Journal of Analysis and Computation (JAC)
(An International Peer Reviewed Journal), www.ijaconline.com, ISSN 0973-2861
Volume XVII, Issue II, July-Dec. 2023

SUPERVISED MACHINE LEARNING ALGORITHMS – A REVIEW

K. Sharmila

Assistant Professor, Department of Computer Science, Thiagarajar College, Madurai, India

ABSTRACT

Machine learning is a type of Artificial Intelligence (AI) that empowers a framework to gain from information programming that has been a key component of digitalization solutions. Machine Learning is a very powerful tool, used to solve real-world problems by processing large data, and manipulating, extracting and retrieving data from large real-world sources. It focuses on the development of computer programs that can read and access data and use it to learn for themselves without human intervention. In this paper, various supervised machine learning algorithms are discussed.

Keywords -Machine learning, Supervised, Regression, Classification

[1] INTRODUCTION

Machine learning (ML) is an application of Artificial Intelligence (AI) which uses algorithms and statistics to find patterns in large amounts of data. Data can be in any format such as text, numbers, images, etc. Machine learning is used to teach machines how to handle large data more efficiently. Thus the machine takes decisions and does forecasting or predictions based on data. ML software parses these data and then learns and make predictions from it by applying patterns. When the computer is given a completely new set of images, it will be able to predict each correct label based on previously acquired experience. In medical field, Machine learning helps doctors to detect diseases at its earliest stages. Medical professionals can detect somatic mutations easily with the help of machine learning tools. In preventive genetics, Machine learning plays an important role. Scientists and researchers rely on ML algorithms to determine how environmental factors, drugs and chemicals influence the human genome.

Journal of Analysis and Computation (JAC)
(An International Peer Reviewed Journal), www.ijaconline.com, ISSN 0973-2861
Volume XVII, Issue II, July-Dec. 2023

Machine Learning is applied in wide variety of fields namely: Image Recognition, Speech Recognition, Traffic Prediction, finance sector, products recommendation, self-driving cars, health care industry, medical diagnosis, Stock market trading, fraud detection, travel industry, social media, etc. Machine Learning (ML) relies on various algorithms to solve large data problems. Data scientists and analyst says that various type of algorithm can be used to solve various data problems that is best suitable to solve a particular problem. Here is a quick review at some of the commonly used machine learning (ML) algorithms.

[2] SUPERVISED LEARNING

Supervised learning is a form of machine learning in which the algorithm is trained on labelled data to make decisions or predictions based on the data inputs. In supervised learning, the algorithm tries to learn the relationship between the input and output data so that it can make accurate predictions on unseen new data. Supervised learning is a ML technique that is widely used in various fields such as healthcare, finance, image classification, spam filtering**,** risk assessment, fraud detection, spam filtering and more. The model can predict the output on the basis of prior experiences with the help of supervised learning.

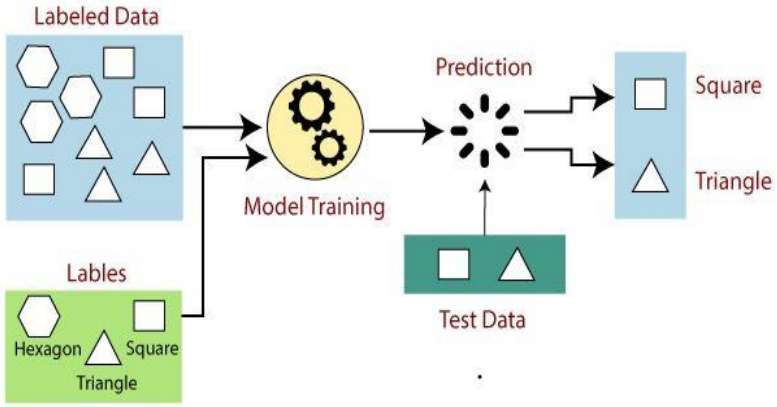

Figure 1: Example for Supervised Learning

Steps involved in Supervised Machine Learning:

Step 1: First determine the type of training dataset.
Step 2: Collect the labelled training data.
Step 3: Split the dataset into training dataset, test dataset, and validation dataset.
Step 4: Determine the input features of the training dataset that should have enough knowledge so that the model can accurately predict the output.
Step 5: Determine the suitable algorithm for the model, such as Linear Regression, Logistic Regression, Support Vector Machine, Decision Tree, etc.
Step 6: Execute the algorithm on the training dataset.
Step 7: Evaluate the accuracy of the model by providing the test set. Our model is accurate, if the model predicts the correct output.

