A Quick Review of Machine Learning Algorithms

Susmita Ray
Department of Computer Science & Technology
Manav Rachna University
Faridabad, India
susmita@mru.edu.in

***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. In this paper author intends to do a brief review of various machine learning algorithms which are most frequently used and therefore are the most popular ones. The author intends to highlight the merits and demerits of the machine learning algorithms from their application perspective to aid in an informed decision making towards selecting the appropriate learning algorithm to meet the specific requirement of the application.**

Keywords*—Gradient Descent, Logistic Regression, Support Vector Machine, K Nearest Neighbor, Artificial Neural Network, Decision Tree, Back Propagation Algorithm, Bayesian Learning, Naïve Bayes.

I. INTRODUCTION

A good start point for this paper will be to begin with the fundamental concept of Machine Learning. In Machine Learning a computer program is assigned to perform some tasks and it is said that the machine has learnt from its experience if its measurable performance in these tasks improves as it gains more and more experience in executing these tasks. So the machine takes decisions and does predictions / forecasting based on data. Take the example of computer program that learns to detect / predict cancer from the medical investigation reports of a patient. It will improve in performance as it gathers more experience by analyzing medical investigation reports of wider population of patients. Its performance will be measured by the count of correct predictions and detections of cancer cases as validated by an experienced Oncologist. Machine Learning is applied in wide variety of fields namely : robotics, virtual personal assistants (like Google), computer games, pattern recognition, natural language processing, data mining, traffic prediction, online transportation network (e.g. estimating surge price in peak hour by Uber app), product recommendation, share market prediction, medical diagnosis, online fraud prediction, agriculture advisory, search engine result refining (e.g. Google search engine), BoTs (chatbots for online customer support), E-mail spam filtering, crime prediction through
video surveillance system, social media services(face recognition in facebook). Machine Learning generally deals with three types of problems namely: classification, regression and clustering. Depending on the availability of types and categories of training data one may need to select from the available techniques of "supervised learning", "unsupervised learning", "semi supervised learning" and
"reinforcement learning" to apply the appropriate machine learning algorithm. In the next few sections, some of the most widely used machine learning algorithms will be reviewed..

II. GRADIENT DESCENT ALGORITHM

Gradient Descent is an iterative method in which the objective is to minimize a cost function. It should be possible to compute the partial derivative of the function which is slope or gradient. The coefficients are computed at each iteration by taking the negative of the derivative and by reducing the coefficients at each step by a learning rate (step size) multiplied by derivative so that the local minima can be achieved after a few iterations. So eventually the iterations are stopped when it converges to minimum value of the cost function after which there is no further reduction in cost function. There are three different types of this method: "Stochastic Gradient Descent" (SGD), "Batch Gradient Descent"(BGD),. and "Mini Batch Gradient Descent" (MBGD)

In BGD error is computed for every example within the training dataset, but the model will be updated only after the evaluation of all training examples are completed.

The main advantage of BGD algorithm is computational efficiency. It produces a stable error gradient and a stable convergence. However the algorithm has the disadvantage that the stable error gradient can sometimes result in a state of convergence that is not the best which the model can achieve. Also the algorithm requires the entire training dataset to be in memory and available to it.

In SGD error is calculated for each training example within the dataset and parameters are updated for every training example. This might result in SGD to be faster than BGD, for the specific problem. SGD has the advantage that the frequent updates result in a detailed rate of improvement. However the frequent updates are more computationally expensive as compared to the BGD approach. The frequency of those updates can also result in noisy gradients, which may cause the error rate to jump around, instead of decreasing slowly. An example application of SGD will be to evaluate performance contribution of employees to the organization which can help in creating an employee incentivisation scheme.

Approach of MBGD is obtained by combining the concepts of SGD and BGD. In this approach the training dataset is split into small batches and an update is performed for each of these batches. Therefore it creates a balance between the robustness of SGD and the efficiency of BGD. This algorithm can be used to train a neural network and so this algorithm is mostly used in deep learning. The approach of Gradient Descent optimization is used in Backpropagation algorithm wherein the gradient of loss function is computed to adjust the weight of neurons.

