IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Estimating Most Effective Methods Of Machine Learning

Mr. Shiva Gupta
Dept. of Information Tech.
Krishna Engineering College

Mr. Aakash Shukla
Dept. of Information Tech.
Krishna Engineering College

Mr. Madhukar Yadav
Prof. Dept. of Information Tech.
Krishna Engineering College

Abstract:

In this review paper we are going to conduct review research on Machine Learning methods with various input parameters and conclude which method is useful or most effective for a particular problem or which method with a particular characteristic is useful/effective for a given problem. We are going to analyze and conclude the most effective methods of Machine Learning. Here we will also configure which methods to use so as to give us a appropriate result of estimation.

Keywords: *Machine Learning, Linear Regression, Logistic Regression, Decision Trees, Support Vector Machines, Gaussian Naive Bayes, K-Nearest Neighbors, K-means.*

INTRODUCTION

Machine learning is a branch of artificial intelligence (AI) focused on building applications that learn from data and improve their accuracy over time without being programmed to do so. The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly. Simply, Machine Learning is based on the idea that machines can learn from past data, identify patterns, and make decisions using algorithms.

Further, Machine learning can be subdivided into 3 types:

1. Supervised Learning - The computer is presented with example inputs and their desired outputs, given by a " teacher" and the goal is to learn a general rule that maps inputs to outputs.

2. Unsupervised learning - No labels are given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or a means towards an end (feature learning).

3. Reinforcement learning - A computer program interacts with a dynamic environment in which it must perform a certain goal (such as driving a vehicle or playing a game against an opponent). As it navigates its problem space, the program is provided feedback.

PROPOSED METHODOLOGY

Methods of Machine Learning we will be using:

1. Linear Regression To understand the working functionality of this algorithm, imagine how you would arrange random logs of wood in increasing order of their weight.

There is a catch; however – you cannot weigh each log. You have to guess its weight just by looking at the height and width of the log (visual analysis) and arrange them using a combination of these visible parameters. This is what linear regression.

2. Logistic Regression Logistic Regression is used to estimate discrete values (usually binary values like 0/1) from a set of independent variables. It helps predict the probability of an event by fitting data to a logit function. It is also called logit regression.

3. Decision Tree It is one of the most popular machines learning algorithms in use today; this is a supervised learning algorithm that is used for classifying problems. It works well

classifying for both categorical and continuous dependent variables.

4. SVM (Support Vector Machine) SVM is a method of classification in which you plot raw data as points in an ndimensional space (where n is the number of features you have). The value of each feature is then tied to a particular coordinate, making it easy to classify the data.

5. Naive Bayes a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

6. KNN (K- Nearest Neighbors) This algorithm can be applied to both classification and regression problems. It's a simple algorithm that stores all available cases and classifies any new cases by taking a majority vote of its k neighbours. The case is then assigned to the class with which it has the most in common.

7. K-Means It is an unsupervised algorithm that solves clustering problems. Data sets are classified into a particular number of clusters (let's call that number K) in such a way that all the data points within a cluster are homogenous and heterogeneous from the data in other clusters.

8. Random Forest A collective of decision trees is called a Random Forest. To classify a new object based on its attributes, each tree is classified, and the tree "votes" for that class. The forest chooses the classification having the most votes (over all the trees in the forest).

9. Dimensionality Reduction Algorithms In today's world, vast amounts of data are being stored and analysed by corporates, government agencies, and research organizations. As a data scientist, you know that this raw data contains a lot of information - the challenge is in identifying significant patterns and variables.

10. Gradient Boosting & AdaBoost These are boosting algorithms used when massive loads of data have to be handled to make predictions with high accuracy. Boosting is an ensemble learning algorithm that combines the predictive power of several base estimators to improve robustness.

LITERATURE REVIEW

1. Athey, Susan, and Guido W. Imbens.

In applications, our method provides a data-driven approach to determine which subpopulations have large or small treatment effects and to test hypotheses about the differences in these effects

2. Padberg, Frank, Thomas Ragg, and Ralf Schoknecht.

The goal is to learn from empirical data the relationship between certain observable features of an inspection and the number of defects actually contained in the document. We show that some features can carry significant nonlinear information about the defect content.

