International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 1 Jan. 2023, pp: 54-60 www.ijaem.net ISSN: 2395-5252

A Comprehensive Overview of Machine Learning Algorithms and It's Real Time Applications

Bojja Vani
Department of CS & AI, SR University, Warangal, Telangana

Date of Submission: 01-01-2023 Date of Acceptance: 08-01-2023

ABSTRACT: Machine Learning is an interesting research area in computer science which changes the thoughts of human to machines. It is mainly used for data analysis that plays a vital role in many application areas where more data is generated and need to turn that data into useful information. Machine Learning (ML) is multidisciplinary field, a combination of statistics and computer science algorithms which is widely used in predictive analyses and classification. However, in recent years, as a result of various technological advancements and research efforts, new data has become available, resulting in new domains in which machine learning can be applied. This paper mainly emphases on explaining the perception and evolution of Machine Learning, some of the prevalent Machine Learning algorithms and try to associate three most popular algorithms based on some basic notions. We also discuss the application of machine learning in different sectors.

KEYWORDS: Machine Learning, Data, Statistics

I. INTRODUCTION

Machine Learning (ML) is the branch of AI wherein it builds a model based on the data which is known as training data and the model is checked based on testing data or validation data and the accuracy of the model is improved by tuning its parameters and the process. Machine Learning Technique is trained using a training data set to create a model. When new input data is introduced to the ML technique, it makes a prediction on the basis of the model. The prediction is evaluated for accuracy and if the accuracy is acceptable, the Machine Learning technique is deployed. If the accuracy is not acceptable, the Machine Learning technique is trained again and again with an augmented training data set.[1]

Learning by doing (ML) is automating the system's learning based on previous experimental data. The goal of machine learning is to make predictions based on the data that already exists. Machine learning is a critical component of artificial intelligence. Nearly all fields of science and technology use it. We leave a digital trail of almost everything we do today, which define human activities, identifies our locality, as well provides a wealth of other message about what we say, what we buy, and so forth. The majority of devices, machines, and everything we use provide information, thanks to data storage capacity and the digitalization of the general public. Examples of data sources include pay stations and parking lots; intelligent phones; social networks; videos; photos; etcetera. All of the collected data must be used to its fullest potential and given a purpose.

The development of ML as a division of AI is accelerating at the moment. As a result, it is now used in diverse domains like learning machines, which are used in intelligent manufacturing, medical science, pharmacology, agriculture and archaeology, as well as games, business, and so on. Large amounts of data of various sizes, types and speeds have recently been the focus of researchers. Researchers must put an awful lot of attention to organize this multitude of information from different sources, such as publications, research and news pages [2] [3]. ML, is a subfield of AI which has grown in popularity in latest years in the aspects of analyzing data as well as computing, letting applications to handle data intelligently [4]. ML typically enables system to learn and progress on their own without being explicitly programmed and is frequently referred to as the most well-known new technology in the 4th industrial revolution [5] [6]

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 1 Jan. 2023, pp: 54-60 www.ijaem.net ISSN: 2395-5252

II. CLASSIFICATION OF MACHINE LEARNING

The machine Learning Algorithms can be classified into four categories; they are Unsupervised Learning, Supervised Learning, Reinforcement Learning and SemiSupervised Learning.

2.1. Supervised Learning

Supervised Learning is the most popular paradigm for performing machine learning operations. It is widely used for data where there is a precise mapping between input-output data. The dataset, in this case, is labeled, meaning that the algorithm identifies the features explicitly and carries out predictions or classification accordingly. [7] As the training period progresses, the algorithm is able to identify the relationships between the two variables such that we can predict a new outcome. Resulting Supervised learning algorithms are task-oriented. As we provide it with more and more examples, it is able to learn more properly so that it can undertake the task and yield us the output more accurately. Some of the algorithms that come under supervised learning are as follows: Linear regression, random forest, support vector machine, artificial intelligence [8], etc.

There are two main types of supervised learning problems: they are classification that involves predicting a class label and regression that involves predicting a numerical value.[8]

- Classification: Supervised learning problem that involves predicting a class label.
- Regression: Supervised learning problem that involves predicting a numerical label.

Both classification and regression problems may have one and more input variables and input variables may be any data type, such as numerical or categorical.

