

B.Tech, Sem -7th

Subject : Machine Learning

Unit 1 : Introduction

Computer Science & Engineering

Pragati Mishra(Assistant Professor, PIET-CSE)



Outline

- Introduction to Machine Learning
- Learning Paradigms
- PAC learning
- Basics of Probability
- Version Spaces.

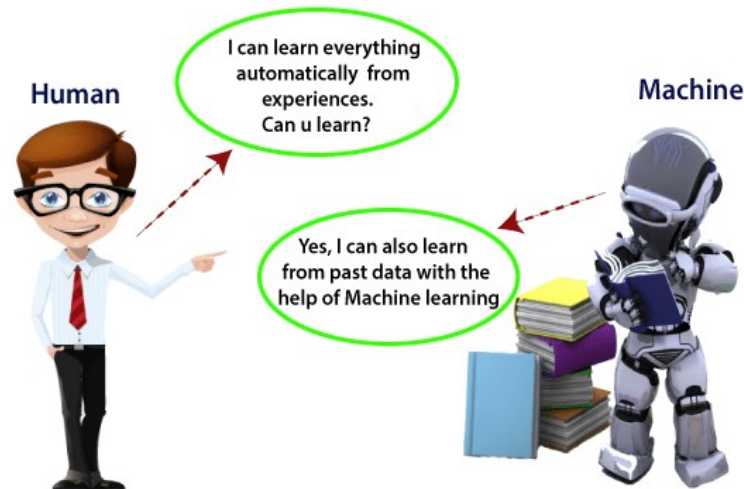
Introduction to Machine Learning

- Machine learning is a subset of artificial intelligence
- It enables the machine to automatically learn from data, improve performance from past experiences, and make predictions.
- Machine learning contains a set of algorithms that work on a huge amount of data.
- Data is fed to these algorithms to train them, and on the basis of training, they build the model & perform a specific task.
- Machine learning uses various algorithms for **building mathematical models and making predictions using historical data or information.**
- It is being used for various tasks such as image recognition, speech recognition, email filtering, Facebook auto-tagging, recommender system, and many more.

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**.

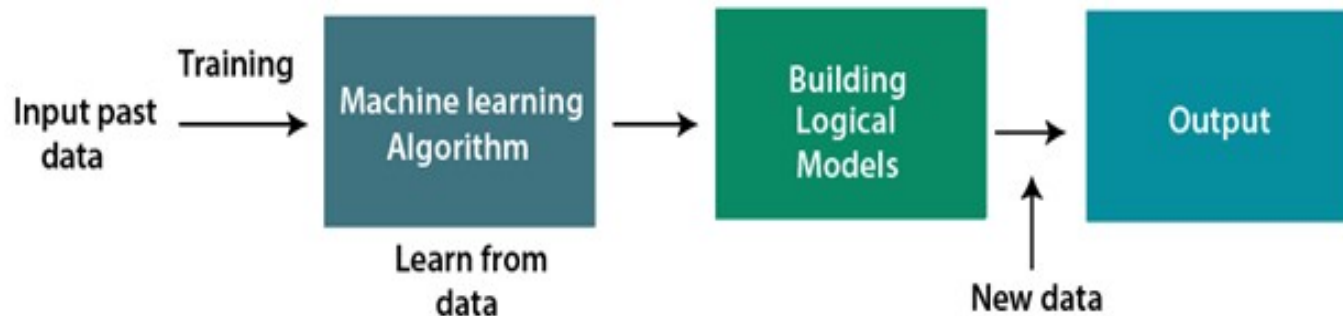


Introduction to Machine Learning

- “Machine learning enables a machine to automatically learn from data, improve performance from experiences, and predict things without being explicitly programmed.”
- With the help of sample historical data, which is known as **training data**, machine learning algorithms build a **mathematical model** that helps in making predictions or decisions without being explicitly programmed. Machine learning brings computer science and statistics together for creating predictive models. Machine learning constructs or uses the algorithms that learn from historical data. The more we will provide the information, the higher will be the performance.
- A machine has the ability to learn if it can improve its performance by gaining more data.

