CHAPTER 1

UNIT I

Introduction to Machine Learning

Syllabus

Classic and adaptive machines, Machine learning matters, Beyond machine learning-deep learning and bio inspired adaptive systems, Machine learning and Big data.

Important Elements of Machine Learning-Data formats, Learnability.

Without being explicitly programmed, machine learning helps to learn from experience and examples. It builds logic based on the data you feed to a generic algorithm besides writing code for it.

Some of examples of machine learning are :

1. Face detection, which identifies faces among many images and determines if the person's face is present or absent
2. Email filtering filters or classification of emails received into spam and not-spam.
3. Medical diagnosis helps to detect that patient has suffered of some disease or not.
4. Weather prediction for prediction of coming storm, rain etc.

1.1 Classic and Adaptive Machines

The classical system receives some input values and produces output as a result of processing them.

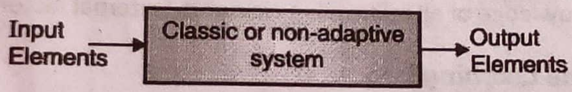


Fig. 1.1.1 : Environment of Classical System

Adaptive system has the ability to adapt its behavior to external signals like datasets or real time input to predict the future.

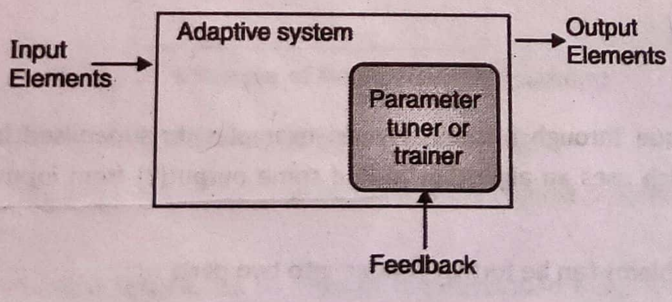


Fig. 1.1.2 : Adaptive System

Adaptive system isn't based on static or permanent structures (model parameters and architectures) but rather on a continuous ability to adapt its behavior to external signals (datasets or real-time inputs) and, like a human being, to predict the future using uncertain and fragmentary pieces of information.

1.2 Machine Learning Matters

Machine Learning

Machine Learning is a subset of Artificial Intelligence, that provides systems ability to learn automatically. It is a concept which allows machines to learn from their experience and examples and improve from experience without being explicitly programmed. Machines learn from their experience and makes predictions based on its experience. A machine learning algorithm is trained on a set of data to create a model, this data is called as training data. When an unseen data is given to the algorithm, it makes prediction using the model built.

Machine learning can be of following types as follows : Supervised learning, Unsupervised learning and Reinforcement learning

Examples of Machine learning in our day to day lives

(A) Online Shopping Recommendation

Recommendation of similar product while searching/shopping a online product. After buying one product, suggestion of people who bought this product also bought another product (combination of product)

(B) Target marketing using the concept of clustering in machine learning

Call centre people calling up and offering a bank loan or credit card. This call is made only to few selected customers who they think will purchase their product

As per the definition given by Giuseppe Bonaccorso, machine learning is an engineering approach that gives maximum importance to every technique that increases or improves the propensity for changing adaptively.

Therefore, the main goal of machine learning is to study, engineer, and improve mathematical models, which can be trained (once or continuously) with context-related data (provided by a generic environment), to infer the future and to make decisions without complete knowledge of all influencing elements (external factors).

Common Approaches to Machine Learning

Some of common approaches to machine learning are as follows :

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

1. Supervised Learning

- It is a learning technique through previously given examples. In supervised learning both input and output variables are given which uses an algorithm to find some output(Y) from input(X) by deriving some mapping function like $\boxed{Y = f(X)}$
- Supervised learning problems can be further divided into two parts :

  (a) Classification : In this case the output variable is a category or a group

  For example "red" or "yellow" or "spam" and "not spam".

