A Study on Machine Learning and Supervised Machine Learning Algorithms

[1]Dr.Ch.Smitha Chowdary, [1]A Kavitha
[1,2] Assistant Professor
[1,2] P.B.Siddhartha College of Arts & Science,Vijayawada,India

Abstract : This paper focuses on machine learning concepts. The recent advancement in technology, machine learning its types specifically, supervised machine learning and unsupervised machine learning their perspectives, essentials and classification of supervised machine learning algorithms wherein it can be applied in data analytics and artificial intelligence.

***IndexTerms* - Component,formatting,style,styling,insert.**

I. INTRODUCTION

Yesterday today or tomorrow data has its own prominence. As the time advanced advancements took place in data analysis also. Currently Machine learning has begun to make its mark in the field of Information Technology to analyze the data smartly. Machine learning is a core sub-area of artificial intelligence. SAS, a North Carolina-based analytics software developer, uses this definition: “Machine learning is a method of data analysis that automates analytical model building.”[1] Machine learning learns from data by using algorithms generates programs to analyze the data. Hence it is a self learning process. These programs operate by construction of a model from example inputs in order to make data driven predictions or choices rather than following the instructions of a static program. When any data is entered newly the programs learn, change and develop by themselves. With the ever increasing amounts of data becoming available there is a good reason to believe that smart data analysis will become even more pervasive as a necessary ingredient for technological progress. In recent years, the fields of machine learning and mathematical programming are increasingly intertwined. Machine Learning algorithm is the hypothesis set that is taken at the beginning before the training starts with real-world data.

II. TYPES OF MACHINE LEARNING

There are four types of machine learning. They are Supervised Learning, Unsupervised Learning, Semi-supervised Learning and Reinforced Learning.

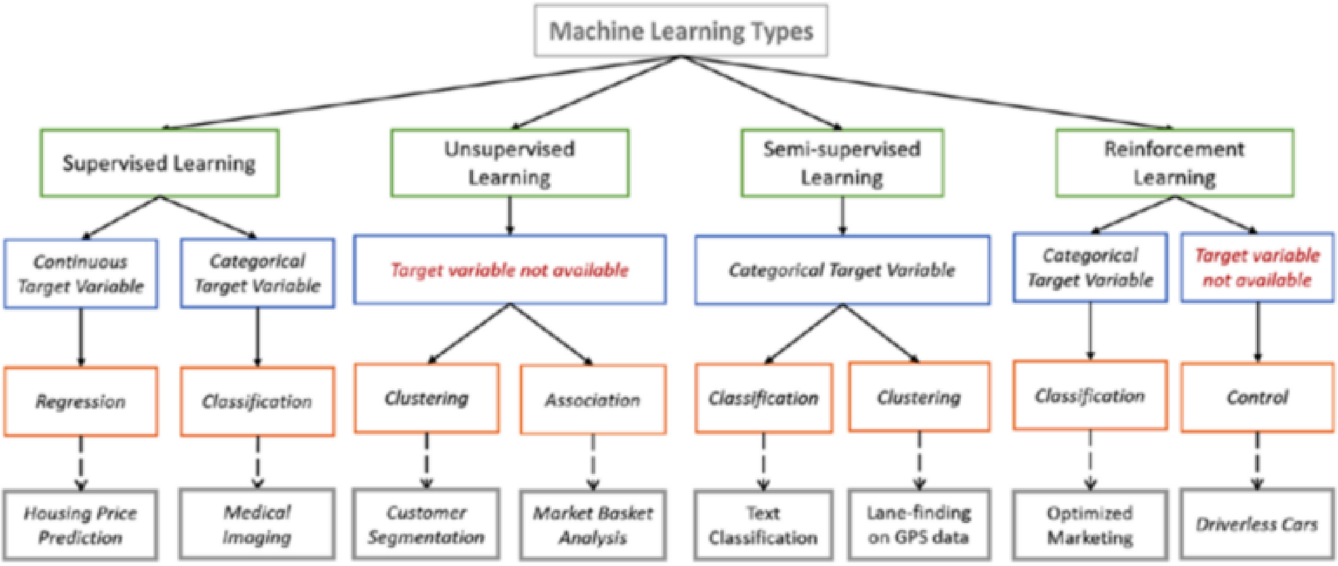
Supervised Learning: Supervised learning is full fledged type of learning widely used by most machine learning algorithms. Learning with supervision is much easier than learning without supervision. Also known as Inductive Learning.
Inductive Learning is where we are given examples of a function in the form of data (x) and the output of the function (f(x)). The goal of inductive learning is to learn the function for new data (x) [2]. The techniques to categorize data are Classification-when the function being learned is discrete. Regression-when the function being learned is continuous and Probability Estimation-when the output of the function is a probability. Some practical examples of induction are Credit risk assessment, Disease diagnosis, Face recognition, Automatic steering. Student admission, fuel for car etc

