APPROACH BASED ANALYSIS OF MACHINE LEARNING ALGORITHMS

Nisha Choudhary

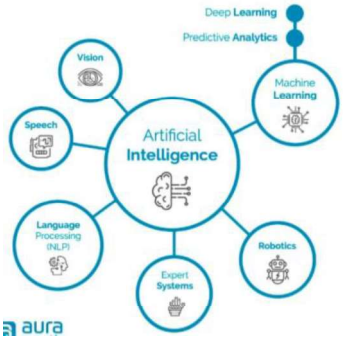
ABSTRACT

In the previous epoch, "artificial intelligence" and "machine learning" have become more often used terms. These phrases are frequently used reciprocally in a variety of fields, including education, research, and industry. One of the most alluring applications of artificial intelligence is machine learning. Many commonplace applications make use of learning algorithms. These learning algorithms improve how well Google and Bing and other online search engines function. Every time you use Facebook and recognise a friend's photo, a machine learning system is at work. Email spam filters are based on a learning algorithm and spare consumers from having to sift through a deluge of spam messages.

Keywords: Artificial intelligence, machine learning, machine learning algorithms

1. Introduction

1.1 Artificial Intelligence



Artificial intelligence refers to machines that mimic human intelligence. In other words, it is a science that tries to create systems that can reason and learn like humans do, as well as learn from experience, identify problems in specific scenarios, verify information, and carry out logical activities.

With the start of the new century and the remarkable technological advancement, AI has emerged as one of the unstoppable trends, and milestones in this field are starting to be very numerous, from expert systems capable of defeating a human in any intellectual activity to virtual assistants capable of managing day-to-day affairs.

A wide range of technologies and domains are included in AI, each requiring its own set of mathematical and technological study. [1].

1.2 AI & ML

Artificial intelligence (AI) and machine learning (ML) are the two most widely used technologies for creating intelligent systems. Machine learning (ML) is a branch of artificial intelligence that enables computers to learn from data without explicit training.

The goal of artificial intelligence (AI), a broader concept, is to create intelligent computers that can mimic human thought processes and conduct.As a result, AI refers to a computer-assisted decision-making system that makes decisions independently and learns from experience. For AI to work, algorithms that can learn from experience are necessary. Here is when AI comes into play. A few examples include automatic language translation, spam and virus filtering in emails, image recognition, automatic friend tag suggestions, medical diagnosis, online fraud detection, product recommendations, self-driving cars, stock market trading, speech recognition, traffic prediction, and virtual own subordinate. Customer service, e-commerce, and finance have all benefited from AI and ML. By 2020, 85% of client contacts will be handled without the involvement of a human.

Assistant Professor, Dept. Of Comp. Sc., Aggarwal College Ballabgarh

1.3 Machine Learning

The study of machine learning (ML) in AI aims to make a computer system understand and connect information the same way a human would. ML models can produce precise forecasts that are valuable to a business. When anticipating the likelihood of a particular outcome, such as whether or not a customer would stop using a product or service, prediction is the output of a machine learning algorithm that has been trained on historical data and applied to current data. It helps with both finding the best model to describe our data and predicting how well that model will function in the future [3].

Data scientists utilise machine learning to solve challenging business problems. They employ big data [4] to produce practical and meaningful insights. Machine learning is utilised in a wide range of applications, including speech recognition, banking software that can identify anomalous transactions, spam filters in email, customised website suggestions, and internet search engines.

1.4 Types of ML algorithms

There is four types of ML algorithms. These are reinforcement , supervised, unsupervised and semi-supervised algorithms [5] [6].

- **Reinforcement learning**

Reinforcement learning is described as a continuous repetition based on "trial and error" that machines can carry out quickly under precise conditions (for instance, game rules) and with a particular objective known as "reward" (winning a game of chess is a classic example). As a result, based on prior experiences, the computer produces unique outcomes, patterns, correlations, paths, and conclusions. The learning model is demonstrated by the Alpha Zero chess AI from Deep Mind [7].

