iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** X **Month of publication:** October 2020

DOI: https://doi.org/10.22214/ijraset.2020.32019

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue X Oct 2020- Available at www.ijraset.com

Regression Techniques in Machine Learning &Applications: A Review

Vidya S. Kadam[1], Shweta Kanhere[2], Shrikant Mahindrakar[3]
[1, 2, 3] *School of Engineering & Technology, D Y Patil International University, Akurdi, Pune, India.*

Abstract: ***Machine learning is one of the most exciting technology& it gives the computer that makes it more similar to humans .It is actively being used today, perhaps in many more places than one would expect. Learning is a natural human behaviour which has been made an essential aspect of the machines as well. There are various techniques used for the same. Traditional machine learning algorithms have been applied in many application areas. This paper presents the most commonly used machine learning algorithms such as supervised, unsupervised, reinforcement. In each of our study we found that the results of various machine-learning algorithm depends on application areas on which they are applied. Our review of study not only suggests that these techniques are competitive with traditional estimators on one data set, but also illustrate that these methods are sensitive to the data on which they are trained.***

Keywords: Regression, Supervised, Unsupervised, Reinforcement, Machine learning

I. INTRODUCTION

Machine Learning provides machines with the ability to learn autonomously based on experiences, observations and analyzing patterns within a given data set without explicitly programming. [1] It is branch of computer science in which machine is trained to perform specific task. Nowadays machine learning algorithms are applied for classification, regression, clustering. Machine learning has proved that it has abilities in various fields such as self-driving car, web searches, fraud detection, email /spam filtering. The machine has to train on some data sets, and then, the algorithms are applied, so that the machine can make predictions and learn, respectively, on the given data sets Machine learning is the way to make programming scalable. In traditional Programming, Data and program is run on the computer to produce the output. In Machine Learning, Data and output is run on the computer to create a program. This program can be used in traditional programming. Machine learning is like farming or gardening. Seeds is the algorithms, nutrients is the data, the Gardner is you and plants is the programs.

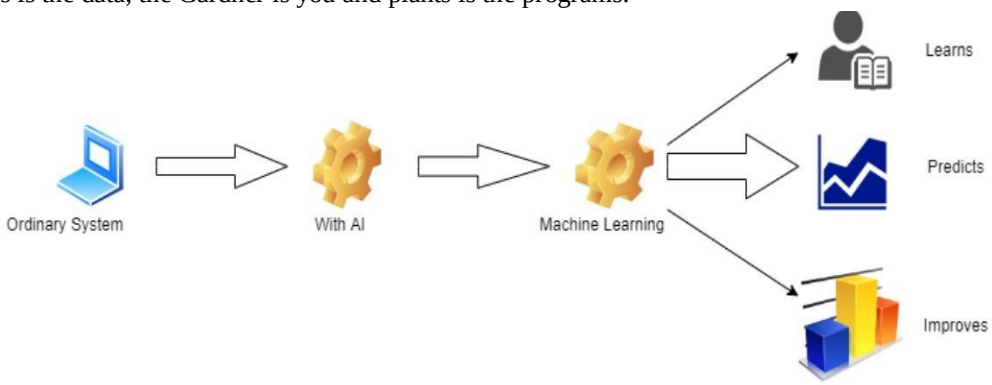


Fig:1 Machine Learning Model

II. TYPES OF LEARNING

A machine learning system learns from past experiences to improve the performances of intelligent application programs. Machine learning system is category into three types.

A. Supervised Learning

In supervised learning, we use known or labeled data for the training data. The input data goes through the Machine Learning algorithm and is used to train the model. Once the model is trained based on the known data, you can use unknown data into the model and get a new response. In this case, the model tries to figure out whether the data is cat or another pet animal. Once the model has been trained well, it will identify that the data is cat and give the desired response.

