A BRIEF REVIEW OF MACHINE LEARNING ALGORITHMS

Nurullayev Ye.Ye
Saparbaev R.K
Omonov I.I.
Urgench branch of the Tashkent University of Information Technologies named after Muhammad Al-Khwarizmi.
elamannurullayev@gmail.com, saparbayevraxmonbergan@gmail.com, ibratbekomonov@gmail.com

Abstract: *Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without being explicitly programmed. Learning algorithms, in which many of the applications we use on a daily basis. Every time, when searching engine, like Google, is used to search the web, one of the reasons it performs so well is because of the learning algorithm that has learned to rank web pages. That is, we use these popular algorithms on a daily basis in services, such as Google. These algorithms are used for various purposes like data mining, image processing, predictive analytics, etc. to name a few. The main advantage of using machine learning is that, once an algorithm learns what to do with data, it can do its work automatically. In this paper, a brief review of various machine learning algorithms has been done which are most frequently used and, therefore, are the most popular ones. Also has been highlighted the merits and demerits of the machine learning algorithms to meet the specific requirement of the application.*

I. Introduction

The fundamental idea of machine learning will be a good place to start for this paper. In machine learning, a computer program is given a set of tasks to complete. If the program's measurable performance improves as it gets more experience with the tasks, it is said to have learned from that experience. As a result, the machine uses data to make decisions and forecasts. Consider a computer program that learns to identify and predict cancer based on a patient's medical examination reports. As it analyzes the medical investigation reports of a larger number of patients, it will gain more experience, which will lead to improvements in performance. The number of cancer cases that it correctly predicts and finds, as verified by an experienced oncologist, will be used to evaluate its performance. There are many applications for machine learning, including: Pattern recognition, natural language processing, data mining, robotics, computer games, traffic prediction, the online transportation network (such as the Uber app's estimate of the surge price during peak hours), product recommendation, share market prediction, medical diagnosis, online fraud prediction, agriculture advisory, BoTs (chatbots for online customer support), e-mail spam filtering, crime prediction through video surveillance, and social media services (face

recognition in Facebook). In general, dealing with these updates in machine learning can also result in noisy gradients, which may cause the error rate to fluctuate rather than gradually decrease. The evaluation of three kinds of problems, which are as follows, is one example of how SGD can be used: regression, classification, and clustering. To apply the appropriate machine learning algorithm, one may need to select from the available techniques of "supervised learning," "unsupervised learning," "semi supervised learning," and "reinforcement learning" depending on the types and categories of training data that are available. A few of the most widely used machine learning algorithms will be discussed in the following sections.

II. **Linear regression algorithm**

One method of supervised learning is regression. It can be used to predict and model continuous variables. The following are some examples of how the linear regression algorithm can be used: forecasting the price of real estate, sales, student exam scores, and stock exchange price movements are all examples of forecasting techniques. In regression, we use labeled datasets and the supervised learning approach because the value of the output variable is determined by the values of the input variables. The most basic type of relapse is direct relapse where the endeavor is made to fit a straight line (straight hyperplane) to the dataset and it is conceivable when the connection between the factors of dataset is direct. [1]

The advantage of linear regression is that it is simple to comprehend and that regularization makes it simple to avoid overfitting. SGD can also be used to add new data to linear models. If the linear relationship between covariates and response variable is known, then linear regression is a good fit. It focuses on data analysis and preprocessing rather than statistical modeling. A good way to learn about the process of data analysis is through linear regression. However, due to its oversimplification of real-world issues, it is not recommended for the majority of practical applications.

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.

III. **Logistic regression**

A classification problem can be solved with the help of logistic regression. Based on the values of the input variables, it returns the binomial result, which is the probability that an event will take place or not (in terms of 0 and 1) The prediction of a tumor's malignancy or benignity, or whether an email is considered spam or not, are examples of Logistic Regression's binomial outcomes. Logistic Regression can also produce multinomial results, such as a prediction of the preferred cuisine: Arabic or

Italian, and so on There are also ordinal outcomes, such as: rating of one to five, etc. Therefore, categorical target variable prediction is the focus of logistic regression. Whereas Linear regression, on the other hand, deals with the prediction of values for continuous variables like estimation of the value of real estate over a three-year period. 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.

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 multicollinearity. [1]

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. [1]

IV. **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 Trees can be used in applications like predicting future use of library books and tumor prognosis problems.

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. [1]

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. [1]

V. **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. [1] It can scale up with high dimensional data. It does not get stuck in local optima.

The SVM has the following disadvantages: Due to the increase in training time, its performance decreases with large data sets. Finding the appropriate kernel function will be difficult. When the dataset is noisy, SVM fails. [1] Estimates of probabilities are not provided by SVM. It is hard to understand the final SVM model. The diagnosis of cancer, the detection of credit card fraud, the recognition of handwriting, the detection of faces, and the classification of text, among other applications, are all practical uses for the Support Vector Machine. Therefore, among the three methods, Logistic Regression, Decision Tree, and SVM, Logistic Regression should be tried first, followed by Decision Trees (Random Forests) to see if there is a significant improvement. SVM can be tested when there are a lot of features and observations.

VI. **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 can be used for applications like medical diagnosis and disaster victim identification etc.

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. [2] It can be computationally intensive.

VII. **Naive 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. [6] 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 can be used in applications such as Recommendation System and forecasting of cancer relapse or progression after Radiotherapy.

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.[6] It handles continuous and discrete data. [2] 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 features 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. [2] 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. [5] 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.

VIII. **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. 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.

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. [4]

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 does not do any generalization on the training data and keeps all of them. [4] It handles large data sets and hence expensive calculation. Higher dimensional data will result in decline in accuracy of regions.

IX. **K means clustering algorithm**

K Means Clustering Algorithm is frequently used for solving clustering problem. It is a form of unsupervised learning. 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

K Means Clustering Algorithm 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. [3]

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. [3]

X. **Conclusion**

The most frequently used machine learning algorithms for solving classification, regression, and clustering problems are reviewed in this paper. The benefits and drawbacks of these algorithms have been discussed, and whenever possible, they have been compared to one another in terms of performance, learning rate, and other factors. In addition, examples of these algorithms' actual uses have been discussed. There has been discussion of supervised learning, unsupervised learning, and semi-supervised learning as types of machine learning techniques. It is anticipated that it will provide readers with the information they need to make an educated choice when it comes to choosing the best machine learning algorithm for a given problem-solving situation.

REFERENCES:

[1] 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.

[2] 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

[3] 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.

[4] “Prediction of Crop Yield using Machine Learning”, International Research Journal of Engineering & Technology, Vol 5, Issue 2, Feb2018

[5]. Nurullayev Ye.Ye. Characteristics of using wireless sensor networks for remote monitoring systems. VII International scientific and practical conference “International forum: problems and scientific solutions”, Melbourne, Australia, April 25-26, 2021.

[6]. Nurullayev Ye.Ye. Iva software as a public safety system. ICISCT 2022, Session 1, "Cybersecurity", September 28-30, 2022.

[7]. Djurayev R. K., Jabborov S. Y., Omonov I. I. Analysis of the open queuing system model of diagnostics of data transmission network elements //itn&t-2022. – c. 48.

[8]. Djurayev, R. K., Jabborov, S. Y., Omonov, I. I., & Rajabov, J. R. (2022). Research on the Model of Malfunction and Diagnostics of Digital Devices of Data Transmission Equipment. International Journal of Innovative Analyses and Emerging Technology, 2(3), 31-35.