Original Research Paper

Volume - 11 | Issue - 07 | July - 2021 | PRINT ISSN No. 2249 - 555X | DOI : 10.36106/ijar

Mathematics

MATHEMATICS FOR DATA SCIENCE

Anjali Nair* *Corresponding Author

Ambika Biradar

Dr. Radhika Menon

ABSTRACT In this paper we will discuss about role of Mathematics in Data Science. All Mathematical concept help us to understand the mechanism of the algorithm like what is happening, why it's happening, and how we can optimize it to obtain the required result.

KEYWORDS :

INTRODUCTION

The knowledge of Mathematical concepts is essential particularly for all professional fields where data is the core part. In this paper we will discuss all the concept in detail. In section 1 we will discuss multi variable calculus in section 2 linear algebra, and in section 3 Probability and Statistics which are needed to understand the behaviour of any of these techniques theoretically.

1. Calculus In Data Science

Machine Learning algorithms aims at performing well on the training the data and developing the model and using the test data for validation of the model which ensures that model best fits the data. It deals with minimizing a cost function (also called an objective function), which is a scalar function of several variables that typically measures how accurately our model fits the data. Calculus and optimization play an important role in these algorithms.

Calculus is a branch of Mathematics which deals with the rate of change of quantities. The calculus is divided into two parts differential calculus and integral calculus. In differential calculus we cut a large piece into small parts to see how it changes and in integral calculus we join the small pieces together to find out how much there. Database can be seen as calculus, because relational calculus, also named asrelational algebra, is the foundation for relational database systems. [1]

In machine learning deals with minimizing a cost function that measures the performance of a model for any given data and also it quantifies the error between predicted values and expected values and presents it in the form of a single real number. After making a hypothesis with initial parameters, we calculate the Cost function to reduce the cost function, we modify the parameters by using the Gradient descent method which is used to find local minima of the cost function.

2. Linear Algebra

Linear Algebra is the branch of Mathematics which deals with linear equations, linear functions and their representations through matrices and vector spaces. It helps to understand geometric terms in higher dimensions. Linear algebra ia one of the foundational block of data science.

Linear algebra is used in a lot of algorithms. We have used linear algebra in principal component analysis which is used to reduce the dimensionality of our data. In PCA we should know how many principal components to preserve and this can be done effectively with the help of concepts of Linear algebra see [2]. All neural network algorithms use linear algebra techniques to represent and process network structures and learning operations.

Vectors and matrices are key part of data analysis. When we build a model, we fit a matrix of features and if we want to take distance between elements in our data, we are usually finding geometric distance between vectors of features.\\ Most of the machine learning algorithms perform matrix algebra during training and predicting phases. If we know how matrix algebra works it will help to understand algorithms. Matrix algebra is used in a number of areas machine learning such as regression, Neural networks, Classification algorithms, Dimension's reduction algorithms heavily depend on the matrix algebra. Eigen vector and eigen values live in the heart of the data science. We can represent a large set of information in a matrix. One eigen value and eigen vector are used to find key information that is stored in a large matrix. The technique of eigenvalues and eigen values are used to compress the data. Many algorithms such as PCA rely on eigen values and eigenvectors to reduce the dimensions.

3. Probability And Statistics

Statistics is branch of Mathematics that deals with collection, analysing and interpreting large amounts of data. Statistics allow us to drive knowledge from large data sets and then this knowledge can be used to make predictions and decisions. Statistics is the use to perform technical analysis of data. A basic visualization such as a bar chart might give some high-level information but with statistics, we get to operate on the data in a much more information driven and targeted way the math involved helps us form concrete conclusions about our data. To explore data set we have to use many statistical parameters like mean, variance and percentiles and many others. In a basic box plot, percentile values give minimum and maximum values also represent the upper and lower ends of data range. The position of median whether it is close to bottom or top we know that the data has lower values or higher values. Standard deviation and variance gives the directions of data variation and correlation coefficients measures the degree of relationship between the data variables.

Some of the Correlation techniques are

- **Covariance:** Mathematical formula for the covariance of x,y is where E(X) & E(Y) are mean values of X and Y.The sign of the covariance indicates the degree of the linear relationship between the variable.

$$Cov(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - E(x))\,(y_i - E(y)$$

- **Pearson Correlation Coefficient:** It is a statistic that measure the linear correlation between two features.

$$r(x,y) = \frac{Cov(x,y)}{\sigma(x)\sigma(y)}$$

It has a value between -1 and +1.

- **Spearman Rank Correlation Coefficient:** The Spearman rank correlation coefficient between the two sample is equal to the Pearson correlation coefficient between the rank values of those two samples. In Spearman rank correlation,
- 1 is a perfect positive correlation.
- 0 is no relation.
- -1 is a perfect negative correlation.

To understand different models and various techniques better these concepts are essential. Also the challenges and opportunities of statistics in data science has been discussed in detail in [3][4]

Sampling is a statistical analysis tool used to select, manipulate and analyse a representative subset of the data points to identify patterns and trends in the larger data set under observation. Commonly used sampling techniques are

Simple Random Sampling, Stratified Sampling, Cluster Sampling and Systematic Sampling

Simple random sampling. In a simple random sample, every member of the population has an equal chance of being selected.

In Systematic sampling Individual values are selected at regular intervals from the sampling fram.

In Stratified sampling population is divided into different subgroup and we can measure the interest to vary between the different subgroups. In Cluster sampling the subgroups of the population are used as sampling unit rather than individual values. Probability means is the percent chance that the event will occur. In data science '0' means the event will not occur and 1 means will occur. The probability distributions is then a function which represents the probabilities of all possible values in the event. Common probability distributions are Uniform, Gaussian and Poisson distributions.

CONCLUSION

In this paper we conclude that mathematics plays an important role in field of data science. It presents how calculus, Linear algebra and Statistics are important part of any Machine learning algorithm.

In the future, we will elaborate these concepts in detail with respect to some of the commonly used Machine learning algorithms.

REFERENCES

[1]. Sun Z (2020), The Calculus of Intelligent Analytics: The State of the Art, PNG UoT BAIS 5(3): 1-9.
[2]. D.N.Punith Kumar, Akram Pasha Insights of Mathematics for Big Data, International Journal of Engineering and Advanced Technology (IJEAT)ISSN: 2249 – 8958, Volume-8, Issue-5S, May 2019.
[3]. He, X., & Lin, X. (2020). Challenges and Opportunities in Statistics and Data Science: Ten Research Areas. Harvard Data Science Review. https://doi.org/10.1162/99608f92.95388fcb
[4]. By Tamara Kolda (2020) Mathematics: The Tao of Data Science Harvard Data Science Review. https://doi.org/10.1162/99608f92.66fb00a2.