JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Role of Mathematics and Statistics in Data Science

[1]Dr. Suvarna Hindole, [2] Dr. Sudhir Anakal, [3] Dr. Satish Uplaonkar

[1] Assistant Professor, [2]Associate Professor, [3]Assistant Professor
[1]Dept. of Mathematics, [2]Faculty of Computer Applications, [3]Dept. of Management Studies
[1]PDA Engineering College, [2]Sharnbasva University, [3]VTU's CPGS, Kalaburagi, Karnataka, India

Abstract: Data Science is a hot cake in the IT market today. Due to the high popularity of Data Science which includes Machine Learning, Artificial Intelligence, and Deep Learning many business are able to find new ways to succeed in the business. A successful Data scientist requires good knowledge of Mathematics and Statistics. In the paper we shall throw light of the importance of Mathematics and Statistics. The importance of Linear algebra, calculus etc. with respect to Mathematics and with respect to Statistics the importance of inferential statistics and descriptive statistics are mandatory to master or work with Data Science. As there is a huge requirement of skilled Data Scientist therefore it is very much important to master the skills of Mathematics and Statistics.

IndexTerms - **Data Science, Mathematics, Statistics, Machine Learning, Artificial Intelligence and Deep Learning.**

I. INTRODUCTION

Significance of Mathematics for Data Science.

Data science requires mathematics to function. A strong foundation in particular mathematical subjects is required for any data science practitioner or someone considering a career in the field. Since machine learning algorithms, data analysis, and producing insights from data all require mathematics, data science careers necessitate mathematical studies. In the subject of data science, mathematics plays a crucial role because its ideas help with pattern recognition and algorithm development. The application of such algorithms in data science requires a thorough understanding of numerous statistical concepts and probability theory.

II. NEED OF MATHEMATICS FOR DATA SCIENCE

What specific mathematics skills are needed for data science?
Maximum Likelihood Estimation, the knowledge of distributions (Binomial, Bernoulli, Gaussian (Normal)), and the Bayes' Theorem are all related to regression. The field of machine learning is concerned with how computers may learn and function without being explicitly taught to do so. The aforementioned mathematical ideas are crucial for comprehending and using machine learning algorithms.
The most typical sorts of arithmetic that you will utilize in the field of data science are listed below.

Linear Algebra

The development of machine learning algorithms depends critically on understanding how to construct linear equations. These are what you'll use to look at and study data sets. Loss functions, regularization, covariance matrices, and support vector machine classification all involve linear algebra in machine learning.

Calculus

Multivariate calculus is employed in algorithm training and gradient descent. You will learn about quadratic approximations, curvature, divergence, and derivatives.

Statistics

When using classifications like logistic regression, discrimination analysis, hypothesis testing, and distributions, machine learning relies heavily on this.

Probability

This is essential for testing hypotheses and for distributions like the probability density function and Gaussian distribution.

The value of statistics

Statistics connect data to the challenges that businesses across all disciplines face, such as how to grow revenue, reduce spending, develop efficiencies, and maximize communications, etc. Data scientists should make an effort to learn statistics. Statistics provide the foundation for many machine learning performance measurements, including precision, accuracy, recall, root mean

squared error, f-score, and others. The first and most important phase in the data analysis process is data exploration. To better comprehend the nature of the data, data analysts employ statistical tools and data visualization to characterize dataset characteristics like size, number, and correctness.

- Determine the significance of features using a range of statistical tests.
- Determining how features are related in order to rule out the potential of duplicate features.
- Changing the characteristics.