3. Seyedzadeh, Saleh, et al.

Ever growing population and progressive municipal business demands for constructing new buildings are known as the foremost contributor to greenhouse gasses. The energy efficiency of the building sector has become an essential target to reduce the amount of gas emission as well as fossil fuel consumption.

4. Qasem, Sultan Noman, et al.

The ability of three data-driven methods of Gene Expression Programming (GEP), M5 model tree (M5), and Support Vector Regression (SVR) were investigated in order to model and estimate the dew point temperature (DPT) at Tabriz station, Iran.

5. Baskeles, Bilge, Burak Turhan, and Ayse Bener.

The main objective of this research is making an analysis of software effort estimation to overcome problems related to it: budget and schedule extension.

6. Stojanova, Daniela, et al.

High quality information on forest resources is important to forest ecosystem management. Traditional ground measurements are labor and resource intensive and at the same time expensive and time consuming.

7. Gleason, Colin J., and Jungho Im.

During the past decade, procedures for forest biomass quantification from light detection and ranging (LiDAR) data have been improved at a rapid pace. The scope of these methods ranges from simple regression between LiDAR-derived height metrics and biomass to methods including automated tree crown delineation, stochastic simulation, and machine learning approaches.

8. Song, Xinyu D., et al.

Pure-tone audiometry has been a staple of hearing assessments for decades. Many different procedures have been proposed for measuring thresholds with pure tones by systematically manipulating intensity one frequency at a time until a discrete threshold function is determined.

9. Biazar, Seyed Mostafa, et al.

Selection of optimal model inputs is a challenging issue particularly for non-linear and dynamic systems. In this study, a new input selection method, procrustes analysis (PA), was implemented and compared with gamma test (GT) for estimating daily global solar radiation (Rs).

10. Pillonetto, Gianluigi, et al.

Most of the currently used techniques for linear system identification are based on classical estimation paradigms coming from mathematical statistics. In particular, maximum likelihood and prediction error methods represent the mainstream approaches to identification of linear dynamic systems, with a long history of theoretical and algorithmic contributions.

11. Braga, Petrônio L., Adriano LI Oliveira, and Silvio RL Meira.

The precision and reliability of the estimation of the effort of software projects is very important for the competitiveness of software companies. Good estimates play a very important role in the management of software projects. Most methods proposed for effort estimation, including methods based on machine learning, provide only an estimate of the effort for a novel project.

Learning Method/ Algorithm	**Generative or Discriminative ?**	**Decision Boundary or Regression Function Shape**	**Model Complexity Reduction**
Linear Regression		Linear	
Logistic Regression	Discriminative	Linear	L2 regularization
Decision Trees	Discriminative	Axis-aligned partition of feature space	Prune tree or limit tree depth
Support Vector Machines (with slack variables, no kernel)	Discriminative	linear (depends on kernel)	Reduce C
Gaussian Naive Bayes	Generative	Equal variance: linear boundary. Unequal variance: quadratic boundary.	Place prior on parameter and us e MAP estimator
K-Nearest Neighbors	Discriminative	Arbitrarily complicated	Increase K
K-means			

12. Kruppa, Jochen, et al.

Probability estimation for binary and multicategory outcome Using logistic and multinomial logistic regression has a longstanding tradition in biostatistics. However, biases may occur if the model is mis specified.

13. Musil, Felix, et al.

We present a scheme to obtain an inexpensive and reliable estimate of the uncertainty associated with the predictions of a machine-learning model of atomic and molecular properties. The scheme is based on resampling, with multiple models being generated based on subsampling of the same training data.

14. Behravan, Iman, and Seyed Mohammad Razavi.

Every year a huge amount of money is invested by the football clubs in the transfer window period to hire or release players. Estimating players' value in the transfer market is a crucial task for the managers of the clubs. Also, it is one of the attractive aspects of football for fans.

15. Mocanu, Elena, et al.

The increasing number of decentralized renewable energy sources together with the grow in overall electricity consumption introduce many new challenges related to dimensioning of grid assets and supply-demand balancing. Approximately 40% of the total energy consumption is used to cover the needs of commercial and office buildings.

