International Journal of Soft Computing and Engineering (IJSCE)
ISSN: 2231-2307 (Online), Volume-12 Issue-5, November 2022

A Review on Machine Learning & It's Algorithms

Nipun Jain, Rajeev Kumar

Abstract*: Machine learning is important because it gives us accurate predictions based on data. It can teach computers to perform complex tasks without any human intervention. Machine learning can analyze complex blocks of data. Machine learning enables entrepreneurs and businesses to quickly recognize potential business opportunities and risks. Businesses that rely solely on large amounts of data are using machine learning as the best way to analyze data and build models. Machine learning is not only considered as the backbone of artificial intelligence, but machine learning also plays a significant role in the development and advancement of artificial intelligence. Using algorithms to solve classification problems with different sets of parameters yields dramatically different classification accuracies. The machine learning challenge of finding the most appropriate parameter values for algorithms that best solve technical problems related to performance metrics. In this paper, the author discussed various types of machine learning such as supervised, unsupervised and reinforcement machine learning. The main emphasis is on supervised machine learning such as classification and regression using various machine learning algorithms such as Decision Tree, Naïve Bayes, K-Nearest Neighbor, Random Forest and SVM Classifier. The author explains all classification-based algorithms well with examples and diagrams. The authors also mention applications or domain areas where these classification algorithms can be used.

***Keywords*: Supervised Learning, K-Nearest Neighbor, SVM, Random Forest, Decision Tree, Naïve Bayes Classifier.**

I. INTRODUCTION

Machine Learning is a sub-field of Artificial Intelligence. Machine learning algorithms help computers to make decisions without needing explicit coding. These algorithms fed historical data and then make predictions about future events. This makes them faster and more powerful than other methods that require hand-coding rules. For example, the recommendation system is a common use case. Other common uses include fraud detection, spam, malware threats, business processes, and predictive maintenance. Machine learning helps develop algorithms or programs useful for understanding, managing, solving, and analyzing big and complex data problems as well as problems that cannot be solved manually.

Machine learning is gaining elevation due to the increasing volume and variety of data, the accessibility and affordability of computing power, and the availability of high-speed Internet. These digital transformation tools enable rapid and automated development of models capable of quickly and accurately analyzing very large and complex datasets. There are many applications of machine learning such as speech recognition, natural language processing, recommendation systems, data analytics, virtual assistants, audio and video monitoring, credit fraud detection, analysis stock market analysis, etc. [1].

II. CATEGORIES OF MACHINE LEARNING

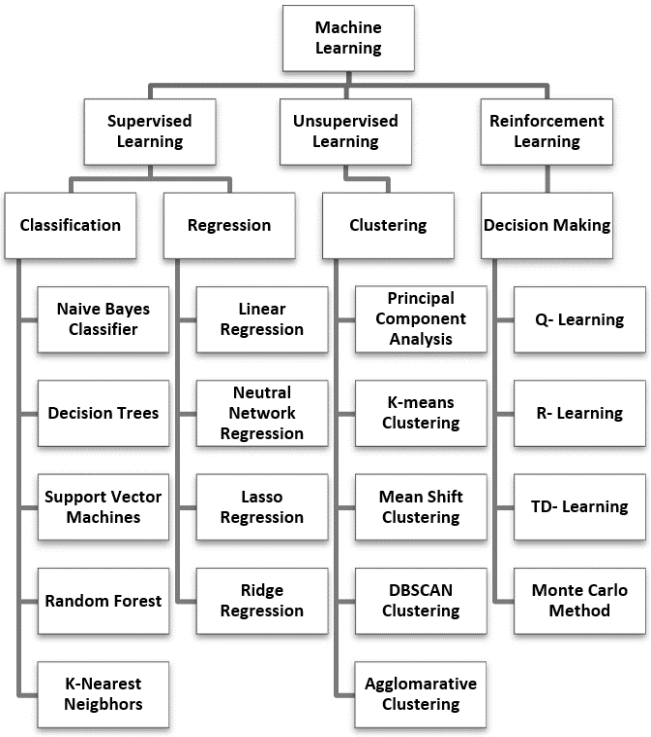

Fig. 1. Categories of Machine Learning

Machine Learning is generally classified into three categories such as supervised machine learning, unsupervised machine learning and reinforcement learning as shown in the Fig. 1. The select of machine learning type depends on the learning process to be adopted.