Gradient Descent algorithm has the following disadvantage: if the learning rate for gradient descent is too fast, it is going to skip the true local minimum to optimize for time. If it is too slow, the gradient descent may never converge because it is trying really hard to find a local minimum exactly.

The learning rate can affect which minimum is reached and how quickly it is reached. A good practice is to have a changing learning rate, that slows down as the error starts to decrease.

III. LINEAR REGRESSION ALGORITHM

Regression is an approach of supervised learning. It can be used to model continuous variables and do the predictions. Examples of application of linear regression algorithm are the following : prediction of price of real-estate, forecasting of sales, prediction of students' exam scores, forecasting of movements in the price of stock in stock exchange. In Regression we have the labeled datasets and the output variable value is determined by input variable values - so it is the supervised learning approach. The most simple form of regression is linear regression where the attempt is made to fit a straight line (straight hyperplane) to the dataset and it is possible when the relationship between the variables of dataset is linear.

Linear regression has the advantage that it is easy to understand and it is also easy to avoid over fitting by regularization. Also we can use SGD to update linear models with new data. Linear Regression is a good fit if it is known that the relationship between covariates and response variable is linear. It shifts focus from statistical modeling to data analysis and preprocessing. Linear Regression is good for learning about the data analysis process. However, it is not a recommended method for most practical applications because it oversimplifies real world problems.

Disadvantage of Linear regression is that it is not a good fit when one needs to deal with non-linear relationships. Handling complex patterns is difficult. Also it is tough to add the right polynomials appropriately in the model. Linear Regression over simplifies many real world problems. The covariates and response variables usually do not have a linear relationship. Hence fitting a regression line using OLS will give us a line with a high train RSS. In real world problems there may not be relationship between mean of dependent and independent variables which linear regression expects.

IV. MULTIVARIATE REGRESSION ANALYSIS

A simple linear regression model has a dependent variable guided by a single independent variable. However real life problems are more complex. Generally one dependent variable depends on multiple factors. For example, the price of a house depends on many factors like the neighborhood it is situated in, area of it, number of rooms, attached facilities, distance of nearest station / airport from it, distance of nearest shopping area from it, etc. In summary in simple linear regression there is a one-to-one relationship between the input variable and the output variable. But in multiple linear regression, there is a many-to-one relationship, between a number of independent (input/predictor) variables and one dependent (output/response) variable. Adding more input variables does not mean the regression will be better, or will offer better predictions. Multiple and simple linear regression have different use cases and one is not superior than the other. In some cases adding more input variables can make things worse as it results in over-fitting. Again as more input variables are added it creates relationships among them. So not only are the input variables potentially related to the output variable, they are also potentially related to each other, this is referred to as multicollinearity. The optimal scenario is for all of the input variables to be correlated with the output variable, but not with each other

Multivariate technique has the following merits : it gives a deep insight to the relationship between the set of independent variables and dependent variables. It also gives insight to relationship among the independent variables. This is achieved through multiple regression, tabulation techniques and partial correlation. It models the complex real world problems in a practical and realistic way.

Multivariate technique has the following demerits : complexity of this technique is high and it requires knowledge and expertise on statistical techniques and statistical modeling. The sample size for statistical modeling needs to be high to get a higher confidence level on analysis outcome. Also it often gets too difficult to do a meaningful analysis and interpretation of the outputs of statistical model.

This Regression Analysis technique involving multiple variables can be used in property valuation, car evaluation, forecasting electricity demand, quality control, process optimization, quality assurance, process control and medical diagnosis etc.

V. LOGISTIC REGRESSION

Logistic regression is used to deal a classification problem. It gives the binomial outcome as it gives the probability if an

event will occur or not (in terms of 0 and 1) based on values of input variables. For example, predicting if a tumor is malignant or benign or an e-mail is classified as spam or not are the instances which can be considered as binomial outcome of Logistic Regression. There can be multinomial outcome of Logistic Regression as well e.g. prediction of type of cuisine preferred : Chinese, Italian, Mexican etc. There can be ordinal outcome as well like : product rating 1 to 5 etc. So Logistic Regression deals with prediction of target variable which is categorical. Whereas Linear Regression deals with prediction of values of continuous variable e,g. prediction of real estate price over a span of 3 years.