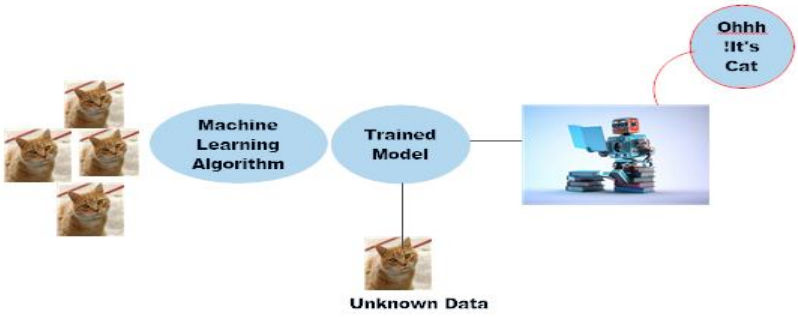


Fig. 2 Supervised Algorithm mechanism

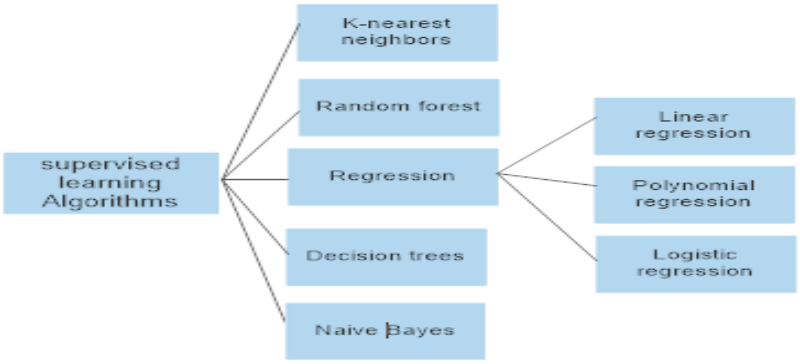


Fig. 3 Types Supervised Algorithm

1) *Regression:* Regression analysis is performed to determine correlation between two or more variables having cause effect relations &to make prediction for the topic by using relations. The regression using single independent variable is called univariate regression analysis while the analysis using more than two independent variable is called multivariate Regression analysis. Linear Regression consists of first Loading the Data& then Exploring the Data. After that we have to do Slicing the Data then Train and Split Data to Generate the Model. Finally evaluate the accuracy. The main goal of regression is the construction of an efficient model to predict the dependent attributes from a bunch of attribute variables. A regression problem is when the output variable is either real or a continuous value i.e. salary, weight, area, etc. We can also define regression as a statistical means that is used in applications like housing, investing, etc. It is used to predict the relationship between a dependent variable and a bunch of independent variables. Let us take a look at various types of regression techniques.

2) *Types Of Regression*

a) *Simple Linear Regression:* A regression technique in which the independent variable has a linear relationship with the dependent variable. The straight line in the diagram is the best fit line. The main goal of the simple linear regression is to consider the given data points and plot the best fit line to fit the model in the best way possible. Linear Regression, when the data is plotted on the graph, there was a linear relationship between both the dependent and independent variables. Thus, it was more suitable to build a linear model to get accurate predictions. The Linear regression always tends to make an error however hard it tries to fit in the data SLR=Y=B0+B1X

b) *Multiple Linear Regression:* In many applications, there is more than one factor that influences the response. Multiple regression models thus describe how a single response variable Y depends linearly on a number of predictor variables. Examples: The selling price of a house can depend on the desirability of the location, the number of bedrooms, the number of bathrooms, the year the house was built, the Square footage of the lot and a number of other factors. The height of a child can depend on the height of the mother, the height of the father, nutrition, and environmental factors. Example: A multiple linear regression model with k predictor variables X1, X2, ..., Xk and a response Y, can be written as $y = \beta0 + \beta1x1 + \beta2x2 + \cdots \beta kxk$

c) *Polynomial Regression:* In this regression technique, we transform the original features into polynomial features of a given degree and then perform regression on it. the Polynomial Regression graph manages to fit the data points onto the line more accurately. $y=w1x+w2x^2+..+b$

d) *Support Vector Regression:* For support vector machine regression or SVR, we identify a hyper plane with maximum margin such that the maximum number of data points are within those margins. It is quite similar to the support vector machine classification algorithm.

e) *Decision Tree Regression:* A decision tree can be used for both regression and classification. In the case of regression, we use the ID3 algorithm (Iterative Dichotomiser 3) to identify the splitting node by reducing the standard deviation.

f) *Random Forest Regression:* In random forest regression, we ensemble the predictions of several decision tree regressions.