Let's discuss each type of machine learning:

A. Supervised Machine Learning

Supervised machine learning is the most popular and widely used machine learning technique.

In supervised learning, the machine learns under some guidance or supervision. All types of activities must be performed or carried out under supervision.

We provide the input dataset and known labels according to the data and the dataset can include any formats like image data, text data, etc. We then build a model by separating the dataset into training and testing datasets. There is an output variable in the train dataset that needs to be predicted or stored.

Manuscript received on 10 October 2022 | Revised Manuscript received on 14 October 2022 | Manuscript Accepted on 15 November 2022 | Manuscript published on 30 November 2022.

* Correspondence Author (s)

Nipun Jain*, Department of Electrical Engineering, IIT Roorkee, Roorkee, India. Email: nipunjain1305@gmail.com

Rajeev Kumar, Department of Electrical Engineering, IIT Roorkee, Roorkee, India. Email: rajeev.kumar@ee.iitr.ac.in, rajeevkumar.rke@gmail.com

All algorithms learn some kind of trend from the training dataset and apply it to the test dataset for prediction or classification [2]. With the explanation of supervised learning, let's discuss the categories of supervised learning. Supervised learning is categorized into two types: Classification & Regression.

1) Classification:

Classification is a technique or process of defining a given dataset into different classes. In other terms we can also say that classification is a technique which learns from a given dataset on how to assign or categorize a class. These classes are also known as labels, targets, etc. Classification learning can be applied to both linear & non-linear data. In classification, the training classes can be in the form of Yes or No, 0 or 1. Classification is applied when the output class has finite and discrete values. Let's understand classification learning with an example i.e. tuberculosis problem can be considered as classification problem. As in this problem, the tuberculosis dataset can only be categorized into two classes i.e. one class can show that it has tuberculosis and the other class can show that it does not have tuberculosis. When given input data or patient data are mapped with related classes and once the classifier is trained accurately, then it can detect the patient with tuberculosis or not. We can categorize classification learning into following types:

a) Binary Classification

As the name suggests, Bi means two, so binary classifier classifies data into two classes. Classes can have the form Yes/No or True/False or 1/0[3].This means we can say that one class represents the normal state and another class represents the abnormal state.

The example discussed above, i.e. the tuberculosis dataset can only be classified into two classes, i.e. one class shows patients having tuberculosis and the other class shows patient not having tuberculosis. Other examples might be spam detection in emails, unsubscribe prediction, etc. Let us understand this classification with a graphic form. In Fig.2., the red diagonal is also known as the classifier and the points above and below the line are two classes, i.e. class A and class B. In each class, the distance between the most recent point and the red line corresponding to the classification error. This dataset represents a binary classifier with only two classes.

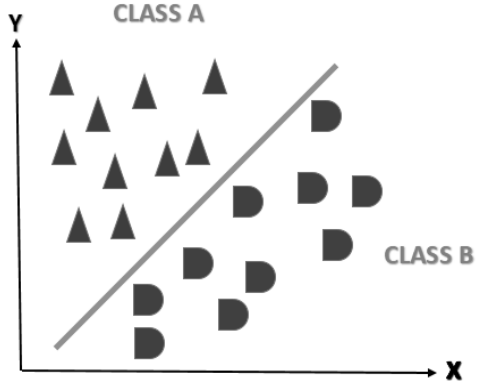

Fig. 2. Binary Classification

b) Multiclass Classification

Multi-class classification classifies data into more than two classes. These classes are not in Yes/No or True/False or 1/0 form. Multilayer data can be classified into special categories like sports, entertainment and politics, sentiment analysis, etc. and so on Let's understand multi-class classification with graph form.

In Fig.3., the (red, green and blue) diagonal lines are also called as classifier and the points above and below the lines are three classes i.e. Class A, Class B and Class C. This dataset represents for this classification.