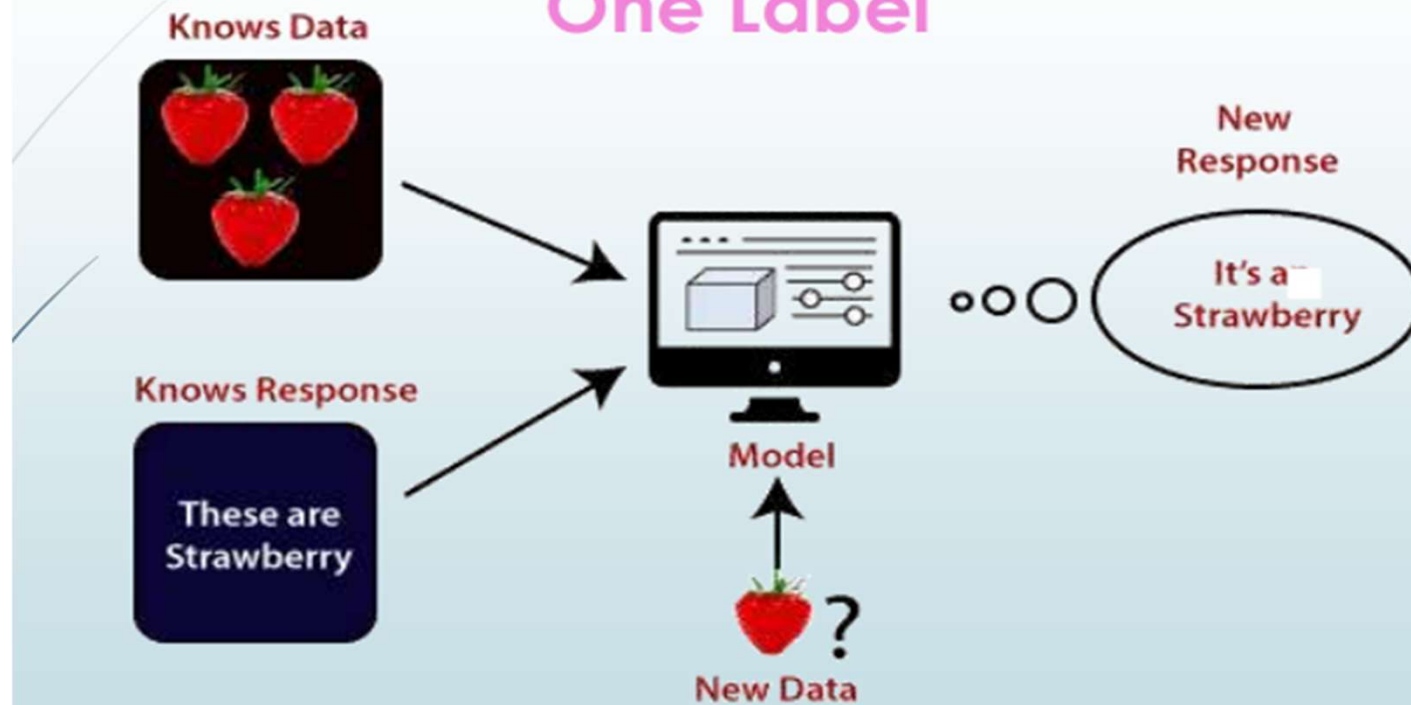
How does Machine Learning work

- 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.



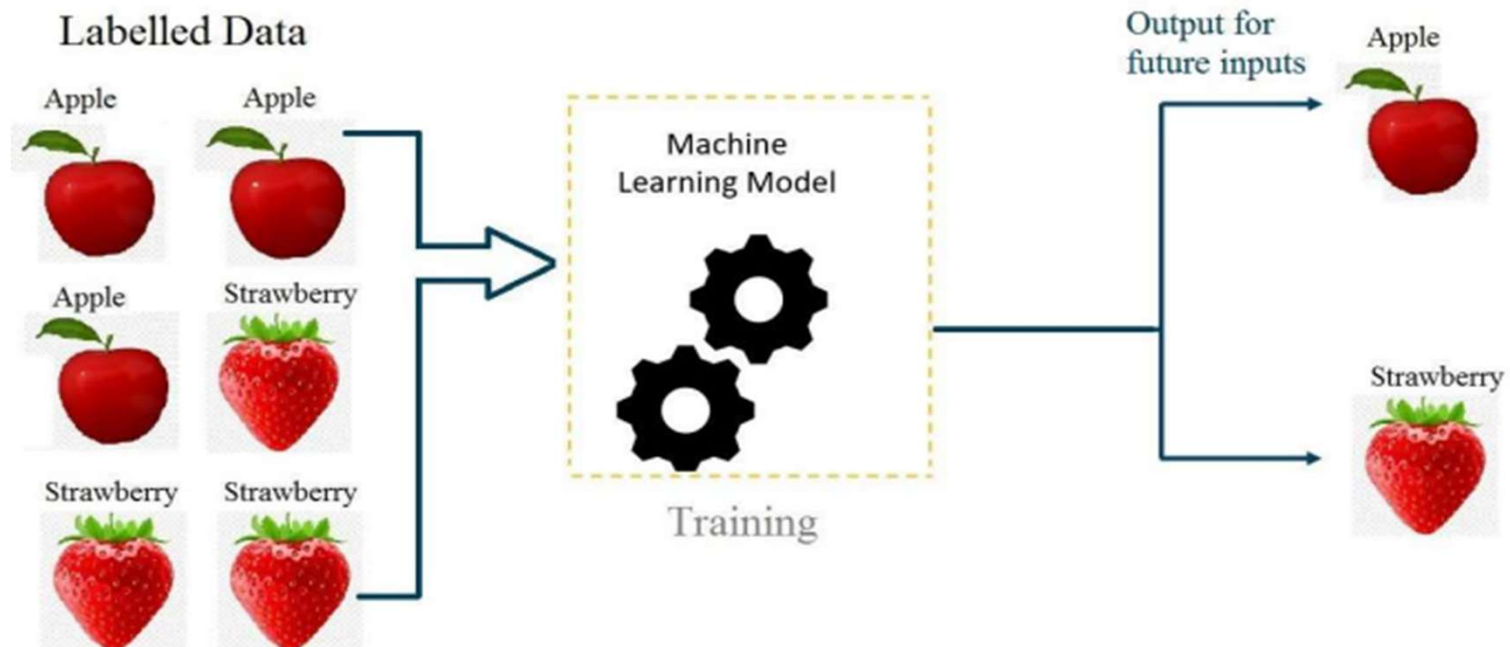
Can a machine predict which fruit is this?

One Label



Can a machine predict which fruit is this?

Two Labels



Machine Learning

Machine Learning

- **Herbert Alexander Simon:**
“Learning is any process by which a system improves performance from experience.”
- “Machine Learning is concerned with computer programs that automatically improve their performance through experience. “



Herbert Simon
[Turing Award](#) 1975
[Nobel Prize in Economics](#) 1978

How does Machine Learning work

- 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:

How machine learning work?



