DOI: 10.56452/2021SP-8-001

ISSN: 0974-5823 Vol. 6 (Special Issue 3, November 2021)

International Journal of Mechanical Engineering

Review of Machine Learning Algorithms

Sonia Bhukra[1] and Anchit[2]

[1]Department of Electronics and Communication Engineering, Chandigarh Engineering College, Jhanjeri, Mohali-140307, Punjab, India

[2]Chandigarh School of Business, Jhanjeri-140307, Punjab, India

Abstract: Machine learning is predominantly an area of Artificial Intelligence which has been a key component of digitalization solutions that has caught major attention in the digital arena**.** This paper provides an extensive review of studies related to expert estimation of software development using Machine-Learning Techniques (MLT). Machine learning in this new era, is demonstrating the promise of producing consistently accurate estimates. Machine learning system effectively "learns" how to estimate from training set of completed projects. The main goal and contribution of the review is to support the research on expert estimation, i.e. to ease other researchers for relevant expert estimation studies using machine-learning techniques. This paper presents the most commonly used machine learning techniques such as neural networks, case based reasoning, classification and regression trees, rule induction, genetic algorithm & genetic programming for expert estimation in the field of software development.

Keywords: Machine Learning Techniques (MLT), Neural Networks (NN), Case Based Reasoning (CBR), Classification and Regression Trees (CART), Rule Induction, Genetic Algorithms and Genetic Programming.

Introduction

Since their evolution, humans have been using many types of tools to accomplish various tasks in a simpler way. The creativity of the human brain led to the invention of different machines. These machines made the human life easy by enabling people to meet various life needs, including travelling, industries, and computing. And

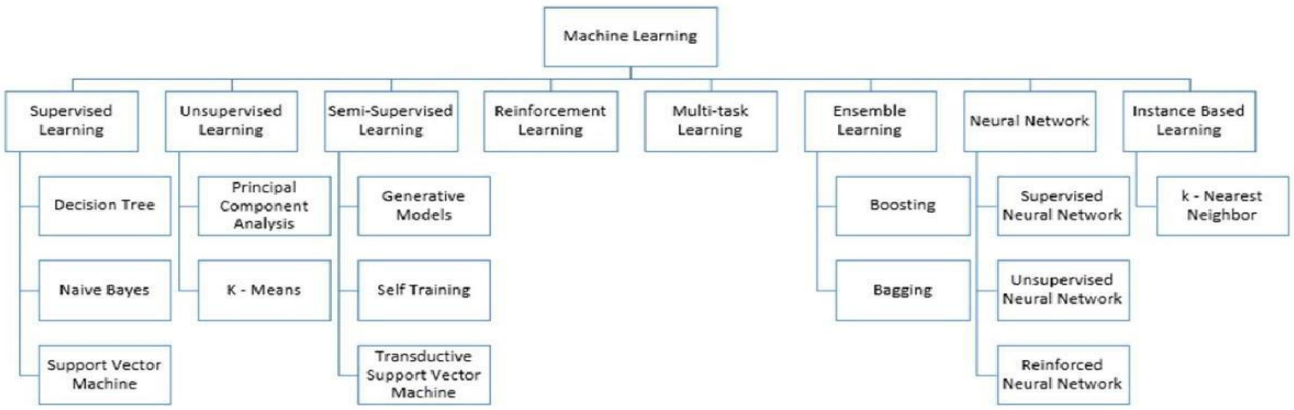

Figure:1 Types of Learning

Machine learning is the one among them. The purpose of machine learning is to learn from the data. Many studies have been done on how to make machines learn by themselves. Machine learning is used to teach machines how to handle the data more efficiently. Sometimes after viewing the data, we cannot interpret the extract information from the data. In that case, we apply machine learning. With the abundance of datasets available, the demand for machine learning is in rise. Many industries apply machine learning to extract relevant data. The purpose of machine learning is to learn from the data. Many studies have been done on how to make machines learn by themselves without being explicitly programmed. Many mathematicians and programmers apply several approaches to find the solution of this problem which are having huge data sets.