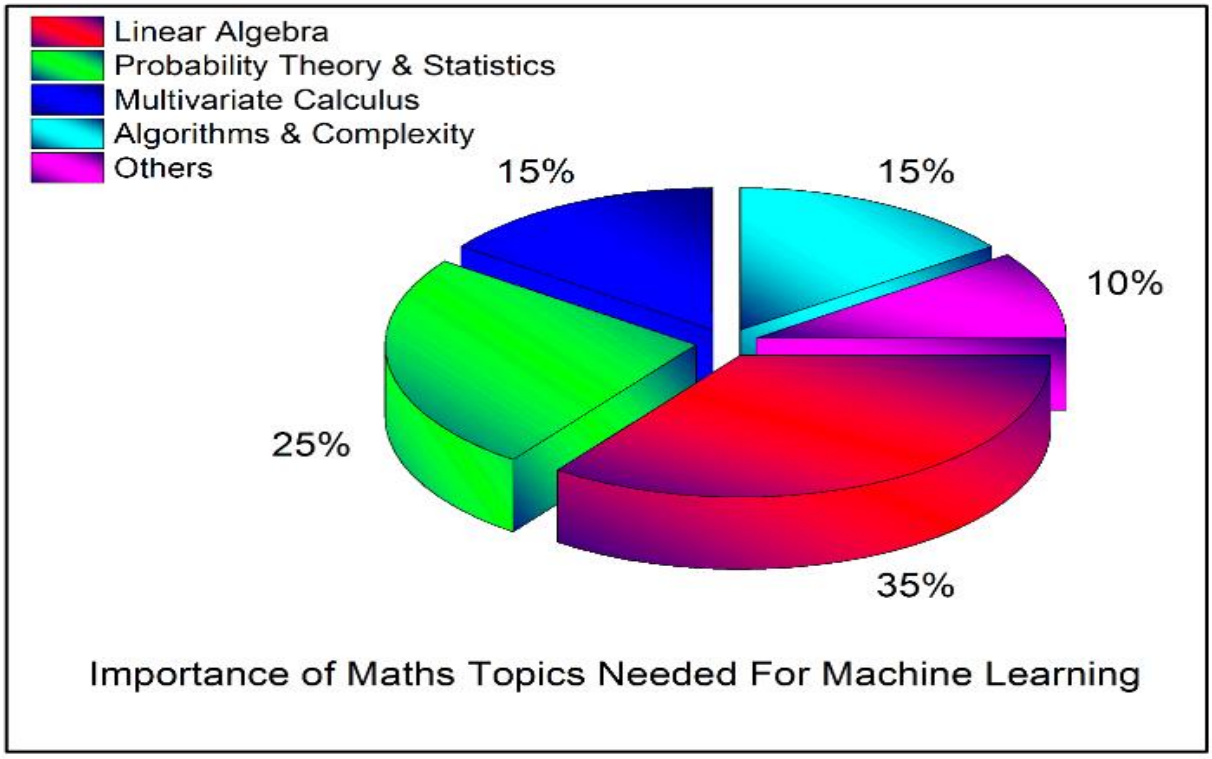


Fig 1. Importance of Math for Machine Learning

III. NEED OF STATISTICS FOR DATA SCIENCE

What specific statistics skills are needed for data science?
Data science relies on statistics to capture and transform data patterns into usable evidence through the use of complex machine learning techniques. Data scientists apply quantitative mathematical models to the right variables and gather, assess, analyse, and derive conclusions from data using statistics. The following are a few essential ideas that are necessary to comprehend the foundations of statistics for data science:

It is used to explain the fundamental characteristics of data that give a summary of the given data collection, which may represent the full population or a sample of the population. Descriptive statistics it comes from calculations that comprise:
• Mean: Also referred to as the arithmetic average, the mean is the central number.
• Mode: This term describes the value that appears in a data set the most frequently.
• Median: The middle value in an ordered set that divides it exactly in half is its median value.

One of the most used statistical methods for determining the relationship between two variables is correlation. The strength of the linear link between two variables is indicated by the correlation coefficient.
• A correlation coefficient greater than zero denotes a link with a positive direction.
• A correlation coefficient less than zero denotes a conflicting relationship.
When the correlation coefficient is zero, there is no correlation between the two variables.
Data Distribution and Tendency: One of the most crucial factors is data distribution. The Normal Distribution, a well-known distribution, is crucial.
Probability Distribution: This concept describes the probability of any conceivable event. An event is just the outcome of an experiment, like throwing a coin. Dependent and independent events fall into two categories.

Testing hypotheses can help determine if a situation warrants taking action or not depending on the outcomes that it will elicit. Other tests with comparable importance include A/B testing, Z testing, T testing, and null hypothesis.

When we discuss various differences in data, we are referring to variations. We discuss data distortion, error, and shift. Along with the range of the data and relationships within the data, there are differences in the data. These factors all contribute to the

data's unpredictability. Variance, Range, Standard Deviation, Error Deviation, Covariance, Correlation, Causality, etc. are some of the crucial terminology to comprehend in this context.

Finding a link between the independent and dependent variables is the simple definition of regression. There are two main types of regression: linear regression and multilinear regression.

In terms of statistics, bias refers to the extent to which a model captures the entire population. To achieve the intended result, this must be minimized.

The following are the top three forms of bias:
• Selection bias: This phenomenon occurs when a group of data is chosen for statistical analysis in a way that prevents randomization, rendering the data unrepresentative of the entire population.
• Confirmation bias: This happens when the statistician conducting the analysis has some preconceived notions.
• Time interval bias: This bias results from selecting a specific time frame to favor a particular result.

IV. PRACTICAL USES OF MATHEMATICS IN DATA SCIENCE

Let's examine some practical applications of mathematics in current data science and machine learning tools and technologies being used by top businesses:
Data engineer or data scientist? You're Option!
Data Scientist or Data Engineer? Data Scientist Master's Program EXPLORE PROGRAM Your Option!
Natural Language Processing (NLP)
For word embedding's and unsupervised learning methods like topic modelling and predictive analytics, linear algebra is employed in NLP. Chabot, language translation, speech recognition, and sentiment analysis are a few applications of NLP.
Computer Vision
Additionally, computer vision applications like image processing and representation employ linear algebra. Companies like Tesla come to mind when people think of computer vision because of their self-driving vehicles. In order to increase yields in areas like agriculture or healthcare, computer vision is routinely used to categories ailments and make better diagnoses.
Marketing and Sales
When evaluating the success of marketing campaigns, such as when testing hypotheses, marketing and sales statistics are helpful. It is also utilized in approaches like causal impact analysis or survey design, as well as personalization recommendations through predictive modelling or clustering, to analyze customer behavior, such as why consumers are buying from a particular brand.