2.2 Unsupervised Learning

Unsupervised machine learning holds the advantage of being able to work with unlabeled data. This means that human labor is not required to make the dataset machine-readable, allowing much larger datasets to be worked on by the program. The model learns through observation and finds structures in the data. Once the model is given a dataset, it automatically finds patterns and relationships in the dataset by creating clusters in it.[9]

In supervised learning, the labels allow the algorithm to find the exact nature of the relationship between any two data points. However, unsupervised learning does not have labels to work off of, resulting in the creation of hidden structures. Relationships between data points are perceived by the algorithm in an abstract manner, with no input required from human beings. The creation of these hidden structures is what makes unsupervised learning algorithms versatile. Instead of a defined and set problem statement, unsupervised learning algorithms can adapt to the data by dynamically changing hidden structures.[9] This offers more post-deployment development than supervised learning algorithms. What it cannot do is add labels to the cluster, like it cannot say this a group of apples or mangoes, but it will separate all the apples from mangoes. Suppose we presented images of apples, bananas and mangoes to the model, so what it does, based on some patterns and relationships it creates clusters and divides the dataset into those clusters. Now if a new data is fed to the model, it adds it to one of the created clusters. The example of unsupervised learning is k-mean clustering, principle component analysis, SVD, FP-growth etc.

There are many types of unsupervised learning, although there are two main problems that are often encountered by a practitioner: they are clustering that involves finding groups in the data and density estimation that involves summarizing the distribution of data.[9]

- Clustering: Unsupervised learning problem that involves finding groups in data
- Density Estimation: Unsupervised learning problem that involves summarizing the distribution of data.

2.3. Reinforcement Learning

Reinforcement learning directly takes inspiration from how human beings learn from data in their lives. It features an algorithm that improves upon itself and learns from new situations using a trial-and-error method. Favorable outputs are encouraged or 'reinforced', and non-favorable outputs are discouraged or 'punished'. Based on the psychological concept of conditioning, reinforcement learning works by putting the algorithm in a work environment with an interpreter and a reward system. In every iteration of the algorithm, the output result is given to the interpreter, which decides whether the outcome is favorable or not.[10]

In case of the program finding the correct solution, the interpreter reinforces the solution by providing a reward to the algorithm. If the outcome is not favorable, the algorithm is forced to reiterate until it finds a better result. In most cases, the reward system is directly tied to the effectiveness of the result.[10]

In typical reinforcement learning use-cases, such as finding the shortest route between two points on a map, the solution is not an absolute

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 1 Jan. 2023, pp: 54-60 www.ijaem.net ISSN: 2395-5252

value. Instead, it takes on a score of effectiveness, expressed in a percentage value. The higher this percentage value is, the more reward is given to the algorithm. Thus, the program is trained to give the best possible solution for the best possible reward.[10] This simple feedback reward is known as a reinforcement signal.

2.4 Semi-Supervised Learning

In this type of learning, the given data are a mixture of classified and unclassified data. This combination of labeled and unlabeled data is used to generate an appropriate model for the classification of data. In most of the situations, labeled data is scarce and unlabeled data is in abundance (as discussed previously in unsupervised learning description).[11] The target of semi-supervised classification is to learn a model that will predict classes of future test data better than that from the model generated by using the labeled data alone. The way we learn is similar to the process of semi-supervised learning.

III. MACHINE LEARNING ALGORITHMS

Depending upon resemblance in rapports of their purpose how the algorithm will work algorithms are of different types. For example, tree-based methods, and neural network inspired methods. One of the most valuable method to cluster procedures and it is the technique that we are going to use here. This is a suitable grouping technique, but these are not perfect. There are also some algorithms that can fair as effortlessly fit into many categories like Learning Vector Quantization(LVQ) that is both a neural network inspired (NN) method and an instance-based method. There are also some classes that have the same name that define the problem and the class of algorithm such as Regression and Clustering.

3.1 Decision Tree Based Classification

Decision tree algorithm is a type of classification that is primarily used to build a model in the way of a structure that resembles a tree like having (root, branch and leaf), that is based on (inferred from) earlier information to classify/predict class or target variables of future(new data) that we can get with the help of decision rules or decision trees. The main usage of Decision Trees is in the numerical as well as categorical data. The algorithm works on greedy search approach that is it will start from top to bottom.

Advantages:

- It Easy to implement
- It Can classify and predict categorical as well as numerical data
- Is Less data pre-processing
- Statistical test can be done to validate the tree model.
- Resembles human decision-making technology.
- Tree structure is easily understandable through visualization

Limitations:

- Low Prediction Accuracy
- Complexity in calculations if class labels are huge
- Need of redrawing for every addition of data to the data set.
- Probability of over-fitting in the decision tree is high.

