A Review on Data Science Techniques

Ruta Kulkarni[1*], Tanvi Pardhi[2]
[1,2]Department of Electronics and Telecommunication Engineering, MKSSS's Cummins College of Engineering for Women, Pune, India

***Abstract*: The term data science has garnered a huge attention in the past few decades and various research is being conducted in this field. There are a variety of techniques and technologies that are being developed and used in the field of data science. Data science is a blend of various techniques, technologies and theories – machine learning, statistics, data mining, mathematics and many other domains. It mostly deals with an aim of using the data strategically and such that we can gain insights from that data. This domain has experienced a boom because of the huge quantity of data that we are trying to collect and process in the last few decades. This review is about the different techniques related to data science and how they come into play according to the type of data that we handle.**

***Keywords*: Data science, decision tree, linear regression, clustering, machine learning, support-vector machine.**

1. Introduction

There are many different techniques that are regularly used in data science and it is important to understand which technique suits the best for a particular type of data. This review paper is an introduction to the different techniques used in data science. The techniques mentioned in the review paper are: Linear regression, Decision Tree, SVM, Clustering and Neural Networks.

2. Data Science and Techniques

A. What is Data Science?

The industry and academia related to computer science are majorly revolving around the concepts of data science, data analytics, big data, data mining, etc. Data science is more inclined towards the term 'science', it's about the science that is used for strategic use of the data available. A few decades ago data science was often interchangeably used with statistics, then the word slowly started having its own identity. It was in the paper: 'What is Data Science? Fundamental Concepts and a Heuristic Example' [1] that data science gained a new definition. Hayashi Chikio explained it as a combination of three aspects: collection, design and analysis of the data. Ten years later, the term 'data scientist' was coined and used by D J Patil and Jeff Hammerbacher in 2008. In 2002, the Data Science Journal was launched and data science started becoming a buzzword. The objective behind this is to find a new way to deal with the big data and finding alternatives to the traditional ways to handle data. The aim is to make use of the data in various domains, from medical to social sciences, from businesses to governments. As the data is getting larger and larger in quantity, the impact of data science is also increasing in the world [2].

B. Techniques in Data Science

1) Linear Regression

This approach is one of the earliest and most used due to the fact that the regression uses a linear style for modelling. The relationship between the target and the predictor is linearly modelled and this is advantageous because they are easier to fit. It is a statistical test that is used with data to get the relationship between the parameters. There exist two types of linear regression: simple linear regression and multiple linear regression.

Mathematically it can be defined as:

$$y= a_0+a_1x+ \varepsilon$$

Where y is the target variable/dependent variable, a0 is the intercept of the line, a1 is the linear regression coefficient, x is the predictor variable/independent variable and ε is the random error.

Linear regression assumed that there is always a linear relationship between the dependent and independent variables.

2) Decision Tree

A decision tree comprises of nodes and which are used to quantify the values. The topmost node in a decision tree is known as a root node. Each link between the nodes also known as branches, represents a condition. The leaf node is the outcome based on the previous node attribute and branch condition.

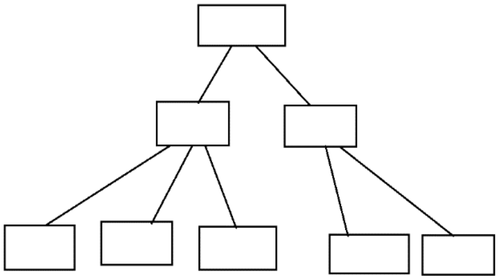

Fig. 1. Decision Tree

It is a predictive modelling approach and is used to represent decisions and decision making. There are decision nodes, end

[*]Corresponding author: rutakulkarni99@gmail.com

nodes and chances nodes that come together to form a tree like structure. They can grow very large depending on the nodes that are formed. The most common practice is the top-down induction of decision trees (TDIDT) [3]. They are formed by if and then conditions and very often they are too complicated to draw manually and the calculations too can get complicated. It can help to explores all possibilities till the end point in one go and easy to understand. Decision Trees can be classified as Categorical Variable decision tree or Continuous variable decision tree that processes datasets that are discrete or continuous in nature respectively. They split the attribute into possible outcomes by using splitting algorithms. Some of the splitting algorithms includes Classification and Regression Trees (CART), (Iterative Dichotomiser 3 (ID3), C4.5 – upgraded version of ID3, C5.0 – upgraded version of C4.5, Scalable Parallelizable Induction of decision Tree algorithm (SPRINT), Supervised Learning In Ques (SLIQ), Chi-square Automatic Interaction Detector (CHAID) and Multi-adaptive regression splines (MARS).