Classification

Classification is a type of supervised machine learning that categorizes input data into predefined labels. It involves training a model on labelled examples to learn patterns between input variables and output variable. Classification models are used to predict categorical discrete values such as green or yellow, yes or no, default or no default and so on. If the outcome can take two possible values, it is known as Binary Classification. When the outcome contains more than two possible values, such as classifying the web text into one of the following: entertainment, sports or technology, it is known as Multiclass Classification. In classification, the target variable is a categorical value. For example, predicting whether it will rain or not on a particular day. The goal of this model is to generalize this learning to make accurate predictions on unseen, new data. Algorithms like Logistic Regression, Support Vector Machines, Decision Tree Classifier, Random Forest Classifier, K Nearest Neighbor Classifier and Neural Networks are commonly used for classification tasks.

Regression

Regression is a type of supervised machine learning technique used to predict continuous numerical values such as price, salary, sales, temperature based on input features. A dataset containing features of the pre-owned cars such as specification details, condition of the car, and kilometre used, and the price of the pre-owned car, a Regression algorithm can be trained to learn the relationship between the features and the price of the pre-owned car. The goal of regression is to minimize the difference between predicted and actual values using algorithms like Linear Regression, K Nearest Neighbor Regressor, Decision Trees, Neural Networks or Random Forest Regressor ensuring the model captures some patterns in the data. There are some common Supervised Machine Learning Algorithm that can be used for both classification and regression task.

A. Linear Regression

Linear regression is one of the simplest and most popular Machine Learning algorithms that is used to predict a continuous output value. Linear regression algorithm shows a linear relationship between one or more independent (x) variables and a dependent (y) variable. It is a method that is used for predictive analysis. Linear regression makes predictions for numeric or continuous variables such as age, sales, price of a product, etc. The linear regression model provides a sloped straight line representing the relationship between the dependent and independent variables.

Journal of Analysis and Computation (JAC)
(An International Peer Reviewed Journal), www.ijaconline.com, ISSN 0973-2861
Volume XVII, Issue II, July-Dec. 2023

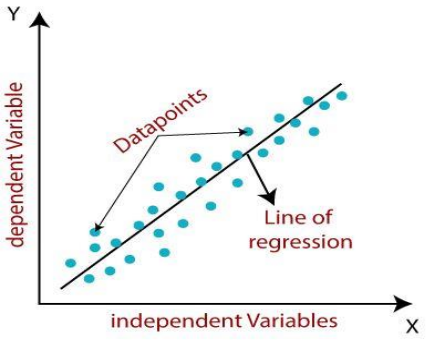

Figure 2: Linear Regression

Linear regression can be represented as

$$\mathbf{Y=a_0+a_1x+\varepsilon}$$

Where,

Y is the Dependent Variable (Target Variable)

X is the Independent Variable (predictor Variable)

a_0 is the intercept of the line

a_1 is the Linear regression coefficient

ε is the random error

B. Logistic Regression

Logistic regression is one of the most popular supervised machine learning technique mainly used for classification task where the aim is to predict the probability that an instance of belonging to a given class. In logistic regression, the algorithm tries to find a linear relationship between the independent variables and the output (dependent) variable. Then the output variable is transformed using a logistic function to generate a probability value between 0 and 1. It is commonly used for decision making in machine learning applications where the output variable is either yes or no, email spam or not, etc. Regression problems can be solved using linear regression, whereas classification problems can be solved using Logistic regression. In Logistic regression, we fit an S shaped logistic function that predicts two maximum values (0 or 1).

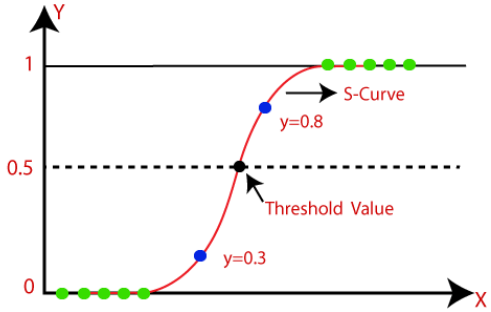

Figure 3: Logistic Regression

The S shaped curve from the logistic function (sigmoid function) indicates the likelihood of something such as whether a person is obese or not based on its weight, the cells are cancerous or not, etc. In logistic regression, the concept of the threshold value is used, that defines the

Journal of Analysis and Computation (JAC)
(An International Peer Reviewed Journal), www.ijaconline.com, ISSN 0973-2861
Volume XVII, Issue II, July-Dec. 2023

probability of either 0 or 1, values below the threshold value tends to 0, and a value above the threshold values tends to 1.