B. *Unsupervised Learning*

In unsupervised learning, the training data is unknown and unlabelled - meaning that no one has looked at the data before. Without the aspect of known data, the input cannot be guided to the algorithm, which is where the unsupervised term originates from. This data is fed to the Machine Learning algorithm and is used to train the model. The trained model tries to search for a pattern and give the desired response. In this case, it is often like the algorithm is trying to break code like the Enigma machine but without the human mind directly involved but rather a machine. In this case, the unknown data consists of cat and dog which look similar to each other. The trained model tries to put them all together so that you get the same things in similar groups.

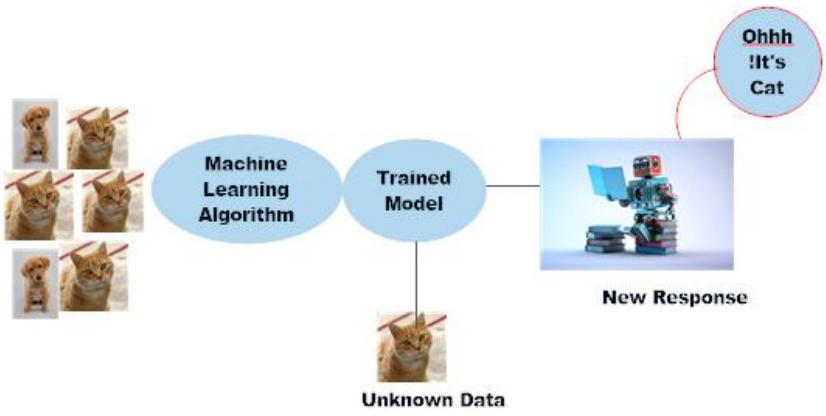


Fig.4 Unsupervised Algorithm mechanism

1) *K-means Clustering:* K-Means algorithm K-Means algorithm based on dividing is a kind of cluster algorithm. This algorithm which is unsupervised is usually used in data mining and pattern recognition. Aiming at minimizing cluster performance index, square-error and error criterion are foundations of this algorithm. To seek the optimizing outcome, this algorithm tries to find K divisions to satisfy a certain criterion. Firstly, choose some dots to represent the initial cluster focal points (usually, we choose the first K sample dots of income to represent the initial cluster focal point); secondly, gather the remaining sample dots to their focal points in accordance with the criterion of minimum distance, then we will get the initial classification, and if the classification if unreasonable, we will modify it(calculate each cluster focal points again), iterate repetitively till we get a reasonable classification. K-Means algorithm based on dividing is a kind of cluster algorithm, and has advantages of briefness, efficiency and celerity. However, this algorithm depends quite much on initial dots and the difference in choosing initial samples which always leads to different outcomes. [5]

2) *Apriori Algorithm:* Apriori algorithm is easy to execute and very simple, is used to mine all frequent itemsets in database. The algorithm makes many searches in database to find frequent itemsets where k itemsets are used to generate k+1-itemsets. Each k-itemset must be greater than or equal to minimum support threshold to be frequency. Otherwise, it is called candidate itemsets. In the first, the algorithm scan database to find frequency of 1-itemsets that contains only one item by counting each item in database. The frequency of 1-itemsets is used to find the itemsets in 2- itemsets which in turn is used to find 3-itemsets and so on until there are not any more k-itemsets If an itemset is not frequent, any large subset from it is also non-frequent this condition prune from search space in database.[6]

3) *Hierarchical Clustering:* Hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters. The quality of a pure hierarchical clustering method suffers from its inability to perform adjustment, once a merge or split decision has been executed. Then it will neither undo what was done previously, nor perform object swapping between clusters. Thus merge or split decision, if not well chosen at some step, may lead to some-what low-quality clusters. One promising direction for improving the clustering quality of hierarchical methods is to integrate hierarchical clustering with other techniques for multiple phase clustering. So in this paper, we describe a few improved hierarchical clustering algorithms that overcome the limitations that exist in pure hierarchical clustering algorithms.[7]
4) *Principal Component Analysis:* Principal component analysis (PCA) is a multivariate technique that analyzes a data table in which observations are described by several inter-correlated quantitative dependent variables.[8] Principal Component analysis also known as PCA is such a feature extraction method where we create new independent features from the old features and from combination of both keep only those features that are most important in predicting the target. New features are extracted from old features and any feature can be dropped that is considered to be less dependent on the target variable. CA is such a technique which groups the different variables in a way that we can drop the least important feature. All the features that are created are independent of each other. PCA is also used for reducing the dimensions. According to the respective eigenvalues arrange the eigenvectors in descending order. The main advantages of PCA is Lack of redundancy of data given the orthogonal components as well as Reduction of noise since the maximum variation basis is chosen and so the small variations in the background are ignored automatically. It is difficult to evaluate the covariance in a proper way. Even the simplest invariance could not be captured by the PCA unless the training data explicitly provides this information. It is difficult to evaluate the covariance in a proper way. Even the simplest invariance could not be captured by the PCA unless the training data explicitly provides this information.

