National Conference on Recent Advances in Computer Science and IT (NCRACIT)
International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 4 | Issue 1 | ISSN : 2456-3307

A Study and Analysis of Machine Learning Algorithms and Its Applications

Dr. Archana Sharma[1], Prof. Vibhakar Mansotra[2]

[1]Department of Computer Science, Institute of Management Sciences (IMS), Jammu, Jammu & Kashmir, India
[2]Department of Computer Sciences and IT, University of Jammu, Jammu, Jammu & Kashmir, India

ABSTRACT

Machine learning is a subfield of artificial intelligence (AI). Deep understanding of data inputs would help in taking output as optimized decisions and also help them to work in more accurate and in efficient manner. Designing and implementing the algorithm and using it in most appropriate way is, the real challenge for the developers and scientists. Machine learning algorithms allow computers to train inputs data and use statistical analysis for optimum decision values. Based on data inputs, machine learning facilitates computers in building models from dataset in order to get automatic decision-making processes. Today, many technical users has benefitted from machine learning. In this paper, we will discuss the machine learning methods, and explore various algorithmic approaches in machine learning providing with some of the positive and negative attributes of each algorithm and most efficient use to make decisions and complete the task in more optimized form.

Keywords: Artificial Intelligence, Machine Learning, Decision-Making Process, Applications.

I. INTRODUCTION

According to a recent study, machine-learning algorithms are expected to replace 25% of the jobs across the world, in the next 10 years. With the rapid growth of big data and availability of programming tools. Machine learning is gaining mainstream presence for data. Machine learning applications are highly automated and self-modifying which continue to improve over time with minimal human intervention as they learn with more data. Machine Learning algorithms are classified as

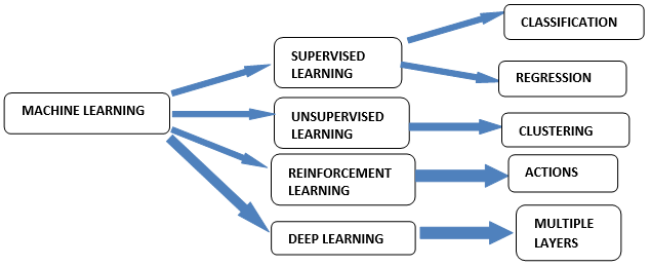

Figure 1. Machine Learning Techniques

A. Supervised Machine Learning Algorithms

Machine learning algorithms that make predictions on given set of samples. Supervised machine learning algorithm searches for patterns within the value labels assigned to data points.

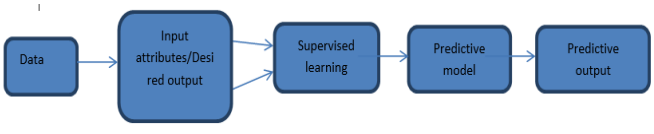

Figure 2. Supervised Learning

B. Unsupervised Machine Learning Algorithms

There are no labels associated with data points. These machine-learning algorithms organize the data into a group of clusters to describe its structure and make complex data look simple and organized for analysis.

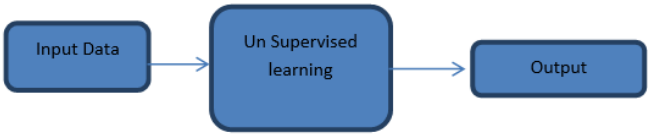

Figure 3. Unsupervised Learning

C. Reinforcement Machine Learning Algorithms

These algorithms choose an action, based on each data point and later learn how good the decision was. Over time, the algorithm changes its strategy to learn better and achieve the best reward.

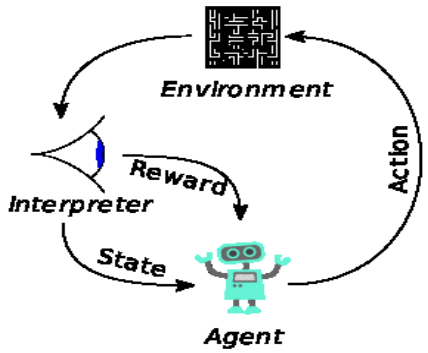

Figure 4. Reinforcement Learning

D. Deep Machine Learning Algorithms

In deep learning, a computer model learns to perform classification tasks directly from images, text, or sound. Deep learning models can achieve state-of-the-art accuracy, sometimes exceeding human-level performance. Models are trained by using a large set of labeled data and neural network architectures that contain many layers.

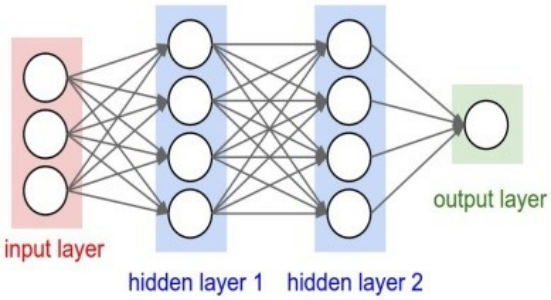

Figure 5. Deep Learning

II. Which Machine Learning Algorithm to Use?

There are dozens of supervised, unsupervised, reinforcement and deep machine learning algorithms, and each takes a different approach to learning.. Finding the right algorithm is partly just trial and error. But algorithm selection also depends on the size and type of data working with, the insights we want to get from the data, and how those insights will be used.

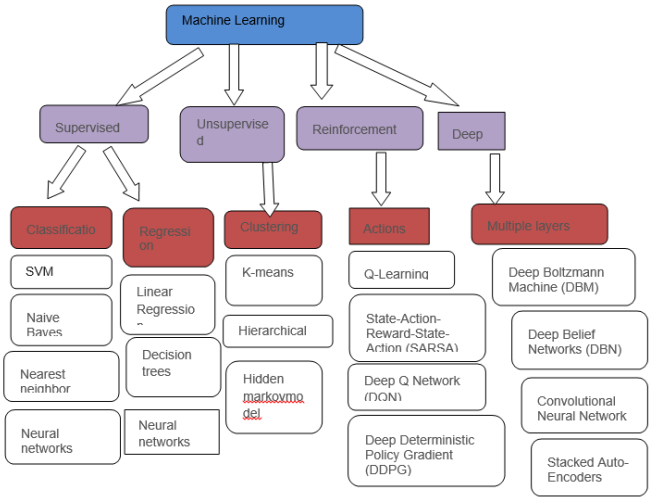

Figure 6. Classification of Machine Learning Algorithms

Some of the machines learning algorithms are explained below:

A. Naïve Bayes Classifier Machine Learning Algorithm

The Bayesian Classification represents a supervised learning method as well as a statistical method for classification. It is a probabilistic model and it allows us to capture uncertainty about the model in a principled way by determining probabilities of the outcomes. It can solve diagnostic and predictive problems. Bayesian classification provides practical learning algorithms and prior knowledge and observed data can be combined. Bayesian Classification provides a useful perspective for understanding and evaluating many learning algorithms. It calculates explicit probabilities for hypothesis and it is robust to noise in input data.

B. Support Vector Machine Learning Algorithm

Support Vector Machine is a supervised machine-learning algorithm for classification or regression problems where the dataset teaches SVM about the classes so that SVM can classify any new data. It works by classifying the data into different classes by finding a line (hyperplane) which separates the training data set into classes. As there are many such linear hyperplanes, SVM algorithm tries to maximize the distance between the various classes that are involved and this is referred as margin maximization. If the line that maximizes the distance between the classes is identified, the probability to generalize well to unseen data is increased.

SVM's are classified into two categories:

- Linear SVM's – In linear SVM's the training data i.e. classifiers are separated by a hyperplane.
- Non-Linear SVM's- In non-linear SVM's it is not possible to separate the training data using a hyperplane.