DOI: 10.56452/2021SP-8-001

Machine Learning relies on different algorithms to solve data problems. Data scientists like to point out that there's no single one- size-fits-all type of algorithm that is best to solve a problem. The kind of algorithm employed depends on the kind of problem you wish to solve, the number of variables, the kind of model that would suit it best and so on. Here's a quick look at some of the commonly used algorithms in machine learning (ML).

Supervised Neural Network*:*

Supervised learning is the machine learning task of learning a function that maps an input to an output based on example input- output pairs. It infers a function from labelled training data consisting of a set of training examples. The supervised machine learning algorithms are those algorithms which needs external assistance. In the supervised neural network, the output of the input is already known. The predicted output of the neural network is compared with the actual output. Based on the error, the parameters are changed, and then fed into the neural network again. Figure 2 will summarize the process. Supervised neural network is used in feed forward neural network.

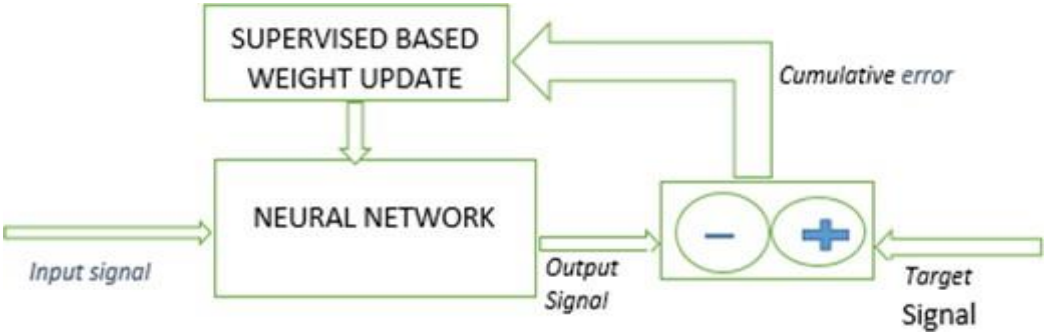

Figure: 2 Process of Supervised Neural Network

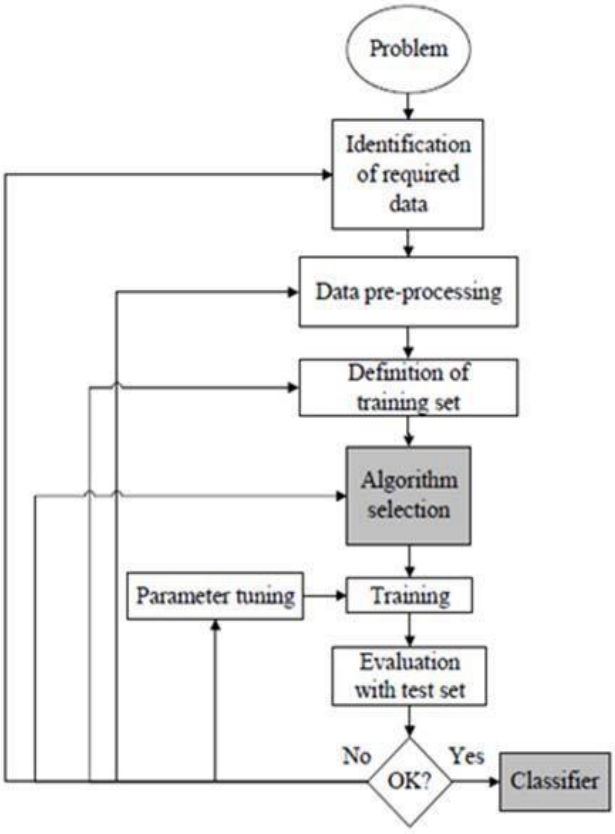

Figure: 3 Work flow of supervised machine learning algorithm

i) Decision Tree : Decision tree is a graph to represent choices and their results in form of a tree. The nodes in the graph represent an event or choice and the edges of the graph represent the decision rules or conditions. Each tree consists of nodes and branches. Each node represents attributes in a group that is to be classified and each branch represents a value that the node can take**.** Decision Tree has the following advantages : it is suitable for regression as well as classification problem, ease in interpretation, ease of handling categorical and quantitative values, capable of filling missing values in attributes with the most probable value, high performance due to efficiency of tree traversal algorithm.