Applications:

- Agriculture
- Medicine
- Financial analysis

3.2. Support Vector Machines

The key objective of this is to discover a hyper plane that is used to divide the classes into two types. Depending upon the values obtained as hyper values the obtained data set is placed into the data set to which the similarities resemble. To draw a hyper plane we must take two rules into consideration. First is that the hyper plane needs to been chosen in such a way that it should separate the two classes and best hyper separator plane should be chosen as maximum-margin hyper.
SVM can be classified into two different types: a) Linear SVM b) Non-Linear SVM

Advantages:

- Robust Classifier for prediction problem
- Efficient if n(D) > n(S) where n(D) the number of dimensions is greater than the n(S) number of samples
- Suitable for high dimensional data spaces
- Memory efficient.

Disadvantages

- Very slow in test phase
- Not suitable for large and noisy data sets
- Classification error percentage will be increased if wrong kernel is selected
- Memory consumption is high

Applications

- Image classification
- Bioinformatics
- Face detection

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 1 Jan. 2023, pp: 54-60 www.ijaem.net ISSN: 2395-5252

3.3 K-Nearest Neighbors

This type of algorithm is used to place the data set into the predefined set by measuring the distance with the predefined data set by taking into consideration k values. The distance that it finds least puts the data set into that data set . the methods that is used to calculate the distance between two points uses k variable values among 0 an d1 normally. Most commonly used algorithm used in this are Euclidean distance , hamming distance. This method is used in such problems where we have to do classification of data set. For continuous variables we are using Euclidean distance formula and for categorical variables we are using hamming distance

Advantages
- Supple with traits and distance functions
- Supports multi class data

Disadvantages
- Finding appropriate K value is difficult
- Needs large sample for high accuracy
- Requires large storage space

3.4. Naive Bayes Algorithm

This algorithm is used in the classification of data set if the data set is in large quantity and has much more records as the multiclass and binary class related classification problem. This can be used in the classification job in the field related to machine learning. The main objective of this is to analyze the text and natural language processing. For naive bayes algorithm one must have the concept related to bayes theorem (based on the conditional probability) Conditional probability can be defined as an event will happen with conditioned (based) on an event already occurred. This will help us to merge the different algorithms to form a naïve bayes by using a common principle.

Types of Naive Bayes:
- Gaussian Naive Bayes
- Multinomial Naive Bayes
- Bernoulli Naive Bayes

Advantages
- Fast
- Scalable
- Efficient for multinomial distributed and also for binary attribute values

Disadvantages:
- Not suitable for regression problems
- Cannot find relationships among attributes

Applications:
- Text, e-mail, and symbol analysis
- Recommendation Systems

3.5. Linear Regression

This type of algorithm is used to find the link between an independent variable also known as (predictor (X)) and a dependent variable (criterion (Y)) variable that can be implemented to predict the future values of the dependent variable. In Simple regression one independent variable is used and in case of multiple regressions we can use two or more independent variables depending upon the data set to predict the future. Dependent variable are those that have a continuous and independent variable are those variables that have discrete or dis-continuous values. regression models are of two kinds. One is linear and other one is non-linear. The linear regression model is one that uses straight line and non-linear regression model is a type of model that uses curved line relationships among dependent and independent variables.

Advantages:
- Shows relationship between dependent and independent variables
- Simple and easy to understand

Disadvantages:
- Not applicable for non-linear data
- Only predicts numerical output
- Data must be independent

Applications:
- Observational Astronomy
- Finance

IV. REAL TIME APPLICATIONS

4.1. Artificial Intelligence Applications

4.1.1. Marketing

Movies Recommendation system, for example: Netflix a popular online application to watch films uses a high predictive technology to recommend movies based on the past reviews, it parses across million records to provide an accurate prediction.

4.1.2. Banking

Banks provides a user support tool which will be available throughout the day to answer queries from the user. Example: HDFC bank has launched an Electronic Virtual Assistant (EVA) which address user queries.

4.1.3. Agriculture

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 1 Jan. 2023, pp: 54-60 www.ijaem.net ISSN: 2395-5252

- See and Spray Robot: This robot has been launched by Blue River Technology which sprays herbicide accurately on plants.
- Plantix: This application developed by PEAT, detects the fault plant which has less nutrient deficiency.

4.1.4. Health Care

A decision support system will assist doctors for patient monitoring. An organization named Cambio Health Care developed a stroke prevention decision support system which provides doctor with an alert when the patient is at risk.

4.1.5. Space

The rover sent to the MARS by NASA is AI based which has to perform independent targeting of cameras to obtain information of MARS.

4.1.5. Automobile

Self – Driving Cars is the automated car which uses AI system that gathers data and produce signals which aids the operation of the vehicle. For Example: Tesla's Car which automatically detect objects without human involvement.