C. Apriori Machine Learning Algorithm

Apriori algorithm is an unsupervised machine-learning algorithm that generates association rules from a given data set. Association rule implies that if an item A occurs, then item B also occurs with a certain probability. Most of the association rules generated are in the IF_THEN format.

Basic principle on which Apriori Machine Learning Algorithm works:

- If an item, set occurs frequently then all the subsets of the item set also occur frequently.
- If an item set occurs infrequently then all the supersets of the item set have infrequent occurrence.

D. Linear Regression Machine Learning Algorithm

Linear Regression algorithm shows the relationship between two variables and how the change in one variable impacts the other. The algorithm shows the impact on the dependent variable on changing the independent variable. The independent variables are referred as explanatory variables, as they explain the factors the impact the dependent variable. Dependent variable is often referred to as the factor of interest or predictor.

E. Decision Tree Machine Learning Algorithm

A decision tree is a graphical representation that makes use of branching methodology to exemplify all possible outcomes of a decision, based on certain conditions. In a decision tree, the internal node represents a test on the attribute, each branch of the tree represents the outcome of the test and the leaf node represents a particular class label i.e. the decision made after computing all of the attributes. The classification rules are represented through the path from root to the leaf node.

Types of Decision Trees

Classification Trees- These are considered as the default kind of decision trees used to separate a dataset into different classes, based on the response variable. These are generally used when the response variable is categorical in nature.

Regression Trees-When the response or target variable is continuous or numerical, regression trees are used. These are generally used in predictive type of problems when compared to classification.

Decision trees can also be classified into two types, based on the type of target variable- Continuous Variable Decision Trees and Binary Variable Decision Trees. It is the target variable that helps decide what kind of decision tree would be required for a particular problem.

F. Random Forest Machine Learning Algorithm

Random Forest is the machine-learning algorithm that uses a bagging approach to create a bunch of decision trees with random subset of the data. A model is trained several times on random sample of the dataset to achieve good prediction performance

from the random forest algorithm.In this ensemble learning method, the output of all the decision trees in the random forest, is combined to make the final prediction. The final prediction of the random forest algorithm is derived by polling the results of each decision tree or just by going with a prediction that appears the most times in the decision trees.

G. Logistic Regression

Logistic Regression machine learning algorithm is for classification tasks and not regression problems. The name 'Regression' here implies that a linear model is fit into the feature space. This algorithm applies a logistic function to a linear combination of features to predict the outcome of a categorical dependent variable based on predictor variables.

- **Binary Logistic Regression** – The most commonly used logistic regression when the categorical response has two possible outcomes i.e. either yes or not. Example –Predicting whether a student will pass or fail an exam, predicting whether a student will have low or high blood pressure, predicting whether a tumour is cancerous or not.
- **Multi-nominal Logistic Regression** - Categorical response has three or more possible outcomes with no ordering. Example- Predicting what kind of search engine (Yahoo, Bing, Google, and MSN) is used by majority of US citizens.
- **Ordinal Logistic Regression** - Categorical response has three or more possible outcomes with natural ordering. Example- How a customer rates the service and quality of food at a restaurant based on a scale of 1 to 10.

Table 1. shows the uses, applications, advantages and drawbacks of various machine learning algorithms.

Machine learning algorithms	Uses	Applications	Advantages	Drawbacks
Naïve Bayes Classifier	In moderate or large training data set, Used Several attributes, instances should be conditionally independent.	Sentiment Analysis, Document Categorization-Classification of news articles. Email Spam Filtering, Hybrid Recommender System and Collaborative Filtering.	input variables are categorical, faster and requiring little training data, performed multi class ons.	Need big data for reliable estimations, precision and recall in small data set will keep very low, none class label and a certain attribute value together then the frequency-based probability estimate will be zero this will affect the posterior probability estimate.
K Means Clustering Algorithm	used by search engines like Yahoo, Google to cluster web pages by similarity and identify the 'relevance rate' of search results.	Behavioral segmentation, Inventory categorization, Sorting sensor measurements, Detecting bots or anomalies.	its simplicity and speed and run on large datasets. It minimizes intra-cluster variance.	it does not yield the same result with each run, resulting clusters depend on the initial random assignments, does not ensure that the result has a global minimum of variance. mean to be definable which is not always the case.