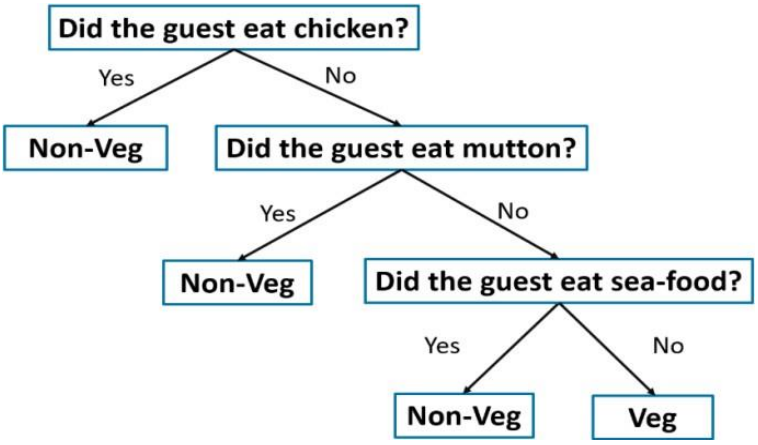

Figure 3: Decision Tree

Disadvantages of decision tree is that it can be unstable, it may be difficult to control size of tree, it may be prone to sampling error and it gives a locally optimal solution- not globally optimal solution. Decision Trees can be used in applications like predicting future use of library books and tumor prognosis problems.

Naive Bayes: It is a classification technique based on Bayes Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. The "probability table" is based on its feature values where one needs to look up the class probabilities for predicting a new observation. The basic assumption is of conditional independence and that is why it is called "naive". In real world context the assumption that all input features are independent from one another can hardly hold true.

Naive Bayes (NB) have the following advantages: implementation is easy, gives good performance , works with less training data, scales linearly with number of predictors and data points, handles continuous and discrete data, can handle binary and multi- class classification problems, make probabilistic predictions. It handles continuous and discrete data. It is not sensitive to irrelevant features.

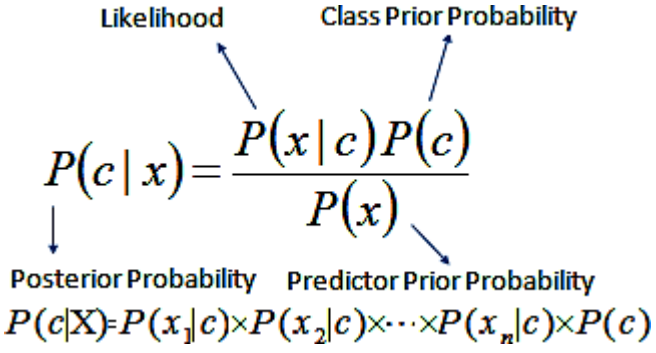

Figure 4: Navie Bayes

Support Vector Machine: Support Vector Machines (SVM) can handle both classification and regression problems. In this method hyperplane needs to be defined which is the decision boundary. When there are a set of objects belonging to different classes then decision plane is needed to separate them. The objects may or may not be linearly separable in which case complex mathematical functions called kernels are needed to separate the objects which are members of different classes. SVM aims at correctly classifying the objects based on examples in the training data set. SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces. It basically, draw margins between the classes. The margins are drawn in such a fashion that the distance between the margin and the classes is maximum and hence, minimizing the classification error.

