ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 2, February 2018

Analysis of Machine Learning Algorithms over Different Data Sets in R Programming

[1] R Akshara, [2] S Kranthi Reddy
[1] Assistant Professor, Vignan Institute of Technology and Science, Hyderabad, Telangana.
[2] Assistant Professor, Vignan Institute of Technology and Science, Hyderabad, Telangana.

Abstract: **- From last two decades onwards Machine Learning algorithms are used in many areas like Banking sector, online sites, social network sites, Health sector etc. By applying Machine learning algorithms on past data of organization, one can take certain decision for the benefits of organization. This paper mainly focuses on explaining the concept and applying different Machine Learning algorithms in R programming over different datasets.**

I. INTRODUCTION

Machine learning enables analysis of massive quantities of data. While it generally delivers faster, more accurate results in order to identify profitable opportunities or dangerous risks for organizations, it may also require additional time and resources to train it properly.

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves.

The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly.

Machine learning algorithms are often categorized as supervised learning or unsupervised learning

II. SUPERVISED & UNSUPERVISED LEARNING

The main goal in supervised learning is to learn a model from labeled training data that allows us to make predictions about unseen or future data. Here, the term supervised refers to a set of samples where the desired output signals (labels) are already known.

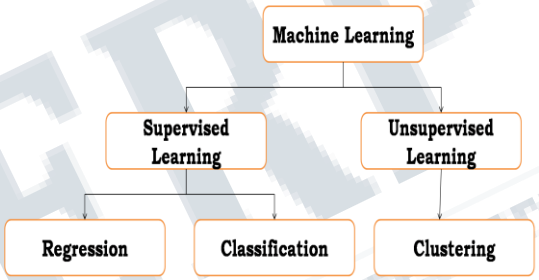


Figure: Types of Machine Learning Algorithms

labeled. Unsupervised learning studies how systems can infer a function to describe a hidden structure from unlabeled data. The system doesn't figure out the right output, but it explores the data and can draw inferences from datasets to describe hidden structures from unlabeled data.

Classification and regression comes under supervised learning algorithms whereas clustering algorithm comes under unsupervised learning algorithms.

III. LINEAR REGRESSION

It is an approach for modeling relationship between dependent variable and one or more explanatory variables. one explanatory variable mean simple linear regression. More than one explanatory variable mean multiple linear regression. Let us consider the following data sets which contain size and price of houses, here size is explanatory variable whereas y is dependent variable.

Size	1120	112	2102	2230	2600	3200	3409	3689
Price	1523	345	198	305	372	550	302	420

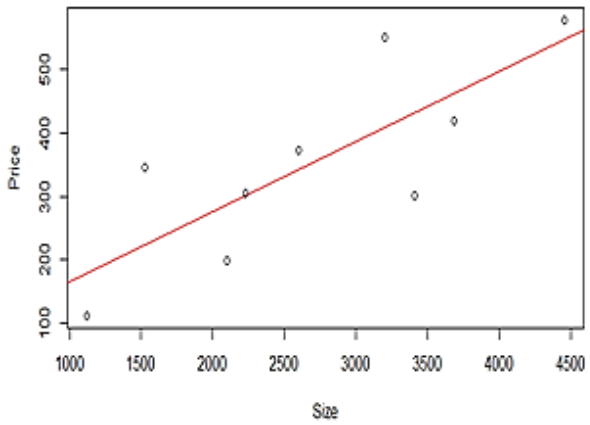
After applying linear regression to above data set in R programming the following graph is generated in which red line represents model i.e., y=a+bx

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 2, February 2018

Where c is constant and b is slope.

With help of above graph now we can predict price of future house. For example in R – Programming there is a function predict() used to predict the future price by giving size as




```
> predict(l,data.frame(Size=1600))
       1
231.5237
```

For the input house of size is 1600 linear regression model predicting the value as 231.523.

Advantages of Linear Regression:

Linear regression gives a statistical model that predicts the value of dependent variable based on independent variable, it is useful when relationships between the independent variables and the dependent variable are linear, and generates results.

Disadvantages of Linear Regression:

There is straight line relationship dependent and independent variables sometimes the results generated by linear regression are not correct.

Linear regression is not appropriate model for non-linear relationships.

IV. LOGISTIC REGRESSION

Logistic Regression is a classification not a regression algorithm. It is used to estimate discrete values (Binary values like 0/1, yes/no, true/false) based on given set of independent variable(s). In simple words, it predicts the probability of occurrence of an event by fitting data to a logit function. Hence, it is also known as logit regression. Since, it predicts the probability, its output values lies between 0 and 1 (as expected).

In logistic regression the probability function model is given as

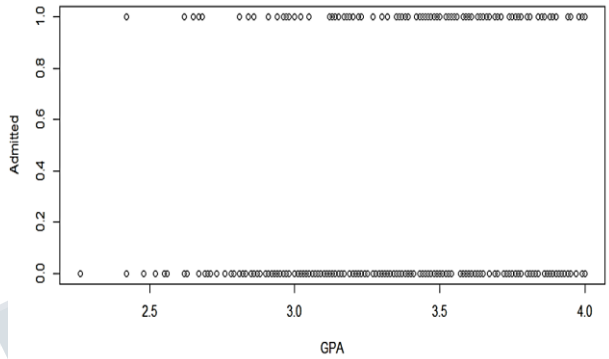
$p=ey/(1-ey)$

where $y = \beta 0+ \beta 1x1 +\beta 2x2+ \beta 3x3+\ldots..+ \beta nxn$

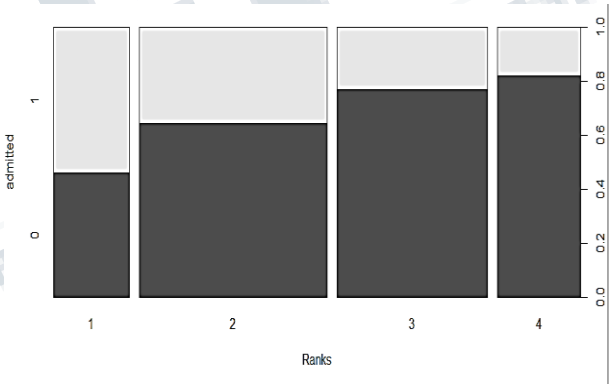
Let us consider student data set which contains information as admit, gre, gpa, rank. If student admitted into the college it contains 1 otherwise 0. Dividing the dataset into two parts training set as 80% and testing test as 20%.

After applying logistic regression on student data set, the model says that to predict the response variable the gre having least significant. We may or may not use that.

The following graph shows relation between admitted and GPA of students



The following graph shows relation between admitted and ranks of students.



On given train data misclassification that we got is about 27% where on test data misclassification we got about 30%.

For new input data the probability can be predicted as follows

```
newdata=data.frame(gpa = 4, rank = 1)
newdata$rank=as.factor(newdata$rank)
str(newdata)
s1<-predict(m, newdata , type="response")
s2=ifelse(s1>0.5,1,0)
s2
```

```
> s1<-predict(m, newdata , type="response")
> s2=ifelse(s1>0.5,1,0)
> s2
1
1
```

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 2, February 2018

V. K-NEAREST NEIGHBOR CLASSIFIER

KNN is supervised learning algorithm and it is very simple method for classification and regression. K-nearest neighbors [3] classifiers can classify examples by assigning them the class of the most similar labeled examples. KNN is a type of instance-based or lazy learning. KNN algorithm identifies K elements in the training dataset that are nearest in similarity. The unlabeled test example is assigned to the class of the majority of the K nearest neighbors. To classify the new data or unlabelled we need a distance function i.e., Euclidean-distance. Identifying the nearest neighbors i.e. determining the value of K is very important to KNN model. Taking high value for K has benefits which include reducing the variance due to the noisy data. Consider the iris data set which is divided into training set and test set. Training set contains 90% whereas test set contains 10%. After applying KNN on iris data the by considering k value as 13 we got 100% accuracy for test data as follows:

```
          m
test_ir_t  A  B  C
        A  1  0  0
        B  0 10  0
        C  0  0 10
```

Where A,B,C are classes m is predicted value
Test_ir_t is training set.

