

Role and Importance of Computational Statistics in Machine Learning

Rahul Kulkarni ^[1]

PG Student

Department of MCA

Dayanand Sagar College of
Engineering
Bengaluru

rahulkulkarni375@gmail.com

Ravi ^[2]

PG Student

Department of MCA

Dayananda Sagar College of
Engineering
Bengaluru

ravibangundi1999@gmail.com

Dr. Vibha M B ^[3]

Assistant Professor

Department of MCA

Dayanand Sagar College of
Engineering
Bengaluru

Vibha-mcavtu@dayanandasagar.edu

ABSTRACT

Computational statistics plays a crucial role in the field of machine learning by providing robust and efficient methods for data analysis and model inference. In this abstract, we explore the intersection of computational statistics and machine learning, highlighting the key contributions and advancements in this rapidly evolving field. We begin by discussing the fundamental principles of computational statistics, emphasizing the importance of statistical modeling, inference, and optimization. These principles form the backbone of machine learning algorithms, enabling data-driven decision making and predictive modeling.

Next, we delve into the challenges and complexities associated with applying computational statistics techniques. These challenges include handling high-dimensional data, dealing with large-scale datasets, and over fitting. Various approaches, such as regularization, dimensionality reduction, and scalable algorithms, are explored as solutions to these challenges.

Keywords—computational statistics, machine learning, data, decision making, optimization.

I. INTRODUCTION

Statistics, as we already know that it is a branch of mathematics and basically it is the study of numbers and analysing these numbers. As per the textbook or wikipedia it is defined as “Statistics is the discipline that concerns the collection, organization, analysis, interpretation, and presentation of data. In applying statistics to a scientific, industrial, or social problem, it is conventional to begin with a statistical population or a statistical model to be study”, and there are some topics like interpolation and extrapolation, statistical inference, statistical quality control and others. These methods used to help, analyse, organise and categories the raw data. The computational statistics is the sub-branch of statistics, and it is defined as, “Computational statistics, or statistical computing, is the bond between statistics and computer science. It means statistical methods that are enabled by using computational methods. It is the area of computational science specific to the mathematical science of statistics. One of the latest technologies called “Machine Learning” it is completely uses statistical methods for the algorithms and techniques.

II. LITERATURE REVIEW

By fusing statistical and computational methodologies, computational statistics plays a significant part in machine learning. A succinct summary of significant advancements in computational statistics for machine learning is given in this literature review.

These chosen publications, which cover a range of techniques, models, and applications, offer a foundation for understanding computational statistics in machine learning. They are useful tools for researchers and professionals who want to use computational and statistical methods in their machine learning projects.

The use of computer tools in statistical application is known as computational statistics. On the other hand, machine learning focuses on creating models and algorithms that can recognize patterns in data and make predictions. Computational statistics in machine learning, which combines these two topics, tries to use statistical methods to improve the efficiency and effectiveness of machine learning algorithms.

III. PROBLEM STATEMENT

The main drawback of computational statistics in machine learning is that there is lack of knowledge about the computational statistics and its methods in machine learning techniques. The basis for computational statistics is numerical statistics and they are inter related to one another.

So they study wrong or improper computational statistics. They use average statistical methods so to get more efficiency now we can get to know the importance of statistics.

IV. METHODOLOGY

Machine learning and statistics are two fields that have a lot in common. In truth, the distinction between the two can occasionally be very hazy. However, there are techniques that unmistakably fall under the category of statistics that are not only helpful but essential while working on a machine learning project. It would be accurate to argue that statistical techniques are necessary to carry out a machine learning predictive modeling project successfully.

Now we will go through the methods or processes in computational statistics which will be using in the data preprocessing and what its stages

- **Data understanding** Understanding data requires a deep understanding of both the distributions of the variables and the connections between the variables. Some of this knowledge might be derived from or require domain knowledge to interpret. However, handling genuine observations from the area will be useful for both specialists and beginners in the subject of research. To help comprehend data, two major branches of statistical approaches are used.

- *Summary Statistic

- *visualization statistics

- **Data cleaning**

Data cleaning, in simple words, refers to the process of identifying and fixing problems or errors in a dataset to improve its quality and reliability. It involves tasks like removing missing values, handling duplicates, correcting inconsistencies, and ensuring data is formatted consistently. Data cleaning ensures that the data is accurate, complete, and suitable for analysis, enabling more reliable insights and decision making.

- *Data corruption
- *Data errors
- *Data loss

• Data selection

Data selection, in simple terms, refers to the process of choosing specific data from a larger dataset for analysis or a particular task. It involves identifying and extracting relevant subsets of data that are necessary and useful for achieving the desired objectives for analysis, enabling more reliable insights and decision-making.