Semi Supervise Learning: Semi-supervised machine learning is a combination of supervised and unsupervised machine learning methods. It can be fruit-full in those areas of machine learning and data mining where the unlabeled data is already present and getting the labeled data is a tedious process.

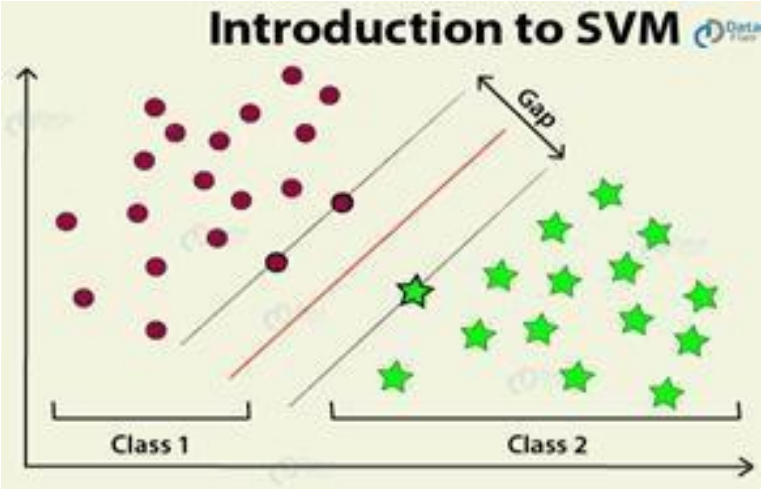

Figure 5: Support Vector Machine

Unsupervised Learning: These are called unsupervised learning because unlike supervised learning above there is no correct answers and there is no teacher. Algorithms are left to their own devises to discover and present the interesting structure in the data. The unsupervised learning algorithms learn few features from the data. The unsupervised learning algorithms learns few features from the data. When new data is introduced, it uses the previously learned features to recognize the class of the data. It is mainly used for clustering and feature reduction.