- **Supervised ML**

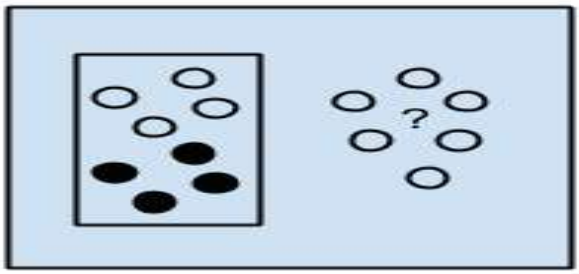
Underpinning supervised learning are predictive models built on training data. The system should be able to produce a certain output given a known set of data, therefore the model is adjusted (trained) until appropriate results are obtained. Consider autonomous vehicles. By categorising all of the data, the algorithms are taught to anticipate the result from the input. In the great majority of real-world machine learning applications, supervised learning is used. [8]. If both the input variable, X, and the output variable, Y, are provided, a relationship between the two will be discovered using the supervised learning technique via the mapping function..

$$Y = f(X)$$

The objective is to correctly estimate the mapping function so you can predict the output variables, or Y, with fresh input data (x). (x).

In supervised learning, an algorithm is learned from a training dataset in an iterative manner, much how a teacher iteratively oversees the learning process. Additionally, the right responses are known. Learning is complete when the algorithm performs to a reasonable degree. Regression and classification are the two subcategories of supervised learning tasks.

- ❖ **Classification**: This method gives result in the form of category in case of output variablesuch as "red" or "blue" or "illness" and "no disease," the problem is called a classification problem. Support vector machine is one of the supervised machine learning methods used for classification problems.
- ❖ **Regression:** This method gives result in the form of a real value, such as "dollars" or "weight," a regression problem exists. Linear regression is one of the supervised machine learning methods used for egression problems.



- ❖ Inspite of problems dedicated to classification and regression, some collective sort of problems fabricated on the pinnacle of classification as well as regression include recommendation and time series prediction respectively. Random forest algorithm is one is one of the supervised machine learning methods used for such types of problems.

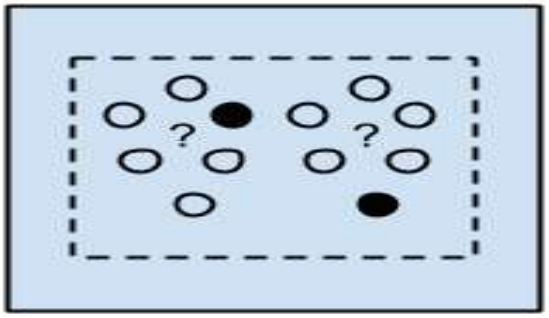
- **Unsupervised ML**

Algorithms for machine learning (ML) that are unsupervised learn information without supervision. These unsupervised ML methods are similar to supervised learning algorithms, with the exception that they only alter their models in response to new data. Simply put, the algorithm self-trains without assistance from outside sources [9]. Because none of

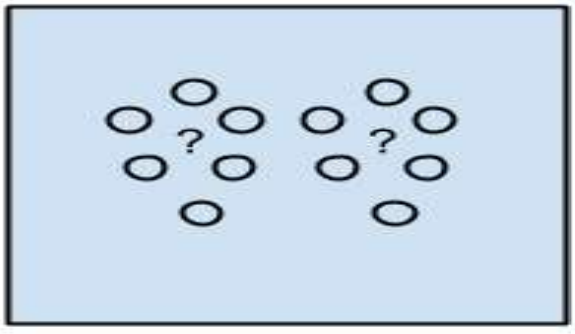
the data is labelled, the computer learns the inherent structure from the incoming data. Unsupervised learning uses X as the only input data; there are no ensuing output variables.

The goal of unsupervised learning is to simulate the distribution or underlying structure of the data. Because there are no right or wrong answers and no teacher in unsupervised learning, it gets its name. The intriguing structure of the data is left to algorithms to uncover and convey on their own..Clustering and association tasks are connected to unsupervised learning.

- **Clustering**: A clustering problem is one in which the data's intrinsic groupings are uncovered such as categorising clients based on their shopping habits. K-means is one of the unsupervised learning method used for clustering problems.



