International Journal of Advanced Research in Science Management and Technology
ISSN(Online)-2454-4159 Volume 3, Issue 4, April 2017

"Survey on Essentials of Machine Learning Algorithms with Python and R codes"

Shailendra Madansing Pardeshi[1],
Assistant professor, Department of IT, R. C. Patel institute of Technology, Shirpur, Maharashtra, India[1]

Abstract— Machine learning is a core sub-area of artificial intelligence as it enables computers to get into a mode of self-learning without being explicitly programmed. When exposed to new data, computer programs, are enabled to learn, grow, change, and develop by themselves. We are probably living in the most defining period of human history. The period when computing moved from large mainframes to PCs to cloud. What makes this period exciting for someone is the democratization of the tools and techniques, which followed the boost in computing. Today, as a data scientist, people can build data crunching machines with complex algorithms for a few dollors per hour.

Keywords— Python, Machine Learning, Programming with R, Decision tree, etc.

I.INTRODUCTION

In machine learning and statistics, classification is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known.

Types of Machine Learning Algorithms:

1.1.Supervised Learning: This algorithm consists of a target / outcome variable (or dependent variable) which is to be predicted from a given set of predictors (independent variables). Using these set of variables, we generate a function that map inputs to desired outputs. The training process continues until the model achieves a desired level of accuracy on the training data. Examples of Supervised Learning: Regression, Decision Tree, Random Forest, KNN, Logistic Regression etc.

1.2. Unsupervised Learning: In this algorithm, we do not have any target or outcome variable to predict / estimate. It is used for clustering population in different groups, which is widely used for segmenting customers in different groups for specific intervention. Examples of Unsupervised Learning: Apriori algorithm, K-means.

1.3 Reinforcement Learning: Using this algorithm, the machine is trained to make specific decisions. It works this way: the machine is exposed to an environment where it trains itself continually using trial and error. This machine learns from past experience and tries to capture the best possible knowledge to make accurate business decisions. Example of Reinforcement Learning: Markov Decision Process [1] [3].

II.LITERATURE SURVEY

List of Common Machine Learning Algorithms- Here is the list of commonly used machine learning algorithms. These algorithms can be applied to almost any data problem:

1.Linear Regression

2.Logistic Regression

3.Decision Tree

4.SVM

5.Naive Bayes

6.KNN

7.K-Means

8.Random Forest

9.Dimensionality Reduction Algorithms

10. Gradient Boost & Adaboost

2.1. Linear Regression

It is used to estimate real values (cost of houses, number of calls, total sales etc.) based on continuous variable(s). Here, we establish relationship between independent and dependent variables by fitting a best line. This best fit line is known as regression line and represented by a linear equation Y= a *X + b.

The best way to understand linear regression is to relive this experience of childhood. Let us say, you ask a child in fifth grade to arrange people in his class by increasing order of weight, without asking them their weights! What do you think the child will do? He / she would likely look (visually analyze) at the height and build of people and arrange them using a combination of these visible parameters. This is linear regression in real life! The child has actually figured out that height and build would be correlated to the weight by a relationship, which looks like the equation above. In this equation:

- Y – Dependent Variable
- a – Slope
- X – Independent variable
- b – Intercept

These coefficients a and b are derived based on minimizing the sum of squared difference of distance between data points and regression line. Look at the below example. Here we have identified the best fit line having linear equation=0.2811x+13.9. Now using this equation, we can find the weight, knowing the height of a person. In figure 2.1 it can mention properly.

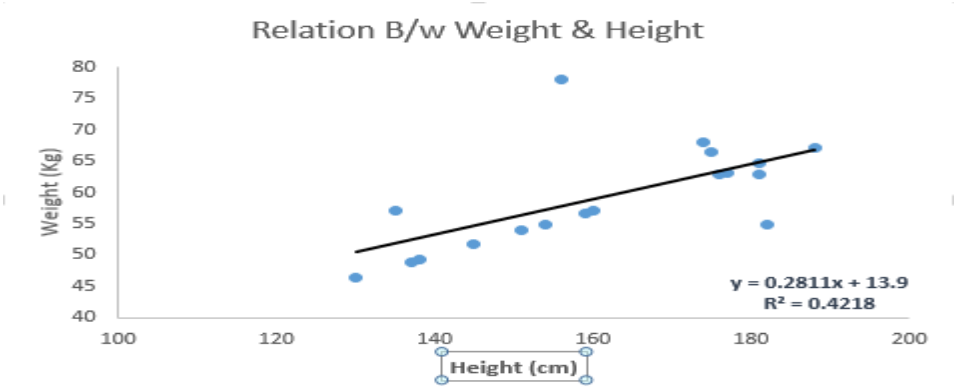

Fig.1 Linear Regression

Linear Regression is of mainly two types: Simple Linear Regression and Multiple Linear Regression. Simple Linear Regression is characterized by one independent variable [2] [3]. And, Multiple Linear Regression (as the name suggests) is characterized by multiple (more than 1) independent variables. While finding best fit line, you can fit a polynomial or curvilinear regression. And these are known as polynomial or curvilinear regression.