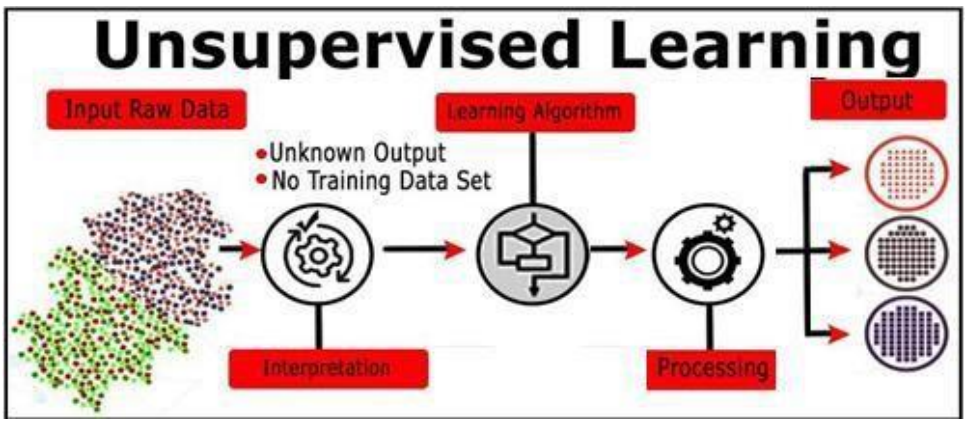

Figure 6: Unsupervised Learning

DOI: 10.56452/2021SP-8-001

An example of workflow of unsupervised learning is given in Fig. 7

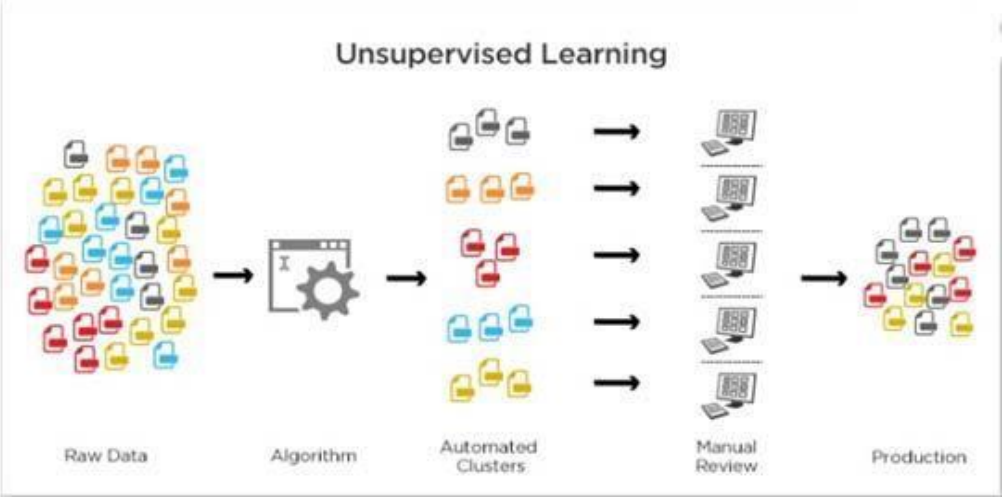

Figure 7: Example of Unsupervised Learning

Principal Component Analysis In Principal Component Analysis or PCA, the dimension of the data is reduced to make the computations faster and easier. Principal component analysis is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. In this the dimension of the data is reduced to make the computations faster and easier. It is used to explain the variance-covariance structure of a set of variables through linear combinations. It is often used as a dimensionality-reduction technique.

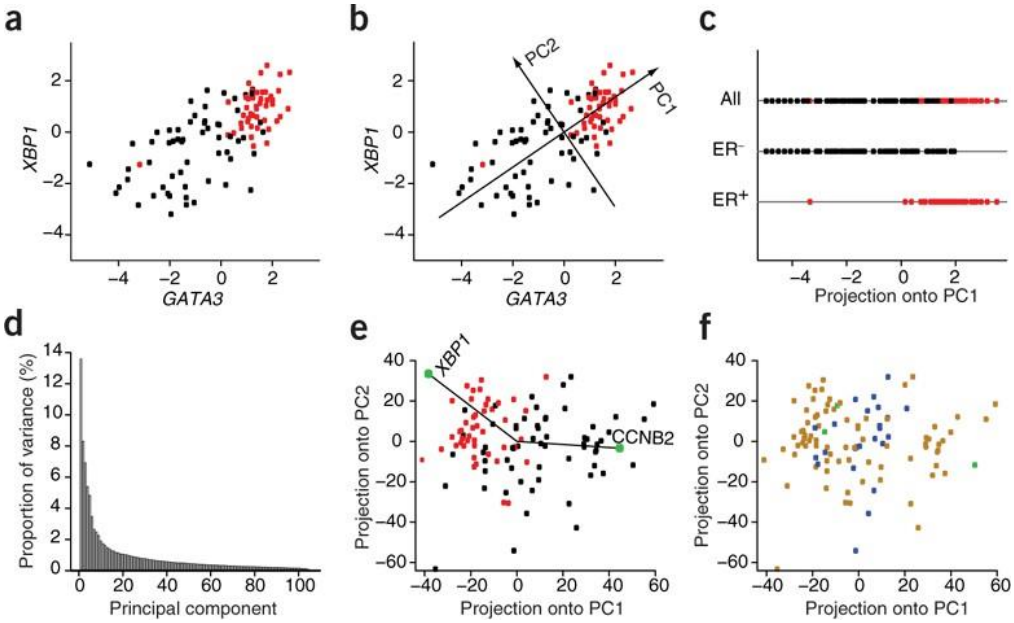

Figure 8: Example of Unsupervised Learning

K-Means Clustering: 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. The main idea is to define k centers, one for each cluster. These centers should be placed in a cunning way because of different location causes different result. So, the better choice is to place them is much as possible far away from each other.