EXPERIMENT RESULT COMPARISION BETWEEN DIFFERENT ALGORITHMS

In this problem, we will review the important aspects of the algorithms we have learned about. Here we will be comparing different algorithms on different traits in a tabular form: Traits:

1.) Generative or Discriminative –
Choose either "generative" or "discriminative"; you may write "G" and "D" respectively to save some writing. Generative models are a wide class of machine learning algorithms which make predictions by modelling joint distribution P(y, x). Discriminative models are a class of supervised machine learning models which make predictions by estimating conditional probability P (y | x).

2.) Decision Boundary / Regression Function Shape –

Describe the shape of the decision surface or regression function, e.g., "linear". If necessary, enumerate conditions under which the decision boundary has different forms.

3.) Model Complexity Reduction –

Name a technique for limiting model complexity and preventing overfitting.

4.) Number of Clusters –

Choose either "predetermined" or "data-dependent"; you may write "P" and "D" to save time.

5.) Loss Function – Write either the name or the form of the loss function optimized by the algorithm

LEARNING METHOD/ ALGORITHM	**Number of clusters**	**Loss Function**
Linear Regression		square loss: (Y ˆ − Y)2
Logistic Regression		− log P(Y |X)
Decision Trees		Either − log P(Y |X) or zero-one loss
Support Vector Machines (with slack variables, no kernel)		hinge loss: |1−y(wT x)|+
Gaussian Naive Bayes		− log P(X, Y)
K-Nearest Neighbors		zero-one loss
K-means	Predetermined	Within-class squared distance from mean

(e.g., “exponential loss”).

CONCLUSION

On comparing above methods on various parameters we concluded that for every different method they offer different characteristic that make them more effective than others. Thus, we can say that for every problem type they provide different features and one can choose what type of solution or feature of a particular method will be effective for them based on their problem type.

REFERENCES

- Stojanova, Daniela, et al. "Estimating vegetation height and canopy cover from remotely sensed data with machine learning." *Ecological Informatics* 5.4 (2010): 256-266.

- Gleason, Colin J., and Jungho Im. "Forest biomass estimation from airborne LiDAR data using machine learning approaches." *Remote Sensing of Environment* 125 (2012): 80-91.

- Song, Xinyu D., et al. "Fast, continuous audiogram estimation using learning." *Ear and hearing* 36.6 (2015): e326.

- Biazar, Seyed Mostafa, et al. "New input selection procedure for machine learning methods in estimating daily global solar radiation." *Arabian Journal of Geosciences* 13 (2020): 1-17

- Pillonetto, Gianluigi, et al. "Kernel methods in system identification, machine learning and function estimation: A survey." Automatica 50.3 (2014): 657-682.

- Braga, Petrônio L., Adriano LI Oliveira, and Silvio RL Meira. "Software effort estimation using machine learning techniques with robust confidence intervals." 7th international conference on hybrid intelligent systems (HIS 2007). IEEE, 2007.

- Kruppa, Jochen, et al. "Probability estimation with machine learning methods for dichotomous and multicategory outcome: theory." Biometrical Journal 56.4 (2014): 534-563.

- Musil, Felix, et al. "Fast and accurate uncertainty estimation in chemical machine learning." Journal of chemical theory and computation 15.2 (2019): 906-915.

- Behravan, Iman, and Seyed Mohammad Razavi. "A novel machine learning method for estimating football players’ value in the transfer market." Soft Computing 25.3 (2021): 2499-2511.

- Mocanu, Elena, et al. "Comparison of machine learning methods for estimating energy consumption in buildings." 2014 international conference on probabilistic methods applied to power systems (PMAPS). IEEE, 2014

- https://www.ibm.com/cloud/learn/machine-learning – Blog, Machine learning focuses on applications that learn from experience and improve their decision-making or predictive accuracy over time.
- https://www.edureka.co/blog/machine-learning-algorithms/ - Upasana -- Research Analyst, Analysis of Machine Learning.
- https://en.wikipedia.org/wiki/Machine_learning - the free encyclopaedia, Machine Learning.
- http://www.cs.cmu.edu/~aarti/Class/10701/ MLAlgo_Comparisons – Aarti (Research Analyst), Comparative Analysis of ML methods.