Supervised learning assists in the problems like where people are not aware of answer they cannot write a program to solve it, where humans can do things that computer cannot perform well like riding a bike or driving a car, where changes are frequent as in stock market, where it is not cost effective to write a custom program for each user. Example is recommendations of movies or books on Netflix or Amazon. We can write program that works perfectly for the data that we have. But we cannot be sure how good it works for new data. Practically it would be brave to predict anything we like but we should obtain accurate approximation of the function. By trying classes of Hypothesis results in form the solution may be or represented. The viewpoints on inductive learning are it eliminates uncertainty to good extent and guess a good and small hypothesis by using trail and error process.

Unsupervised Learning: It is the training of an artificial intelligence (AI) algorithm using information that is neither classified nor labeled and allowing the algorithm to act on that information without guidance. In unsupervised learning, an AI system may combine disordered information according to similarities and differences even though there are no categories provided. AI systems capable of unsupervised learning adopt generative learning models, along with retrieval-based approach.
Examples for unsupervised learning are Chatbots, self-driving cars, facial recognition programs, expert systems and robots are among the systems that may use either supervised or unsupervised learning approaches. In unsupervised learning, an AI system is presented with unlabeled, uncategorised data and the system’s algorithms act on the data without prior training. The output is dependent upon the coded algorithms. Subjecting a system to unsupervised learning is one way of testing AI. Compared to supervised learning algorithms difficult tasks can be done by using Unsupervised learning algorithms. Moreover they are more unpredictable.

Semi-supervised Learning: It is in-between that of Supervised and Unsupervised Learning. Where the combination is used to produce the desired results and it is the most important in real-world scenarios where all the data available are a combination of labelled and unlabeled data.
Reinforced Learning: The machine is exposed to an environment where it gets trained by trial and error method, here it is trained to make a much specific decision. The machine learns from past experience and tries to capture the best possible knowledge to make accurate decisions based on the feedback received.

Fig1: Types of Machine Learning



III. CLASSIFICATION OF SUPERVISED MACHINE LEARNING ALGORITHMS

The supervised machine learning algorithms which deals more with classification includes Regression, Logistic Regression, Naïve Bayes Classifier, Perceptron, Support Vector Machine, K-Means Clustering, Boosting, Decision Tree, Random Forest (RF), Neural networks, Bayesian Networks and so on.

1)Regression: A regression problem is when the output variable is a real or continuous value, such as "salary" or "weight". Many different models can be used, the simplest is the linear and multivariate regression. Linear models for classification separate input vectors into classes using linear decision boundaries. It is to group items that have similar feature values and depending on the margin they are quantified into groups. It is the fastest classifier and performs well for large dimensions[3].
***2)Logistic regression*:It** is used to predict the class (or category) of individuals based on one or multiple predictor variables (x). It is used to model a binary outcome, that is a variable, which can have only two possible values: 0 or 1, yes or no, diseased or non-diseased.It allows us to estimate the probability (p) of class membership. The probability will range between 0 and 1. You need to decide the threshold probability at which the category flips from one to the other. By default, this is set to p = 0.5, but in reality it should be settled based on the analysis purpose[4].Most commonly used tools for applied statistics and discrete data analysis.
3) Naive Bayesian (NB) Networks: These are very simple Bayesian networks , that explicitly apply Bayes' Theorem for problems such as classification and regression which are composed of directed acyclic graphs with only one parent (representing the unobserved node) and several children (corresponding to observed nodes) with a strong assumption of independence among child nodes in the context of their parent[5] . Bayes classifiers are usually less accurate that other more sophisticated learning algorithms. It is found that sometimes superior to the other learning schemes, even on datasets with substantial feature dependencies. To achieve maximum prediction accuracy NB may need a relatively small dataset. Naive Bayes (NB) requires little storage space during both the training and classification stages: the strict minimum is the memory needed to store the prior and conditional probabilities. Naive Bayes is naturally robust to missing values since these are simply ignored in computing probabilities and hence have no impact on the final decision.[6]
4) Multi-layer Perceptron: They are a class of pattern matching that are commonly used for regression and classification problems but are really an enormous subfield comprised of hundreds of algorithms and variations for all manner of problem types. This is a classifier in which the weights of the network are found by solving a quadratic programming problem with linear constraints, rather than by solving a non-convex, unconstrained minimization problem as in standard neural network training. Perceptron algorithm is used for learning from a batch of training instances by running the algorithm repeatedly through the training set until it finds a prediction vector which is correct on all of the training set. This prediction rule is then used for predicting the labels on the test set.
5) *Support Vector Machines (SVMs):*A Kernel Method which is really a constellation of methods in and of itself. Kernel Methods are concerned with mapping input data into a higher dimensional vector space where some classification or regression problems