DOI: 10.56452/2021SP-8-001

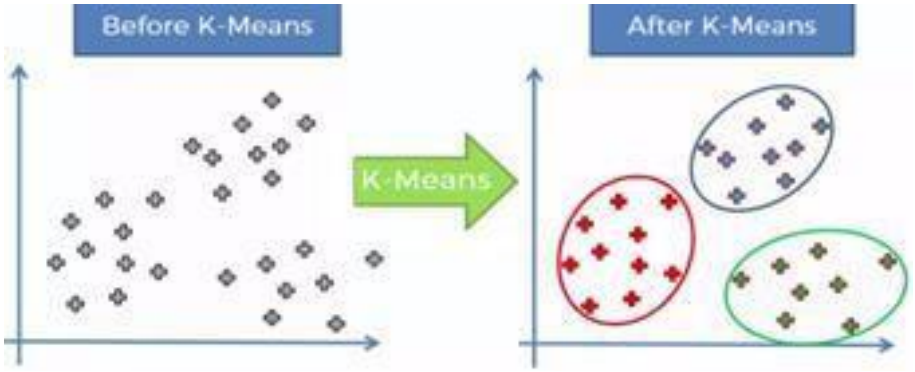

Figure 9: K-Means Clustering

Semi Supervise Learning: Semi-supervised machine learning is a combination of supervised and unsupervised machine learning methods. It can be fruit-full in those areas of machine learning and data mining where the unlabeled data is already present and getting the labeled data is a tedious process. Semi – supervised learning algorithms is a technique which combines the power of both supervised and unsupervised learning. It can be fruit-full in those areas of machine learning and data mining where the unlabeled data is already present and getting the labeled data is a tedious process. There are many categories of semi- supervised learning . Some of which are discussed below:

Generative Models: A Generative model is the one that can generate data. It models both the features and the class (i.e. the complete data). If we model P(x,y): I can use this probability distribution to generate data points - and hence all algorithms modeling P(x,y) are generative. One labeled example per component is enough to confirm the mixture distribution.

Self-Training In self-training, a classifier is trained with a portion of labeled data. The classifier is then fed with unlabeled data. The unlabeled points and the predicted labels are added together in the training set. This procedure is then repeated further. Since the classifier is learning itself, hence the name self-training.

Transductive SVM Transductive support vector machines (TSVM) has been widely used as a means of treating partially labeled data in semisupervised learning. Around it, there has been mystery because of lack of understanding its foundation in generalization. It is used to label the unlabeled data in such a way that the margin is maximum between the labeled and unlabeled data. Finding an exact solution by TSVM is a NP-hard problem.

Reinforcement Learning Reinforcement learning is an area of machine learning concerned with how software agents ought to take actions in an environment in order to maximize some notion of cumulative reward. Reinforcement learning is one of three basic machine learning paradigms, alongside supervised learning and unsupervised learning.

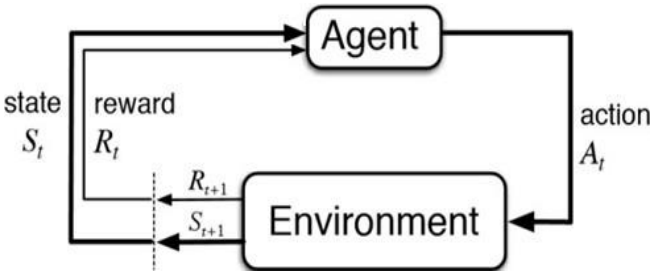

Figure 10: Reinforcement Learning

Ensemble Learning Ensemble learning is the process by which multiple models, such as classifiers or experts, are strategically generated and combined to solve a particular computational intelligence problem. Ensemble learning is primarily used to improve the performance of a model, or reduce the likelihood of an unfortunate selection of a poor one. Other applications of ensemble learning include assigning a confidence to the decision made by the model, selecting optimal features, data fusion, incremental learning, non-stationary learning and error-correcting.

