The Role of Mathematics and Statistics in the Field of Data Science

Neethumol K.G[1] ,
Asst. Professor, Department Of IT/ CS
Pillai HOC College of Arts, Science and and Commerce, Rasayani

Priya Prakash[2] ,
Asst. Professor, Department Of IT/ CS
Pillai HOC College of Arts, Science and Commerce, Rasayani

Abstract— Mathematics is the solid foundation of any contemporary discipline of science and also which is the very important in the field of data science as concepts within mathematics aid in identifying patterns and assist in creating algorithms. Statistics and math are the very important in Data Science but only because of the concepts they surface and the tools they make possible. Beyond the basics of calculus, discrete mathematics and linear algebra there is a certain kind of mathematical thinking which is required to understand data. The understanding of various concepts of Statistics and Probability Theory are key for the implementation of such algorithms in data science. Almost all the techniques of modern data science, including machine learning, have a deep mathematical foundation. In this paper, the evidence to support our premise that math and statistics are the most important disciplines to provide tools and methods to find structure into data. The knowledge of math is very important for newcomers arriving at data science from other professions.

Index Terms— Algorithm, data science, Statistics and Probability

1 INTRODUCTION

Data Science as a scientific discipline is influenced by Mathematics, Statistics, computer science, operations research, and informatics as well as the applied sciences.

Mathematics and Statistics are essential in the field of Data Science because these disciples form the basic foundation of all the Machine Learning Algorithms. Mathematics is behind everything around us, from shapes, patterns, sequences and colors to the count of petals in a flower. Mathematics is embedded in each and every aspect of our real life. Although having a good understanding and making of programming languages, Machine Learning algorithms and following a data-driven approach is necessary to become a Data Scientist, Data Science isn't all about these fields. This paper is divided into two parts, first will see the impact of Mathematics in Data science and the second section will discussed about Statistics and Data science.

Nowadays, these ideas are combined with Data Science, leading to different definitions. One of the most comprehensive definitions of Data Science was recently given by Cao as the formula as follows [1]:

Data science = (statistics + informatics + computing + communication + sociology + management) | (data + environment + thinking), analysis, and visualization).

2 RESEARCH METHODOLOGY

The nature of research is descriptive and analytical research. Secondary data from books, newspapers and websites have been used to study and analyze the issue on hand

3. MATHEMATICS FOR DATA SCIENCE

Mathematics has already created an impact on every discipline. The significance of the usage of mathematics varies according to the disciplines. There are two main components of mathematics that contribute to Data Science they are namely Linear Algebra and Calculus. In this section of mathematics for data science, will briefly overview these two fields and learn how they contribute towards Data Science.

3.1 Linear Algebra

Linear algebra is the branch of mathematics concerning linear equations such as the study of lines and planes, vector spaces and mappings that are required for linear transforms. Linear algebra is central to all areas of mathematics and which is the most important math skill in machine. It is one of the branch of mathematics, but mainly concerning that linear algebra is the mathematics of data where matrices and vectors are the language of data. Linear Algebra is designed for solving problems of linear equations. These equations can sometimes

contain variables of higher dimension. These higher dimension variables cannot be visualized or manipulated directly. Therefore it is necessary to use the power of matrices to help in manipulating data of n-dimensions and linear algebra most commonly used in machine learning to understand how algorithms work under the hood [4]. When coming to basics, there are three kinds of matrices as follows.

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3n} \\ \vdots & \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \cdots & a_{mn} \end{bmatrix}$$

← second row

↑ third column

eans hav-

and have
ally m-

Linear Algebra Techniques for Data Science

There are some important linear algebra techniques that are used in data science are follows.

a) Single value decomposition - In linear algebra, the single value decomposition is a factorization of a real matrix or a complex matrix that generalizes the Eigen decomposition of a square normal matrix to any other matrix. Singular Value Decomposition allows to manipulate matrices by dividing them into three different matrices. These matrices will be a product of scaling, rotation, and shearing.

b)Eigen value Decomposition - Eigen value decomposition allows to reduce matrices in order to improve the operations of matrix. It helps to generate new vectors that are in the same direction as the former ones. In advance that decompose the matrix into Eigen value and eigenvectors.

c)Principal Component Analysis - In order to reduce higher dimensions, it is essential to use the Principal Component Analysis. It is most generally used for dimensionality reduction, which is that the processing of reducing the amount of variables or dimensions without losing strongly correlated labels.