V. EXAMPLES OF REGRESSION AND CLASSIFICATION

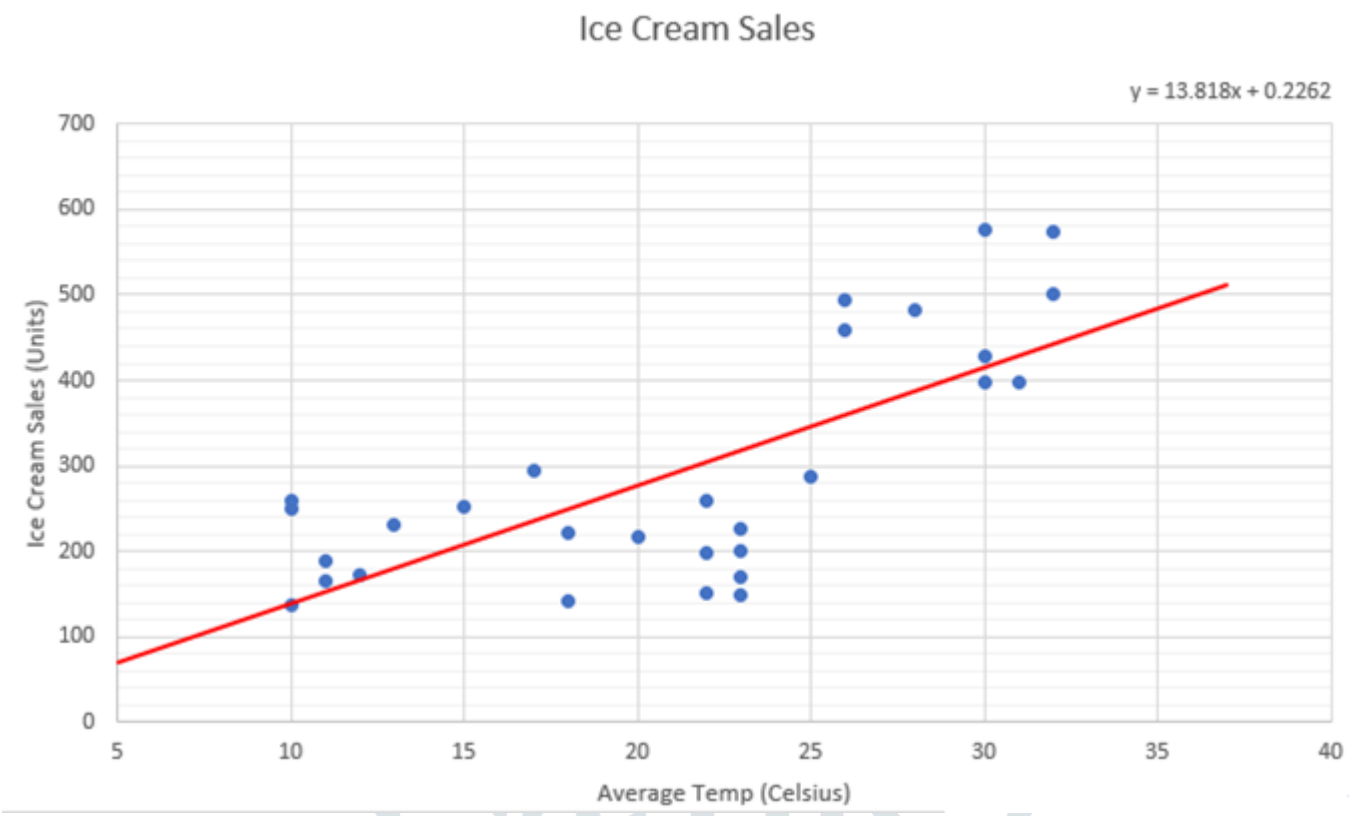
1. Regression

A statistical technique called regression can be used to make predictions for a given dataset. Regression techniques include multiple linear, polynomial, logistic, and simple linear.
I could be interested in establishing a link between a student's test results and the amount of time I spend teaching them each day. I might also be interested in learning how much my income influences my spending. Regression can be used to address problems.
Let's examine a straightforward linear regression example. By fitting a line that would most accurately depict the relationship between the dependent and independent variables, the statistical approach of linear regression can be used to predict a response variable.
Consider receiving a data set (training set) that displays ice cream sales y depending on the typical temperature on a specific day x over a specified time period. Regression learns weights w that best suit the training set of data, then uses those weights to predict y.

The objective of learning regression line weights is to minimize the error function:

$$E(\mathbf{w}) = \sum_{i=1}^{N} \left(y_i - f(\mathbf{x}_i; \mathbf{w})\right)^2$$

The closed form approach, which essentially involves finding the derivative of E(w) and solving for zero derivative, can be used to minimize E(w). The weights that will minimize the gap between the regression line and the training set of data will be shown to us as a result.



As you can see from the graph, there is a connection between daily ice cream sales and the average temperature. Thus, a high average temperature can be used to anticipate high ice cream sales per unit.
The learned weights in this case are 13.818 and 0.2262, resulting in the equation y = 13.818x + 0.2262 for the regression equation. This can now be used to forecast unit sales for a given day at a specific average temperature.

