithms:

1. **Clustering**
2. **Association**

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.

MACHINE LEARNING EXPLAINED

- Machine learning is an important sub-branch of Artificial Intelligence (AI).
- A frequently quoted definition of machine learning was by Arthur Samuel, one of the pioneers of Artificial Intelligence.
- He stated that “Machine learning is the field of study that gives the computers ability to learn without being explicitly programmed.”
- The key to this definition is that the systems should learn by itself without explicit programming.
- How is it possible? It is widely known that to perform a computation, one needs to write programs that teach the computers how to do that computation.
- In conventional programming, after understanding the problem, a detailed design of the program such as a flowchart or an algorithm needs to be created and converted into programs using a suitable programming language.
- This approach could be difficult for many real-world problems such as puzzles, games, and complex image recognition applications.
- Initially, artificial intelligence aims to understand these problems and develop general purpose rules manually.

- Then, these rules are formulated into logic and implemented in a program to create intelligent systems.
- This idea of developing intelligent systems by using logic and reasoning by converting an expert's knowledge into a set of rules and programs is called an expert system.
- An expert system like MYCIN was designed for medical diagnosis after converting the expert knowledge of many doctors into a system.
- However, this approach did not progress much as programs lacked real intelligence.
- The word MYCIN is derived from the fact that most of the antibiotics' names end with 'mycin'.
- The above approach was impractical in many domains as programs still depended on human expertise and hence did not truly exhibit intelligence.
- Then, the momentum shifted to machine learning in the form of data driven systems. The focus of AI is to develop intelligent systems by using data-driven approach, where data is used as an input to develop intelligent models.
- The models can then be used to predict new inputs.
- Thus, the aim of machine learning is to learn a model or set of rules from the given dataset automatically so that it can predict the unknown data correctly.
- As humans take decisions based on an experience, computers make models based on extracted patterns in the input data and then use these data-filled models for prediction and to take decisions. For computers, the learnt model is equivalent to human experience. This is shown in Figure 1.2.

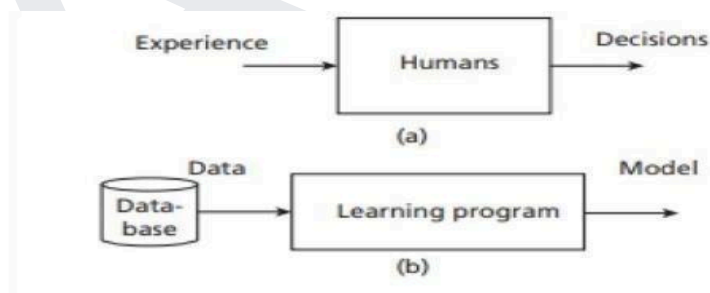


Figure 1.2: (a) A Learning System for Humans (b) A Learning System for Machine Learning

MACHINE LEARNING IN RELATION TO OTHER FIELDS

Machine learning uses the concepts of Artificial Intelligence, Data Science, and Statistics primarily. It is the resultant of combined ideas of diverse fields.

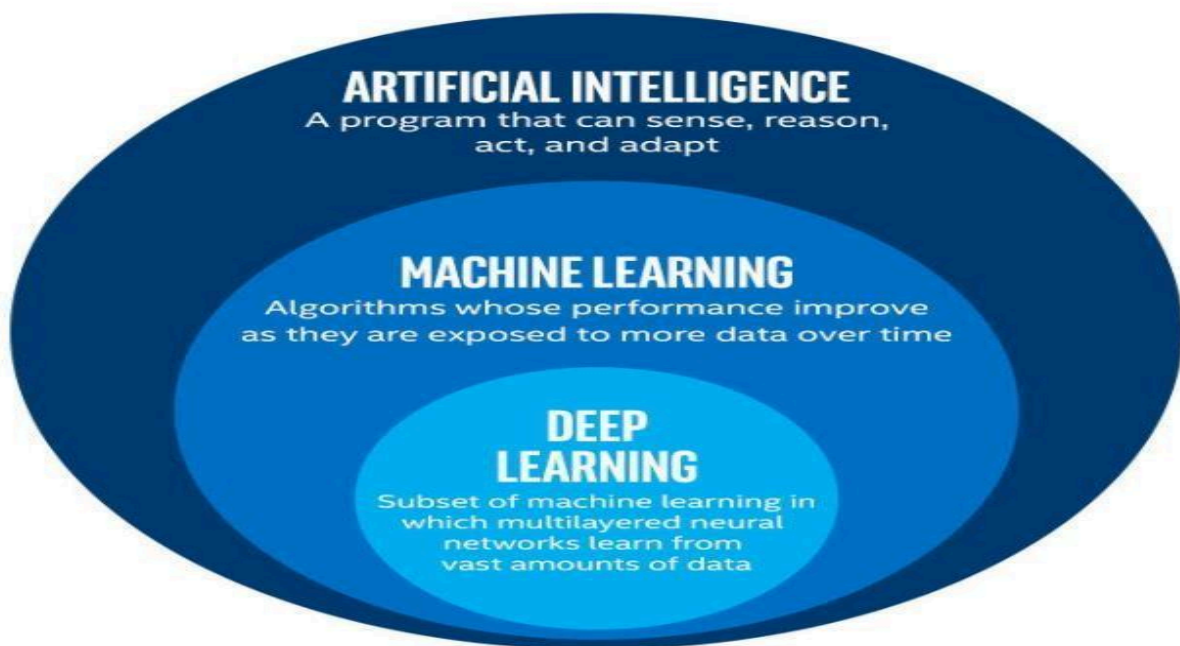
Machine Learning and Artificial Intelligence