4.1.6. Chat bots

Chat bots are virtual Assistant and it is available on mobile phones, laptop. The popular chat bots are Siri, Cortona and Amazon Echo. The technology used in these applications are Speech Recognition and Natural Language Processing (NLP) which has to translate the human language into set of instructions to be performed like playing music, a making a call etc.

4.2. Unsupervised Machine Learning Algorithms Applications

4.2.1. Medical Field – Neurological Disorder

The Clustering based algorithm is applied to Neurological dataset mainly Alzheimer Disorder. The different Clustering methods applied are K means, Hierarchical Agglomerative Clustering, Multi-Layer clustering. [12]

4.2.2. Networking and Security

Intrusion in a network is a serious issue. The Clustering based and Expectation Maximization Algorithm were used to detect the Network Intrusion detection system.[13]

4.3. Supervised Machine Learning Applications

4.3.1. Covid 19 Prediction

The standard Models created for the Covid 19 prediction were Multi-Layer Perceptron and Adaptive Network based Fuzzy Inference System. [14].

4.3.2. .Optical Networks and Systems

Autoregressive Integrated Moving Average (ARIMA) is a supervised learning method used for Traffic Prediction in optical Networks.[15]

4.4.Reinforcement Learning Applications

4.4.1. Human Level Video Game Play

Multi-Layer Artificial Neural Network and Deep Q Network (DQN) has been used for functioning the Video Game Play.[16]

4.4.2. Watson's Daily-Double Wagering

This is the system which uses Back propagation Artificial Neural Network developed by the team of developers of IBM.[17]

4.5. Semi-Supervised Learning Applications

4.5. 1.Customer Behavior Modeling

The Multilayer Perceptron with Back Propagation Algorithm is used to predict the Customer Behavior. The training data of the Algorithm consists of set of Labeled data and set of unlabeled data.[18]

4.5. 2. Video Recommendation System

A graph- based Algorithm Known as the Adsorption Algorithm has been used for Video Recommendation.[19]

V. JEOPARDIZES IN MACHINE LEARNING

5.1. Data Poisoning/Destroying

For the security of any ML system data plays a vital role. The reason for this is that the ML system learns to do what it does right from data that we have given to it. If a mugger can deliberately handle the data that we have used in ML system in a synchronized fashion, the whole ML system can be conceded. A special attention is needed for Data poisoning attacks. ML engineers must take those measures into consider that in what fraction of the training data an attacker can control and to what extent.

5.2. Data Privacy/Secrecy

In ML Data protection is problematic ample without flinging into the combination. One of the most sole challenge in ML is the caring subtle or private data that we have given to the model, through training, are built right into a model. Understated but actual extraction attacks against an ML system's data are an important category of risk.

5.3. Online System Manipulation

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 1 Jan. 2023, pp: 54-60 www.ijaem.net ISSN: 2395-5252

When a ML system continues to learn during operational use, modifying its behaviour over time this is said to be online system. In this situation, a cunning attacker can use system input to deliberately nudge the still-learning system in the wrong direction, gradually retraining the ML system to perform the wrong thing. It's worth noting that such an attack might be both subtle and simple to carry out. To adequately handle this risk, ML engineers must take into account data provenance, algorithm selection, and system operations.

5.4. Over-fitting

When a model learns the information and noise in the training data to the point where it degrades the model's performance on fresh data, this is known as over fitting. This means that the model picks up on noise or random fluctuations in the training data and learns them as ideas. The difficulty is that these notions do not apply to fresh data, causing the model's ability to generalise to be harmed. Nonparametric and nonlinear models, which have more flexibility when learning a target function, are more prone to over fitting. As a result, many nonparametric machine learning algorithms incorporate parameters or approaches that limit and constrain the amount of detail learned by the model.

VI. BENEFITS OF MACHINE LEARNING

- Autonomous Vehicles Machine learning algorithms are an essential component of self-driving automobiles, and they will play an increasingly crucial role in their capacity to operate. These learning systems are widely employed for tasks such as image identification and scheduling, but they are challenging to train in chaotic real–world contexts. It's difficult to collect, analyze, and combine a lot of different sorts of data. So far, the solution has been to do extensive road testing in order to catch as many of these unusual instances as possible. Algorithms are employed to create replies based on this data, which are subsequently tested in simulations.
- Medicine, Healthcare When it comes to diagnosis or decision making, machine learning algorithms are not a good replacement for clinicians - at least in most situations. A good diagnosis must take into account structured (e.g. diagnosis codes, medications), unstructured data (clinical notes), image data (X-rays) and even subtle visual cues from the patient (do they look ill, how did they answer family history questions) in a very short time frame.
- Governmental machine learning Machine learning is now being used by a small number of government agencies, such as the Government Digital Services (GDS), which is using it to anticipate page views in order to find nomalies11, and the HMRC, which is using clustering algorithms to segment VAT clients. The adoption rate is still low, and there is a lot of untapped potential. GDS has thus far concentrated on proving machine learning algorithms' capabilities on a variety of goods and prototype services. The development of a 'data first' mindset at a much earlier level in the policy process is one of the first steps toward increased exploitation of these learning systems.