Boosting: Boosting is a technique in ensemble learning which is used to decrease bias and variance. Boosting creates a collection of weak learners and convert them to one strong learner. A weak learner is a classifier which is barely correlated with true classification. On the other hand, a strong learner is a type of classifier which is strongly correlated with true classification.

Bagging*:* Bagging or bootstrap aggregating is applied where the accuracy and stability of a machine learning algorithm needs to be increased. It is applicable in classification and regression. Bagging also decreases variance and helps in handling overfitting

Neural Networks A neural network is a series of algorithms that endeavors to recognize underlying relationships in a set of data through a process that mimics the way the human brain operates. In this sense, neural networks refer to systems of neurons, either organic or artificial in nature. Neural networks can adapt to changing input; so the network generates the best possible result without needing to redesign the output criteria.

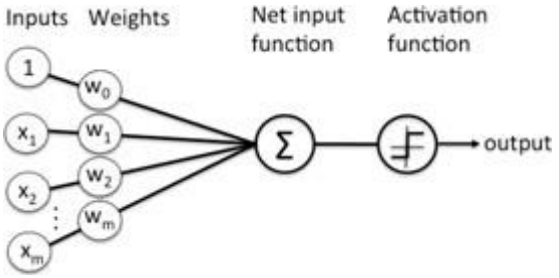

Figure 11: Neural Networks

Supervised Neural Network In the supervised neural network, the output of the input is already known. The predicted output of the neural network is compared with the actual output. Based on the error, the parameters are changed, and then fed into the neural network again. Supervised neural network is used in feed forward neural network.

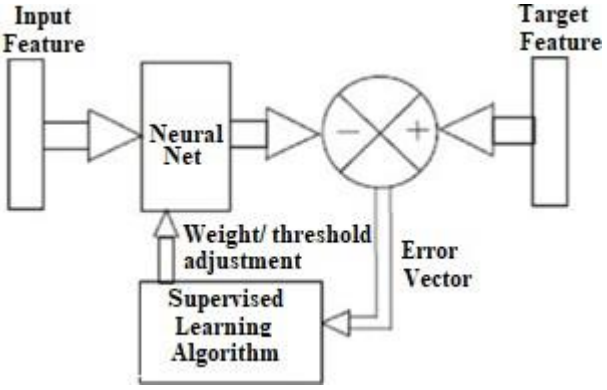

Figure 12: Supervised Neural Network

DOI: 10.56452/2021SP-8-001

Unsupervised Neural Network

The neural network has no prior clue about the output the input. The main job of the network is to categorize the data according to some similarities. The neural network checks the correlation between various inputs and groups them.

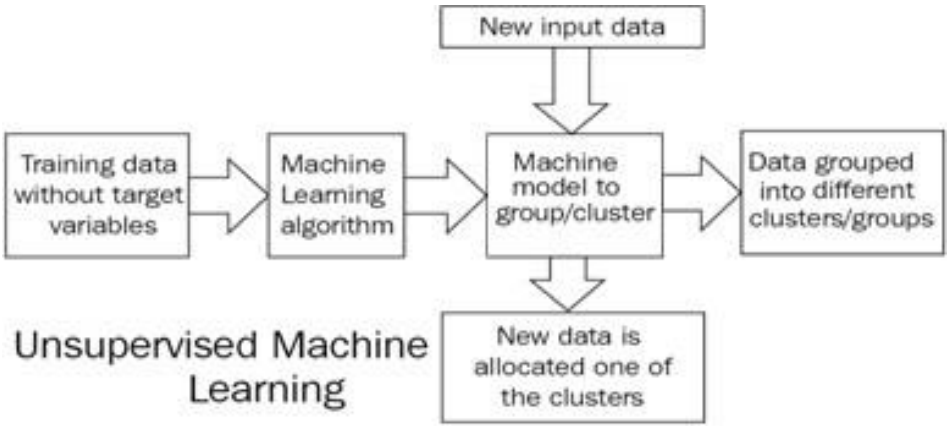

Figure 13: Unsupervised Neural Network

Reinforced Neural Network*:* In reinforced neural network, the network behaves as if a human communicates with the environment. From the environment, a feedback has been provided to the network acknowledging the fact that whether the decision taken by the network is right or wrong. If the decision is right, the connections which points to that particular output is strengthened. The connections are weakened otherwise. The network has no previous information about the output.