C. Decision Tree

Decision Tree is a Supervised learning technique that can be applied for both regression and classification problems, but mostly it is suitable for solving Classification problems. Decision tree is a tree structured classifier that is used to model decisions and their possible consequences. Each internal node in the tree represents features of dataset, branches represent the decision rules, while each leaf node represents a possible outcome. The tests (decisions) are performed on the basis of features of the given dataset. The decision tree algorithm can be utilized to model complex relationships between input (independent) features and output (dependent) variables. In machine learning, the decision tree as classifier is utilized to train the model based on the categorical label, while the Decision Tree as regressor is utilized to train the model based on a non-categorical label.

Steps involved in Decision Tree:

Step 1: Begin the tree with the root node, DT that contains the complete dataset.
Step 2: Find the best variable in the dataset.
Step 3: Divide the DT into subsets that contains possible values for the best variables.
Step 4: Generate the decision tree node that contains the best variable.
Step 5: Recursively make new decision trees, by using the subsets of datasets that are obtained in step 3. Continue this process until a stage is reached where the tree cannot be further classified.

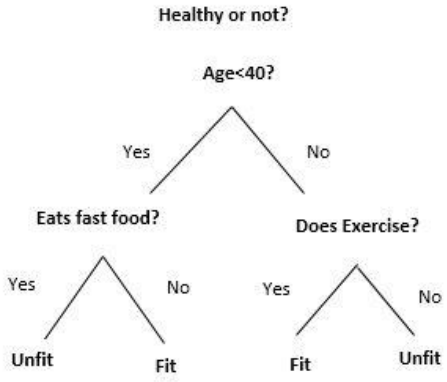

Figure 4: Decision Tree

D. Random Forest Algorithm

Random Forest Algorithm is a supervised machine learning algorithm which is very popular and is used for regression and classification problems in Machine Learning. Random forests algorithms are made up of multiple decision trees which work together to make predictions. Each tree in the random forest is trained on a different subset of the input features and data. The final prediction is made by combining the predictions of all the trees in the random forest. Random forests are an ensemble learning technique which is used for both regression and classification tasks. *Random Forest Regression* combines multiple decision trees to reduce

overfitting and improve prediction accuracy. *Random Forest Classifier* contains several decision trees on various subsets of the given dataset and takes the average to improve the prediction accuracy of that dataset while minimizing overfitting.

Steps involved in Random Forest Algorithm:

Step 1: Random samples are selected from a given data or training set.
Step 2: Random Forest algorithm will build a decision tree for every training data.
Step 3: Then Voting will take place by averaging the decision tree.
Step 4: Final prediction result will be selected based on the majority voting.

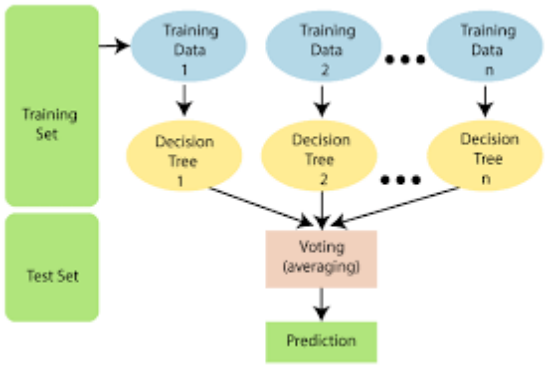

Figure 5: Random Forest

E. Naïve Bayes Classifier Algorithm

Naïve Bayes algorithm is a supervised machine learning technique that is based on Bayes theorem and it is used for solving classification problems. It is mainly utilized in *text classification* which includes a high-dimensional training dataset. Naïve Bayes is a probabilistic classifier that means it predicts on the basis of the probability of an object**.** Some popular examples of Naïve Bayes Algorithm are Sentimental analysis, spam filtration, etc**.** It is named Naïve since it assumes that the occurrence of a certain feature is independent of the occurrence of other features. It is named Bayes since it depends on the principle of Bayes' Theorem. Bayes' theorem is also known as *Bayes' law* or *Bayes' rule* that is used to determine the probability of a hypothesis with prior knowledge. It depends on the conditional probability.

The formula for Bayes' theorem is:

$$P(A|B)=\frac{P(B|A)P(A)}{P(B)}$$

F. K-Nearest Neighbors (K-NN)

K-Nearest Neighbor (K-NN) is one of the simplest supervised Machine Learning technique. KNN algorithm assumes the similarity between the new case or data and available cases and put the new case into the category, which is most similar to the available categories. K-Nearest Neighbor algorithm stores all the available data and based on the similarity, it classifies a new data point. K-NN algorithm can be used for classification as well as for regression but mostly

K-NN algorithm is used for the classification problems. K-NN works by finding k training examples closest to a given input, then the class or value is predicted based on the majority class or average value of these neighbors. The choice of k and the distance metric used to measure proximity influence the performance of K-NN. A K-Nearest Neighbors (K-NN) is a type of algorithm which is used for both regression and classification tasks. *K-Nearest Neighbor (K-NN) Regression* predicts the continuous values by averaging the outputs of the k closest neighbors. In *K-Nearest Neighbors (K-NN) Classification*, data points are classified based on the majority class of their k closest neighbors.