Scanned with CamScanner

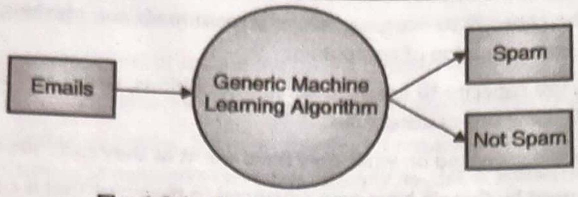


Fig. 1.2.1 : Example of Classification

(b) Regression: In this type of problem, the output variable is a real value

For example "dollars" or "weight"

2. **Unsupervised Learning**

- It is a learning technique where the system discovers the patterns or structures directly from the example given.
- In unsupervised learning, only input data (X) is given to find corresponding output variables.
- Unsupervised learning problems can be further divided into parts :

(a) Association : This type of problem discovers rules to describe large data.

For example, "If a person buys an item 'A' then also tends to buy item 'B'."

(b) Clustering : This type of problem discovers groups of data based on similarities.

For example, "Customers' grouping based on their buying habit".

3. **Reinforcement Learning**

- It is a learning technique where dynamic environment is given and computer program interacts with it to perform a particular task. The program is also provided rewards and punishments as a feedback to navigate its problem space.
- Here, the machine is trained to make particular decisions and it continuously trains using trial and error methods.

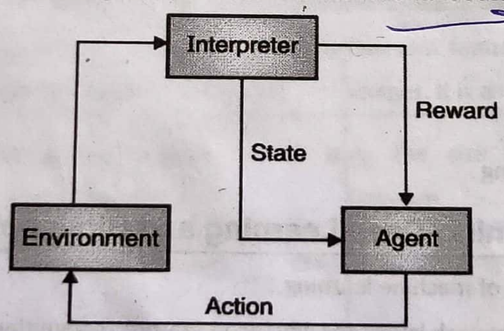


Fig. 1.2.2 : Example of Reinforcement Learning

1.2.1 Role of Machine Learning Algorithms in Spam Filtering

- Email is the way to exchange information and increase in spam is the biggest problem as it may get generated from anywhere in the world wide web.
- Most spam filtering methods use text techniques so most of the developed models for minimizing spam have been machine learning algorithms.
- To effectively handle the threat posed by email spams, leading email providers such as Gmail, Yahoo mail and Outlook have employed the combination of different machine learning (ML) techniques such as Neural Networks in its spam filters.

- These ML techniques have the capacity to learn and identify spam mails and phishing messages by analyzing loads of such messages throughout a vast collection of computers.
- Since machine learning have the capacity to adapt to varying conditions, Gmail and Yahoo mail spam filters do more than just checking junk emails using pre-existing rules.
- They generate new rules themselves based on what they have learnt as they continue in their spam filtering operation.
- The machine learning model used by Google have now advanced to the point that it can detect and filter out spam and phishing emails with about 99.9 percent accuracy.
- The implication of this is that one out of a thousand messages succeed in evading their email spam filter.

1.2.2 Role of Machine Learning Algorithms in Natural Language Processing

- Machine learning algorithms used for natural language processing (NLP) to improve, accelerate and automate text analytics functions and NLP features that turn unstructured text into useable data and insights.
- A machine learning model is the sum of the learning that has been acquired from its training data. The model changes as more learning is acquired.
- Machine learning for NLP and text analytics involves a set of statistical techniques for identifying parts of speech, entities, sentiment, and other aspects of text.
- The techniques can be expressed as a model that is then applied to other text, also known as **supervised machine learning.**
- It also could be a set of algorithms that work across large sets of data to extract meaning, which is known as **unsupervised machine learning.**
- Text data requires a special approach to machine learning. This is because text data can have hundreds of thousands of dimensions (words and phrases) but tends to be very sparse.
- The most popular supervised NLP machine learning algorithms are :
  - Support Vector Machines
  - Bayesian Networks
  - Maximum Entropy
  - Conditional Random Field
  - Neural Networks/Deep Learning

1.3 Beyond Machine Learning-Deep Learning and Bio-Inspired Adaptive Systems

- Deep learning is a specialized form of machine learning.
- Deep learning uses many neural network layers for advanced feature recognition and prediction. So it is also called as deep neural network.
- Traditional neural networks only contain 2-3 hidden layers, while deep networks can have as many as 150.
- In deep learning, classification can be performed directly from a dataset of images, sound or text.
- It can achieve excellent accuracy as compared to human performance.
- Deep learning model needs large amount of labelled data and many layered neural network architecture.