2.2. Logistic Regression

It 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 alogit function. Hence, it is also known as logit regression. In figure 2.2 it predicts the probability, its output values lies between 0 and 1.

Again, let us try and understand this through a simple example. Let's say your friend gives you a puzzle to solve. There are only 2 outcome scenarios – either you solve it or you don't. Now imagine that you are being given wide range of puzzles / quizzes in an attempt to understand which subjects you are good at. The outcome to this study would be something like this

– if you are given a trigonometry based tenth grade problem, you are 70% likely to solve it. On the other hand, if it is grade fifth history question, the probability of getting an answer is only 30%. This is what Logistic Regression provides you [4].

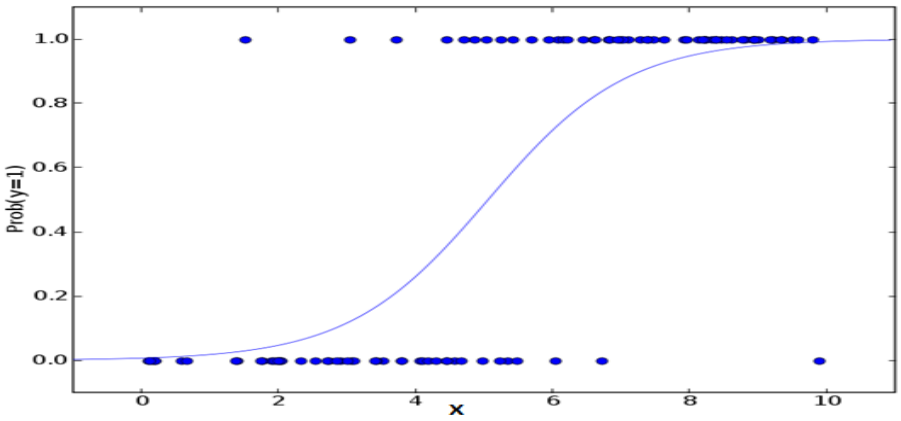

Fig.2 Logistic Regression

2.3. Decision Tree

This is one of my favorite algorithms and I use it quite frequently. It is a type of supervised learning algorithm that is mostly used for classification problems. Surprisingly, it works for both categorical and continuous dependent variables. In this algorithm, we split the population into two or more homogeneous sets. This is done based on most significant attributes/ independent variables to make as distinct groups as possible.

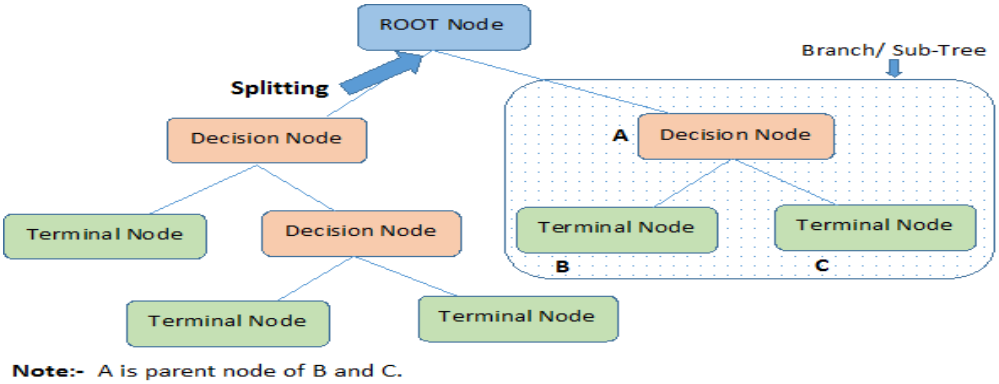

Fig.3 Decision Tree

In the figure 2.3, you can see that population is classified into four different groups based on multiple attributes to identify 'if they will play or not'. To split the population into different heterogeneous groups, it uses various techniques like Gini, Information Gain, Chi-square, entropy [3] [5]. Important Terminology related to Decision Trees is

