International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

COPY RIGHT

IJIEMR Transactions, online available on 23rd Dec 2019. Link :http://www.ijiemr.org/downloads.php?vol=Volume-08&issue=ISSUE-12

Title: **MACHINE LEARNING IN A NUT SHELL: A COMPREHENSIVE STUDY**

Volume 08, Issue 12, Pages: 392–397.

Paper Authors

DR. SHAMEENA BEGUM, T.BHAVANI

RAMACHANDRA COLLEGE OF ENGINEERING, ELURU

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

MACHINE LEARNING IN A NUT SHELL: A COMPREHENSIVE STUDY

DR. SHAMEENA BEGUM[1], T.BHAVANI[2]

Professor, Department of CSE, Ramachandra College of Engineering
Assistant Professor, Department of CSE, Sasi Institute of Technology & Engineering

ABSTRACT

Today's biggest trend in the market is a Machine Learning, an application of Artificial Intelligence that focuses mainly on designing the machines thereby allowing them to learn and make predictions based on data or experience from the environment exclusive of being overtly programmed. This paper presents an overview on the role of machine learning in today's world; the basic machine learning models that are in use to train the machine for better prediction; characteristic to choose the model; validating and improving the model performance.

I. INTRODUCTION

Machine Learning is a subfield of Artificial Intelligence, where a machine is taught how to learn on the basis of input data. Scientific models can be built using Machine Learning for the principle of prediction and classification.

International Data Corporation predicts that, by 2022 the global expenditure on cognitive and Artificial Intelligence systems will reach $77.6B [1].

II. FEATURES OF MACHINE LEARNING

Machine Learning is a kind of data analysis that automates analytical model building.

Features of Machine Learning include i) Detection of patterns in a dataset ii) Develop programs to make them learn to act accordingly when exposed to new data and iii) Enables machines to find hidden insights using iterative algorithms.

III. APPLICATIONS OF MACHINE LEARNING

Facial Recognition[2], Speech Recognition, Traffic Prediction, Self-Driving Cars, Medical Diagnosis, E-mail Spam filtering, Product Recommendations, Automatic Language Translation, Online Fraud Detection, Stock Market Trading, DNA Pattern Recognition [3] etc.,

IV. MACHINE LEARNING ALGORITHM AND HOW MACHINE LEARNS

Machine Learning Algorithm is the program fed into the machine that learns automatically from the data provided.

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

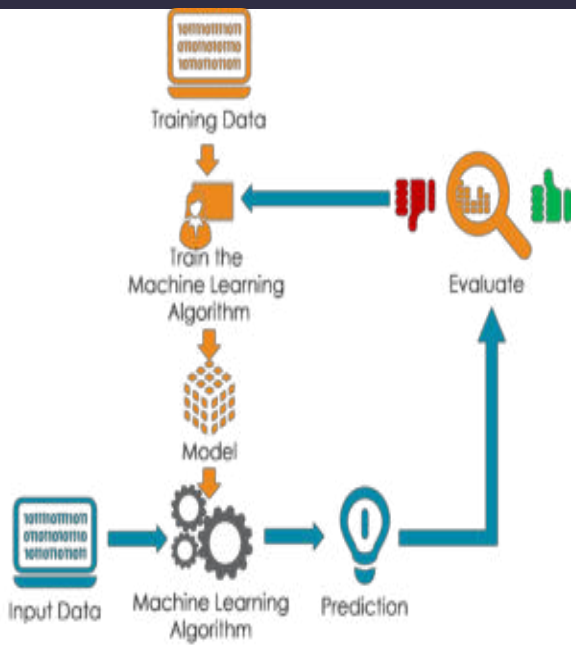

Figure 1: Building a Model

A labeled or unlabeled training data set is used to train the Machine Learning algorithm to produce a model. New input data is initiated to the ML algorithm and it makes a prediction based on the model.

The prediction of an algorithm is evaluated for accuracy and if the accuracy is acceptable, the ML algorithm is deployed. If the accuracy is not acceptable, the ML algorithm is trained again with an augmented training data set. The steps to build a machine learning model is shown in Figure 1 [4].

V. DATASETS

The dataset provided is a set of observations or records and each record is a set of variables or features. Datasets can be tagged into labeled and unlabelled [10] as shown in Figure 2 and Figure 3. If any class (output) is used to label the features (inputs), it is labeled data, otherwise unlabelled.

Country	Age	Salary	Purchased
France	44	72000	No
Spain	27	48000	Yes
Germany	38	56000	No
France	42	69000	Yes

Figure 2: Labeled Dataset

CustID	Gender	Age	Annual Income(K$)	Spending Score
1	Male	29	15	39
2	Male	31	15	81
3	Female	27	16	06
4	Female	34	16	77

Figure 3: Unlabeled Dataset

VI. DEPENDENT AND INDEPENDENT VARIABLES

All the features except the class label are called Independent Variables (X) and a Class label is termed as Dependent Variable (Y) [11]. The dataset consists of features like country, age and salary which is a matrix of independent variables and the class label purchased is a vector of dependent variable. Unlabelled dataset will only contain a dataset with just independent variables and no dependent variable.

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

VII. TYPES OF MACHINE LEARNING

Machine learning is categorized into three types. Supervised learning, Unsupervised learning and Reinforcement learning [5]. Supervised learning [6] is where the algorithm uses the input variables (X) and output variable (y) to learn the mapping from input to output. Here, the labeled dataset is provided to the machine to learn. Aim of Supervised Learning is to forecast the outcomes. If an unlabelled dataset is provided to the machine, the learning is unsupervised. Its aim is to discover underlying patterns and if no dataset is provided, it is reinforcement learning. Reinforcement learning automatically adapts and learns series of actions from the environment by producing actions or discovering errors.

VIII. MACHINE LEARNING MODELS

The three basic categories of machine learning models are Classification Model, Regression Model and Clustering Model. Model classification is depicted in Figure 4.

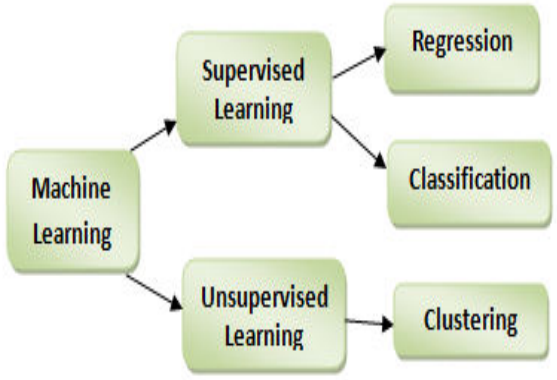

Figure 4: Machine Learning Model Classification

A) REGRESSION MODEL

Regression model is used to predict a continuous variable, an outcome (y) given the set of independent variables (X). The goal of regression model is to build a mathematical equation that defines y as a function of the x variables [13].

$$y=f(X)$$

Few Regression Models include Linear Regression, Logistic Regression and Polynomial Regression etc.

B) CLASSIFICATION MODEL

Classification model is the model where the data is categorized into a specified number of classes. This model uses the labeled data and identifies the category or the class to which a new data belongs to. Algorithms used for classification modeling include Support Vector Machine, K-Nearest Neighbors, Naïve-Bayes, Random Forest algorithms etc., [12].

C) CLUSTERING MODEL

As there will be no specific outcomes or class labels in the dataset, Clustering model focuses on identifying clusters of related records and labeling them according to the cluster to which they belong. This model learns from the patterns in the data being trained to the model [7]. Few clustering algorithms include K-means clustering, Hierarchical Clustering etc.,

IX. WHICH MODEL TO CHOOSE – MODEL SELECTION

Supervised learning uses both Regression and Classification models. Regression model is adapted only if the dataset contains

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

a dependent variable (class label) which is continuous. The classification model is adapted when the dependent variable is categorical [8].

Clustering model is case of unsupervised learning where the dataset with input features has no outcome or dependent variable.

X. MODEL EVALUATION

Dataset is usually split into Training Set and Test Set. The model is trained using training set and the test set is used to validate it on new data. The classic approach for split is 80%-20% split, 70%-30% or 90%-10%.

Performance of a classification model can be illustrated with an error matrix often termed as a confusion matrix [9]. It is a matrix that helps to visualize the performance of an algorithm. It is easy to identify the number of mislabeled classes.

	Class A Predicted	Class B Predicted
Class A Actual	**TP**	**FN**
Class B Actual	**FP**	**TN**

Figure 5: Confusion Matrix

The confusion matrix (2X2) is shown below in Figure 5 with four grids, namely True Positive, True Negative for correct number of predictions and True Negative, False Positive and False Negative for incorrect number of predictions. Here, the Classification Accuracy is given by

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Accuracy of the model cannot be measured based on performance of one test set. As test set changes, accuracy of the model may change. To improve the variance in accuracy, k-fold cross validation technique is used.

A) k-fold cross validation

In k-fold cross validation, more than one split (say k) on the dataset is done. These splits are termed as Folds.

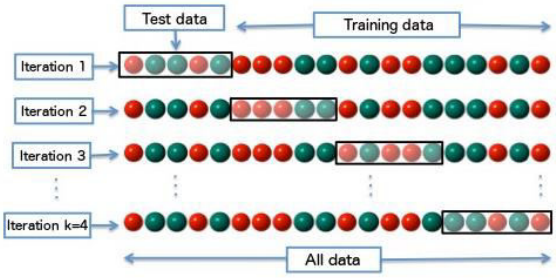

Figure 6: k-fold cross validation where k=4

Figure 6 shows the k-iterations for different combinations of (k-1) training and one test set. In k-fold cross validation [14], the model is trained with (k-1) folds and test the model on one remaining fold. Thus k different combinations for k-folds are used to train the model. To evaluate the variance in accuracies, Standard Deviation is computed by taking an average of accuracies obtained from k different combinations.

XI. MODEL PERFORMANCE

A model can be improved by finding the optimal values of hyperparameters. Any Machine Learning model is composed of two types of parameters. First type of parameters is the one that are learnt through machine learning algorithm. Second type of

parameters is the one we choose for the model. For example: 'kernel' in SVM Model. Lots of parameters are not learnt by the model. Selecting optimal values for hyper parameters will improve the model performance [15].

A) Grid Search Technique

For particular problem, Grid Search will help to choose linear model like SVM or non linear model like kernel SVM. It helps to find the optimal hyperparameters of a model for accurate prediction. The Grid Search technique [16] will map the key parameters with the optimized values.

XI. CONCLUSION

Machine Learning is a practice of training machines in distinct areas to carry out the activities that a human brain can do and that in a faster way. Machine Learning can be a Supervised or Unsupervised. If lesser amount of data with clear labels is available for training, opt for Supervised Learning. Unsupervised Learning would generally give better performance and results for large data sets [17].

REFERENCES

[1]. Lee, I., & Shin, Y. J. (2019). Machine learning for enterprises: Applications, algorithm selection, and challenges. Business Horizons.

[2]. Mullainathan, S., & Spiess, J. (2017). Machine learning: an applied econometric approach. Journal of Economic Perspectives, 31(2), 87-106.

[3]. Libbrecht, M. W., & Noble, W. S. (2015). Machine learning applications in genetics and genomics. Nature Reviews Genetics, 16(6), 321-332.

[4]. Song, C., Ristenpart, T., & Shmatikov, V. (2017, October). Machine learning models that remember too much. In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (pp. 587-601). ACM.

[5]. Rosman, B. (2019). Benjamin Rosman gives us an introductory overview. Quest, 15(3), 8-9.

[6]. El Naqa, I., & Murphy, M. J. (2015). What is machine learning?. In Machine Learning in Radiation Oncology (pp. 3-11). Springer, Cham.

[7]. Heung, B., Ho, H. C., Zhang, J., Knudby, A., Bulmer, C. E., & Schmidt, M. G. (2016). An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. Geoderma, 265, 62-77.

[8]. Valletta, J. J., Torney, C., Kings, M., Thornton, A., & Madden, J. (2017). Applications of machine learning in animal behaviour studies. Animal Behaviour, 124, 203-220.

[9]. Sumesh, A., Rameshkumar, K., Mohandas, K., & Babu, R. S. (2015). Use of machine learning algorithms for weld quality monitoring using acoustic signature. Procedia Computer Science, 50, 316-322.

[10]. Liu, H., & Cocea, M. (2018). Semi-supervised Learning Through Machine Based Labelling. In Granular Computing Based Machine Learning (pp. 23-28). Springer, Cham.

[11]. Mittal, M., Goyal, L. M., Sethi, J. K., & Hemanth, D. J. (2019). Monitoring the Impact of Economic Crisis on Crime in India Using Machine Learning. Computational Economics, 53(4), 1467-1485.

[12]. Wei, Y., Zhang, X., Shi, Y., Xia, L., Pan, S., Wu, J., ... & Zhao, X. (2018). A review of data-driven approaches for prediction and classification of building energy consumption. Renewable and Sustainable Energy Reviews, 82, 1027-1047.

[13]. Milosevic, N., Dehghantanha, A., & Choo, K. K. R. (2017). Machine learning aided Android malware classification. Computers & Electrical Engineering, 61, 266-274.

[14]. Grimm, K. J., Mazza, G. L., & Davoudzadeh, P. (2017). Model selection in finite mixture models: A k-fold cross-validation approach. Structural Equation Modeling: A Multidisciplinary Journal, 24(2), 246-256.

[15]. Mehmani, A., Chowdhury, S., Meinrenken, C., & Messac, A. (2018). Concurrent surrogate model selection (COSMOS): optimizing model type, kernel function, and hyper-parameters. Structural and Multidisciplinary Optimization, 57(3), 1093-1114.

[16]. Syarif, I., Prugel-Bennett, A., & Wills, G. (2016). SVM parameter optimization using grid search and genetic algorithm to improve classification performance. Telkomnika, 14(4), 1502.

[17]. Alpaydin, E. (2009). Introduction to machine learning. MIT press.