are easier to model.These are the most recent supervised machine learning technique .Support Vector Machine (SVM) models are closely related to classical multilayer perceptron neural networks.[7] SVMs revolve around the notion of a margin either side of a hyperplane that separates two data classes.[SVMs perform well when multi-collinearity is present and a nonlinear relationship exists between the input and output features. SVMs construct a hyperplane that separates two classes. Essentially, the algorithm tries to achieve maximum separation of the classes. Supervised machine learning techniques are applicable in numerous domains. SVMs perform much better when dealing with multi-dimensions and continuous features.

6)*K-means:* K-means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. K-Means algorithm is be employed when labeled data is not available. General method of converting rough rules of thumb into highly accurate prediction rule. Given weak learning algorithm that can consistently find classifiers (rules of thumb) at least slightly better than random, say, accuracy 55%, with sufficient data, a boosting algorithm can provably construct single classifier with very high accuracy, say, 99% .

7)*Decision Trees:* Decision Trees (DT) are trees that classify instances by sorting them based on feature values. Each node in a decision tree represents a feature in an instance to be classified, and each branch represents a value that the node can assume. Instances are classified starting at the root node and sorted based on their feature values. Decision tree learning, used in data mining and machine learning, uses a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value. More descriptive names for such tree models are classification trees or regression trees. A validation set is employed to prune the performance of decision trees. Any node can be removed and assigned the most common class of the training instances that are sorted to it.

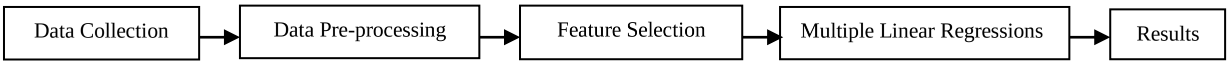
8) *Neural Networks:* Neural Networks (NN) that can actually perform a number of regression and/or classification tasks at once. Neural networks have three layers: an input, hidden, and output layer. Each layer is made up of nodes. The layers are connected by vectors. Neural networks were one of the first machine learning models to be created. Neural networks require complete records to do their work.

9)Random Forest: Random Forest an ensemble of decision trees. To classify a new object based on attributes, each tree gives a classification and we say the tree “votes” for that class.[8] The forest chooses the classification having the most votes (over all the trees in the forest).Each tree is planted & grown as if the number of cases in the training set is N, then sample of N cases is taken at random but with replacement. This sample will be the training set for growing the tree. If there are M input variables, a number m<<M is specified such that at each node, m variables are selected at random out of the M and the best split on these m is used to split the node. The value of m is held constant during the forest growing. Each tree is grown to the largest extent possible. There is no pruning.

IV. Methodology Considered

In here, the dataset is collected from the regressit[9] . This dataset was taken from the StatLib library which is maintained at Carnegie Mellon University. This dataset relates to multiple regression. The Proposed layout is shown in Fig. 2.

Fig.2: Flow of Data in Multiple Regressions.



Data Collection: The dataset is collected from the regress it. A data science site that contains a variety of externally contributed interesting datasets. The dataset selected is Automobile fuel economy [9]

Data Pre-processing: refers to the transformations applied to our data before feeding it to the algorithm. Data Preprocessing is a technique that is used to convert the raw data into a clean data set.