3.2 Calculus

Another important requirement for Data Science is Calculus. Calculus is employed essentially in optimization techniques. Without calculus, it's very difficult to urge a deep knowledge of machine learning within the field of data science. Using calculus, you'll perform mathematical modeling of artificial neural networks and also increase their accuracy and performance.
Calculus can be classified into two areas, namely - Differential calculus and Integral calculus.

Differential Calculus

Differential Calculus studies the speed at which the quantities change. Derivatives are most generally used for locating the maxima and minima of the functions and they are utilized in optimization techniques where we have to seek out the minima so as to attenuate the error function. Another important concept of derivate is to realize that the partial derivatives that are used for designing back propagation in neural networks. Chain Rule is another important concept want to compute back propagation. Aside from minimizing error functions and back propagation, we utilize differential theory of games for Generative Adversarial Neural networks.

Integral Calculus

Integral Calculus is the mathematical methods to find the area, volume ,central points and many useful things.Integrals are divided into definite integrals and indefinite integrals.Integration in data science is used for computing probability density functions and variance of the random variable.

4. STATISTICS FOR DATA SCIENCE

Statistics is a Mathematical Science concerning data collection, analysis, interpretation and presentation. Statistics is employed to process complex problems within the world in order that Data Scientists and Analysts can search for meaningful trends and changes in Data. In simple words, Statistics are often wont to derive meaningful insights from data by performing mathematical computations. Several Statistical functions, principles, and algorithms are implemented in the field of researching data, build a Statistical Model and infer or predict the result. The field of Statistics has an influence over all domains of life; the stock exchange , life sciences, weather, retail, insurance, and education are but to call a couple of . It is important to know some of the basic terminologies in Statistics while dealing with Data Science, they are follows [3]:

a) Population is the set of sources from which data has to be collected.
b) A Sample is a subset of the Population
c) A Variable is any characteristics, number, or quantity that can be measured or counted. It also be called a data item.
d) Also known as a statistical model, A statistical Parameter or population parameter is a quantity that indexes a family of probability distributions. For example, the mean, median, etc of a population.

4.1.Types of Analysis in Statistics

An analysis of any event can be done in one of two ways:

a) Quantitative Analysis:
Quantitative Analysis or the Statistical Analysis is the science of collecting and interpreting data with numbers, letters and

graphs to identify patterns and trends.

b)Qualitative Analysis
: Qualitative or Non-Statistical Analysis gives general information and uses text, sound and other forms of media to do so.

For example, if there is a need of purchasing a coffee from Starbucks, it is available in Short, Tall and Grande. This is an example of Qualitative Analysis. But if a store sells 80 regular coffees a week, it is Quantitative Analysis because we have a number representing the coffees sold per week. Even if the purpose of both these analyses is to provide results, the Quantitative analysis provides a clearer picture hence making it crucial in analytics.

4.2.Categories In Statistics

There are two important categories in Statistics, namely:

- **Descriptive Statistics**
- **Inferential Statistics**

Descriptive Statistics

Descriptive Statistics uses the information to supply descriptions of the population, either through numerical calculations or graphs or tables and it helps organize data and focuses on the characteristics of knowledge providing parameters. Suppose you would like to review the typical height of scholars during a classroom, in descriptive statistics you'd record the heights of all students within the class then you would find out the maximum, minimum and average height of the class. Descriptive Statistics or summary statistics is used for describing the data. It deals with the quantitative summarization of data. This summarization is performed through graphs or numerical representations. In order to have a full grasp of descriptive statistics, you must possess some of the following key concepts [2]:

a) Normal Distribution

In a normal distribution, we need to represent a large number of data samples in a plot. So in Gaussian distribution , we represent an outsized number of knowledge samples during a plot. Using Gaussian distribution , we represent large values of variables during a normal curve which is additionally referred to as a normal curve . This bell curve is symmetric in nature, meaning that the values further far away from the mean taper off equally in both the left and right directions. For undertaking inferential statistics, it's mandatory that the info is generally distributed.