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 need for machine learning is increasing day by day. The reason behind the need for machine learning is that it is capable of doing tasks that are too complex for a person to implement directly. As a human, we have some limitations as we cannot access the huge amount of data manually, so for this, we need some computer systems and here comes the machine learning to make things easy for us.
- ❑ We can train machine learning algorithms by providing them the huge amount of data and let them explore the data, construct the models, and predict the required output automatically. The performance of the machine learning algorithm depends on the amount of data, and it can be determined by the cost function. With the help of machine learning, we can save both time and money.

Need for Machine Learning

- ❑ The importance of machine learning can be easily understood by its uses cases, currently, machine learning is used in self-driving cars, cyber fraud detection, face recognition, and friend suggestion by Facebook, etc. Various top companies such as Netflix and Amazon have built machine learning models that are using a vast amount of data to analyze the user interest and recommend product accordingly.

Applications of Machine Learning

Sample applications of machine learning:

- ☐ **Web search:** ranking page based on what you are most likely to click on.
- ☐ **Computational** biology: rational design drugs in the computer based on past experiments.
- ☐ **Finance:** decide who to send what credit card offers to. Evaluation of risk on credit offers. How to decide where to invest money.
- ☐ **E-commerce:** Predicting customer churn. Whether or not a transaction is fraudulent
- ☐ **Space exploration:** space probes and radio astronomy.

Applications of Machine Learning

- ❑ **Robotics:** how to handle uncertainty in new environments.
Autonomous. Self-driving car.
- ❑ **Information extraction:** Ask questions over databases across the web.
- ❑ **Social networks:** Data on relationships and preferences. Machine learning to extract value from data.
- ❑ **Debugging:** Use in computer science problems like debugging. Labor intensive process. Could suggest where the bug could be.

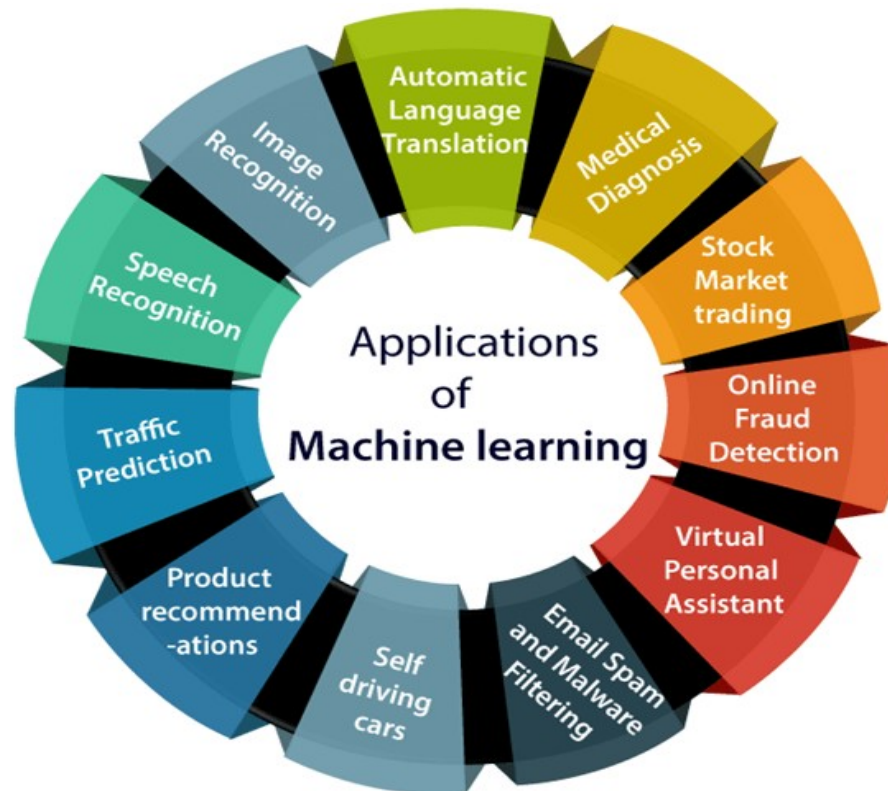
Applications of Machine Learning

But there are much more examples of ML in use

- ❑ **Prediction** — Machine learning can also be used in the prediction systems. Considering the loan example, to compute the probability of a fault, the system will need to classify the available data in groups.
- ❑ **Image recognition** — Machine learning can be used for face detection in an image as well. There is a separate category for each person in a database of several people.
- ❑ **Speech Recognition** — It is the translation of spoken words into the text. It is used in voice searches and more. Voice user interfaces include voice dialing, call routing, and appliance control. It can also be used a simple data entry and the preparation of structured documents.
- ❑ **Medical diagnoses** — ML is trained to recognize cancerous tissues.
- ❑ **Financial industry and trading** — companies use ML in fraud investigations and credit checks.