VI. DECISION TREE

Decision tree is another supervised learning algorithm which is used for classification problems. It works for both categorical and continuous data. In Decision Tree algorithm, we split the total sample into two or more sub sets based on most significant splitter on input data. Decision tree identifies the most significant variable and it's value that gives best homogeneous sub sets of sample. In order to identify the most significant variable decision uses various algorithms like ID3, C4.5, CART etc.

All three algorithms are used to find root node and splitting of decision node.

ID3: The ID3[6] algorithm was invented by Ross Quilan to create decision tree from data sets. By calculating entropy for every attribute in the dataset. The attribute with highest gain can be considered as significant node or root node.

The entropy for class attribute can be calculated as

$$E(c) = -p \log_2(p/p+n) - q \log_2(n/n+p)$$

Where p is probability of successes, q is probability of failure.

After calculating entropy for class. For each attribute we need to calculate gain.

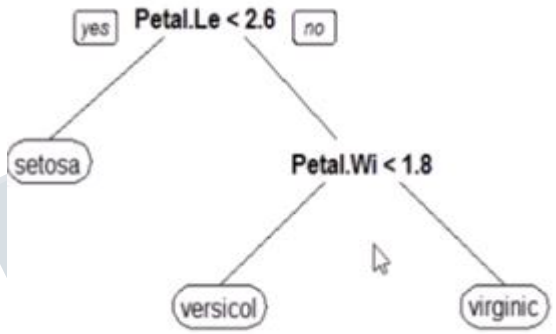
Gain(attribute)=Entropy of class-Entropy of attribute.

The attribute with highest gain can be considered as root node.

C4.5: It is extension of ID3 algorithm. C4.5 algorithm [3] handles both continuous and discrete values. It is widely used because of classification and high precision. C4.5 uses gain ration to select construct decision tree. The gain ratio is given as follows:

Gain ratio=information gain/entropy.

Now let us apply decision tree algorithm on iris data set in R programming. The Following tree is generated.



The above graph generated from R programming when we apply decision tree on iris data set.

VII. CONCLUSION

This paper specifies basic machine learning algorithms used in many area such linear regression, logistic regression, KNN. Linear regression is applicable whenever there is single explanatory variable on the other hand logistic regression gives probability of success or failure and KNN classes unlabelled data to a specific class. All the three algorithms are applied on different datasets in R programming and different accuracy got for all three algorithms among all three KNN given accuracy more.

REFERENCES

[1] Seema Sharma, Jitendra Agrawal, Shikha Agarwal, Sanjeev Sharma "Machine Learning Techniques for Data Mining: A Survey" 2013 IEEE International Conference on Computational Intelligence and Computing Research.

[2] Kajaree Das1, Rabi Narayan Behera "A survey on machine learning : concept, Algorithms and Applications" international journal of innovative research in computer and communication engineering, Vol. 5, Issue 2, February 2017.

[3] Balagatabi, Z. N., & Balagatabi, H. N. (2013). Comparison of Decision Tree and SVM Methods in Classification of Researcher's Cognitive St.

[4] Han, J., Kamber, M., & Pei, J. (2006). Data mining: concepts and techniques. Morgan kaufmann.

[5] Jason Bell, Wiley Machine Learning for Big Data.

[6] Mathur, N., Kumar, S., Kumar, S., & Jindal, R. The Base Strategy for ID3 Algorithm of Data Mining Using Havrda and Charvat Entropy Based on Decision Tree.

[7] Referredwebsites: https:// www. analyticsvidhya. Com /blog /2015 / https:// www.udemy. Com /machine-learning- in- r