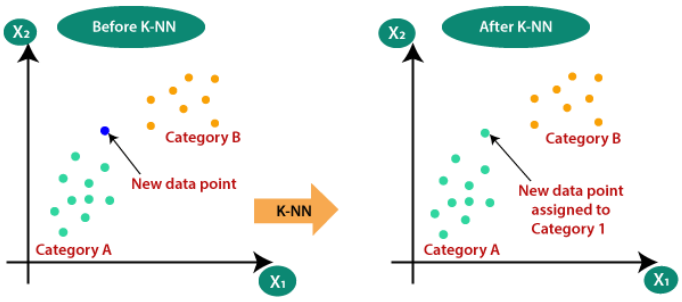

Figure 6: K-Nearest Neighbors

The steps involved in K-NN algorithm:

Step 1: Select the k number of neighbors.
Step 2: Calculate the Euclidean distance (distance between two points) of k number of neighbors.
Step 3: Select the k nearest neighbors as per the Euclidean distance.
Step 4: Count the number of the data points in each category among these k neighbors.
Step 5: Assign the new data points to that category for which has the maximum number of neighbors.
Step 6: K-NN model is ready.

G. Support Vector Machine (SVM)

Support Vector Machine (SVM) is one of the most popular Supervised Learning technique that is used for regression as well as classification problems. However, it is mainly used for classification problems in Machine Learning. The goal of SVM algorithm is to create hyperplane to segregate n-dimensional space into classes and to identify the correct category of new data points. Support Vector Machine chooses the extreme points or vectors that help in creating the hyperplane. These extreme cases are termed as support vectors, and hence the technique is named as Support Vector Machine (SVM). *Support Vector Regression* is used for predicting continuous values. *Support Vector Classifier* aims to find the best hyperplane which maximizes the margin between data points of different classes.

Consider the figure in which there are two different categories which are classified using a decision boundary or hyperplane:

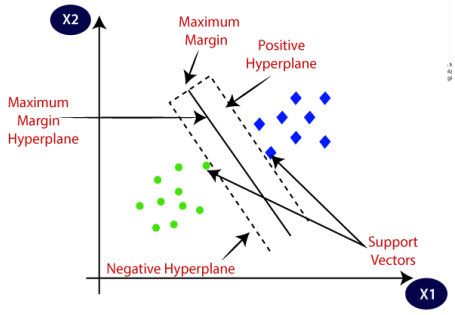

Figure 7: Support Vector Machine

Support Vector Machine can be of two types, *Linear SVM* and *Non-linear SVM*. Linear SVM is utilized for linearly separable data, that means if a dataset can be classified into two classes by using a single straight line, then such data is called as linearly separable data, and classifier used is termed as Linear SVM classifier. Non-Linear SVM is utilized for non-linearly separated data that means if a dataset cannot be classified by using a straight line, then such data is called as non-linear data and classifier used is termed as Non-linear SVM classifier.

[3] CONCLUSION

In this paper an attempt was made to review most frequently used supervised machine learning algorithms to solve classification and regression problems. Today everyone is using machine learning knowingly or unknowingly, from getting a recommended product in online shopping to updating or posting photos in social networking sites. This paper gives an introduction to most of the popular supervised machine learning techniques. It is expected that it will give insight to the readers and then selecting the appropriate supervised machine learning algorithm in the specific problem solving context.

REFERENCES

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

[2] Batta Mahesh, "Machine Learning Algorithms – A Review", International Journal of Science and Research (IJSR) ISSN: 2319-7064, Volume 9 Issue 1, January 2020.

[3] https://www.geeksforgeeks.org/supervised-machine-learning

[4] https://www.javatpoint.com/supervised-machine-learning

[5] https://www.simplilearn.com/tutorials/machine-learning-tutorial

[6] https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm

[7] Susmita Ray, "A Quick Review of Machine Learning Algorithms", 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (Com-IT-Con), India, 14th -16th Feb 2019

[8] A. Kavitha, "A Review on Machine Learning Algorithms and their Application, International

Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056, p-ISSN: 2395-0072, Volume: 07 Issue: 03 | Mar 2020