Applications of Machine Learning

But there are much more examples of ML in use

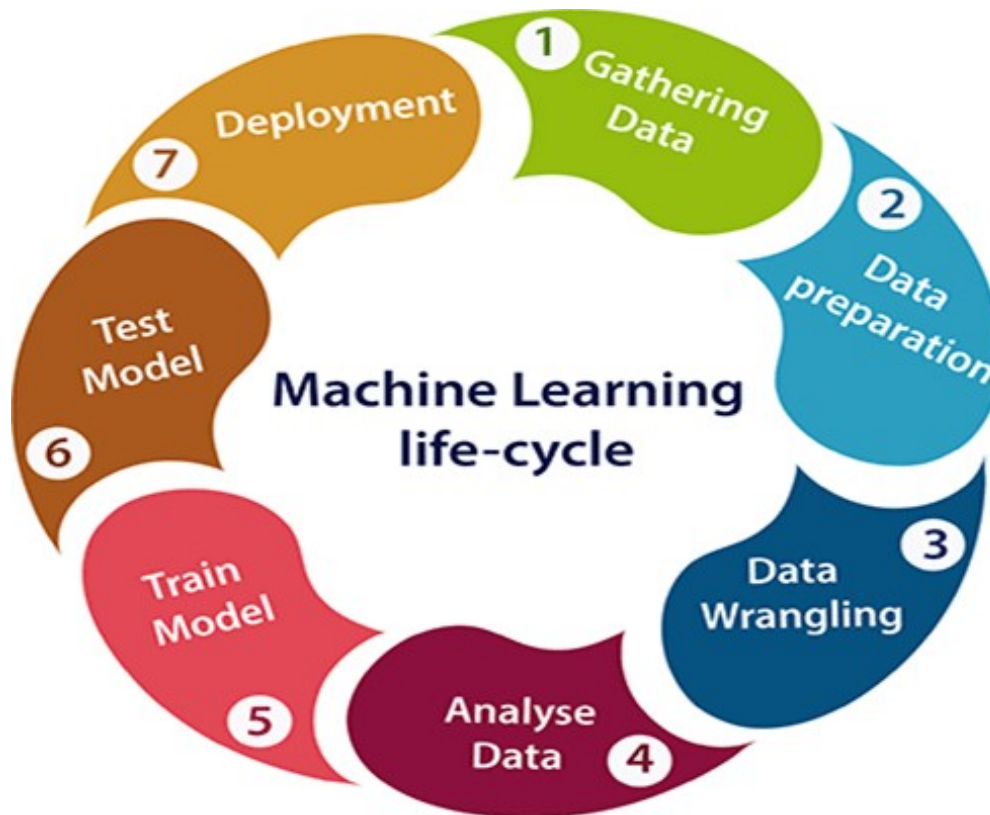


Applications of Machine Learning

5. Self-driving cars:

One of the most exciting applications of machine learning is self-driving cars. Machine learning plays a significant role in self-driving cars. Tesla, the most popular car manufacturing company is working on self-driving car. It is using unsupervised learning method to train the car models to detect people and objects while driving.

Machine learning Life cycle



Machine learning Life cycle

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.



Machine learning Life cycle

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.

Machine learning Life cycle

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.



Machine learning Life cycle

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.



Machine learning Life cycle

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.



Learning Paradigms in Machine Learning

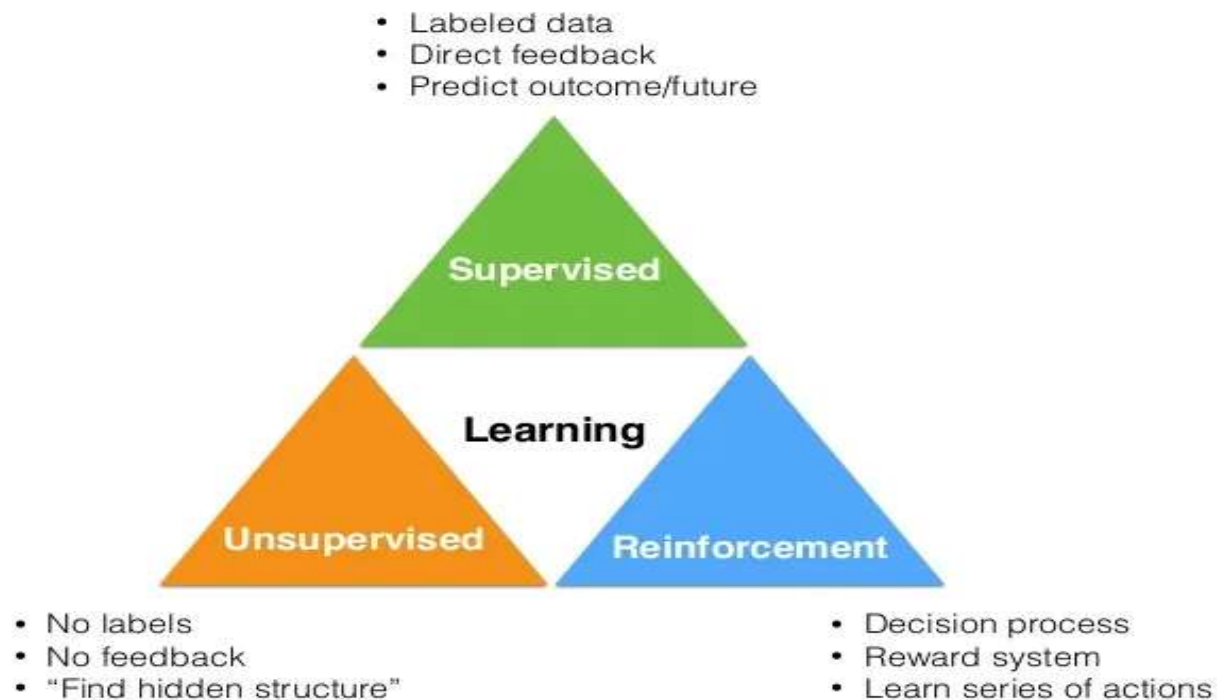
Learning Paradigms basically states a particular pattern on which something or someone learns.

Learning Paradigms related to machine learning, i.e how a machine learns when some data is given to it, its pattern of approach for some particular data.

Machine learning is commonly separated into three main learning paradigms:

1. **Supervised Learning**
2. **Unsupervised Learning**
3. **Reinforcement Learning**

Learning Paradigms in Machine Learning



Learning Paradigms in Machine 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 system creates a model using labeled data to understand the datasets and learn about each data, once the training and processing are done then we test the model by providing a sample data to check whether it is predicting the exact output or not.
- The goal of supervised learning is to map input data with the output data. 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**.



Learning Paradigms in Machine Learning

- The main goal of the supervised learning technique is to map the input variable(x) with the output variable(y). Some real-world applications of supervised learning are Risk Assessment, Fraud Detection, Spam filtering, etc.

Learning Paradigms in Machine Learning

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 characteristics.
- In unsupervised learning, we don't have a predetermined result. The machine tries to find useful insights from the huge amount of data.



Learning Paradigms in Machine Learning

Advantages of Unsupervised Learning

- Unsupervised learning is used for more complex tasks as compared to supervised learning because, in unsupervised learning, we don't have labeled input data.
- Unsupervised learning is preferable as it is easy to get unlabeled data in comparison to labeled data.

Disadvantages of Unsupervised Learning

- Unsupervised learning is intrinsically more difficult than supervised learning as it does not have corresponding output.
- The result of the unsupervised learning algorithm might be less accurate as input data is not labeled, and algorithms do not know the exact output in advance.



Learning Paradigms in Machine Learning

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.

Learning Paradigms in Machine Learning

Advantages and Disadvantages of Reinforcement Learning

Advantages

- It helps in solving complex real-world problems which are difficult to be solved by general techniques.
- The learning model of RL is similar to the learning of human beings; hence most accurate results can be found.
- Helps in achieving long term results.

Disadvantage

- RL algorithms are not preferred for simple problems.
- RL algorithms require huge data and computations.
- Too much reinforcement learning can lead to an overload of states which can weaken the results.



PAC - Learning

- **PAC** stand for **Probably approximately correct**
- **Probably approximately correct (PAC) learning** is a framework for mathematical analysis of machine learning algorithm.
- In the other words PAC learning is a theoretical framework for analyzing the generalization performance of machine learning algorithms.

Goal: With High Probability ("Probably"), the select hypothesis will have lower error ("Approximately Correct")

In the PAC model, we specify two small parameters, ϵ (epsilon) and δ (delta) and require that with probability at least $(1-\delta)$ a system learn a concept with error at most ϵ .



PAC - Learning

ϵ and δ parameters:

- ϵ gives an upper bound on the error in the accuracy with which h approximated (Accuracy : $1-\epsilon$)
- δ gives the probability of failure in the achieving this accuracy (Confidence : $1-\delta$)



PAC - Learning

- ❑ A good learner will learn with high probability and close approximation to the target concept
- ❑ With high probability, the selected hypothesis will have lower the error ("Approximately Correct") with the parameter ϵ and δ

PAC - Learning

- PAC learning, requires
 - small parameters ϵ and δ ,
 - with probability at least $(1 - \delta)$, a system learn the concept with error at most ϵ .
- ϵ is upper bound on the error in accuracy, i.e. the hypothesis with error less than ϵ

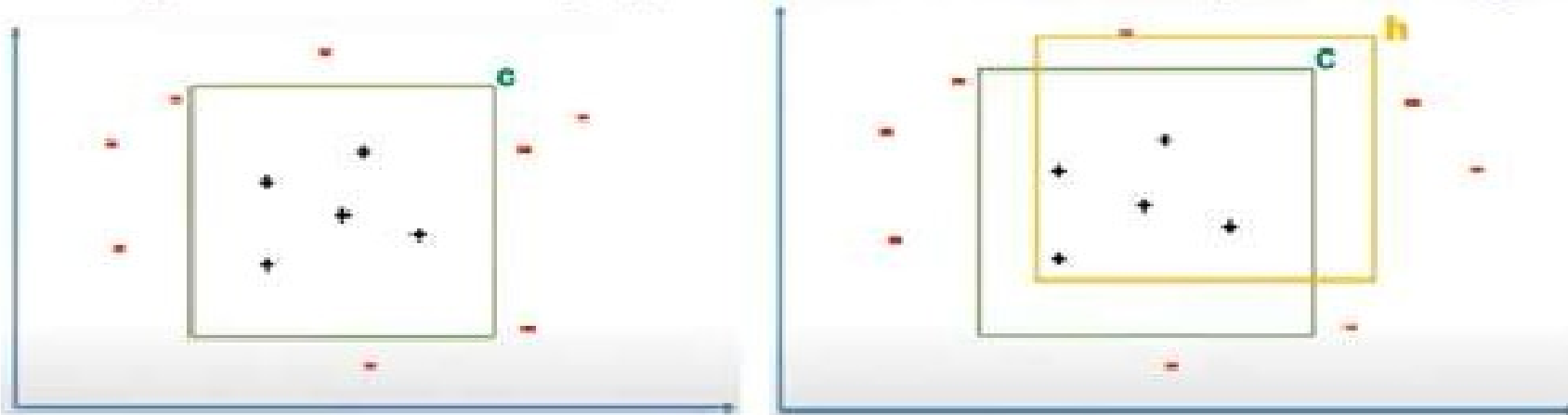
Accuracy: $1 - \epsilon$

- δ give the probability of failure in achieving this accuracy δ , ($0 < \delta \leq 1$), the hypothesis generated is approximately correct at least $1 - \delta$ of the time.