- **Association**: An association problem is one in rules are identified to describe considerable hunks of the data. E.g. the customers who purchasean item 'X' aso purchases the item 'Y'. These rules are called association rules and the problems belong to this category are called association rule learning problems. For such type of problems, Apriori algorithm is used.
- **Semi-Supervised ML**



In the real world, there is a plentiful supply of information. Data can either be labelled or not. However, most of the data is unlabeled. Thus, a combination of supervised and unsupervised algorithms can be utilised to handle mixed data (labelled and unlabelled). Issues with semi-supervised learning occur when there is a sizable amount of input data (X), but only some of it is tagged (Y).

This category includes a lot of real-world machine learning problems. This is because data labelling can be costly or time-consuming, as it may necessitate the involvement of domain experts. While, unlabeled data is inexpensive and simple to obtain and keep. In this scenario, unsupervised learning techniques are used to discover and learn the structure in the input variables, while supervised learning techniques are used to make best guess predictions for unlabeled data, feed that data back into the supervised learning algorithm as training data, and use the model to predict new unseen data.

2. **Considerations to be pondered during choosing a ML**

- **Size of the training data:** To achieve solid predictions, it is normally recommended to collect a wide range of data. However, data availability is frequently a limitation. As a result, whether the training data is smaller or the dataset contains fewer observations but more features, such as genetics or textual data, the algorithm is chosen accordingly.
- **Accuracy and/or Interpretability of the output:** The accuracy of a model is defined as the function's ability to predict a response value for a given observation that is near to the genuine response value. A highly interpretable method (restrictive models like Linear Regression) means that each individual predictor can be clearly understood, whereas flexible models provide more accuracy at the tradeoff of low interpretability. Restrictive algorithms can only generate a small number of different mapping function shapes, but flexible algorithms can generate a much larger number of different mapping function shapes. If inference is the goal, restrictive models are preferable since they are more interpretable, but if accuracy is the goal, flexible models are better. Interpretability is inversely related to a method's flexibility.
- **Speed or Training time:** Training time is directly proportional to accuracy. So higher will be training time, higher will be the accuracy. A large amount of data leads to longer time for algorithms to learn. Thus, these two features are mostly responsible to select an algorithm in real-world applications.

- **Linearity:** If data can be divided by a straight line, resulting in the formation of classes, this is known as linear data; otherwise, non linear data exists when classes cannot be separated.
- **Number of features:** There could be a lot of features in the dataset, but not all of them are useful and significant. These characteristics differ based on the type of data, such as genetics or textual data.

3. Analysis of ML Algorithms

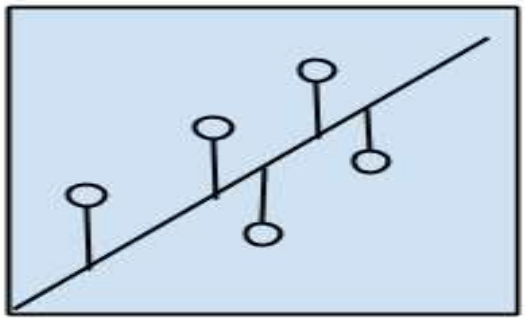
A number of ML algorithms have been proposed. Sometimes, the naïve user get confused choosing the relevant algorithm. An effort has been done in the direction to approach the ML algorithm on the basis of two criteria i.e. the grouping of algorithms by their learning style and the grouping of algorithms by their similarity in form or function.

Approach 1: Grouping of Algorithms on the basis of Learning Style

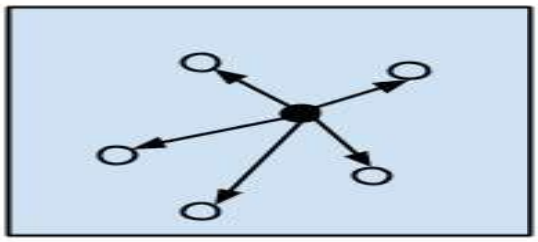
Table 1: Comparative anlaysis of different machine learning methods (Learning style)