Instance-Based Learning

In instance-based learning, the learner learns a particular type of pattern. It tries to apply the same pattern to the newly fed data. Hence the name instance-based. It is a type of lazy learner which waits for the test data to arrive and then act on it together with training data. The complexity of the learning algorithm increases with the size of the data.

K-Nearest Neighbor

The k-nearest neighbors (KNN) algorithm is a simple, supervised machine learning algorithm that can be used to solve both classification and regression problems.

Conclusion

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 surveys various machine learning algorithms. Today each and every person is using machine learning knowingly or unknowingly. From getting a recommended product in online shopping to updating photos in social networking sites. This paper gives an introduction to most of the popular machine learning algorithms.

References

[1] W. Richert, L. P. Coelho, "Building Machine Learning Systems with Python", Packt Publishing Ltd., ISBN 978-1-78216-140

[2] J. M. Keller, M. R. Gray, J. A. Givens Jr., "A Fuzzy K-Nearest Neighbor Algorithm", IEEE Transactions on Systems, Man and Cybernetics, Vol. SMC-15, No. 4, August 1985

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

[4] S. Marsland, Machine learning: an algorithmic perspective. CRC press, 2015.

[5] M. Bkassiny, Y. Li, and S. K. Jayaweera, "A survey on machine learning techniques in cognitive radios," IEEE Communications Surveys & Tutorials, vol. 15, no. 3, pp. 1136–1159, Oct. 2012.

[6] https://en.wikipedia.org/wiki/Instance-based_learning

[7] R. S. Sutton, "Introduction: The Challenge of Reinforcement Learning", Machine Learning, 8, Page 225-227, Kluwer Academic Publishers, Boston, 1992

[9] S. S. Shwartz, Y. Singer, N. Srebro, "*Pegasos: Primal Estimated sub -Gradient Solver for SVM*", Proceedings of the 24th International Conference on Machine Learning, Corvallis, OR, 2007

[10] http://www.simplilearn.com/what-is-machine-learning-and-why-itmatters-article

[11] P. Harrington, "*Machine Learning in action*", Manning Publications Co., Shelter Island, New York, 2012

[12] http://pypr.sourceforge.net/kmeans.html

[13] K. Alsabati, S. Ranaka, V. Singh, "*An efficient k-means clustering algorithm*", Electrical Engineering and Computer Science, 1997

[14] M. Andrecut, "*Parallel GPU Implementation of Iterative PCA Algorithms*", Institute of Biocomplexity and Informatics, University of Calgary, Canada, 2008

[15] X. Zhu, A. B. Goldberg, "*Introduction to Semi – Supervised Learning*", Synthesis Lectures on Artificial Intelligence and MachineLearning, 2009, Vol. 3, No. 1, Pages 1-130

[16] X. Zhu, "*Semi-Supervised Learning Literature Survey*", Computer Sciences, University of Wisconsin-Madison, No. 1530, 2005

[17] R. S. Sutton, "*Introduction: The Challenge of Reinforcement Learning*", Machine Learning, 8, Page 225-227, Kluwer Academic Publishers, Boston, 1992

[18] L. P. Kaelbing, M. L. Littman, A. W. Moore, "*Reinforcement Learning: A Survey*", Journal of Artificial Intelligence Research, 4, Page 237-285, 1996

[19] R. Caruana, "*Multitask Learning*", Machine Learning, 28, 41-75, Kluwer Academic Publishers, 1997

[20] D. Opitz, R. Maclin, "*Popular Ensemble Methods: An EmpiricalStudy*", Journal