January-February 2020
ISSN: 0193-4120 Page No. 12427 - 12431

Certain Investigations on Machine Learning Algorithms for Multiple Applications

B.Devananda Rao [1]
Associate Professor, Department of Computer Science & Engineering,
Dr.K.V.Subba Reddy College Of Engineering For Women, Kurnool.

D.Lakshmi Renuka Devi [2]
Assistant Professor, Department of Computer Science & Engineering,
Dr.K.V.Subba Reddy College Of Engineering For Women, Kurnool.

Article Info
Volume 82
Page Number: 12427 - 12431
Publication Issue:
January-February 2020

Article History
Article Received*: 18 May 2019*
Revised*: 14 July 2019*
Accepted*: 22 December 2019*
Publication*: 23 February 2020*

Abstract:

Machine learning is an essential area of Artificial Intelligence, which is the main component in solutions of digitalization in the digital domain. I want to review various machine learning algorithms that are most popular, and therefore, they are frequently used. I want to show the merits and demerits of machine learning algorithms from their application perspective to decide on selecting the appropriate learning algorithm.

Keywords:*Machine Learning Algorithms.*

I. INTRODUCTION

In machine learning to perform a task, the computer program is assigned, and if the measurable performance is improved, we can say that the machine has learned from the experience. Based on the experience, the machine takes a decision and makes predictions or forecasting based on the data. Consider an example of a computer program that learns to detect or predict cancer from the patient medical investigation reports. It will improve the performance as it gains more experience by analyzing medical investigation reports of a large population of patients. Experienced oncologists will measure its performance by the number of correct predictions and detections of cancer cases. Machine learning is applied in my fields: robotics, computer games, natural language processing, data mining, online transport network, traffic prediction. Pattern recognition.share market prediction, medical diagnosis, online fraud prediction, agriculture advice, E-mail spam filtering.

II. EXISTING MACHINE LEARNING ALGORITHMS

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 calculated at each iteration by taking the negative of the derivative and by reducing the factors 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 the minimum value of the cost function, after which there is no further reduction in the cost function. There are three different types of this method: "Stochastic Gradient Descent" (SGD), "Batch Gradient Descent"(BGD). and "Mini Batch Gradient Descent" (MBGD).

The BGD compute an error 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 the BGD algorithm is computational efficiency. It produces a stable error gradient and convergence. However, the algorithm has the disadvantage that the

January-February 2020
ISSN: 0193-4120 Page No. 12427 - 12431

stable error gradient can sometimes result in a state of convergence that is not the best which the model can achieve. Also, the algorithm needs the entire training dataset to be in memory and available to it.

In SGD, an 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 regular 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 the performance contribution of employees to the organization which can help in creating an employee incentivization scheme.

The 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 SGD robustness and the BGD efficiency. 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 the Backpropagation algorithm wherein the gradient of the 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 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.

LINEAR REGRESSION ALGORITHM

Regression is an approach to supervised learning. It can be used to model continuous variables and make predictions. Examples of application of linear regression algorithm are the following: prediction of the price of real-estate, forecasting of sales, prediction of students' exam scores, forecasting of movements in the price of a stock in the stock exchange. In Regression, we have the labeled datasets and input variable values determine the output variable value - 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 a dataset is linear.

Linear regression has the advantage that it is easy to understand, and it is also easy to avoid overfitting 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 the response variable is linear. It shifts focus from statistical modeling to data analysis and preprocessing. Linear Regression is useful for learning about the data analysis process. However, it is not a recommended method for most practical applications because it oversimplifies real-world problems.

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

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.

TEST Engineering & Management

January-February 2020
ISSN: 0193-4120 Page No. 12427 - 12431

Generally one dependent variable depends on multiple factors. For example, the house price depends on many factors like the neighborhood it is situated in, area of it, number of rooms, attached facilities, a 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 amount of input variables does not mean the regression will give better predictions. Several and simple linear regression have different use cases and one is not superior to the other. In some cases adding more amount of input variables can make things worse as it results in over-fitting. Again as more input variables are added, it generates relationships among them. So not only are the input variables possibly related to the output variable, they are also possibly related to each other, this is referred to as multicollinearity. The best scenario is for all of the input variables to be correlated with the output variable, but not with each other.