Category of Method	Model Building	Training data/ Input	Result status	Type of problems	Algorithms proposed
Supervised Learning Algorithm	A model is created by going through a training procedure that lasts until the model reaches the appropriate degree of accuracy on the training data.	Labelled Training data	Known	• Classification • regression	• Logistic Regression • Back Propagation Neural Network.
Unsupervised Learning algorithm	Deducing structures from the incoming data is used to create a model. This could be for the purpose of extracting broad rules. It might be done using a mathe-matical procedure to systematically minimise redundancy,or it could be done manually sorting data by similarity.	Unlabelled input data	Un-known	• Clustering • dimensionality reduction • association rule learning.	• Apriori algorithm • K-Means.
Semi-supervised	There is a desired prediction problem but the model must learn the structures to organize the data as well as make predictions i.e. partially trained model	A mixture of labelled & unlabelled	Depends on type of input	• Classification • Regression • clustering	Extensions to other flexible methods that make assumptions about how to model the unlabeled data.
Reinforcement Learning	Models are based on a reward or penalty system, i.e. a strategy that is driven by the environment.	learning agent's experience	Depends on input	•Classification •control	Q-learning

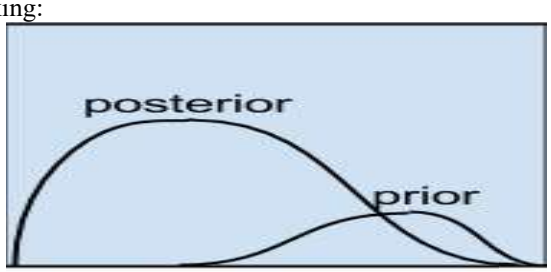
Approach 2: Grouping of Algorithms on the basis of Similarity in function
Following is the ML algorithm grouped in function or the way of working:

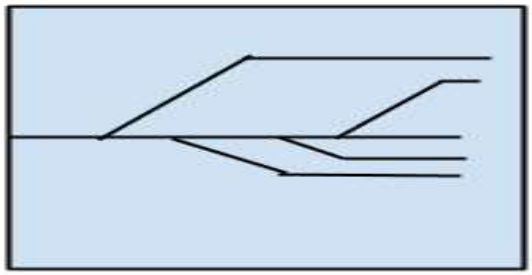


- **Regression Algorithms:**Regression is focused with iteratively refining a model's relationship between variables using a measure of inaccuracy in the model's predictions. Regression methods, which are a workhorse of statistics and have been co-opted into statistical machine learning, are a class of issue and approach. The most commonly used regression methods include Ordinary Least Squares Regression (OLSR), Linear Regression, Logistic Regression, Stepwise Regression, Multivariate Adaptive Regression Splines (MARS), and Locally Estimated Scatterplot Smoothing (LOESS).

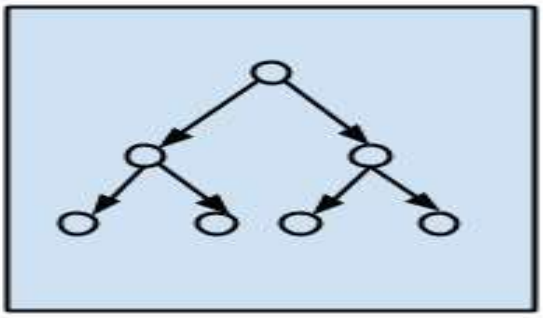


- **Instance-based Algorithms:** The model is created for the decision problems with instances or illustrations of training data that are required by the model is an instance-based learning model. These algorithms In order to discover the superlative match and generate a forecast, such algorithms often fabricate a database using instance data. This database is, then, compared with new data using a similarity measure. As a result, instance-based approaches are also known as winner-take-all and memory-based learning methods. The representation of the stored instances and the similarity measures used between instances are both highlighted. k-Nearest Neighbor (kNN), Learning Vector Quantization (LVQ), Self-Organizing Map (SOM), Locally Weighted Learning (LWL), and Support Vector Machines are the most prominent instance-based techniques (SVM).





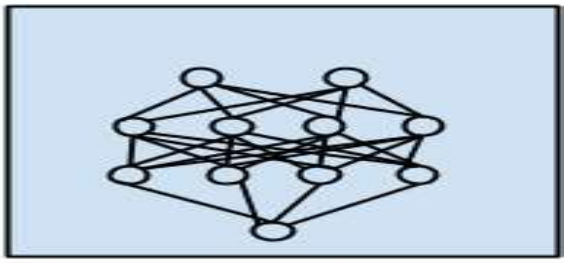
- **Regularization Algorithms:**These algorithms are an extension of another method, particularly regression methods, which penalise models based on their complexity, favouring simpler models that are also better at generalisation. Ridge Regression, Least Absolute Shrinkage &Selection Operator (LASSO), Elastic Net, and Least-Angle Regression are the most used regularisation techniques (LARS).