Evaluation of a decision tree is done by using confusion matrix. They are mostly used in Operations Research. They are easy to represent, understand and interpret and require very less data preparation.

3) Support Vector Machines

They are supervised learning models that are used for regression analysis. They can perform both linear and non-linear classifications. They are based on the principle of hyperplane. The hyperplane separates the data into the required classifications. SVM can be used for both linearly separable and non-linearly separable data; it can be used with labelled and unlabeled data. When the data is unlabeled, unsupervised learning is used which deals with clustering of data and then mapping these groups to the new data. For non-linear classification, kernel trick is used. They are used for classification and regression. They are mostly used for classification like classification of satellite data, classification of images, etc. Vladimir N. Vapnik and Alexey Ya. Chervonenkis invented the first SVM algorithm in 1963 and their colleagues designed a way to create nonlinear classifiers. This was done applying the kernel trick to maximum-margin hyperplanes [4].

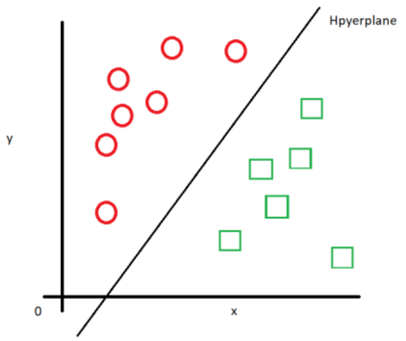

Fig. 2. Support Vector Machine

SVM is used in face recognition, where it can classify a face and other non-face objects. It is also used for handwriting recognition which is helpful with signatures. The disadvantage of SVM is that it is used only where two-class tasks are present but, in these cases, multi-class SVMs can be used. One more disadvantage comes when selecting the correct kernel.

4) Clustering

It is a way of grouping similar type of objects so that an object is more closely related to one group than the other. Different algorithms are used to solve the cluster analysis; the algorithms depend on what kind of data one is handling. Different cluster models have different cluster algorithms, there are various cluster models like the group-based models, neural models, density models, connectivity models, etc. There are also different ways of clustering, mainly hard and soft clustering. In hard clustering, an object either belongs to one cluster or does not; in soft clustering the object belongs to the cluster up to a certain degree. The different algorithms that are used: quantum clustering, Hoshen-Kopelman algorithm, fuzzy clustering, nearest-neighbor chain algorithm, etc.

One of the popular clustering techniques is K-means clustering. Nearby group of observations are partitioned into k clusters based on closest centroid. Clusters are formed recursively using Euclidean distance until centroids become stable. This technique is effective for large datasets.

EUCLIDEAN DISTANCE (X, Y) =
$(|X1- Y1|^{2} + |X2-Y2|^{2} + \ldots + |XN-1-YN-1|^{2} + |XN-YN|^{2})^{½}$

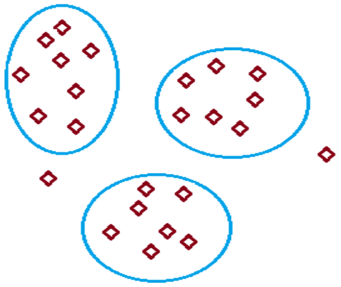

Fig. 3. K-means clustering

5) Neural Networks

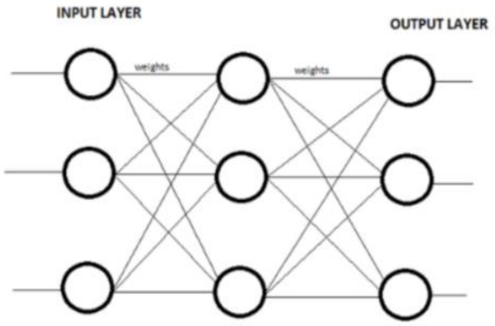

Fig. 4. Layers of Neural Network

A neural network consists of different layers of interconnected neurons. The first and last layers are called input layer and output layers respectively. Each neuron mimics the biological neuron in terms of computation and communication with the succeeding neuron. such connections with next layer neurons have a weight and bias associated.