Machine learning is an important branch of AI, which is a much broader subject. The aim of AI is to develop intelligent agents. An agent can be a robot, humans, or any autonomous systems. Initially, the idea of AI was ambitious, that is, to develop intelligent systems like human beings. The focus was on logic and logical inferences. It had seen many ups and downs. These down periods were called AI winters.

The resurgence in AI happened due to development of data driven systems. The aim is to find relations and regularities present in the data. Machine learning is the subbranch of AI, whose aim is to extract the patterns for prediction. It is a broad field that includes learning from examples and other areas like reinforcement learning.

The relationship of AI and machine learning is shown in Figure 1.3. The model can take an unknown instance and generate results.

Figure 1.3: Relationship of AI with Machine Learning



- Deep learning is a subbranch of machine learning.
- In deep learning, the models are constructed using neural network technology.
- Neural networks are based on the human neuron models.
- Many neurons form a network connected with the activation functions that trigger further neurons to perform tasks.

Topic-02: Understanding Data

WHAT IS DATA?

- All facts are data. In computer systems, bits encode facts present in numbers, text, images, audio, and video.
- Data can be directly human interpretable (such as numbers or texts) or diffused data such as images or video that can be interpreted only by a computer.
- Data is available in different data sources like flat files, databases, or data warehouses. It can either be an operational data or a non-operational data.
- Operational data is the one that is encountered in normal business procedures and processes. For example, daily sales data is operational data, on the other hand, non-operational data is the kind of data that is used for decision making.
- Data by itself is meaningless. It has to be processed to generate any information. A string of bytes is meaningless. Only when a label is attached like height of students of a class, the data becomes meaningful.
- Processed data is called information that includes patterns, associations, or relationships among data. For example, sales data can be analyzed to extract information like which product was sold larger in the last quarter of the year.

Elements of Big Data

Data whose volume is less and can be stored and processed by a small-scale computer is called 'small data'. These

data are collected from several sources, and integrated and processed by a small-scale computer. Big data, on the other hand, is a larger data whose volume is much larger than 'small data' and is characterized as follows:

1. Volume – Since there is a reduction in the cost of storing devices, there has been a tremendous growth of data. Small traditional data is measured in terms of gigabytes (GB) and terabytes (TB), but Big Data is measured in terms of petabytes (PB) and exabytes (EB). One exabyte is 1 million terabytes.
2. Velocity – The fast arrival speed of data and its increase in data volume is noted as velocity. The availability of IoT devices and Internet power ensures that the data is arriving at a faster rate. Velocity helps to understand the relative growth of big data and its accessibility by users, systems and applications.
3. Variety – The variety of Big Data includes:
 - Form – There are many forms of data. Data types range from text, graph, audio, video, to maps. There can be composite data too, where one media can have many other sources of data, for example, a video can have an audio song.
 - Function – These are data from various sources like human conversations, transaction records, and old archive data.
 - Source of data – This is the third aspect of variety. There are many sources of data. Broadly, the data source can be classified as open/public data, social media data and multimodal data.
 - Some of the other forms of Vs that are often quoted in the literature as characteristics of big data are:
4. Veracity of data – Veracity of data deals with aspects like conformity to the facts, truthfulness, believability, and confidence in data. There may be many sources of error such as technical errors, typographical errors, and human errors. So, veracity is one of the most important aspects of data.
5. Validity – Validity is the accuracy of the data for taking decisions or for any other goals that are needed by the given problem.