C. Reinforcement Learning

There are many unsolved problems that computers could solve if the appropriate software existed. Flight control systems for aircraft, automated manufacturing systems, and sophisticated avionics systems all present difficult, nonlinear control problems. Many of these problems are currently unsolvable, not because current computers are too slow or have too little memory, but simply because it is too difficult to determine what the program should do. If a computer could learn to solve the problems through trial and error, that would be of great practical value.

Reinforcement Learning is an approach to machine intelligence that combines two disciplines to successfully solve problems that neither discipline can address individually.

Dynamic Programming is a field of mathematics that has traditionally been used to solve problems of optimization and control. However, traditional dynamic programming is limited in the size and complexity of the problems it can address.[5] reinforcement learning helps you to take your decisions sequentially. It Works on interacting with the environment. In RL method learning decision is dependent. Therefore, you should give labels to all the dependent decisions. It Supports and work well in AI, where human interaction is prevalent. Eg.Chess game. Reinforcement learning is the training of machine learning models to make a sequence of decisions. The agent learns to achieve a goal in an uncertain, potentially complex environment. In reinforcement learning, an artificial intelligence faces a game-like situation.

III. CONCLUSIONS

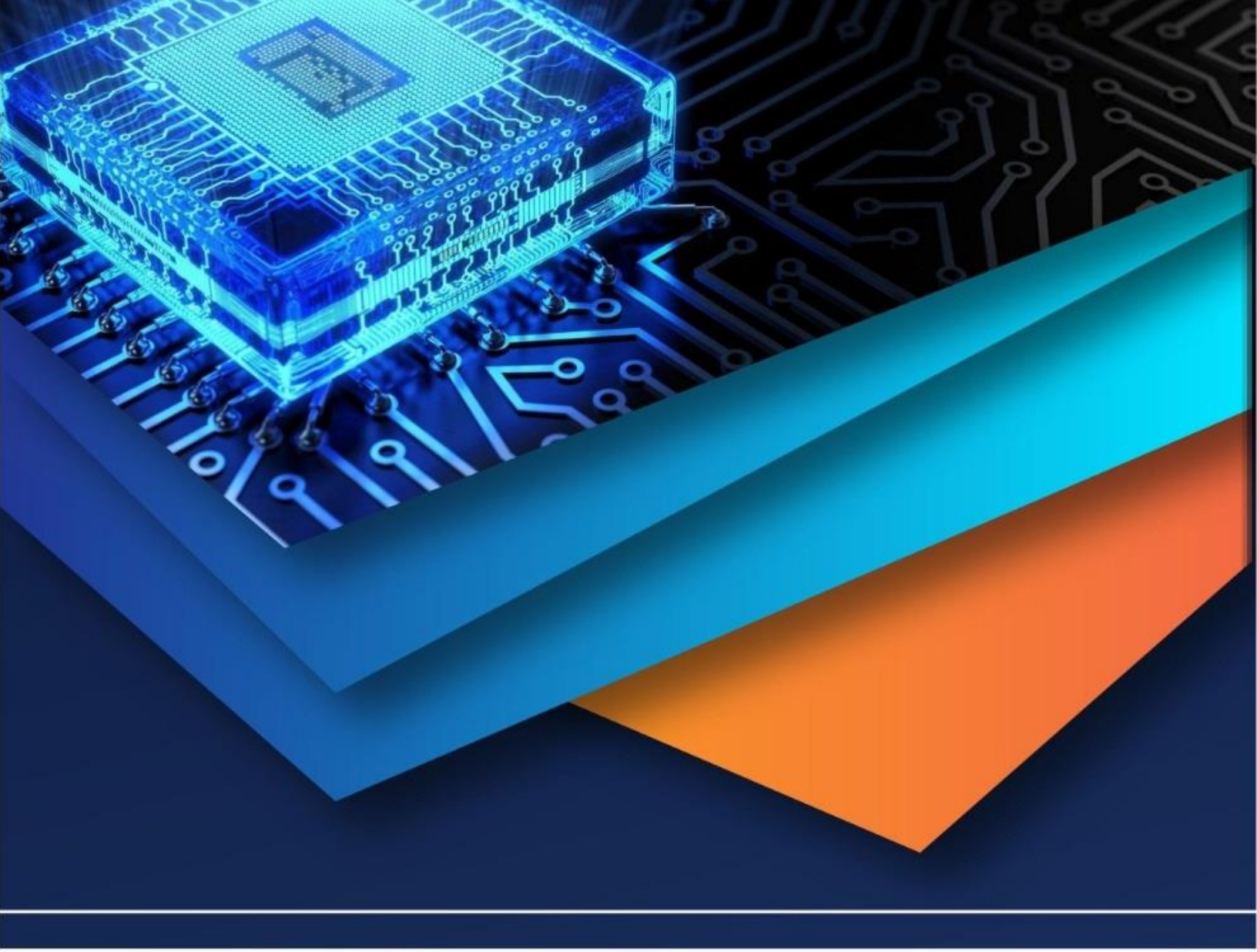
Machine Learning is a technique of training machines to perform the activities a human brain can do, albeit bit faster and better than an average human-being. Machine Learning can be a Supervised or Unsupervised. If you have lesser amount of data and clearly labelled data for training, opt for Supervised Learning. Unsupervised Learning would generally give better performance and results for large data sets. If you have a huge data set easily available, go for deep learning techniques. You also have learned Reinforcement Learning and Deep Reinforcement Learning. This paper gives an introduction to most of the popular machine learning algorithms.