Feature Selection: It is the process where you automatically or manually select those features which contribute most to your prediction variable or output in which you are interested in. Having irrelevant features in your data can decrease the accuracy of the models and make your model learn based on irrelevant features.

Regression: In simple linear regression there is a one-to-one relationship between the input variable and the output variable. But in multiple linear regression, as the name implies there is a many-to-one relationship, instead of just using one input variable, you use several.

Fundamentally multiple linear regression works on OLS (Ordinary least squares) principle. The regression equation will take a shape like:

$Y=B_0+B_1X_1+B_2X_2+B_3X_3$..... Where, B_i is Different coefficients and X_i is various independent variables.

V. Results and Analysis

A case is taken where have to predict a car's fuel consumption from its physical attributes. (Automobile fuel economy dataset)[1]. Excel is used as a platform to run the regression on the dataset. The data set contains 392 rows. The data concerns city-cycle fuel consumption in miles per gallon, to be predicted in terms of 3 multi valued discrete and 5 continuous attributes. There are 9 variables in data out of which GallonsPer100Miles is a dependent variable while rests of them are predictors or independent variables.

Computing parameters:

In multivariate linear regression initially one should focus on selecting the best possible independent variables that contribute well to the dependent variable. For this a correlation matrix will be constructed for all the independent variables and the dependent variable from the observed data. The correlation value gives us an idea about which variable is significant and by what factor. From this matrix we pick independent variables in decreasing order of correlation value and run the regression model to estimate the coefficients by minimizing the error function. We stop when there is no prominent improvement in the estimation function by inclusion of the next independent feature. This method can still get complicated when there are large no.of independent features that have significant contribution in deciding our dependent variable. The calculated correlation between our dependent & independent variables using regression tool is presented below. Here developed a least-squares regression equation to predict car's fuel consumption from its physical attributes. Have assessed how well the regression equation predicts fuel consumption the dependent variable and also assessed the contribution of each independent variable to the prediction. The Regression Equation is

$$\hat{y} = b0 + b1x1 + b2x2 + b3x3 + b4x4 + b5x5 + b6x6 + b7x7 + b8x8 + b9x9$$

Table 1:

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	7.913886844	0.698315638	11.33282203	7.32453E-26	6.540863223	9.286910465	6.540863223	9.286910465
X Variable 1	-0.110900272	0.00666805	-16.63159031	4.10748E-47	-0.124010948	-0.097789597	-0.124010948	-0.097789597
X Variable 2	0.107616263	0.042641093	2.523768854	0.012015099	0.023775626	0.191456901	0.023775626	0.191456901
X Variable 3	-0.053317218	0.099184772	-0.537554476	0.59119767	-0.248333665	0.141699229	-0.248333665	0.141699229
X Variable 4	1.022394203	0.181581386	5.630501146	3.48642E-08	0.665370074	1.379418333	0.665370074	1.379418333
X Variable 5	0.389377182	0.095635845	4.071456486	5.68135E-05	0.201338611	0.577415753	0.201338611	0.577415753
X Variable 6	0.041976688	0.012913334	3.250646767	0.001253622	0.016586575	0.067366801	0.016586575	0.067366801
X Variable 7	-0.050262355	0.008908233	-5.642236042	3.27397E-08	-0.067777665	-0.032747045	-0.067777665	-0.032747045
X Variable 8	-0.195392763	0.096328681	-2.028396538	0.043213131	-0.384793584	-0.005991943	-0.384793584	-0.005991943
X Variable 9	0.065198144	0.037781049	1.725683796	0.085212981	-0.009086706	0.139482995	-0.009086706	0.139482995

Here, we see that the regression intercept (b0) is 7.913886844, the regression coefficient for x variable 1 (b1) is -0.110900272, X Variable 2(b2) is 0.107616263, X Variable 3(b3) is -0.053317218, X Variable 4 (b4) is 1.022394203, X Variable 5 (b5) is 0.389377182, X Variable 6 (b6) is 0.041976688, X Variable 7 (b7) is -0.050262355, X Variable 8 (b8) is -0.195392763and X Variable 9 (b9) is 0.065198144. So the least-squares regression equation can be re-written as:

$$\hat{y}=7.913886844+(-0.110900272)x1+(0.107616263)x2+(-0.053317218)x3+(1.022394203)x4+(0.389377182)x5+ (0.041976688)x6 + (-0.050262355)x7+ (-0.195392763)x8 + (0.065198144)x9$$