Confidence: $1 - \delta$

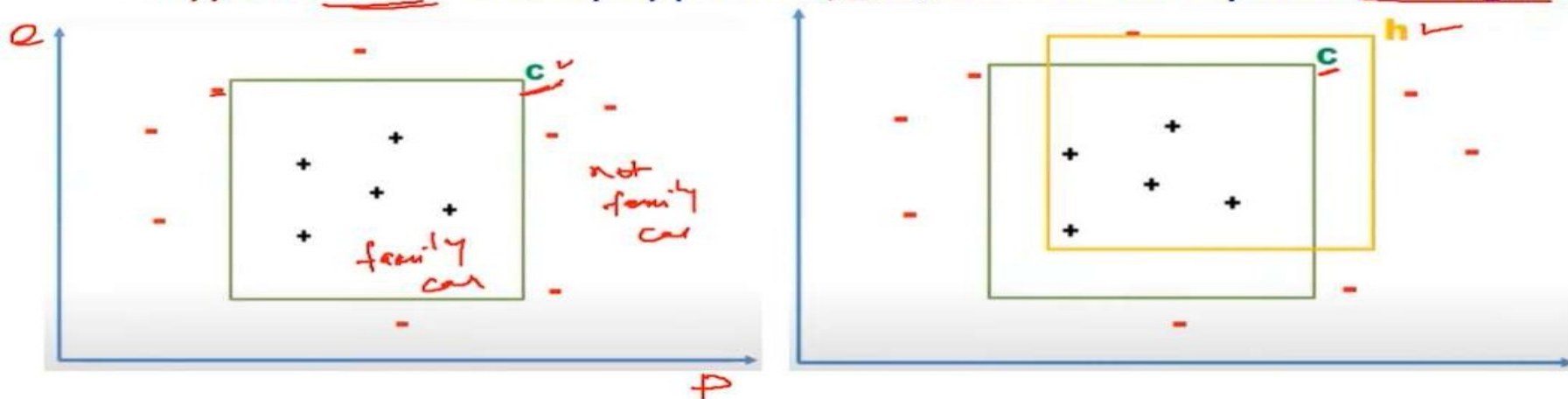
PAC – Learning Example

- N number of Car having Price and Engine power, as training set, (p, e) , find the car is family car or not.
- An algorithm gives answer whether the car is family car or not.
- C – Target function
- Instances within rectangle represents family cars and outside are not family cars
- Hypothesis h – closely approximate C , and there may be error region.



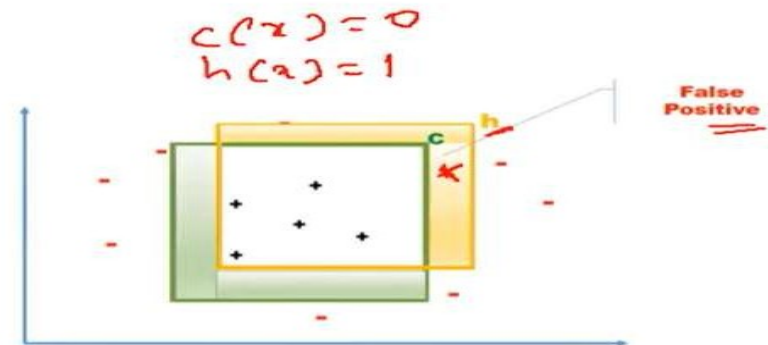
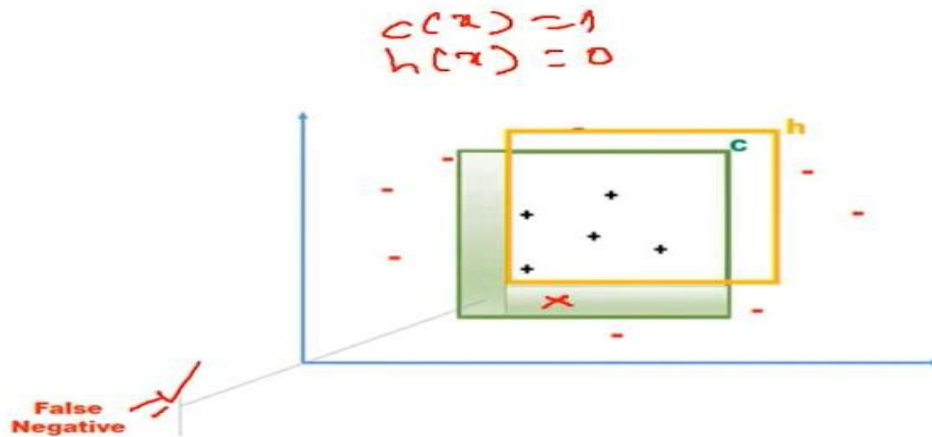
PAC – Learning Example

- N number of Car having Price and Engine power, as training set, (p, e) , find the car is family car or not.
- An algorithm gives answer whether the car is family car or not.
- C – Target function
- Instances within rectangle represents family cars and outside are not family cars
- Hypothesis h – closely approximate C , and there may be error region.



False Negative and False Positive

- Instances lies on shaded region are positive/negative according to our actual function 'C', but those are **negative/positive** based on the **hypothesis h**. Hence it is called as **false negative** or **false positive**

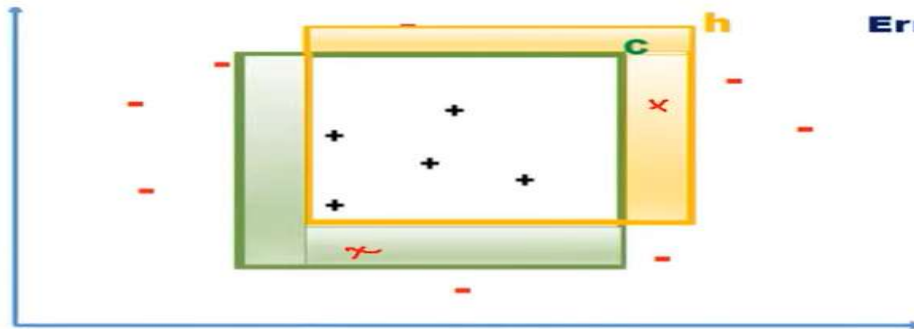


PAC – Learning Example

Error Region

- The probability of error region to be small
- The error region : $P(C \text{ XOR } h) \leq \epsilon$.