Logistic Regression has the following advantages : simplicity of implementation, computational efficiency, efficiency from training perspective, ease of regularization. No scaling is required for input features. This algorithm is predominantly used to solve problems of industry scale. As the output of Logistic Regression is a probability score so to apply it for solving business problem it is required to specify customized performance metrics so as to obtain a cutoff which can be used to do the classification of the target. Also logistic regression is not affected by small noise in the data and multi-collinearity. Logistic Regression has the following disadvantages: inability to solve non-linear problem as its decision surface is linear, prone to over fitting, will not work out well unless all independent variables are identified. Some examples of practical application of Logistic Regression are: predicting the risk of developing a given disease, cancer diagnosis, predicting mortality of injured patients and in engineering for predicting probability of failure of a given process, system or product.

VI. DECISION TREE

Decision Tree is a Supervised Machine Learning approach to solve classification and regression problems by continuously splitting data based on a certain parameter. The decisions are in the leaves and the data is split in the nodes. In Classification Tree the decision variable is categorical (outcome in the form of Yes/No) and in Regression tree the decision variable is continuous. 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. Decision Tree might encounter the problem of over-fitting for which Random Forest is the solution which is based on ensemble modeling approach.

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.

VII. 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. Following are the advantages of SVM : it can handle both semi structured and structured data, it can handle complex function if the appropriate kernel function can be derived. As generalization is adopted in SVM so there is less probability of over fitting. It can scale up with high dimensional data. It does not get stuck in local optima.

Following are disadvantages of SVM : its performance goes down with large data set due to the increase in the training time. It will be difficult to find appropriate kernel function. SVM does not work well when dataset is noisy. SVM does not provide probability estimates. Understanding the final SVM model is difficult. Support Vector Machine finds its practical application in cancer diagnosis, fraud detection in credit cards, handwriting recognition, face detection and text classification etc. So among the three approaches of Logistic Regression, Decision Tree and SVM the first approach to attempt will be the logistic regression approach, next the decision trees (Random Forests) can be tried to see if there is significant improvement. When the number of observations and features are high then SVM can be tried out..

VIII. BAYESIAN LEARNING

In Bayesian Learning a prior probability distribution is selected and then updated to obtain a posterior distribution Later on with availability of new observations the previous posterior distribution can be used as a prior. Incomplete datasets can be handled by Bayesian network. The method can prevent over-fitting of data. There is no need to remove contradictions from data. Bayesian Learning has the following disadvantages : selection of prior is difficult. Posterior distribution can be influenced by prior to a great extent. If the prior selected is not correct it will lead to wrong predictions. It can be computationally intensive. Bayesian Learning can be used for applications like medical diagnosis and disaster victim identification etc.

IX. NAÏVE BAYES

This algorithm is simple and is based on conditional probability . In this approach there is a probability table which is the model and through training data it is updated. 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.

Naïve 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.

Naïve Bayes has the following disadvantages: Models which are trained and tuned properly often outperform NB models as they are too simple. If there is a need to have one of the feature as “continuous variable” (like time) then it is difficult to apply Naive Bayes directly, Even though one can make “buckets” for “continuous variables” it’s not 100% correct. There is no true online variant for Naive Bayes, So all data need to be kept for retraining the model. It won’t scale when the number of classes are too high, like > 100K. Even for prediction it takes more runtime memory compared to SVM or simple logistic regression. It is computationally intensive specially for models involving many variables.

Naïve Bayes can be used in applications such as Recommendation System and forecasting of cancer relapse or progression after Radiotherapy.

X. K NEAREST NEIGHBOUR ALGORITHM

K Nearest Neighbor (KNN) Algorithm is a classification algorithm It uses a database which is having data points grouped into several classes and the algorithm tries to classify the sample data point given to it as a classification problem. KNN does not assume any underlying data distribution and so it is called non-parametric. Advantages of KNN algorithm are the following : it is simple technique that is easily implemented. Building the model is cheap. It is extremely flexible classification scheme and well suited for Multi-modal classes. Records are with multiple class labels. Error rate is at most twice that of Bayes error rate. It can sometimes be the best method. KNN outperformed SVM for protein function prediction using expression profiles.

Disadvantages of KNN are the following: classifying unknown records are relatively expensive. It requires distance computation of k-nearest neighbors. With the growth in training set size the algorithm gets computationally intensive,. Noisy / irrelevant features will result in degradation of accuracy.

It is lazy learner; it computes distance over k neighbors. It does not do any generalization on the training data and keeps all of them. It handles large data sets and hence expensive calculation. Higher dimensional data will result in decline in accuracy of regions.. KNN can be used in Recommendation system, in medical diagnosis of multiple diseases showing similar symptoms, credit rating using feature similarity, handwriting detection, analysis done by financial institutions before sanctioning loans, video recognition, forecasting votes for different political parties and image recognition.

XI. K MEANS CLUSTERING ALGORITHM

K Means Clustering Algorithm is frequently used for solving clustering problem. It is a form of unsupervised learning. It has the following advantages: it is computationally more efficient than hierarchical clustering when variables are huge. With globular cluster and small k it produces tighter clusters than hierarchical clustering. Ease in implementation and interpretation of the clustering results are the attraction of this algorithm. Order of complexity of the algorithm is O(K*n*d) and so it is computationally efficient.

Disadvantages of K-Means Clustering Algorithm are the following: prediction of K value is hard. Performance suffers when clusters are globular. Also since different initial partitions result in different final clusters it impacts performance. Performance degrades when there is difference in the size and density in the clusters in the input data. Uniform effect often produces clusters with relatively uniform size even if the input data have different cluster size. Spherical assumption (i.e. joint distribution of features within each cluster is spherical) is hard to be satisfied as the correlation between features break it and would put extra weights on correlated features. K value is not known. It is sensitive to outliers. It is sensitive to initial points and local optimal, and there is no unique solution for a certain K value - so one needs to run K mean for a K value lots of times(20-100times) and then pick the results with lowest J.

K Means Clustering algorithm can be used for document classification, customer segmentation, rideshare data analysis, automatic clustering of IT alerts, call record details analysis and insurance fraud detection.

XII. BACK PROPAGATION ALGORITHM

This algorithm provides a very simple and efficient way to compute the gradient in a neural network and one can use it in conjunction with stochastic gradient descent which is also quite simple. There are more complex "quasi-Newton" techniques which make a better estimate of the gradient direction and step size, but they don't perform better than backprop and SGD. Back Propagation Algorithm is used in deep learning. Neural Network (NN) has its specific applications in different industry segments and it has its merits and demerits. Scenarios where there are no well defined criteria or rules to find an answer then NN is useful. It gives the solution but it becomes difficult to explain how the solution is arrived at and so it is like a blackbox.. NN finds its application in classification of credit rating and in forecasting market dynamics in financial sector. Here are some of the applications of NN in marketing segment : in product classification, in classification of customer segments i.e. which customers will like and purchase which products, finding new market for specific product category, in associating relationship between customer and company. NN becomes instrumental in increasing revenue of a business house, in increasing the percentage of response to direct marketing. NN finds its application in Post offices for sorting the letters/parcels based on area zip code / postal code. Following are the merits of NN for which it is widely used in industry segments as mentioned above: easy adaption to new scenarios, fault tolerant, ability to handle noisy data. Shortcoming of NN are the following : training time of NN is very long and for training the NN efficiently, the sample sets need to be large. Back Propagation Algorithm encounters the Moving Target Problem which impacts its efficiency. There are number of hidden layers in the Artificial Neural Network (ANN). Every unit within the network contributes to the overall performance of the network. But the complexity increases as all of the units are changing simultaneously and the units in ANN layer are unable to communicate among them. What every ANN unit can see are its inputs and the error signal which is propagated back to it from ANN output. Each ANN unit tries to solve this problem which is defined by the error signal and the complexity comes into picture as this problem is changing all the time. As a result it takes a long time for this dynamics to settle down among all units in ANN. However research has shown that with the increase in number of hidden layers in ANN there is an exponential rate of slowing down in back propagation learning. Herd effect is one common manifestation of the moving-target problem. Other problems with Back Propagation Learning are network paralysis, local minima and slow convergence. The algorithm works in a way to reduce the error by changing the weights and as a result “Local Minima” occurs. But if the error in this process goes up as part of a more general fall, it will “get stuck” (as it can not go uphill) and the error will stop reducing. During training when the weights are adjusted to very large values then these large weights can force most of the units to operate at extreme values, in a region where the derivative of the activation function is very small and it results in Network Paralysis. ANN with multilayer needs many repeated presentations of the input patterns, in which we need to adjust the weights so that an optimal solution is achieved with the network settling down.