IV. ACKNOWLEDGMENT

We would like to acknowledge to all our teachers in the DY Patil International University who helped me into the evolution and conclusion of our research work in machine learning. we would also like to thank staff and my classmates who have helped me indirectly or directly &who have been supportive of my career goals and who worked actively to provide us with the protected academic time to pursue those goals.

REFERENCES

[1] Jonathan Schmidt 1, Mário R. G. Marques 1, Silvana Botti2 and Miguel A. L. Marques "Recent advances and applications of machine learning in solidstate materials science", in npj Computational Materials Journal
[2] M. Welling, "A First Encounter with Machine Learning. Bowles, "Machine Learning in Python: Essential Techniques for Predictive Analytics", John Wiley & Sons Inc., ISBN: 978-1-118-96174-2
[3] R. Vijaya Kumar Reddy1 , Dr. U. Ravi Babu2 "A Review on Classification Techniques in Machine Learning ",IJARSE,Vol-7,special issue 3,March 2018
[4] Reinforcement Learning: A Tutorial , Mance E. Harmon , Stephanie S. Harmon Taiwo Oladipupo Ayodele University of Portsmouth United Kingdom
[5] Youguo Li, Haiyan Wu Department of Computer Science Xinyang Agriculture College Xinyang, Henan 464000, China' A Clustering Method Based on K-Means Algorithm' 2012 International Conference on Solid State Devices and Materials
[6] Mohammed Al-Maolegi1 , Bassam Arkok2 Computer Science, Jordan University of Science and Technology, Irbid, Jordan "an Improved apriori algorithm for association rules "
[7] Yogita Rani¹ and Dr. Harish Rohil2 'A Study of Hierarchical Clustering Algorithm' International Journal of Information and Computation Technology. ISSN 0974-2239 Volume 3, Number 11 (2013),
[8] Sidharth Prasad Mishra , Uttam Sarkar, Subhash Taraphder, Sanjay Datta, Devi Prasanna Swain1 , Reshma Saikhom, Sasmita Panda2 and Menalsh "Multivariate Statistical Data Analysis- Principal Component Analysis (PCA)"



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY