International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 03 | Mar 2020 www.irjet.net p-ISSN: 2395-0072

An Overview of Machine Learning Algorithms for Data Science

Rupesh Deshmukh[1], Milind Kubal[2]

[1,2]*Student, Dept. of Computer Engineering, Terna Engineering College, Nerul, Navi Mumbai, India*

Abstract – *Data Science has become a huge trend over the past few years. Organizations all over the world have realized the true intrinsic value of their data and the demand for data scientists has risen tremendously. Setting up Business Intelligence departments and making data-driven decisions has gained popularity. Uncovering knowledge and hidden patterns from huge chunks of data can prove highly beneficial to an organization in terms of profit or otherwise. But analyzing the data in spreadsheets for this information with the naked eye turns out to be time-consuming and highly inefficient. Various machine learning algorithms have been designed over the past decade to make the data classification and information extraction process effortless. In this paper, we describe some of the basic machine learning algorithms that every data science enthusiast should be familiar with.*

Key Words: Data Science, Machine Learning, Supervised, Unsupervised, Algorithms

1. INTRODUCTION

Data Science is the art of uncovering patterns and information from a huge chunk of data, that can prove beneficial to the business. The first step is to understand the business model and try to comprehend its goals and its vision for its customers. It is possible to find something in data if only we know what we are looking for. Most datasets are never perfect enough in their raw form to be able to extract information from it. The data needs to be cleaned, sorted and even scaled [1]. Missing values are to be handled and inconsistencies and redundancies have to be taken care of. After the data cleaning is complete, it is examined visually. If processing is performed on all the data available, it might take months to complete the analysis. To make it feasible, some important features are selected from the complete dataset, and other features are discarded. This saves a lot of unnecessary calculations and processing time. Finally, when the targeted dataset is ready, it can be processed using various machine learning algorithms.

The insights derived from the data can be in more than one form. They can be patterns, strays and even predictions for future data. These insights are easy to comprehend for data experts, but might not be understandable by ordinary business people. Hence, it needs to be presented visually in a way that could be understood by anyone. This is the final step of the data science process, and this presented information can be used for further business operations.

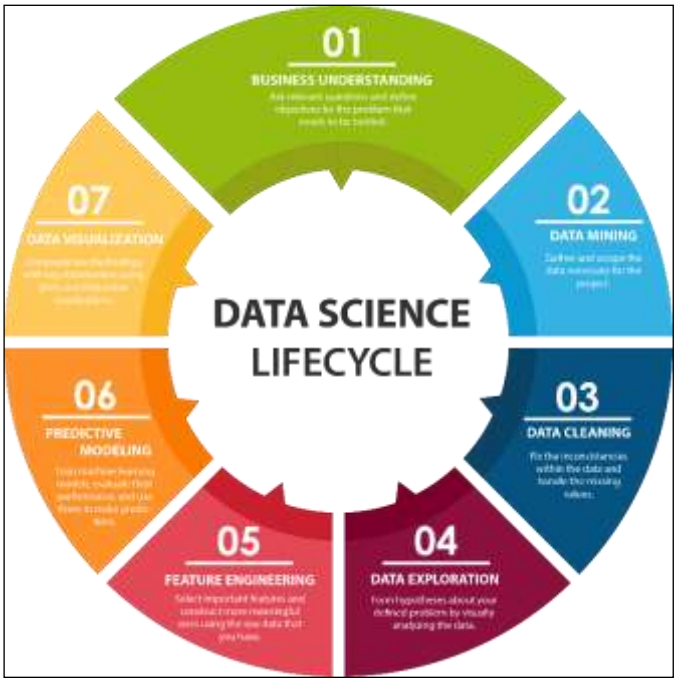

Fig -1: Data Science Lifecycle

Machine Learning has a huge variety of applications and they are growing every day. The algorithms in machine learning learn from the training data and tune the algorithm's parameters accordingly to accurately understand and classify the real-world data. The part of the dataset selected for training always contains the same features as the data to be classified. Machine Learning can be considered as one of the most important tools in a data science professional's toolset. Nowadays, a huge number of organizations rely on decisions backed by these algorithmic findings from the data. As not all the data can be considered useful, it is up to a data scientist to deal with redundant or incomplete data, select the appropriate algorithms and work with them to get the best insights out of the datasets.

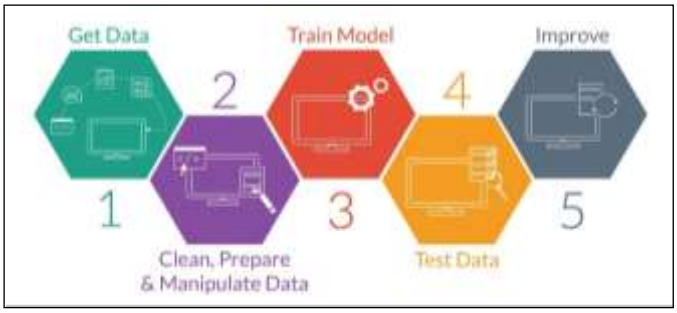

Fig -2: The Machine Learning Process

2. TYPES OF ALGORITHMS

Machine Learning algorithms can broadly be classified into two types – Supervised Learning and Unsupervised Learning.

2.1 Supervised Learning Algorithms

In Supervised Learning, the class labels are already known by the machine learning model. The goal is to learn from the training data and classify future data into the pre-defined classes as accurately as possible. Below are some of the supervised learning algorithms.

2.1.1 Decision Tree

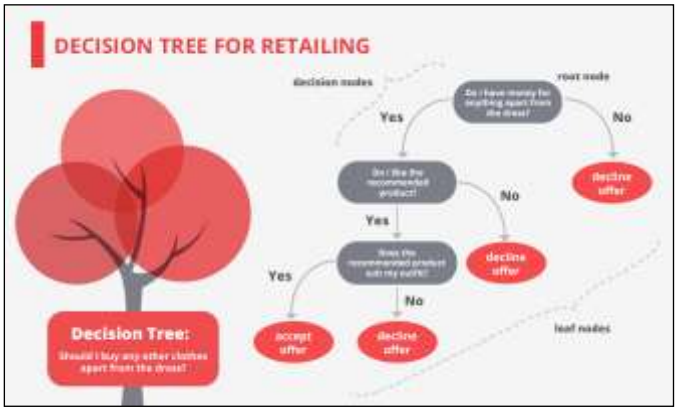

Fig -3: Example of a Decision Tree Model

One of the most classic algorithms, the Decision Tree algorithm consists of a flowchart-like structure that indicates various outputs as a result of different sequences of decisions. The decision tree model analyzes the training dataset and automatically creates a hierarchical tree-like structure [2]. This process is known as "Induction". Future data can be classified into different classes based on this tree. Although this tree doesn't require any expertise to comprehend visually, a huge tree can make it harder for the model to correctly classify new data and runs a risk of overfitting as the new data cannot be expected to be exactly accurate as training data. Hence, unnecessary branches are removed from the tree. This process is known as "Pruning", and helps reduce the complexity of the structure as well as makes it easier to understand.

Decision Trees can be further classified into two types – Categorical Variable Decision Tree and Continuous Variable Decision Tree. This is based on the nature of the target variable. For example, making financial decisions can only have outcomes like 'Yes' and 'No'; there cannot be a 'Maybe' outcome as it does not fit this application. But some qualities, like the willingness of a customer to buy a product, can be visualized on a continuous scale.

Tree-based algorithms empower prognosticative models with high accuracy and easy interpretation. They are used to map non-linear relationships as well. Techniques like decision trees, random forest and gradient boosting are being popularly used in all kinds of data science problems. They are useful to solve classification as well as regression problems.

2.1.2 K-Nearest Neighbours

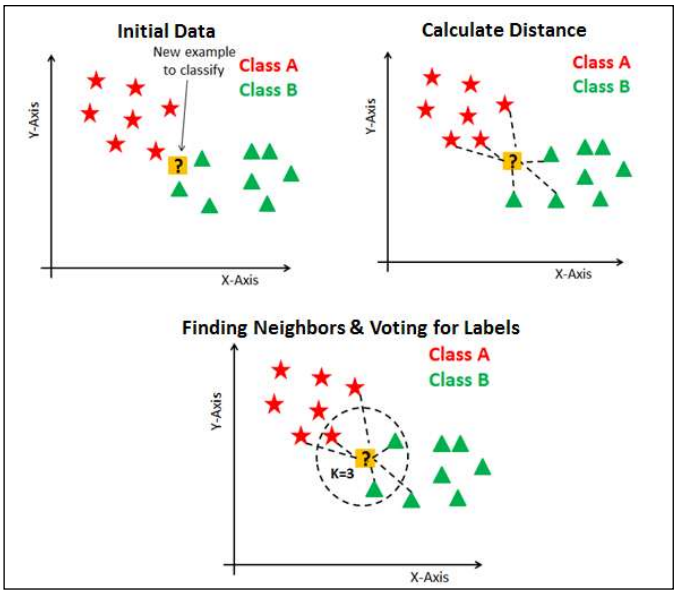

Fig -4: Steps for KNN Algorithm

The K-Nearest Neighbours (KNN) algorithm is a simple, easily interpretable supervised machine learning algorithm that can be used to solve both classification and regression problems. For each new data point in the classified dataset, we calculate the distance between the query point and the neighboring data points. The value of 'K' tells us about the number of closest neighbors to consider. After the neighboring data points are selected, we vote to find out the maximum points of a particular cluster among them. The new data point is assigned to the winning cluster [3].

KNN makes predictions based on the outcome of the 'K' neighbors closest to that point. Usually odd value of 'K' is selected to avoid ties during voting. To make predictions with KNN, we need to define a metric for measuring the distance between the new and the neighboring points from the data set. One of the most popular choices to measure this distance is Euclidean Distance, although Cosine, Chi-Square, and Minkowski measures are also used.

KNN can out-perform many other classifiers in selected applications that cannot entirely rely on features of their datasets. Sometimes it becomes impossible to fill in the missing data into the dataset. Hence, plotting the available features on a graph and classifying new points based on their proximity to original points becomes a better alternative. Applications like text mining and classification, and fingerprint and pattern recognition use KNN to achieve their goals optimally.

2.1.3 Linear Regression

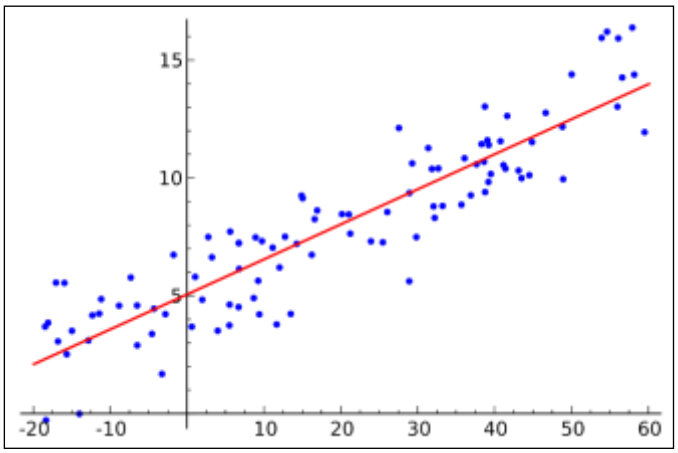

Fig -5: Example of a Linear Regression Model

Linear regression is a supervised machine learning algorithm that plots a linear relationship of a target variable against other variables. The model plots the available points from the dataset on a graph and finds a statistical relationship between the scalar variable with the target variable [4]. A statistical relationship differs from a deterministic relationship in the sense that it cannot be always accurately predicted. The basic idea of this algorithm is to plot a line that can be used to represent the trend of the target variable against other variables with minimum error. An error in this model can be defined as the closest distance from the point to the line. Once the line is plotted, it can be used to predict the value of the target variable for future independent variables.

The equation of Linear Regression line can be given as follows:

$$Y = B_0 + B_1 X$$

Here, 'Y' represents the target variable to be predicted, X is the independent variable or predictor. 'B_0' is the intercept and B_1 is the coefficient which represents the slope of the line. To check the accuracy of our model, the R-square metric can be formulated as follows:

$$R^2 = 1 - \frac{RSS}{TSS}$$

Here, 'RSS' is the Root Mean Square which is the squared difference between the observed value and the predicted value. 'TSS' is the Total Sum of Squares which is defined as the squared difference between the observed value and the mean. Typically, a high R2 value is associated with a good model, but it often depends on what the application considers as a fit model. Linear Regression is popular with applications that have a linear relationship but not exactly predictable. For example, changes in house prices with respect to the area or height of a person concerning his or her age.

2.2 Unsupervised Algorithms

Unsupervised algorithms look for trends in unlabeled classes of data with minimal or no human intervention. They are usually classified into two types – Clustering and Association.

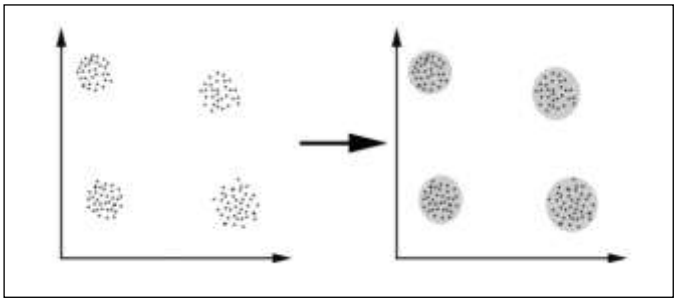

Fig -6: Clustering

Clustering can be defined as grouping data points into clusters based on the similarity between their features. Different datasets might require different similarity measures like Cosine, Jaccard or Euclidean to get the optimum clusters based on the application [5].