Support Vector Machine	used for classification as well as pattern recognition purpose, Speech data, emotions and other such data classes can be used.	used for stock market forecasting, comparison of stocks helps manage investment making decisions based on the classifications.	Best performance(accuracy) on the training data, renders more efficiency for future data, does not make strong assumptions, does not over-fit the data.	it covers the determination of the parameters for a given value of the regularization and kernel parameters and choice of kernel, the problem of over-fitting from optimizing the parameters to model selection.
Apriori Machine Learning Algorithm	Used in market basket analysis.	Detecting Adverse Drug Reactions, .Market Basket Analysis,Auto-Complete Applications	easy to implement and can be parallelized easily, makes use of large item set properties.	Needs iterations, uniform minimum support threshold, Difficulties to find rarely occuring events,alternative approaches focus on partition and sampling.
Linear Regression	used to assess risk in financial services or insurance domain, credit card industry,	Estimating Sales Risk Assessment	most interpretable machine learning algorithms, easy to explain to others, requires minimal tuning,runs fast.	straight-line relationship between them which is incorrect sometimes, very sensitive to the anomalies in the data (or outliers).
Decision Tree	used in statistics, data mining and machine learning. used in operations research, specifically in decision analysis,	use in finance for option pricing, Remote sensing is an application area for pattern recognition based on decision trees. used by banks to classify loan applicants by their probability of defaulting payments.	its ability to assign specific values to problem, decisions, and outcomes of each decision	less accuracy in expected outcome, do not fit for continuous variables and result in instability, time consuming task, consider only one attribute at a time,, Large sized decision trees with multiple branches pose several presentation difficulties.

III. CONCLUSION

Machine learning research spans almost four decades. Much of the research has been to define various types of learning, establishing the relationship among them, and elaborate the algorithms that characterize them. But much less effort has been devoted to bring machine learning to bear on real world applications. Recently researchers have found broader applications of machine to real world problems such as

- Classifying DNA sequences
- Virtual Personal Assistants
- Computer vision, including object recognition
- Game playing
- Natural language processing

- handwriting recognition
- Speech recognition
- Brain-machine interfaces
- Adaptive websites
- Videos Surveillance
- Predictions while Commuting
- Social Media Services
- Online Customer Support
- Online Fraud Detection
- Bioinformatics and there are many more fields where machine learning is used now a days.

IV.REFERENCES

[1] Nahum Shimkin, "Learning in Complex System", Lecture Notes, Spring 2011.

[2] Thomas G. Dietterich, "Machine-Learning Research", AI Magazine Volume 18 Number 4 (1997).

[3] Rob Schapire, "Machine Learning Algorithms for Classification", Princeton University.

[4] S. B. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques", Informatica 31 (2007) 249-268.

[5] Leslie Pack Kaelbling, Michael L. Littman, Andrew W. Moore "Reinforcement Learning: A Survey", Journal of Artificial Intelligence, Research 4 (1996) 237-285, May 1996.

[6] R. Sathya, Annamma Abraham, "Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification", (IJARAI) International Journal of Advanced Research in Artificial Intelligence, Vol. 2, No. 2, 2013.

[7] R. Sathya and A. Abraham, "Unsupervised Control Paradigm for Performance Evaluation", International Journal of Computer Application, Vol 44, No. 20, pp. 27-31, 2012.

[8] Taiwo Oladipupo Ayodele, "Types of Machine Learning Algorithms", University of Portsmouth, United Kingdom.

[9] Alberto Maria Segre, "Applications of Machine Learning", Cornell University.

[10] J. A. Hartigan and M.A. Wong, "A K-Means Clustering Algorithm", Journal of the Royal Statistical Society. Series C (Applied Statistics), Vol. 28, No. 1 (1979), pp. 100-108.

[11] [11]. https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/.