• Data preparation

Data preparation is defined as preparing the data before proceeding further and data preprocessing or data cleaning, refers to the process of transforming given raw data useful data and organized, and analytically ready format for further analysis. It involves various techniques and operations to handle inconsistencies, errors, missing values, outliers, and other data quality issues that may exist in the initial dataset. [1] Now let us deep dive into the methods and techniques used for computation. Statistics provides a lot of concepts and techniques to analyze the problems in the real world. In the real-world tons of data available in the 21st century. So nowadays data plays an important role that is why to solve the problems in the real world, the data's are categorized and analyzed. This is where the computational statistics comes, so the computational statistics helps to solve these kinds of problems.

statistical analysis, or statistics, involves collecting, organizing and analysing data based on established principles to identify patterns and trends. In statistics, we deal with data continuously so we have to analyse data. The statistical analysis defined as the process of collecting, organizing and analysing the given data. [2] [3]

Types:

- * Descriptive statistical analysis
- * Inferential statistical analysis

Descriptive analysis in computational statistics, It mainly deals with the raw data given to it. So descriptive means in-detail. The given raw data is processed and gives or analyses it in readable and usable format. The descriptive analysis is the simplest form in computational statistics. [4]

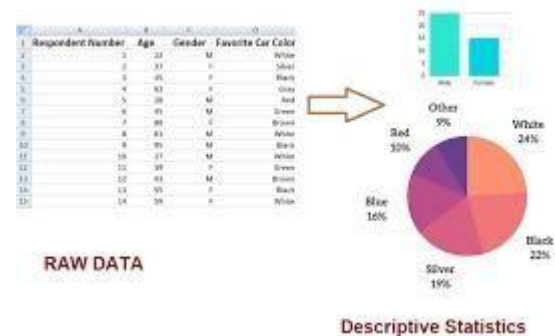


Figure 1:

Diagrammatic representation of descriptive analysis[5]

Inferential statistical analysis, this analysis is used to draw conclusions or draw conclusions about a larger population based on the findings of a sample group within it. This can help researchers find differences between sample groups. Inferential statistics are also used to confirm generalizations based on a sample, as it can account for errors in inferences based on a segment of a larger group.[6]

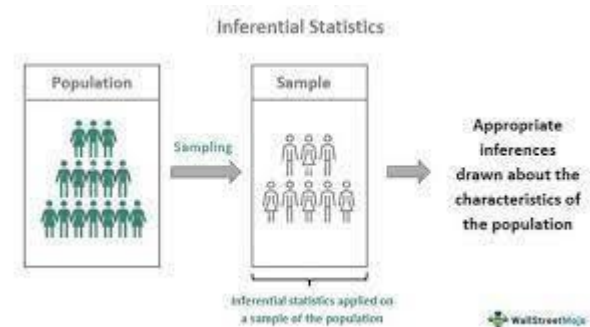


Figure 2:

Diagrammatic representation of inferential analysis[7]

In machine learning there are mainly classified into 3 types and there as follows:

*Reinforcement learning *Unsupervised learning

*Supervised learning

The most common supervised ML models in machine learning are classification and regression In machine learning, the supervised learning algorithm mainly deals with two techniques or the supervised learning algorithm can be divided into two types and they are as follows [8]

*Classification

*Regression

In classification the given data which is preprocessed is classified or divide data according their nature. But in regression the preprocessed data is used to predict the variable, In regression there is one dependent and independent variable

$$Y = aX + b$$

Where, Y is dependent variable X is independent variable a and b are the linear coefficients [9]

Before going through the techniques of techniques used in supervised learning (i.e., regression and classification) there are several common methods to preprocess and prepare the data. Some of them are listed here

- Data cleaning
- Feature selection
- Data transformation
- Model selection and evaluation

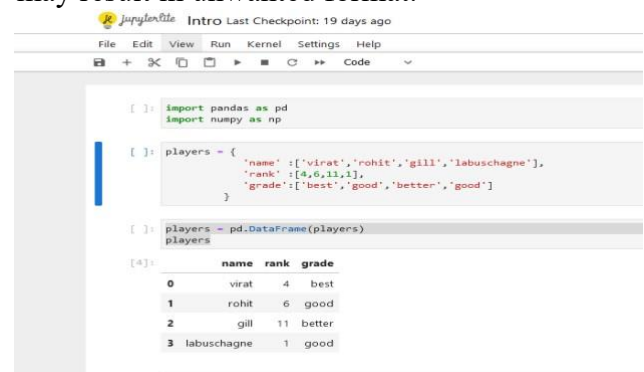
These methods used here may cause some disadvantages like under fitting, lack of interpretability, volumes of data handling and sampling.