The multivariate technique has the following merits: it gives a deep insight into the relationship between the set of independent variables and dependent variables. It also provides insight into the relationship between the independent variables. This is achieved through multiple regression, tabulation techniques and partial correlation. It models the complex real-world problems practically and realistically.

The multivariate technique has the following demerits: the complexity of this technique is high and it requires knowledge and expertise on statistical methods and statistical modeling. The sample size for statistical modeling needs to be elevated 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 the 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.

LOGISTIC REGRESSION

Logistic regression is used to deal with a classification problem. It gives the binomial outcome as it provides the probability of an event that will occur or not (in terms of 0 and 1) based on the 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 that can be considered as a binomial outcome of Logistic Regression. There can be the multinomial outcome of Logistic Regression as well e.g. prediction of the type of cuisine preferred: Chinese, Italian, Mexican etc. There can be the ordinal outcome as well as: product rating 1 to 5 etc. So Logistic Regression deals with the prediction of the target variable which is categorical. Whereas Linear Regression deals with the prediction of values of continuous variable e,g. prediction of real estate prices over 3 years.

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

January-February 2020
ISSN: 0193-4120 Page No. 12427 - 12431

injured patients and engineering for predicting the probability of failure of a system or product.

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 the Regression tree the decision variable is continuous. Decision Tree has the following advantages: it is suitable for regression as well as a 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 that is based on an ensemble modeling approach.

Disadvantages of the decision tree are that it can be unstable, it may be difficult to control the size of a 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.

K MEANS CLUSTERING ALGORITHM

K Means Clustering Algorithm is frequently used for solving the 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 clusters 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. The order of complexity of the algorithm is O(K*n*d) and so it is computationally efficient.

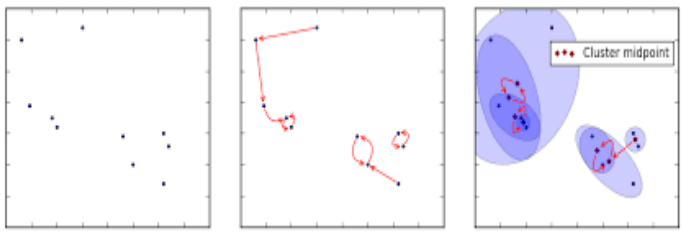
The disadvantages of the K-Means Clustering Algorithm are the following: prediction of K value is hard. Performance suffers when clusters are globular. Also since different initial partitions result in different final clusters it impacts performance. Performance degrades when there is a difference in the size and density in the clusters in the input data. The unvarying effect often produces clusters with relatively uniform size even if the input data have different cluster sizes. Spherical assumption (i.e. the joint distribution of features within each cluster is spherical) is hard to be satisfied as the correlation between features breaks 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.

III. PROPOSED METHODOLOGY

In this section, the cross clustering algorithm is proposed, which performs autonomous partitioning by a similar distance. It is a clustering algorithm which need not specify the number of categories in advance and allows crossover among clusters. In other words, an instance object can belong to multiple categories simultaneously

The visualization of the cross clustering process.



IV. CONCLUSION

In this paper, an attempt was made to review the most frequently used machine learning algorithms to solve classification, regression and clustering problems. The advantages, disadvantages of these algorithms have been discussed along with a comparison of different algorithms (wherever possible) in terms of performance, learning rate etc. Along with that, examples of practical applications

January-February 2020
ISSN: 0193-4120 Page No. 12427 - 12431

of these algorithms have been discussed. Types of machine learning techniques namely supervised learning, unsupervised learning, semi-supervised learning, have been considered. It is expected that it will give insight to the readers to make 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. RushikaGhadge, 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, Feb2018.
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 Battle, 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. KedarPotdar, RishabKinnerkar, "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.