1.3.1 Deep Learning

Deep learning is a smaller element, it is a subset of ML and more specific application of AI. Deep learning refers to a technique for creating an AI-powered layered neural network, much like a simplified replica of the human brain. It solves more complex problem than ML algorithm

- For example during a game of chess, a neural network is trained predominantly.
- Deep learning builds a more accurate model or reduce the time it takes to create it.
- Most deep learning methods use neural network architectures, so it is also referred to as deep neural networks.
- The term "deep" means the number of hidden layers in the neural network. Though the traditional neural networks only contain 2-3 hidden layers but deep networks can have as many as 150. It is trained by using huge sets of labelled data and neural network architectures to learn features directly without manual feature extraction.
- One of the most popular types of deep neural networks is known as convolutional neural networks (CNN or ConvNet)

Applications of Deep Learning

1. Deep learning is used in cancer detection to automatically detect cancer cells.
2. In driverless car to detect objects such as stop signs and traffic lights, but its development desires huge number of images and thousands of hours of videos.
3. In automated industry, it detects the unsafe distance between machines and human or objects.
4. It is used in speech translation like home appliances that respond to voice.

Drawbacks of Deep Learning

1. large computing power is require to get accurate result.
2. Difficulty in interpreting the resulting models.
3. It requires large amounts of labelled data.

1.3.2 Difference between Machine Learning and Deep Learning

Sr. No.	Machine Learning	Deep Learning
1.	A model is created by relevant features which are manually extracted from images to detect an object in the image.	Relevant features are automatically extracted from images. It is an end-to-end learning process.
2.	When more examples or training data is given, then it is consistent at a certain level of performance.	As the size of data increases, it continues to improve.
3.	Machine learning is less complex as compare to deep learning.	Deep learning is generally more complex.
4.	Without a high-performance GPU and lots of labelled data, machine learning techniques can be used based on application.	To get reliable results in less time one should have a high-performance GPU and lots of labelled data.

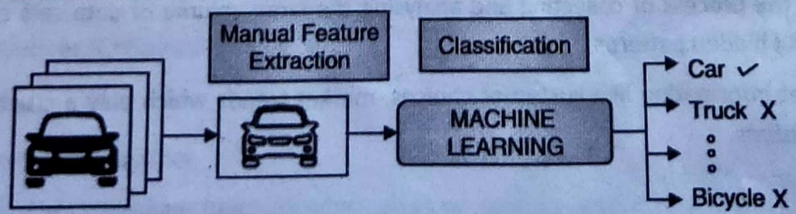
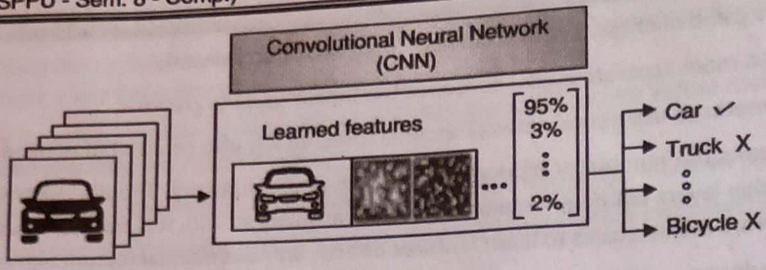


Fig. 1.3.1(a) : Machine Learning



(b) Deep Learning

Fig. 1.3.1 : Comparing a Machine Learning approach (above diagram) with Deep Learning (below diagram) for automated car driving

1.3.3 Bio-Inspired Adaptive Systems

- Bio-inspired Computing is the combination of computational intelligence and collective intelligence.
- It is used to solve complex problems.
- It is adaptive, reactive and distributed system.
- The goal of the system is to produce computational tools which is robust, scalable and flexible with effective human interface.