Association can be defined as grouping objects that are most frequently grouped based on past data. This does not depend on attributes of the object but rather on the frequent associations of objects with one another.

2.2.1 K-Means

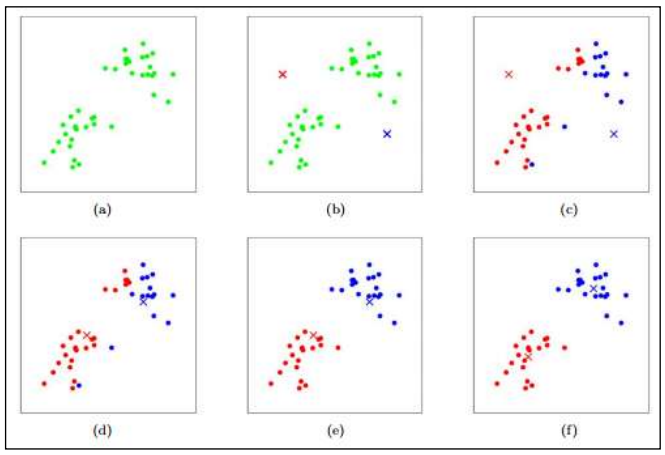

Fig -7: Steps for K-Means Algorithm

K-Means is the most basic algorithm used for partitioning data into the required number of clusters. For 'k' clusters, 'k' representatives are selected from the given data. These are called as centroids. The proximity of every data element with every centroid is calculated, and the data elements are assigned to a cluster with the lowest proximity from its centroid. The classical version of the K-Means algorithm used Euclidean distance as a proximity measure among data points plotted on a two-dimensional plane. Once all the points are allotted to some cluster, new centroids are calculated as the barycenter of each cluster, and the process is repeated until a fixed criterion is met,

usually until two consecutive iterations do not show a significant difference in the clusters [6].

Although the algorithm can be proved to terminate in every case, it does not guarantee optimal partitioning of data into clusters. Also, there are no guides for selecting the number of clusters and the initial representatives which lead to varied results with different values.

2.2.2 Hierarchical Clustering

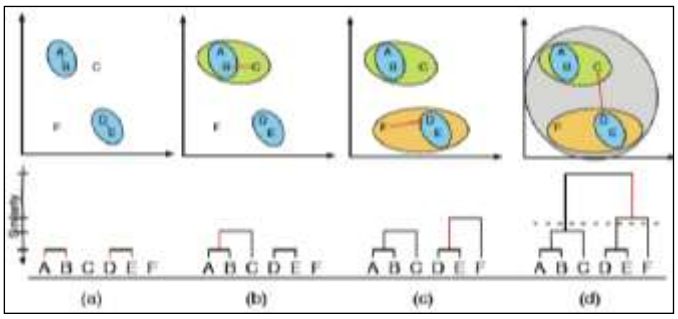

Fig -8: Steps for Hierarchical Clustering Algorithm

Some applications require datasets to be partitioned into nested clusters that might be arranged to form a tree-like structure. The leaf or the smallest clusters contain a single data element whereas the root or the largest cluster contains the whole data set. There are two approaches to creating the hierarchy – Agglomerative and Divisive [7].

The 'Agglomerative' method is a bottom-up approach that starts with every cluster containing a single object, and proceeds with merging clusters until one final cluster is obtained. The 'Divisive' method is a top-down approach that begins with the whole dataset in one cluster and splits clusters with each iteration until every cluster contains a single object.

Steps of creating the hierarchy can be given as follows:

1. Assign every object to an individual cluster.

2. After 'n' objects are assigned 'n' individual clusters, a cluster pair-wise similarity matrix of order 'n*n' is created.

3. Based on the matrix, a similar pair of clusters are merged into a single cluster, and the similarity matrix is updated for new clusters.

4. Step 3 is repeated until a single cluster remains or until set criteria are satisfied. Set criteria could be the maximum number of merges or leaves.

Hierarchical Clustering is used in applications that need to find the source or path from one node to another related node. For example, the evolution tree of a species or a possible source of a virus can be tracked using this algorithm.