$$P(C \oplus h)$$



Error Region : $C \text{ XOR } h$

PAC – Learning Example

Approximately Correct

- the hypothesis h , that approximately correct, and error is less than or equal to ϵ .
- Where $0 \leq \epsilon \leq 1/2$ 0 - 0.5
- i.e. $P(C \text{ XOR } h) \leq \epsilon$ ✓

PAC – Learning Example

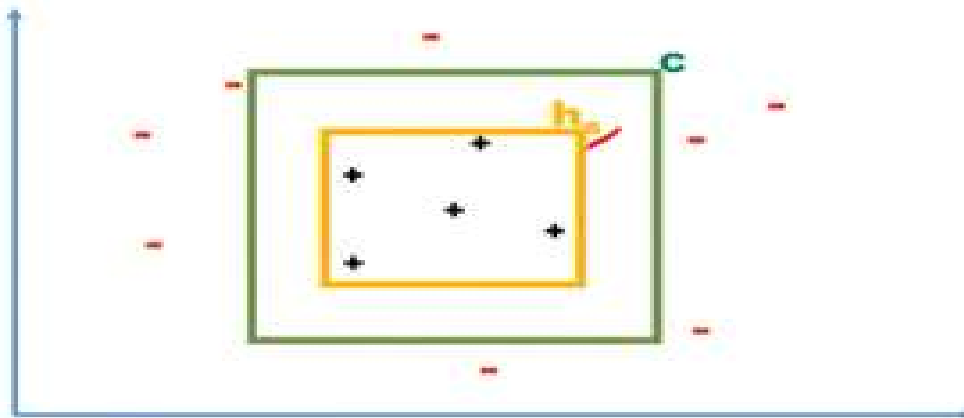
Probably Approximately Correct

- Low generalization error with high probability
- $[P(\text{Error}(h) \leq \epsilon)] \leq 1 - \delta$ ✓
- $P(P(C \text{ XOR } h) \leq \epsilon) \leq 1 - \delta$ ✓

PAC – Learning Example

PAC learnability for axis-aligned rectangle

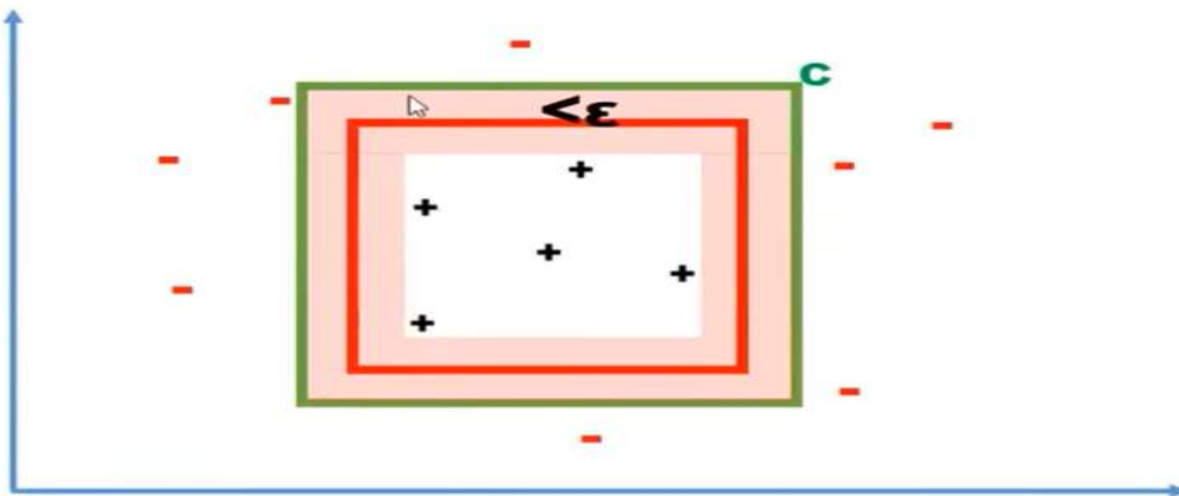
- Specialization:
- h_s is the tightest possible rectangle around a set of positive training examples.
- h_s is subset of C , Hence Error region = $C - h$



PAC – Learning Example

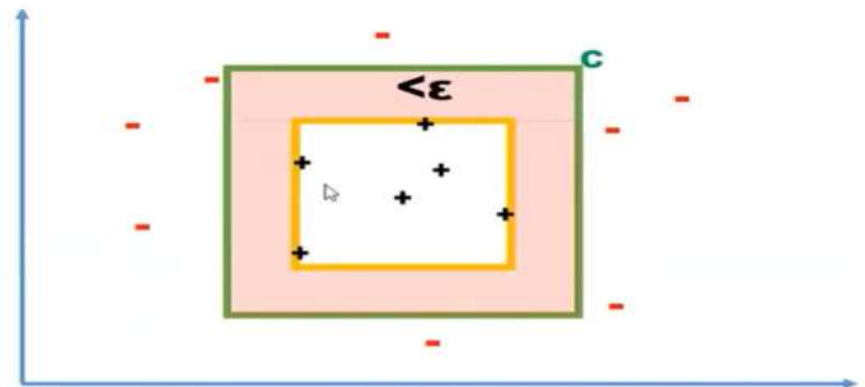
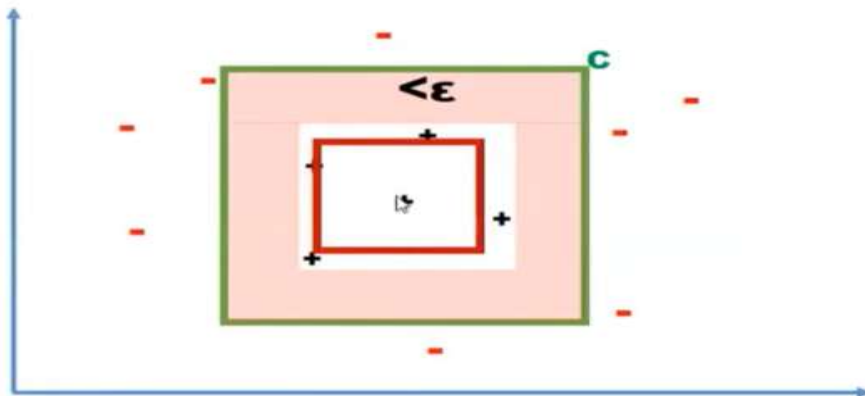
Approximately correct

- If an hypothesis lies between h and c (shaded region) then it is approximately correct.



PAC – Learning Example

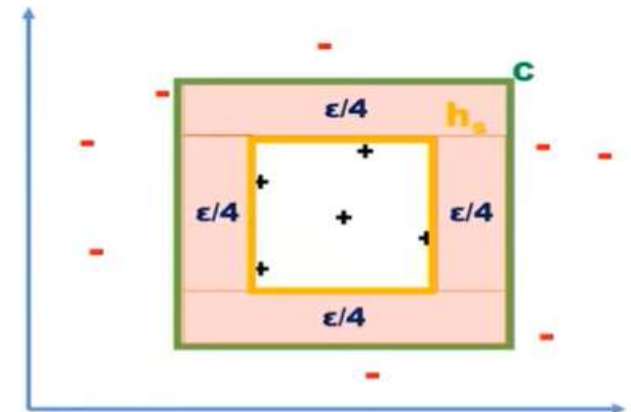
- If the generated hypothesis does not touch any of these region
- Error region is greater than ϵ and not approximately correct, because the error region got increased.
- Atleast one +ve example at each side of the rectangle



PAC – Learning Example


Error Region

- Error Region = sum of four rectangular strips $< \epsilon$
- Each strip is at most $\epsilon/4$
- Probability of positive example falling in any one of the strip (error region = $\epsilon/4$)
- Probability that a randomly drawn positive example misses a strip = $1 - \epsilon/4$
- $P(\text{m instance miss a strip}) = (1 - \epsilon/4)^m$
- $P(\text{m instances miss any strip}) < 4(1 - \epsilon/4)^m$
- Finally we get $m > \frac{4}{\epsilon} \log \frac{4}{\delta}$



PAC – Learning Example

Example Problem 1



Sl.No.	Error(h1)
1	0.001
2	0.025
3	0.07
4	0.003
5	0.035
6	0.045
7	0.027
8	0.065
9	0.012
10	0.036

- Hypothesis h1 generated the errors with respect to price and engine power of given 10 samples,
- *Given, $\epsilon = 0.05$ $\delta = 0.20$*
- *$P(h1) \geq 1 - \delta$*
- $P(h1) = 8/10 = 0.80$ (3rd and 8th values are greater than ϵ)
- Therefore, $0.80 \geq (1 - 0.20)$ i.e. $0.80 = 0.80$
- **Hence h1 is probably approximately correct**



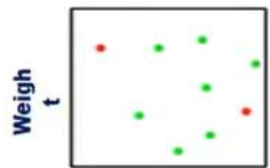
PAC – Learning Example

Example Problem 2

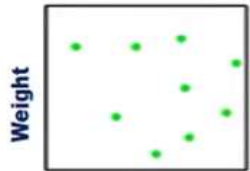
Sl.No.	Error(h2)
1	0.012
2	0.015
3	0.071
4	0.063
5	0.022
6	0.045
7	0.011
8	0.029
9	0.066
10	0.031

- Hypothesis h2 generated the errors with respect to price and engine power of given 10 samples,
- *Given*, $\epsilon = 0.05$ $\delta = 0.20$
- $P(h2) \geq 1 - \delta$
- $P(h2) = 7/10 = 0.70$ (3rd, 4th, 9th values $> \epsilon$)
- Here, $0.70 \geq (1 - 0.20)$ i.e. $0.70 < 0.80$
- Hence h2 is not probably approximately correct

PAC – Learning Example

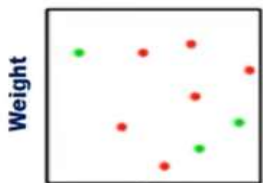


Height



Height

⋮



Height

Error(H1) **Error(h2)**

0.04	0.04
0.03	0.035
0.09	0.039
0.06	0.06
0.025	0.025
0.049	0.059
0.04	0.04
0.03	0.03
0.05	0.55
0.043	0.043

$\epsilon = 0.05$

$\delta = 0.20$

$P(H1) = 8/10 = 0.80$

$P(H1) = 8/10 = 0.80 \geq 1 - 0.20$

Hence H1 is probably approximately correct

$P(H2) = 7/10 = 0.70$

$P(H2) = 7/10 = 0.70 < 1 - 0.20$

Hence H2 is not probably approximately correct

Version Space

- Subset of hypothesis H consistent with training example (D) .

$$VS_{H,D} = \{H_{E_H} \mid \text{Consistent}(h,D)\}$$

H = hypothesis

Consistent

D = training Example

$h(x)=c(x)$

Algorithm to obtain Version Space

List-Then-Eliminate algorithm

1. Version Space \leftarrow a list containing every hypothesis is in H
2. For each training example, $\langle x, c(x) \rangle$ remove from Version Space any hypothesis h for which $h(x) \neq c(x)$
3. Output the list of hypotheses in Version Space



Thank You

Parul[®] University



www.paruluniversity.ac.in