XIII. CONCLUSION

In this paper an attempt was made to review most frequently used machine learning algorithms to solve classification, regression and clustering problems. The advantages, disadvantages of these algorithms have been discussed along with comparison of different algorithms (wherever possible) in terms of performance, learning rate etc. Along with that, examples of practical applications of these algorithms have been discussed. Types of machine learning techniques namely supervised learning, unsupervised learning, semi supervised learning, have been discussed. It is expected that it will give insight to the readers to take an informed decision in identifying the available options of machine learning algorithms and then selecting the appropriate machine learning algorithm in the specific problem solving context.

REFERENCES

[1] D. Pelleg, A. Moore (2000): “X-means: Extending K-means with Efficient Estimation of the Number of Clusters”; ICML ’00 Proceedings of the Seventeenth International Conference on Machine Learning, pp. 727-734.

[2] Rushika Ghadge, Juilee Kulkarni, Pooja More, Sachee Nene, Priya R ,

[3] “Prediction of Crop Yield using Machine Learning”, International Research Journal of Engineering & Technology, Vol 5, Issue 2, Feb-2018.

[4] C. Phua, V. Lee, K. Smith, R. Gayler (2010); “Comprehensive Survey of Data Mining-based Fraud Detection Research”, ICICTA ’10 Proceedings of the 2010 International Conference on Intelligent Computation Technology and Automation Volume 1, pp. 50-53.

[5] S. Cheng, J. Liu, X. Tang (2014); “Using unlabeled Data to Improve Inductive Models by Incorporating Transductive Models”; International Journal of Advanced Research in Artificial Intelligence, Volume 3 Number 2, pp. 33-38.

[6] Sonal S. Ambalkar, S. S. Thorat2, “Bone Tumor Detection from MRI Images using Machine Learning: A Review”, International Research Journal of Engineering & Technology”, Vol. 5, Issue 1, Jan -2018.

[7] Rajat Raina, Alexis Battele, Honglak Lee,Benjamin Packer, Andrew Y. Ng , “Self-taught Learning : Transfer of Learning from Unlabeled Data”, Computer Science Department, Stanford University, CA, USA, Proceedings of 24th International Conference on Machine Learning Corvallis, OR, 2007.

[8] Jimmy Lin, Alek Kolcz, “Large-Scale Machine Learning at Twitter”, Proceedings of SIGMOD ’12, May 20–24, 2012, Scottsdale, Arizona, USA.

[9] Dr. Rama Kishore, Taranjit Kaur, “Backpropagation Algorithm: An Artificial Neural Network Approach for Pattern Recognition”, International Journal of Scientific & Engineering Research, Volume 3, Issue 6, June-2012.

[10] Kedar Potdar, Rishab Kinnerkar, “A Comparative Study of Machine Algorithms applied to Predictive Breast Cancer Data”, International Journal of Science & Research, Vol. 5, Issue 9, pp. 1550-1553, September 2016.