- **Root Node**: It represents entire population or sample and this further gets divided into two or more homogeneous sets.
- **Splitting**: It is a process of dividing a node into two or more sub-nodes.
- **Decision Node**: When a sub-node splits into further sub-nodes, then it is called decision node.
- **Leaf/ Terminal Node**: Nodes do not split is called Leaf or Terminal node.
- **Pruning**: When we remove sub-nodes of a decision node, this process is called pruning. You can say opposite process of splitting.
- **Branch / Sub-Tree**: A sub section of entire tree is called branch or sub-tree.
- **Parent and Child Node**: A node, which is divided into sub-nodes is called parent node of sub-nodes where as sub-nodes are the child of parent node.

2.4. SVM (Support Vector Machine)

It is a classification method. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. For example, if we only had two features like Height and Hair length of an individual, we'd first plot these two variables in two dimensional space where each point has two co-ordinates (these co-ordinates are known as Support Vectors) [6].

2.5. Naive Bayes

It is a classification technique based on Bayes' theorem with an assumption of independence between predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter [7].

2.6. KNN (K- Nearest Neighbors)

It can be used for both classification and regression problems. However, it is more widely used in classification problems in the industry.

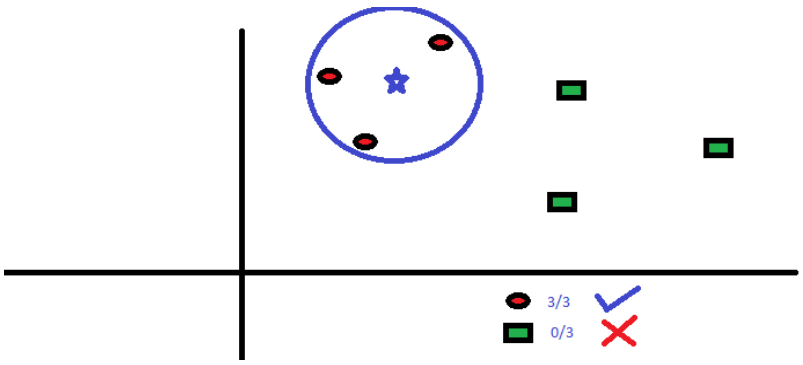

Fig.4 K- Nearest Neighbors

K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases by a majority vote of its k neighbors [8]. The case being assigned to the class is most common amongst its K nearest neighbors measured by a distance function. These distance functions can be Euclidean, Manhattan, Minkowski and Hamming distance. First three functions are used for continuous function and fourth one (Hamming) for categorical variables. In figure 2.4 if K = 1, then the case is simply assigned to the class of its nearest neighbor. At times, choosing K turns out to be a challenge while performing KNN modeling. KNN can easily be mapped to our real lives [9] [10]. If you want to learn about a person, of whom you have no information, you might like to find out about his close friends and the circles he moves in and gain access to his/her information.

2.7. K-Means

It is a type of unsupervised algorithm which solves the clustering problem. Its procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters). Data points inside a cluster are homogeneous and heterogeneous to peer groups [11] [12]. Remember figuring out shapes from ink blots? k means is somewhat similar this activity. You look at the shape and spread to decipher how many different clusters or population is present.

How K-means forms cluster:

- K-means picks k number of points for each cluster known as centroids.
- Each data point forms a cluster with the closest centroids i.e. k clusters.
- Finds the centroid of each cluster based on existing cluster members. Here we have new centroids.
- As we have new centroids, repeat step 2 and 3. Find the closest distance for each data point from new centroids and get associated with new k-clusters. Repeat this process until convergence occurs i.e. centroids does not change [13].

2.8. Random Forest

Random Forest is a trademark term for an ensemble of decision trees. In Random Forest, we've collection of decision trees. To classify a new object based on attributes, each tree gives a classification and each tree is planted & grown as follows:

• If the number of cases in the training set is N, then sample of N cases is taken at random but with replacement. This sample will be the training set for growing the tree.

• If there are M input variables, a number m<<M is specified such that at each node, m variables are selected at random out of the M and the best split on these m is used to split the node. The value of m is held constant during the forest growing.

• Each tree is grown to the largest extent possible [14] [15].