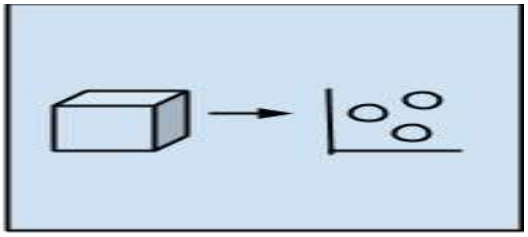
- **Decision Tree Algorithms:** Decision tree approaches create a prototypical of decisions based on the actual values of data characteristics. In tree architectures, decisions fork until a forecast choice is formulated for a specific record. For classification and regression problems, decision trees are trained on data. Decision trees are popular in ML because they are often expeditious and precise. Classification and Regression Tree (CART), C4.5 & C5.0 (various variants of a powerful technique), Iterative Dichotomiser 3 (ID3), Chi-squared Automatic Interaction Detection (CHAID),

Decision Stump, M5 & Conditional Decision Trees are the most prominent decision tree algorithms.

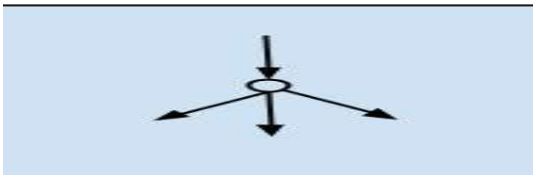
- **Bayesian Algorithms:** Bayesian approaches are those that use Bayes' Theorem explicitly to solve issues like classification and regression. Naive Bayes, Gaussian Naive Bayes, Multinomial Naive Bayes, Averaged One-Dependence Estimators (AODE), Bayesian Belief Network (BBN), and Bayesian Network are the most prevalentBayesian algorithms (BN).



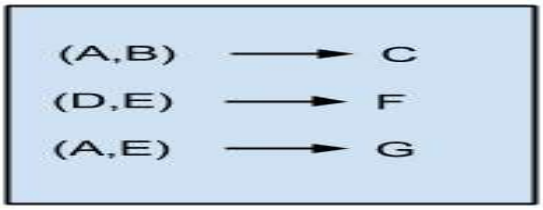
- **Clustering Algorithms:** Clustering, like regression, defines a problem class and a method class. Centroid-based as well Modeling methodologies such as centroid-based and hierarchal clustering methods are commonly used. All approaches are concerned with utilising the data's inherent structures to effectively organise the data into groups with the greatest amount of similarity. k-Means, k-Medians, Expectation Maximization (EM), and Hierarchical Clustering are the most used clustering techniques.



- **Association Rule**
- **Learning Algorithms:** Approaches for extracting rules that best describe observed linksbetween variables in data are known as association rule learning methods. These guidelines can help an organisation uncover meaningful and commercially useful correlations in huge multidimensional datasets. Apriori algorithm and the Eclat algorithm are two of the most widely used association rule learning algorithms.

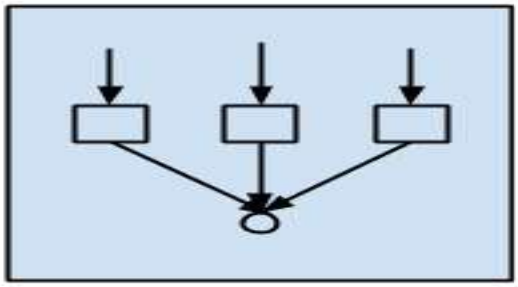


- **Artificial Neural Network Algorithms (ANNs):** ANNs are models based on the structure and/or function of biological neural networks. They're a sort of pattern matching that's often used for regression and classification problems, but they're actually a huge subfield with hundreds of methods and variations for a extensiverange of applications. Perceptron, Multilayer Perceptrons (MLP), Back-Propagation, Stochastic Gradient Descent, Hopfield Network, and Radial Basis Function Network (RBFN) are the most prominent ANN.