6. Value – Value is the characteristic of big data that indicates the value of the information that is extracted

- from the data and its influence on the decisions that are taken based on it. Thus, these 6 Vs are helpful to characterize the big data. The data quality of the numeric attributes is determined by factors like precision, bias, and accuracy.
- Precision is defined as the closeness of repeated measurements. Often, standard deviation is used to measure the precision.
- Bias is a systematic result due to erroneous assumptions of the algorithms or procedures. Accuracy is the degree of measurement of errors that refers to the closeness of measurements to the true value of the quantity. Normally, the significant digits used to store and manipulate indicate the accuracy of the measurement.

Types of Data

In Big Data, there are three kinds of data. They are structured data, unstructured data, and semi structured data.

Structured Data

In structured data, data is stored in an organized manner such as a database where it is available in the form of a table. The data can also be retrieved in an organized manner using tools like SQL. The structured data

frequently encountered in machine learning are listed below:

Record Data A dataset is a collection of measurements taken from a process. We have a collection of objects in a

dataset and each object has a set of measurements. The measurements can be arranged in the form of a matrix.

Rows in the matrix represent an object and can be called as entities, cases, or records. The columns of the dataset

are called attributes, features, or fields. The table is filled with observed data. Also, it is better to note the general

jargons that are associated with the dataset. Label is the term that is used to describe the individual observations.

Data Matrix It is a variation of the record type because it consists of numeric attributes.

The standard matrix

operations can be applied on these data. The data is thought of as points or vectors in the multidimensional space

where every attribute is a dimension describing the object.

Graph Data It involves the relationships among objects. For example, a web page can refer to another web page.

This can be modeled as a graph. The nodes are web pages and the hyperlink is an edge that connects the nodes.

Ordered Data Ordered data objects involve attributes that have an implicit order among them.

The examples

of ordered data are:

- Temporal data – It is the data whose attributes are associated with time. For example, the customer purchasing patterns during festival time is sequential data. Time series data is a special type of sequence data where the data is a series of measurements over time.
- Sequence data – It is like sequential data but does not have time stamps. This data involves the sequence of words or letters. For example, DNA data is a sequence of four characters – A T G C.
- Spatial data – It has attributes such as positions or areas. For example, maps are spatial data where the points are related by location.

Unstructured Data

Unstructured data includes video, image, and audio. It also includes textual documents, programs, and blog data. It is estimated that 80% of the data are unstructured data.

Semi-Structured Data

Semi-structured data are partially structured and partially unstructured. These include data like XML/JSON data, RSS feeds, and hierarchical data.

Data Storage and Representation

Once the dataset is assembled, it must be stored in a structure that is suitable for data analysis. The goal of data storage management is to make data available for analysis. There are different approaches to organize and

manage data in storage files and systems from flat file to data warehouses. Some of them are listed below:

Flat Files These are the simplest and most commonly available data source. It is also the cheapest way of organizing the data. These flat files are the files where data is stored in plain ASCII or EBCDIC format. Minor changes of data in flat files affect the results of the data mining algorithms.

Hence, flat file is suitable only for storing small dataset and not desirable if the dataset becomes larger.

Some of the popular spreadsheet formats are listed below:

- CSV files – CSV stands for comma-separated value files where the values are separated by commas. These are used by spreadsheet and database applications. The first row may have attributes and the rest of the rows represent the data.

- TSV files – TSV stands for Tab separated values files where values are separated by Tab. Both CSV and

TSV files are generic in nature and can be shared. There are many tools like Google Sheets and Microsoft Excel to process these files.

BIG DATA ANALYTICS AND TYPES OF ANALYTICS

- The primary aim of data analysis is to assist business organizations to take decisions. For example, a business organization may want to know which is the fastest selling product, in order for them to market activities.
- Data analysis is an activity that takes the data and generates useful information and insights for assisting the organizations.
- Data analysis and data analytics are terms that are used interchangeably to refer to the same concept.
- However, there is a subtle difference. Data analytics is a general term and data analysis is a part of it. Data analytics refers to the process of data collection, preprocessing and analysis. It deals with the complete cycle of data management. Data analysis is just analysis and is a part of data analytics. It takes historical data and does the analysis.