III. APPLICATIONS OVER MACHINE LEARNING ALGORITHMS

• **Face detection**: The face detection feature in mobile cameras is an example of what machine learning can do. Cameras can automatically snap a photo when someone smiles more accurately now than ever before because of advances in machine learning algorithms.

• **Face recognition**: This is where a computer program can identify an individual from a photo. You can find this feature on Facebook for automatically tagging people in photos where they appear. Advances in machine learning means more accurate auto-face tagging softwares.

• **Image classification**: A good example is the application of deep learning to improve image classification or image categorization in apps such as Google photos. Google photos would not be possible without advances in deep learning.

• **Speech recognition**: Another good example is Google now. Improvements in speech recognition systems has been made possible by, you guessed right, machine learning specifically deep learning.

• **Google**: Google defines itself as a machine learning company now. It is also a leader in this area because machine learning is a very important component to it's core advertising and search businesses. It applies machine learning to improve search results and search suggestions.

• **Anti-virus**: Machine learning is used in Anti-virus softwares to improve detection of malicious software on computer devices.

• **Anti-spam**: machine learning is also used to train better anti-spam software systems.

• Genetics: Classical data mining or clustering algorithms in machine learning such as agglomerative clustering algorithms are used in genetics to help find genes associated with a particular disease.

• **Signal denoising**: Machine learning algorithms such as the K-SVD which is just a generalization of k-means clustering are used to find a dictionary of vectors that can be sparsely linearly combined to approximate any given input signal. Thus such a technique is used in video compression and denoising.

• **Weather forecast**: Machine learning is applied in weather forecasting software to improve the quality of the forecast.

IV. CONCLUSION

The question of how to measure the performance of learning algorithms and classifiers has been investigated. This is a complex question with many aspects to consider One conclusion of the analysis is that classifier performance is often measured in terms of classification accuracy, e.g., with cross validation tests. Some methods were found to be general in the way that they can be used to evaluate any classifier (regardless of which algorithm was used to generate it) or any algorithm (regardless of the structure or representation of the classifiers it generates), while other methods only are applicable to a certain algorithm or representation of the classifier.

REFERENCES

[1] http://www.britannica.com/EBchecked/topic/1116194/machine-learning This tertiary source reuses information from other sources but does not name them.

[2] Ron Kohavi; Foster Provost (1998). "Glossary of terms". Machine Learning. 30: 271–274.

[3] Machine learning and pattern recognition "can be viewed as two facets of the same field."

[4] https://www.analyticsvidhya.com/blog/2015/08/common-machine-learning-algorithms/

[5] Wernick, Yang, Brankov, Yourganov and Strother, Machine Learning in Medical Imaging, IEEE Signal Processing Magazine, vol. 27, no. 4, July 2010, pp. 25–38.

[6] Mannila, Heikki (1996). Data mining: machine learning, statistics, and databases. Int'l Conf. Scientific and Statistical Database Management. IEEE Computer Society.

[7] Friedman, Jerome H. (1998). "Data Mining and Statistics: What's the connection?". Computing Science and Statistics. 29 (1): 3–9.

[8] "Machine Learning: What it is and why it matters". www.sas.com. Retrieved 2016-03-29.

[9] Harnad, Stevan (2008), "The Annotation Game: On Turing (1950) on Computing, Machinery, and Intelligence", in Epstein, Robert; Peters, Grace, The Turing Test Sourcebook: Philosophical and Methodological Issues in the Quest for the Thinking Computer, Kluwer

[10] Russell, Stuart; Norvig, Peter (2003) [1995]. Artificial Intelligence: A Modern Approach (2nd ed.). Prentice Hall. ISBN 978-0137903955.

[11] Langley, Pat (2011). "The changing science of machine learning". Machine Learning. 82 (3): 275–279. doi:10.1007/s10994-011-5242-y.

[12] Le Roux, Nicolas; Bengio, Yoshua; Fitzgibbon, Andrew (2012). "Improving First and Second-Order Methods by Modeling Uncertainty". In Sra, Suvrit; Nowozin, Sebastian; Wright, Stephen J. Optimization for Machine Learning. MIT Press. p. 404.

[13] Alpaydin, Ethem (2010). Introduction to Machine Learning. London: The MIT Press. ISBN 978-0-262-01243-0. Retrieved 4 February 2017.

[14] Cornell University Library. "Breiman: Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author)". Retrieved 8 August 2015.

[15] Gareth James; Daniela Witten; Trevor Hastie; Robert Tibshirani (2013). An Introduction to Statistical Learning. Springer. p. vii.