$$\sum_{n}^{i=1} wixi + bias = w1x1 + w2x2 + w3x3 + \ldots + wnxn + bias$$

$$output = f(x) = 1 \text{ if } \sum_{n}^{i=1} wixi + b >= 0;$$

$$output = f(x) = 0 \text{ if } \sum_{n}^{i=1} wixi + b < 0$$

Input at a node is multiplied with its weight, added together and passed through the activation function. If the output exceeds given threshold, then the out is passed on to next layer neuron. Such passing on of data to succeeding layers is called as feedforward network.

Neural networks can be commonly be classified as perceptron, MLPs, or CNN.

Perceptron neural network is the oldest and simplest form of neural network consisting of a single neuron. Multi-Layer Perceptron (MLP) also known as feedforward neural network comprises on input layer, output layer and one or many hidden layers in between input and output layers. MLP actually makes use of sigmoid neurons to handle nonlinear data. Data is fed to train the network and used to solve other real-life problems. MLPs are applied in NLP, computer vision and forms foundation for other networks.

Convolution Neural Network (CNN) works similar to feedforward network with additional linear algebra, matrices concepts for pattern recognition in images and classification.

Neural network models are trained on a part of data set and tested on rest of the data. Types of training models are supervised – labelled data, unsupervised – unlabelled data and re-enforced - feedback mechanism.

3. Conclusion

Different algorithms and techniques are used for different datasets. It is important to use the correct type of algorithm to get the optimum output. We may have to deal with data that is unlabeled and need to find patterns in the dataset, in these cases supervised learning is the best. For unsupervised learning, we have to use clustering, like the K-means clustering and K nearest neighbors. We can also use neural networks to work with unsupervised learning.

Then we have labelled data where we can use supervised learning to train the model. Techniques like linear regression, non-linear regression, regression trees are best suited for such cases. Other techniques that are used are SVM, decision trees, random forest, and logistic regression. Decision trees could be used on smaller datasets whereas clustering and Neural networks could be a better choice for larger datasets.

Based on the requirement, one can use any of the above techniques to solve the task.

References

[1] Hayashi C., Yajima K., Bock HH., Ohsumi N., Tanaka Y., Baba Y. “What is Data Science? Fundamental Concepts and a Heuristic Example”. Data Science, Classification, and Related Methods. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Tokyo.

[2] Martin, Sophia "How Data Science Will Impact Future of Businesses?". Medium. 2019.

[3] Quinlan, J. R. "Induction of decision trees" (PDF). Machine Learning. 1: 81–106.

[4] Boser, Bernhard E.; Guyon, Isabelle M.; Vapnik, Vladimir N. "A training algorithm for optimal margin classifiers". Proceedings of the fifth annual workshop on Computational learning theory – COLT '92, 1992.

[5] Harsh H. Patel, Purvi Prajapati, “Study and Analysis of Decision Tree Based Classification Algorithms,” in International Journal of Computer Sciences and Engineering, vol. 6, no. 10, Oct. 2018

[6] Anuja Priyama, Abhijeeta, Rahul Gupta, Anju Ratheeb, and Saurabh Srivastava, “Comparative Analysis of Decision Tree Classification Algorithms,” International Journal of Current Engineering and Technology, vol. 3, no. 2, June 2013.

[7] K. Chitra and D. Maheswari, “A Comparative Study of Various Clustering Algorithms in Data Mining,” in International Journal of Computer Science and Mobile Computing, vol. 6, no. 8, August 2017.

[8] Nitin Malik, Artificial Neural Networks and their Application,” in National Conference on ‘Unearthing Technological Developments & their Transfer for Serving Masses’ GLA ITM, Mathura, India 17-18 April 2005.