b) Central Tendency

Using a central tendency, first identify the central point of the given data. Mean, Median and Mode are the most important parts of central tendency. Mean is the arithmetic average of all the values in the given sample data. Whereas, the median is the middle value of the and mode, which is the most frequently occurring value in the given sample.

c) Skewness & Kurtosis

There are often instances of knowledge , where the distribution doesn't exhibit any sort of symmetry. For instance , a normal curve has zero skewness. When more data accumulates to the left side, we observe a positive skew and when data accumulates on the proper side, then there is a negative skew. Kurtosis measures the "tailedness" of the graph. By tailedness, we infer that kurtosis measures the acute values in either tails of the graph. Basically, distributions with an outsized kurtosis have tails that are larger than those exhibited by normal distributions whereas, negative kurtosis has smaller tails than normal distributions.

d)Variability

Variability measures the distance between the data-point and the central mean of the distribution. There are so many various measures of variability such as range, variance, standard-deviation and inter-quartile ranges.

Inferential Statistics

Inferential Statistics is that the procedure of inferring or concluding from the info . In inferential statistics, make a conclusion about the larger population by running several tests and deductions from the smaller sample. For instance , during an election survey, you would like to understand what percentage people support a specific party . so as to try to to this, you merely need to ask everyone about their views. This idea and movement is just not right, because there are billions of individuals in India and surveying every single person is an excruciatingly difficult task. Therefore, select a smaller sample, make deductions from that sample and attribute our observations on the larger population.
There are various techniques in inferential statistics which are useful in the field of data science. Some of these techniques are follows.

a) Central Limit Theorem

In the concept of central limit theorem, the mean of the smaller sample is the same as that of the mean of the larger population. So the resulting standard deviation is equal to the standard deviation of the population. The most important concept of the Central Limit Theorem is the estimation of the population mean and margin error can be calculated by multiplying the standard error of the mean with the z-score of the percentage of confidence level [6].

b)Hypothesis Testing

Hypothesis testing is the measure of assumption of the data. Using hypothesis testing, attribute the results from a smaller sample on a much larger group. There are two hypotheses that require to test against each other – Null Hypothesis and Alternate Hypothesis. A null hypothesis represents the ideal scenario whereas an alternate hypothesis is usually the opposite of that.

c) ANOVA

Using ANOVA, we can test our hypothesis for multiple groups. It is an improvement of another form of an inferential technique that is called t-test. ANOVA ,which performs the testing with a minimal error rate. One metric for measuring ANOVA is called f-ratio. F-ratio is that the ratio of the mean-square internally to the group and mean-square between the groups.

5.CONCLUSION

Following the above assessment of the capabilities and impacts of mathematics and statistics our conclusion is:

The role of Statistics and math are the very important in Data Science but only because of the concepts they surface and the tools they make possible, compared to computer science.

With the help of mathematical methods and computational algorithms and statistical reasoning, particularly for big data, will help to scientifically appropriate results based on suitable approaches.

6.REFERENCES

[1]. Weihs, C., Ickstadt, K. Data Science: the impact of statistics. Int J Data Sci Anal 6, 189–194 (2018). https://doi.org/10.1007/s41060-018-0102-5

[2].. https://data-flair.training/blogs/math-and-statistics-for-data-science/

[3]. https://dzone.com/articles/a-complete-guide-to-math-and-statistics-for-data-s

[4]. https://towardsdatascience.com/mathematics-for-data-science-e53939ee8306

[5]. https://medium.com/s/story/essential-math-for-data-science-why-and-how-e88271367fbd

[6].https://www.dataquest.io/blog/math-in-data-science/

[7]. https://link.springer.com/chapter/10.1007/978-3-662-49851-4_1

[8].file:///C:/Users/Administrator/Downloads/A_Study_on_Role_of_MathematicsStatistics_in_IT_Fields.pdf