2. Classification

Classification is a technique used to categories a set of facts in order to help with precise analysis and forecasting. You can create a predictive model to address the following issue by using classification methods, which expose you to an existing dataset and make you aware of the classes of certain cases. What class does a given instance belong to for each subsequent occurrence in the dataset?
Max Entropy, K-Nearest Neighbor, and Naive Bayes are examples of categorization algorithms.
Max Entropy (Logistic Regression): Rather than learning to predict continuous values, as in the regression notion discussed above, weights are taught to predict categorical values.
K-Nearest Neighbor: New occurrences are categorized according to how closely they resemble historical data points after being compared to historical data points.

The Nave Bayes method, which classifies features based on their relationships with one another rather than their independence from one another, is built on the Bayes' Theorem. Here is a fantastic illustration of the Naive Bayes algorithm.
Among the uses of classification are:

- Determining whether or not a message is spam.
- Identifying a cat or dog in a certain image
- Organizing YouTube video categories.

In a nutshell, data science is used to find/identify patterns. By understanding different mathematical notions (some of which are mentioned in this post), patterns can be portrayed in such a way that can be analysed, which is crucial for developing statistical models, algorithms, and procedures to accurately make decisions.

VI. CONCLUSION

Data science is not complete without mathematics. Anyone working as a data scientist or planning to pursue a career in the field needs to have a solid foundation in a variety of mathematical disciplines.
To be eligible for employment at most companies, you will typically require a B.A., M.A., or Ph.D., depending on your choice of job as a data scientist. Success and proficiency in mathematics play a critical role in your capacity to apply your data science expertise in practical situations.
Mathematical knowledge is necessary for data science careers because machine learning algorithms, data analysis, and insight discovery all depend on it. Although there are other requirements for your degree and employment in data science, math is frequently one of the most crucial. One of the most crucial steps in a data scientist's workflow is locating and comprehending business challenges before translating them into mathematical ones.

REFERENCES

[1]. Ethem Alpaydin, 2 machine learning, statistics, and data analyticsMachine Learning, MIT Press (2021), pp. 35-69.

[2]. N. Ghs Dedić, C. Stanier, Towards differentiating business intelligence, big data, data analytics and knowledge discovery Innovations in Enterprise Information Systems Management and Engineering. Lecture Notes in Business Information Processing, Berlin; Heidelberg, Springer International Publishing (2019), pp. 114-122, 10.1007/978-3-319-58801-8_10 285ISBN 978-3-319-58800-1. ISSN 1865-1356. OCLC 909580101

[3]. P. Subramani, P. BD, Prediction of muscular paralysis disease based on hybrid feature extraction with machine learning technique for COVID-19 and post-COVID-19 patients Person. Ubiquit. Comput. (2021), pp. 1-14

[4]. X. Jin, B. W.Wah, X. Cheng, Y. Wang, Significance and challenges of big data research Big Data Res., 2 (2) (2015), pp. 59-64

[5]. D.N. Tran, T.N. Nguyen, P.C.P. Khanh, D.T. Trana, An iot-based design using accelerometers in animal behavior recognition systems IEEE Sensors J. (2021).

[6]. Adanma. Eberendu, Unstructured Data: an overview of the data of Big Data Int. J. Comput. Trend. Technol., 38 (2016), pp. 46-50, 10.14445/22312803/IJCTT-V38P109

[7]. K. Yu, L. Lin, M. Alazab, L. Tan, B. Gu, Deep learning-based traffic safety solution for a mixture of autonomous and manual vehicles in a 5G-enabled intelligent transportation system IEEE transactions on intelligent transportation systems, 22 (2020), pp. 4337-4347.

[8]. L. Gary, For Whom the Bell Curve Tolls, Harvard Management Review (2001) November

[9]. K. Sankaran, New Extended Uniform Distribution, Int. J. Stat.l Distribut. Appl., 2 (2016), p. 35, 10.11648/j.ijsd.20160203.12

[10]. Mohieddine. Rahmouni, A new generalization of the exponential-poisson distribution using order statistics Int. J. Appl. Math. Stat. Sci. (IJAMSS), 6 (2017), pp. 27-36