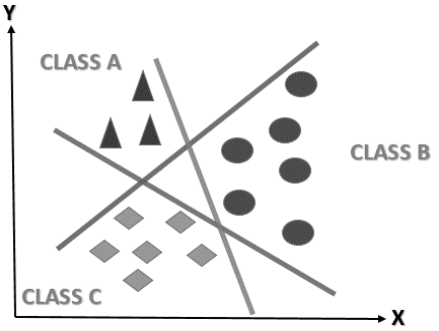

Fig. 3. Multiclass Classification

c) Multilabel Classification

Multi-label classification is a general form of multi-class classification. In this classification, each class can be classified into subclasses or we can say each input variable can be mapped to more than two or more instances or subclasses of a single class. There is no constraint on the number of classes that can be affected in a multi-label problem, the more classes there are, the more complex the problem becomes.

In Fig.4., the (red, green and blue) diagonal lines are also called as classifier and the points above and below the lines are four classes i.e. Class A, B, C, D and E. Here the dependent variable to more than two categories and the observations can be mapped with more than one category i.e. Class C plus Class B. This represents a multi label classification.

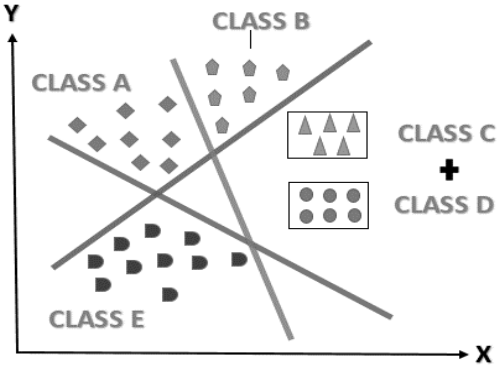

Fig. 4. Multilabel Classification

2) Regression

Regression is a technique that aims to model the relationship between several characteristics and a continuous target variable.

Regression is a machine learning technique in which a model predicts outputs that are continuous numbers.

Domain areas where regression is generally used are finance, investment, forecasting, time series modeling and finding the cause and effect relationship between variables, the relationship between rash driving and the number of road accidents can be better analyzed using regression, etc.

Retrieval Number: 100.1/ijsce.E35831112522
DOI: 10.35940/ijsce.E3583.1112522
Journal Website: www.ijsce.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

B. Unsupervised Machine Learning

Unlike supervised learning, this type of learning has no guidance or supervision. It allows the model to work on its own to uncover previously undetected trends and information.
These are primarily unlabeled data.
For example, suppose there are three trees in an image and the machine has no idea about either tree. Meaning the machine has no hints or teachers to guide it on the data tree. But the machine was able to classify the characteristics of the three trees based on their structure, pattern, colour, shape, size and other similarities.
Unsupervised has many real-world applications like: referral systems, data mining, prepared and direct data visualization, customer segmentation, referral systems and targeted marketing campaigns, etc.

C. Reinforcement Machine Learning

Reinforcement is a technique in which an expert decides to act in an environment by performing operations and visualizing the results of the operations. For each excellent performance, the expert gets either a positive or a good opinion, and for each poor performance, the expert gets either a negative or a bad opinion. It does not need labeled data, rather it automatically learns from feedback.
Thus, this technique is also called as feedback technique. Some real-life examples of Reinforcement machine learning are: Playing games, used in health sector for the treatment, Face Recognition, Text Detection, Speech-to-text conversion, etc.

III. TYPES OF MACHINE LEARNING ALGORITHMS

A. Naive Bayes Classifier

Naïve Bayes classifier contains a collection of many algorithms. Naïve Bayes is a widely popular and easy to build classification algorithm. It is generally used for large types of datasets. This classification algorithm follows a principle that says 'Each feature present in the classifier is unrelated to or independent of any other feature present in the classifier'. It works well with the textual and categorical data compared to the numerical data. It is based on probability logic, it calculates probabilities based on mathematical approach using Bayes theorem which says:

$$P(\text{Class} \mid \text{Features}) = \frac{P(\text{Features} \mid \text{Class}) \times P(\text{Class})}{P(\text{Features})}$$