2.2.3 Apriori Algorithm

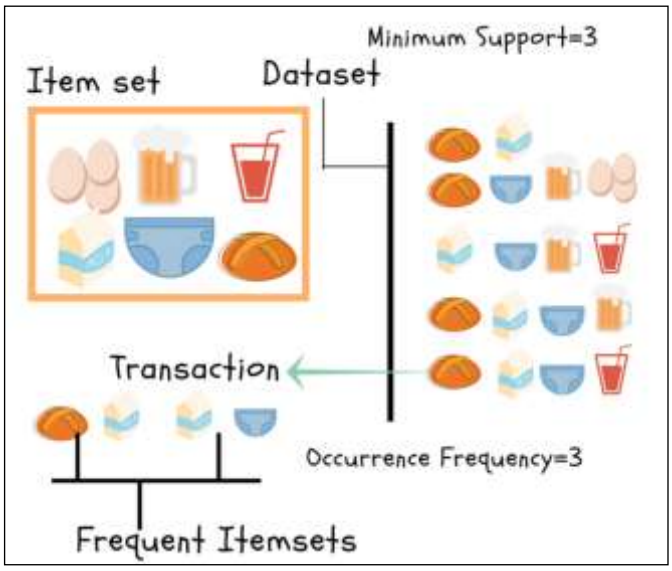

Fig -9: Apriori Algorithm

Apriori Algorithm is an association algorithm which is used to find frequent item-sets in a given dataset. An iterative or a level-wise approach can be used to eliminate the infrequent item-sets. It uses k-frequent item-sets to find (k+1)-frequent item-sets using the 'Apriori' property. The Apriori property states that if an item-set is frequent, all its non-empty subsets will also be considered frequent. This implies that, if an item-set is infrequent, all its supersets will also be infrequent.

To begin, we create a table for each item in the dataset. This table is known as 'Candidate Set'. Then we eliminate the candidates that have frequencies less than a pre-set support level. Once we get a valid candidate set, we create a Level 2 candidate set with two elements each in the item-set. This can be derived from the Level 1 candidate set using Apriori Property. This process is repeated until we get the required level of frequent item-sets. Association Rules can be derived from these item-sets, and these rules can be used to predict future item-sets [8].

In Data Science, these Association Rules can be of huge significance since they can be used to find out patterns in applications that record consumer behavior. For example, supermarket stores and online shopping portals use these insights to recommend similar products and services to customers to increase their sales. Association algorithms can be used in real-time as consumer data is constantly updated every day.

3. SUMMARY

Machine Learning models and algorithms are an important part of the Data Science process. They are useful to extract important insights and patterns from the data, which can be beneficial to the organization to achieve its goals. In the real world, no two applications are the same. The

algorithm that will provide optimal results varies for each application. Being familiar with the above machine learning algorithms will help data science professionals make the right decisions regarding their models.

REFERENCES

[1] Claus Weihs, Katja Ickstadt, "Data Science: The Impact of Statistics", International Journal of Data Science and Analytics, 2018.

[2] Himani Sharma, Sunil Kumar, "A Survey on Decision Tree Algorithms of Classification in Data Mining", International Journal of Science and Research (IJSR), 2016.

[3] Jasmina Novakovic, Alempije Veljovic, Sinisa Ilic, Milos Papic, "Experimental Study of Using the K-Nearest Neighbor Classifier with Filter Methods", Computer Science and Technology, At Varna, Bulgaria, June 2016.

[4] Azizur Rahman, Mayooran Thevaraja, Mathew Ekele Gabriel, "Recent Developments in Data Science: Comparing Linear, Ridge and Lasso Regressions Techniques Using Wine Data", DISP '19, Oxford, United Kingdom, April 2019.

[5] Memoona Khanam, Tahira Mahboob, Warda Imtiaz, Humaraia Abdul Ghafoor, Rabeea Sehar, "A Survey on Unsupervised Machine Learning Algorithms for Automation, Classification and Maintenance, International Journal of Computer Applications 119(13):34-39, June 2015.

[6] Md. Zakir Hossain, Md.Nasim Akhtar, R.B. Ahmad, Mostafijur Rahman, "A Dynamic K-Means Clustering for Data Mining", IJEECS, February 2019.

[7] Fahdah Alalyan, Nuha Zamzami, Nizar Bouguila, "Model-Based Hierarchical Clustering for Categorical Data", IEEE 28th International Symposium on Industrial Electronics (ISIE), June 2019.

[8] Foxiao Zhan, Xiaolan Zhu, Lei Zhang, Xuexi Wang, Lu Wang, Chaoyi Liu, "Summary of Association Rules", IOP Conference Series Earth and Environmental Science 252:032219, July 2019.