Data analytics, instead, concentrates more on future and helps in prediction.

There are four types of data analytics:

1. Descriptive analytics
2. Diagnostic analytics
3. Predictive analytics
4. Prescriptive analytics

BIG DATA ANALYSIS FRAMEWORK

For performing data analytics, many frameworks are proposed. All proposed analytics frameworks have some common factors. Big data framework is a layered architecture. Such an architecture has many advantages such as genericness.

A 4-layer architecture has the following layers:

1. Data connection layer
2. Data management layer
3. Data analytics later

4. Presentation layer

- Data Connection Layer It has data ingestion mechanisms and data connectors. Data ingestion means taking raw data and importing it into appropriate data structures. It performs the tasks of ETL process. By ETL, it means extract, transform and load operations.
- Data Management Layer It performs preprocessing of data. The purpose of this layer is to allow parallel execution of queries, and read, write and data management tasks. There may be many schemes that can be implemented by this layer such as data-in-place, where the data is not moved at all, or constructing data repositories such as data warehouses and pull data on-demand mechanisms.
- Data Analytic Layer It has many functionalities such as statistical tests, machine learning algorithms to understand, and construction of machine learning models. This layer implements many model validation mechanisms too.
- The processing is done as shown in Box 2.1.
- Presentation Layer It has mechanisms such as dashboards, and applications that display the results of analytical engines and machine learning algorithms.
- Thus, the Big Data processing cycle involves data management that consists of the following steps.

1. Data collection

2. Data preprocessing

3. Applications of machine learning algorithm

4. Interpretation of results and visualization of machine learning algorithm

- This is an iterative process and is carried out on a permanent basis to ensure that data is suitable for data mining.
- Application and interpretation of machine learning algorithms constitute the basis for the rest of the book. So, primarily, data collection and data preprocessing are covered as part of this chapter.

Data Collection

The first task of gathering datasets is the collection of data. It is often estimated that most of the time is spent for collection of good quality data. A good quality data yields a better result. It is often difficult to characterize a 'Good data'. 'Good data' is one that has the following properties:

1. Timeliness – The data should be relevant and not stale or obsolete data.
2. Relevancy – The data should be relevant and ready for the machine learning or data mining algorithms. All the necessary information should be available and there should be no bias in the data.
3. Knowledge about the data – The data should be understandable and interpretable, and should be self-sufficient for the required application as desired by the domain knowledge engineer.

Broadly, the data source can be classified as open/public data, social media data and multimodal data.

1. Open or public data source – It is a data source that does not have any stringent copyright rules or restrictions. Its data can be primarily used for many purposes. Government census data are good examples of open data:

- Digital libraries that have huge amount of text data as well as document images • Scientific domains

with a huge collection of experimental data like genomic data and biological data

- Healthcare systems that use extensive databases like patient databases, health insurance data, doctors' information, and bioinformatics information

2. Social media – It is the data that is generated by various social media platforms like Twitter, Facebook, YouTube, and Instagram. An enormous amount of data is generated by these platforms.

Data Preprocessing

In real world, the available data is 'dirty'. By this word 'dirty', it means:

- Incomplete data • Inaccurate data
 - Outlier data • Data with missing values
 - Data with inconsistent values • Duplicate data
- Data preprocessing improves the quality of the data mining techniques. The raw data must be preprocessed to give accurate results. The process of detection and removal of errors in data is called data cleaning.
 - Data wrangling means making the data processable for machine learning algorithms. Some of the data errors include human errors such as typographical errors or incorrect measurement and structural errors like improper data formats. Data errors can also arise from omission and duplication of attributes.
 - Noise is a random component and involves distortion of a value or introduction of spurious objects. Often, the noise is used if the data is a spatial or temporal component. Certain deterministic distortions in the form of a streak are known as artifacts.
 - Consider, for example, the following patient Table 2.1. The 'bad' or 'dirty' data can be observed in this table.

Table 2.1: Illustration of 'Bad' Data

Patient ID	Name	Age	Date of Birth (DoB)	Fever	Salary
1.	John	21		Low	-1500
2.	Andre	36		High	Yes
3.	David	5	10/10/1980	Low	" "
4.	Raju	136		High	Yes