Where, P (Class |Features) is the posterior probability of the class, given that already known features of any classifier,
P (Features | Class) is the likelihood of the prior probability,
P (Class) is the probability of the class,
P (Features) is the probability of the other features
There are three types of Naive Bayes model in the scikit-learn library: Gaussian, Optimal, Bernoulli and Multinomial. Domain areas where Naïve Bayes can be used are real-time prediction, multi-class prediction, text classification, recommended systems, sentiment analysis, spam filtering, etc. Let's discuss Naïve Bayes with one example, suppose we have drawn a black card from a deck of playing cards. What's the probability that it's a four? We apply conditional probability. There are 26 possible black cards and two of them are fours. Thus,
P (four | black) = P (four) × P (black | four)/ P (black)
Bayes Theorem allows us to reformulate the problem as follows:
P (four | black) = P (four) × P (black | four)/ P (black)
= (4/52 × 2/4) / (26/52)
= 1/13

B. Decision Tree

A decision tree classifier is a graphical representation for obtaining all possible outcomes or solutions to a problem with a given condition.

A decision tree classifier is a flowchart divided into two nodes, a decision node and a leaf node. The process of dividing a single node into two or more nodes is called splitting.

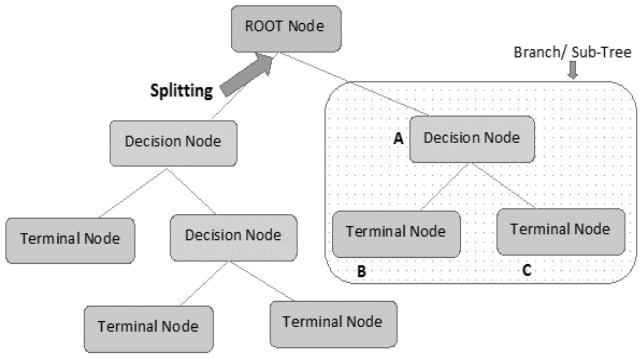

Fig. 5. Decision Tree Classifier

Where each inner node represents the characteristics of a dataset, branches represent the decision rules, and if no further nodes split, then that node is called a leaf node, each leaf node represents the outcome. Fig.5 depicts the structure of a decision tree classifier. It can be used by both classification & regression.
The areas where the Decision Tree Classifier can be used are: Commerce Management, Customer Relationship Management, Fraud Declaration Detection, Energy Consumption, Health Management, Error Diagnosis, etc.

C. Support Vector Machines (SVM) Classifier

Support Vector Machine is one of the most popular learning algorithm in machine learning. SVM can be used for both classification and regression. Primarily SVM is used for solving classifications problems. SVM is generally used to deal with small and complex type of dataset. SVM follows a totally unique approach compared to other machine learning algorithms. The working of SVM is as follows: SVM creates an n-dimensional space called a hyperplane, which segregates the classes present in the dataset and the hyperplanes also help detect existing data points. The decision to create a hyperplane depends on the size of the input features. If the input features are assumed to be two, tthe hyperplane will be treated as a single line and if there are more features than two, the hyperplane will be treated as a two-Dimensional plane. Now the focus is on optimized hyperplane search that can segregate classes well.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

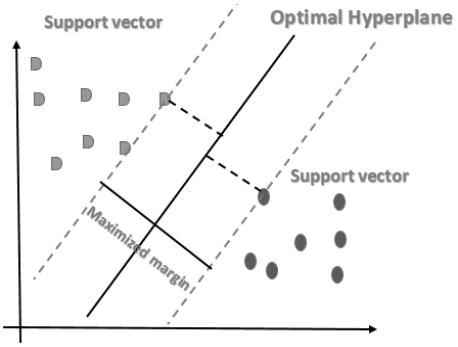

Fig. 6. SVM Classifier

This can be decided using margin gap. The data points closest to the hyperplanes are called support vectors.

The larger the margin gap between the support vectors, the greater the chance of choosing an optimized hyperplane. This will be clearer with the help of Fig.6.

SVM are of two types:

1) Linear SVM Classifier:

When the data is classified into only two classes using a straight-line hyperplane, then the classifier is called a linear SVM.

2) Non-Linear Classifier:

When the data is not classified into two classes, the classifier is called a nonlinear SVM.

D. Random Forest Classifier

The random forest classifier is the most popular and widely used algorithm in machine learning. Random forests (RF) are commonly used in various computer vision and computer applications. Their popularity is mainly due to the high computational efficiency during training and evaluation and the achievement of modern results [4].

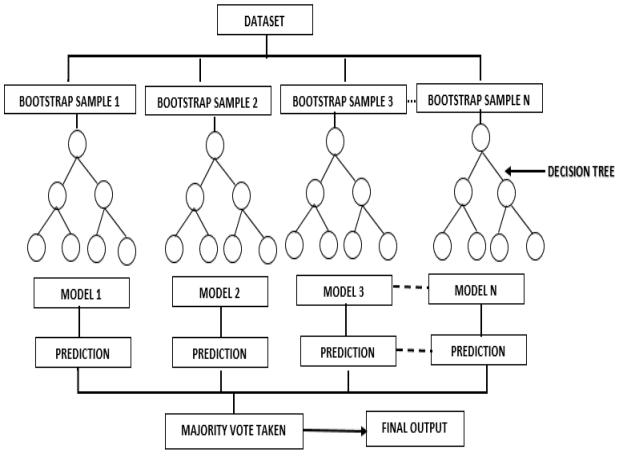

Fig. 7. Random Forest Classifier

Random Forest classifier is based on a ensemble technique, known as Bagging or Bootstrap aggregation. The Ensemble technique chooses a random n no. of samples from the dataset, where samples are called as bootstrap samples. From these samples, n decision trees are created.

The number of bootstrap samples depends on how many samples we want to train. Let's say if we want to create 100 models, it will generate 100 bootstrap templates. Then a decision tree model was created for each sample. Each decision tree then generates final predictions that are combined to get the final result or prediction. Fig.7 illustrates the workflow of the Random Forest classifier.

The areas where the random forest classifier can be used are: Credit Card Fraud Detection, Diabetes and Breast Cancer Prediction, Price Optimization in E-commerce, Stock Market, etc.

E. K-Nearest Neighbour Classifier

The K-Nearest Neighbor classifier is a supervised machine learning algorithm known for its simplicity, flexibility, and efficiency, abbreviated as KNN. This classifier is also known as lazy algorithm, it produces very high accuracy or efficiency. KNN is a non-parametric classifier, i.e. relative to the data set, KNN makes no assumptions.

When storing data points in the learning phase, the KNN learns nothing; rather, it learns in the experimental phase. Its performance depends on choosing the best possible value for K. Here, the nearest neighbors are the data points with the minimum distance in the feature space from our new data point, and K is the number of these data points that we consider when implementing the algorithm. There is no principled method to choose K except through expensive computational techniques such as cross-validation [5]. The purpose of this algorithm is to process the sample whose classes are unclear so as to obtain their intended classes.

Firstly, the distance between the samples to be tested in the test set and the sample in the training set is calculated and sorted in ascending order, where the distance formula is generally Euclidean distance. Then, the first k samples with small distance from each other are selected to determine the class of k samples.

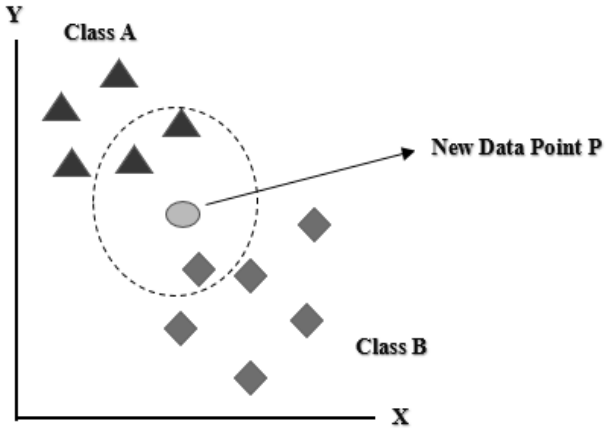

Fig. 8. K-Nearest Neighbour Classifier

The class with the highest occurrence frequency is the prediction class of the samples to be tested [6].

The areas where knn classifier can be used are: audio recognition, face recognition weather & climate forecasting, helps in agriculture by determining soil & water levels, stock market, Handwriting detection, etc.

In Fig.8., consider that we have a data set containing two classes i.e. class A and class B. Now whenever a new input comes in for example a data point p then we choose a value has a value of k. For the purposes of this example, let's say we choose 3 as the value of k. Then we will find the distance of the three nearest values and the distance can be calculated using different methods including these, Euclidean method, Manhattan method, Minkowski method etc. The one with the smallest distance will have more probability to decide which class the new points belong to.

Retrieval Number: 100.1/ijsce.E35831112522
DOI: 10.35940/ijsce.E3583.1112522
Journal Website: www.ijsce.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

IV. CONCLUSION

In this paper, the author discussed the various types of machine learning algorithms, these machine learning algorithms are immensely popular and widely used by researchers. The author also discussed on areas where such machine learning algorithms can be used and implemented. Based upon the reading, user can clearly grasp the overview of all machine learning algorithms. User are able to easily differentiate between ML algorithms and able to choose which algorithm should be used for solving the specific problem.

The machine's performance cannot be judged based on only on its algorithm because the performance depends on the quality of the data. This is why it is necessary to create good data sets. The data sets serve as the sole determining factor for an algorithm.

We can say that every algorithm does not work for every data set.

REFERENCES

1. Y. Singh, P. K. Bhatia, and O. Sangwan, "A review of studies on machine learning techniques," *International journal of computer science and security*, vol. 1, no. 1, pp. 70–84, 1999.
2. R. Konieczny and R. Idczak, "Mössbauer study of Fe-Re alloys prepared by mechanical alloying," *Hyperfine Interact*, vol. 237, no. 1, pp. 1–8, 2016, doi: 10.1007/s10751-016-1232-6. [CrossRef]
3. I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," *SN Computer Science*, vol. 2, no. 3, pp. 1–21, 2021, doi: 10.1007/s42979-021-00592-x. [CrossRef]
4. C. Leistner and H. Bischof, "On-line Random Forests," pp. 1393–1400, 2009.
5. A. Singh, N. Thakur, and A. Sharma, "A review of supervised machine learning algorithms," *Proceedings of the 10th INDIACom; 2016 3rd International Conference on Computing for Sustainable Global Development, INDIACom 2016*, pp. 1310–1315, 2016.
6. H. Xue and P. Wang, "An Improved Sample Mean KNN Algorithm Based on LDA," *Proceedings - 2019 11th International Conference on Intelligent Human-Machine Systems and Cybernetics, IHMSC 2019*, vol. 1, pp. 266–270, 2019, doi: 10.1109/IHMSC.2019.00068. [CrossRef]

AUTHORS PROFILE

Nipun Jain, received his B. Tech degree from the Mangalayatan University, Aligarh, Uttar Pradesh in 2016. He received his school education from Adarsh Bal Niketan Senior Secondary School, IIT Roorkee. He is currently working as a Senior Project Associate in the Department of Electrical Engineering at IIT Roorkee. He has more than two years' experience as a Project Associate in the 'Virtual Labs', project funded by the Ministry of Education. He had also worked as a Python Instructor at an accredited institute in India. He has more than two years' teaching experience. His current research interest includes Artificial Intelligence, Machine Learning, Data Science, Cyber Security, Cyber and Digital Forensics, Web Development, Cloud Computing, etc.

Rajeev Kumar, received his B. Tech and M. Tech degrees from the Uttarakhand Technical University, Dehradun in 2011 and 2015 respectively. He is pursuing his Ph.D. in Electrical Drives from IIT Roorkee. He is currently working as a Senior Research Fellow in the Department of Electrical Engineering at IIT Roorkee. He has more than five years' experience as a Research Fellow in the 'Virtual Labs', project funded by the Ministry of Education. He had also worked as a lecturer at an accredited institute in India. He is the author of 5 research papers in journals and conferences. His current research interest includes machine learning, condition monitoring, electrical machine fault analysis, signal processing, online lab development and digital education etc.