- **Deep Learning Algorithms:** Deep Learning is a recent update on Artificial Neural Networks that takes advantage of available low-cost computation. They are concerned with creating far larger and more complicated neural networks, and many of the methods are concerned with very huge datasets of tagged analogue data, such as image, text, audio, and video, as mentioned previously. Convolutional Neural Networks (CNNs), Long Short-Term Memory Networks (LSTMs), Recurrent Neural Networks (RNNs), , Stacked Auto-Encoders, Deep Boltzmann Machines (DBMs), and Deep Belief Networks are the most prominent deep learning methods (DBN)
- **Dimensionality Reduction Algorithms:** Dimensionality reduction, like clustering approaches, pursueand feats the data's underlying structure, but in this case in an unsupervised way or in order to summarise or describe data with less information. This can be beneficial for visualising dimensional data or simplifying data for use in supervised learning methods. Many of the approaches listed below can be used in classification and regression. Principal Component Analysis (PCA), Principal Component Regression (PCR), Partial Least Squares Regression (PLSR), Sammon Mapping,

Multidimensional Scaling (MDS), Projection Pursuit, Linear Discriminant Analysis (LDA), Mixture Discriminant Analysis (MDA), Quadratic Discriminant Analysis (QDA), and Flexible Discriminant Analysis (FDA) are a few examples (FDA).



- **Ensemble Algorithms:** Some models are weak enough to train independently. These weak models are collaborated with the help of ensemble methods and The data is then combined in some way to produce a final prediction. It takes a lot of thought to figure out which categories of poor learners to mix and how to combine them. This is a very potent group of tactics, and as a result, it is extremely popular. Boosting, Bootstrapped Aggregation (Bagging), Ada Boost, Weighted Average (Blending), Stacked Generalization (Stacking), Gradient Boosting Machines (GBM), Gradient Boosted Regression Trees (GBRT), and Random Forest are some examples of these techniques.

Conclusion

The paper discusses various machine learning algorithms with the help of two approaches i.e. learning style and similarity of functions. In the first approach, algorithms are grouped in broad categories i.e. supervised, unsupervised as well as semi-supervised. As far as second approach is concerned, algorithms are grouped by similarity in terms of their function. This approach can be a expedientgrouping method. However, it isn't ideal.. Some algorithm can be fit in multiple categories. Such algorithms can be chosen from a group of people who are subjectively the "best" fit.

Future scope

The ML methods can be studied on other aspects. One of the most important aspect is nature of data. The data can be numerical data, categorical data and mixed data. In the next research, the machine learning methods will be covered on behalf of nature of data.

REFERENCES

[1] Pooja Agarwal, Pooja Yadav et al, "Research Paper on Artificial Intelligence", CASE STUDIES JOURNAL VOL-2, ISSUE 6 ISSN (2305-509X)] July 1, 2013

[2] The International Congress for global Science and Technology ICGST International Journal on Artificial Intelligence and Machine Learning (AIML), Vol (14), Issue (I), August, 2014 www.icgst.com www.icgst-amc.com www.icgst-ees.com

[3] Batta Mahesh, "Machine Learning Algorithms – A Review", International Journal of Science and Research (IJSR) ISSN: 2319-7064 ResearchGate, Impact Factor (2018): 0.28 | SJIF (2018): 7.426

[4] K. Sree Divya, "Machine Learning Algorithms in Big data Analytics", International Journal of Computer Sciences and Engineering Open Access Review Paper Volume-6, Issue-1 E-ISSN: 2347-2693

[5] Ayon Dey, "Machine Learning Algorithms – A Review", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 7 (3) , 2016, 1174-1179

[6] Özer ÇELİK, "A Research on Machine Learning Methods and Its Applications", Journal of Educational Technology & Online Learning Volume 1, Issue 3, 2018

[7]Ahmad Hammoudeh," A Concise introduction to reinforcement learning", : https://www.researchgate.net/publication/323178749, 2018

[8] Osisanwo F.Y, "Supervised Machine Learning Algorithms: Classification and Comparison", International Journal of Computer Trends and Technology (IJCTT) – Volume 48 Number 3 June 2017

[9]Memoona Khanum, "A Survey on Unsupervised Machine Learning Algorithms for Automation, Classification and Maintenance", International Journal of Computer Applications (0975 – 8887) Volume 119 – No.13, June 2015

[10]Dong Yu, ctive learning and semi-supervised learning for speech recognition: A unified framework using the global entropy reduction maximization criterion", www.sciencedirect.com, 2008