VII. CONCLUSION

Machine learning is widely used technology nowadays which is very helpful for the statistical analysis of data. It also gains the information from the historical data. In this paper we present the comparative analysis of mostly used algorithm such as decision tree, support vector machine, and naïve bayes classifer and it is found that naïve bayes gives more accurate results than the other two. Various applications of machine learning have been used, and this paper depicts that an overview is also provided, assisting readers interested in this domain to use it as a reference and tools effectively for their respective research work in the future.

REFERENCES

[1] https://www.edureka.co/blog/what-is-machine-learning
[2] Yang, Seungwon, TarekKanan, and Edward Fox: "Digital library educational module development strategies and sustainable enhancement by the community." In International Conference on Theory and Practice of Digital Libraries, pp. 514-517. Springer, Berlin, Heidelberg, 2010.
[3] Vandana Korde Sarda rVallabhbhai, Namrata Mahender: “Text Classification and Classifiers: A survey”, National Institute of Technology, Department of Computer Science &IT, International Journal of Artificial Intelligence & Applications (IJAIA), Vol.3, No.2, March 2012
[4] Iqbal H Sarker. Deep cyber security: A comprehensive overview from neural network and deep learning perspective. SN Computer Science, 2021.

[5] Iqbal H Sarker, Mohammed Moshiul Hoque, MdKafil Uddin, and Tawfeeq Alsanoosy. Mobile data science and intelligent apps: Concepts, AI-based modeling and research directions. Mobile Networks and Applications, pages 1–19, 2020
[6] Iqbal H Sarker, ASM Kayes, Shahriar Badsha, Hamed Alqahtani, Paul Watters, and Alex Ng. Cyber security data science: an overview from machine learning perspective. Journal of Big Data, 7(1):1–29, 2020.
[7] Christopher M. Bishop, "Pattern Recognition and Machine Learning (Information Science and Statistics)", 2006. Page-3
[8] Russel, "Artificial Intelligence: A Modern Approach", January 1, 2015
[9] Hastie et al. "The Elements of Statistical Learning: Data Mining, Inference, and Prediction", Second Edition (Springer Series in Statistics) , 2016, pp. 28. [online]. Available:https://www.amazon.com/Elements-StatisticalLearning-Prediction-Statistics.
[10] Richard S. Sutton et al., "Reinforcement Learning: An Introduction (Adaptive Computation and Machine Learning",2018, second edition, pp.-2. [Online]. Available : https://www.amazon.com/Reinforcement-LearningIntroduction-Adaptive-Computation
[11] Mohammed et al., "Machine Learning Algorithms and Applications", 2017, CRC Press Taylor & Francis, London, ISBN-13:978-1-4987-0538-7
[12] Alashwal Hany et al, "The Application of Unsupervised Clustering Methods to Alzheimer's Disease", Frontiers in Computational Neuroscience, 2019.
[13] Uttam Kumar Dey et al, "Network Intrusion Detection with Unlabeled Data using Unsupervised Clustering Approach." International Journal of Engineering Science and Computing, February 2019.
[14] Ardabili, Sina et al "COVID-19 Outbreak Prediction with Machine Learning.", Research Gate ,2020
[15] Francesco Musumeci et al, "An Overview on Application of Machine Learning Techniques in Optical Networks" arXiv,2018.
[16] A. G. Barto et al, "Some Recent Applications of Reinforcement Learning."
[17] E. Chandra Blessie et al, "Exploring the Machine Learning Algorithm for Prediction the Loan Sanctioning Process, International Journal of Innovative Technology and Exploring Engineering (IJITEE), ISSN: 2278-3075, Volume-9 Issue-1, November, 2019
[18] Siavash Emtiyaz et al, "Customers Behavior Modeling by Semi-Supervised Learning in Customer Relationship Management." Advances in information Sciences and Service Sciences (AISS) Volume3, Number9, October 2011.
[19] Baluja, S. et al, "Video suggestion and discovery for youtube: Taking random walks through the view graph." In Proceedings of the 17th international conference on world wide web (pp. 895–904). ACM