This is the only linear equation that satisfies a least-squares criterion. That means this equation fits the data from which it was created better than any other linear equation. The fact that our equation fits the data better than any other linear equation does not guarantee that it fits the data well. So need to know how well does our equation fit the data? Hence we look at the coefficient of multiple determination (R2). The coefficient of multiple determination measures the proportion of variation in the dependent variable that can be predicted from the set of independent variables in the regression equation. When the regression equation fits the data well, R2 will be large (i.e., close to 1); and vice versa.

Table 2: Regression

Regression Statistics	
Multiple R	0.966200614
R Square	0.933543627
Adjusted R Square	0.931977901
Standard Error	0.433973958
Observations	392

A quick glance at the output suggests that the regression equation fits the data pretty well. The coefficient of muliple determination is 0.966200614. For our sample problem, this means 96.6% fuel consumption variation can be explained by independent variables in study. Another way to evaluate the regression equation would be to assess the statistical significance of the regression sum of squares. For that, we examine the ANOVA table produced by Excel:

Table 3: Anova Statistics

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	9	1010.621856	112.2913174	596.2368851	8.6355E-219
Residual	382	71.94335725	0.188333396		
Total	391	1082.565214			

This table tests the statistical significance of the independent variables as predictors of the dependent variable. The last column of the table shows the results of an overall F test. The F statistic (596.2368851) is big, and the p value (8.6355E-219) is small. This indicates that one or both independent variables has explanatory power beyond what would be expected by chance. This concludes that the regression equation fits the data well.

With multiple regressions, there is more than one independent variable; so it is natural to ask whether a particular independent variable contributes significantly to the regression after effects of other variables are taken into account. The answer to this question can be found in the regression coefficients table: 1. The regression coefficients table shows the information of each coefficient: its value, its standard error, a t-statistic, and the significance of the t-statistic.

VI. CONCLUSION:

In this paper we studied about Machine learning a rapidly growing field in computer science. Machine learning approaches applied in systematic reviews of complex research fields such as quality improvement has applications in nearly every other field of study and is already being implemented commercially because machine learning can solve problems too difficult or time consuming for humans to solve. To describe machine learning in general terms, a variety models are used to learn patterns in data and make accurate predictions based on the patterns it observes. Increased reviewer agreement appeared to be associated with improved predictive performance. This paper also studied the best known supervised techniques in relative detail. Prediction is made using regression analysis. Now we are aware of which method performs well under which situation. Which of the algorithms are to be integrated to solve a problem can be identified now as we are aware of the strengths and limitations of the algorithms.

REFERENCES

[1] http://californiadatascience.com/machine-learning/

[2] https://machinelearningmastery.com/basic-concepts-in-machine-learning/

[3] Emerging Artificial Intelligence Applications in Computer Engineering , Ilias G Maglogiannis, Kostas Karpouzis , Manolis Wallace , John Soldatos,IOS press:SB Kostiantis ,Supervised learning:A review of classification techniques.

[4] Machine Learning Essentials: Practical Guide in R1, Alboukadel Kassambara.

[5] Book:Datamining Technique,Practical machine Learning Tools and Techniques.Ian H Witten.Eibe Frank,Mark A Hall.Christopher J Pal

[6] Osisanwo F.Y, Akinsola J.E.T, Awodele O, Hinmikaiye J. O, Olakanmi O, Akinjobi J. “Supervised Machine Learning Algorithms: Classification and Comparison” , International Journal of Computer Trends and Technology (IJCTT), Volume 48, Number 3. ISSN: 2231-2803, June 2017,Pg 128.

[7] Divyansh Khanna,Rohan Sahu, Veeky Baths, and Bharat DeshpandeComparative Study of Classification Techniques (SVM, Logistic Regression and Neural Networks) to Predict the Prevalence of Heart Disease International Journal of Machine Learning and Computing, Vol. 5, No. 5, October 2015,414,DOI: 10.7763/IJMLC.2015.V5.544.

[8] https://www.datasciencecentral.com/profiles/blogs/a-tour-of-machine-learning-algorithms

[9] https://regressit.com/data.html