1.4 Machine Learning and Big Data

Machine Learning

- Machine Learning is a field of Artificial Intelligence, which educates computers on how to perform complex tasks.
- With Machine Learning software applications can learn to increase their accuracy of expected outputs.
- Machine learning applications in everyday life :
  - o Netflix or Amazon recommendation system uses machine learning.
  - o To determine the price and wait time for Uber/Ola.
  - o To determine fraudulent transaction for financial institutions.
  - o Self-driving car.

Other applications of machine learning

(a) Web Search	(e) Spam Filters
(b) Recommender Systems	(f) Ad placement
(c) Credit Scoring	(g) Stock Trading
(d) Computer Vision	(h) Drug Design

Big Data

- **Big Data Analytics** is the process of collecting and analysing the large volume of data sets called as Big Data, which helps to discover useful hidden patterns.
- It is also helpful to get information like customer choices, market trends which play a crucial role to take customer oriented business decisions.

- Big data describe the data in terms of :
  - Huge volume of data
  - Wide range of variations of Data types
  - High Velocity of Data Processing

Difference between Big Data and Machine Learning

Sr. No.	Big Data	Machine Learning
1.	Big data analytics look for emerging patterns by extracting existing information which helps in the decision making process.	It teaches the machine by learning from existing data.
2.	Problem : Dealing with large volumes of data.	Problem : Overfitting.
3.	It stores large volumes of data and finds out patterns from data.	It learns from trained data and predicts future results.
4.	It processes and transforms data to extract useful information.	Machine Learning uses data for predicting output.
5.	It deals with High-Performance Computing.	It is a part of Data Science.

1.5 Important Elements of Machine Learning

1.5.1 Data Formats

- In supervised learning, a dataset is required, which is a finite set of real vector having n features.
- Consider a data set X in which all samples are independent and identically distributed. This means all variables belong to the same distribution D, and considering an arbitrary subset of m values,

$$P\left(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_m\right) = \prod_{i=1}^{m} P(\bar{x}_i)$$

- If the corresponding output values are numerical continuous, then the process is called as Regression and if output is categorical then it is called as **Classification**.
- **Predictor :** A predictor is a function f that maps an input x to an output y. In statistics, y is known as a response, and when x is a real vector it is known as the **covariates**.
- Types of prediction tasks :

  Binary classification (e.g., email ⇒ spam/not spam) :

  $$x \rightarrow \boxed{f} \rightarrow y \in \{+1, -1\}$$

  Regression (e.g., location, year ⇒ housing price) :

  $$x \rightarrow \boxed{f} \rightarrow y \in R$$

  Where f can be a regressor or classifier

- **Generic regressor :** It is a vector-valued function, which gives continuous output.
- **Generic classifier :** It is a vector-values function, which predicts output as categorical or discrete.

1.5.2 Parametric Learning

- To simplify the learning process, assumptions or internal parameter vector can be considered, but this can limit the learning process. This approach is called parametric learning.
- **Parametric model :** It summarizes data with a set of parameters of fixed size, so it doesn't matter how much data is being considered for parametric model.
- Examples of parametric Machine Learning Algorithms :
  (a) Logistic Regression
  (b) Linear Discriminant Analysis
  (c) Perceptron
  (d) Naive Bayes
  (e) Simple Neural Networks
- Benefits of Parametric Machine Learning Algorithms :
  1. **Simpler :** Result analysis or interpretation is easy to understand.
  2. **Speed :** Learning using data is rapid.
  3. **Less Data :** Lesser amount of data is sufficient for training .
- Limitations of Parametric Machine Learning Algorithms :
  1. **Constrained :** By choosing a functional form, these methods are highly constrained to the specified form.
  2. **Limited Complexity :** Only suitable for simple problems.
  3. **Poor Fit :** Not likely to match the primary mapping function.

1.5.3 Non-parametric Learning

- These Types of Algorithms do not Make Strong Assumptions for Mapping Functions.
- As there are no assumptions, any functional form can be learnt from the training data. This type of Non-Parametric learning is applicable when there is a lot of data and no prior knowledge is available.
- A very common Non-Parametric family is called **Instance-Based Learning** and makes real-time predictions (without pre-computing parameter values) based on hypothesis determined only by the training samples (instance set).
- Examples of Non-Parametric Machine Learning Algorithms :
  (a) K-Nearest Neighbors
  (b) Decision Trees like CART and C4.5
  (c) Support Vector Machines
- Benefits of Non-Parametric Machine Learning Algorithms :
  1. **Flexibility :** It is flexible to fit for a number of functional forms.
  2. **Power :** Assumptions about the primary function are not required.
  3. **Performance :** Models for prediction show high performance.
- Limitations of Non-Parametric Machine Learning Algorithms :
  1. **More data :** Large training data are required to evaluate the mapping function.
  2. **Slower :** Training is slow as more parameters are required to train.
  3. **Overfitting :** Difficult to understand why specific predictions are made as there is a risk to overfit the training data.

1.5.4 Multiclass Strategies

- In classification problem, when output classes is greater than one, there are two possibilities:

  1. One-vs-all 2. One-vs-one

- In both cases, the output will be the final value or class. e.g., classify a set of images of vegetables which may be onion, potato or tomato. Multiclass classification makes the assumption that each sample is either an onion or a potato, but not both at the same time.

1. **One-vs-all**

   - Most of algorithms of scikit-learn adopts this strategy.
   - n classifiers will be trained for n output classes to distinguish the actual class and remaining one.
   - Complexity is O(n) and at the most n–1 checks are required to get the correct class.
   - Interpretability is the advantage of this approach.
   - Since one and only one classifier represents each class, it is possible to gain knowledge about the class by inspecting its corresponding classifier.

2. **One-vs-one**

   - Train a model for each pair of classes.
   - The complexity is $O(n^2)$.
   - Majority voting is used to find the right class.
   - If two classes have equal number of votes, it selects the class with the highest aggregate classification confidence by summing over the pair-wise classification confidence levels computed by the underlying binary classifiers.
   - More expensive and should be adopted only when a full dataset comparison is not preferable.
   - This method is usually slower than one-vs-all.

1.6 Learnability

A parametric model can be divided into two parts :

1. **Static structure of parameters**

   It is determined by the choice of specific algorithm and mostly immutable unless model provides some re-modelling functionalities.

2. **Dynamic set of parameters**

   It is the objective of optimization, which considers n unbounded parameters, which generates n-dimensional space.

   The aim of a parametric learning process is :

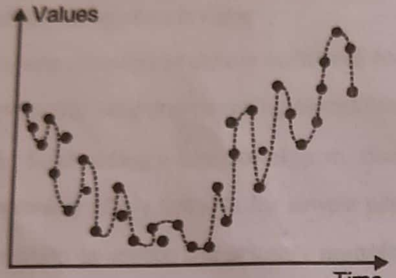
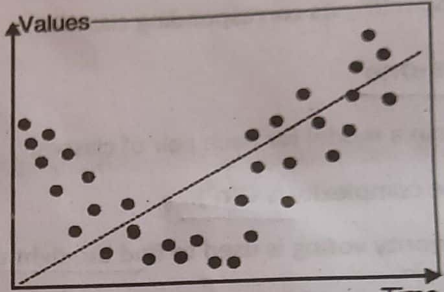
   (i) To find the best hypothesis

   (ii) Minimize the corresponding prediction error

   (iii) Avoid overfitting.

1.6.1 Underfitting and Overfitting

The machine learning model uses training data from the problem domain to generalize well, which helps to predict future unseen data. Poor performance of machine learning algorithms is only because overfitting and underfitting of the model. In statistics, a fit refers to how well a target function is approximated.

Sr. No.	Overfitting	Underfitting
1.	The training data are modelled very well	The training data is not modelled and even no generalization of new data
2.	In this case, learning occurs from detailed data as well as from noise; hence it gives a negative impact on the performance of the model while classifying unseen data.	This is not suitable a model as it gives poor performance on the training data.
3.	Overfitting can be avoided by using linear algorithm for linear data or using parameters like maximal depth for decision trees.	Underfitting can be avoided by using more data and also reducing the features by feature selection.
4.		

1.6.2 Error Measures

1. A non-negative error measure

– A non-negative error measure em is used to calculate total value of error over the whole dataset. Expected and predicted output is the input parameter to compute em .

$$\text{Error}_H = \sum_{i=1}^{n} e_m\left(\tilde{y}_i, y_i\right) \qquad \text{where } e_m \geq 0 \ \forall \ \tilde{y}_i, y_i$$

– Error_H is also implicitly dependent on the specific hypothesis H. To find the optimal hypothesis, optimization of error is also required. Hence, Mean Squared Error (MSE) is useful.

$$\text{Error}_H = \frac{1}{n}\sum_{i=1}^{n}\left(\tilde{y}_i, y_i\right)^2$$

2. Loss function

To find the global minimum the value must be minimized through an optimization problem. It avoids a large number of iterations and risk of overfitting. This measure is known as **loss function**.

(i) **Zero-one-loss :** It is effective for binary classification. It is based on the probability of misclassification and can be adopted in loss function.

$$L_{0/1H}\left(\bar{y}_i, y_i\right) = \begin{cases} 0 & \text{if } \bar{y}_i = y_i \\ 1 & \text{if } \bar{y}_i \neq y_i \end{cases}$$

(ii) **Generic (and continuous) loss function:** It is used for generic and continuous values and expressed in terms of potential energy.

$$\text{Energy}_H = \frac{1}{2}\sum_{i=1}^{n} e_m\left(\bar{y}_i, y_i\right)^2$$

1.6.3 PAC Learning

- Any hypothesis that is consistent with a significantly large set of training examples is unlikely to be seriously wrong: it must be probably approximately correct (PAC).
- Any (efficient) algorithm that returns hypotheses that are PAC is called a **PAC-learning algorithm.**
- A concept is a map $c : X \rightarrow Y$, If $Y=\{0,1\}$ then a concept can trivially be identified with a subset of X (that subset on which the map takes the value 1).
- A concept class is a set of concepts.
- The goal of PAC learning to find an algorithm A which can recognize any concept approximately from given concept class or sample examples from concept.
- Therefore, learning a concept means to minimize the global loss function by minimizing the respective loss function of specific class belongs to the same universe.
- As defined by Michael J. Kearns (Author : An Introduction to Computational Learning Theory)

Definition : Let X be a set, and C be a concept class over X. We say that C is PAC-learnable if there is an algorithm $A(\varepsilon,\delta)$ with access to a query function for C and runtime $O\left(\text{poly}\frac{1}{\varepsilon}, \frac{1}{\delta}\right)$, such that for all $c \in C$, all distributions D over X, and all inputs ε, δ between 0 and 1/2, the probability that A produces a hypothesis h with error at most ε is at least $1-\delta$. In symbols, $P_D\,(P_{x-D}\,(h(x) \neq c(x)) \leq \varepsilon\,) \geq 1-\,\delta)$

Where the first P_D is the probability over samples drawn from D during the execution of the program to produce h. Equivalently, we can express this using the error function, $P_D\,(eTT_{c,D}\,(h) \leq \varepsilon)\;1-\delta$.

Review Questions

Q. 1 Define Machine Learning and its different approaches.

Q. 2 Differentiate between Deep Learning and Machine Learning.

Q. 3 Differentiate between Big Data and Machine Learning.

Q. 4 Explain the important elements of Machine Learning.

Q. 5 Explain the concept of overfitting and underfitting with respect to machine learning.