So now we have some basic and must needed foundation for machine learning and we have discussed that how data or statistical analysis can be done and how simple it is. Now we can implement these basic computational statistical methods to get more or to make more efficiency in supervised learning.[10]

V. IMPLEMENTATION

For models to be more accurate, comprehensible, and robust, statistical methods must be used in machine learning. The machine learning pipeline incorporates statistical techniques at numerous points, from model evaluation through data preprocessing. Statistical techniques like feature selection during the preprocessing step aid in locating the most important variables, lowering dimensionality, and increasing effectiveness. In order to make the data similar across features and handle outliers that can affect the model's performance, techniques like normalization and outlier detection ensure that the data is in an appropriate format.

Now we are going to demonstrate that what we have going to conclude. So the basic data for machine learning has to be numeral. If the data is present in ordinal or theoretical then the prediction may result in unwanted format.



```

import pandas as pd
import numpy as np

players = {
    'name': ['virat', 'rohit', 'gill', 'labuschagne'],
    'rank': [4, 6, 11, 1],
    'grade': ['best', 'good', 'better', 'good']
}

players = pd.DataFrame(players)
players

```

	name	rank	grade
0	virat	4	best
1	rohit	6	good
2	gill	11	better
3	labuschagne	1	good

Figure 3:

Diagrammatic representation of before data processing

So to deal with this we have first the available data which is in ordinal format we have import or

given to the system, then we have accomplished the task of converting ordinal data to numeral data by using a method called “Data mapping”. As show below :

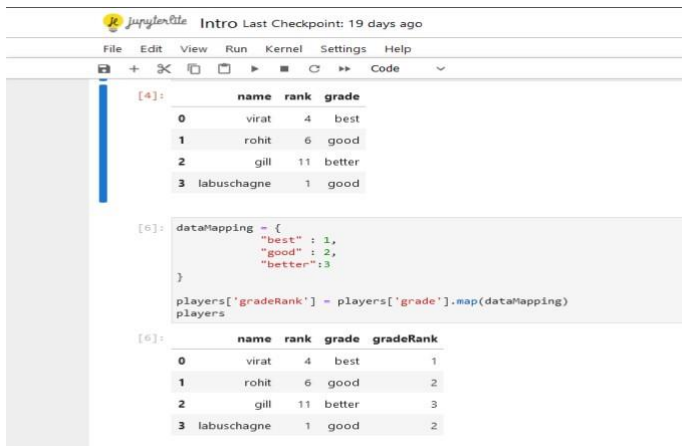


Figure 4:

Diagrammatic representation of after data processing

After dealing with the data mapping we have successfully accomplished the task . This type of data preprocessing helps in solving the problems of classification and regression . The below figure is for simple linear regression. We have converted the both the data using the above technique or model.

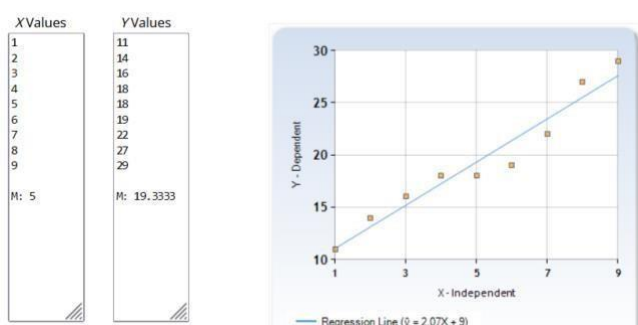


Figure 5: Diagrammatic

representation of linear regression after data processing

For models to be more accurate, comprehensible, and robust, statistical methods must be used in machine learning. The machine learning pipeline

incorporates statistical techniques at numerous points, from model evaluation through data preprocessing. Statistical techniques like feature selection during the preprocessing step aid in locating the most important variables, lowering dimensionality, and increasing effectiveness. In order to make the data similar across features and handle outliers that can affect the model's performance, techniques like normalisation and outlier detection ensure that the data is in an appropriate format.[11] [12]

VI. RESULTS AND FINDINGS

Computational statistics plays a crucial role in machine learning by providing a set of techniques and methods for analysing and interpreting data. It contributes to various aspects of machine learning, including model training, model evaluation, and statistical inference. [13]

Model selection and validation

Computational statistics helps in selecting appropriate models and assessing their performance. Techniques like hypothesis testing, model comparison criteria (e.g., AIC, BIC), and statistical tests for over fitting and generalization enable researchers and practitioners to choose the best model among competing alternative

Optimization

Machine learning often involves optimizing objective functions to train models and find optimal parameter settings. Computational statistics provides optimization algorithms and techniques that can handle complex, high dimensional, and non-linear optimization problems. Methods like gradient descent, stochastic gradient descent, and evolutionary

algorithms are commonly employed in training machine learning models. [14]

Sampling methods

Computational statistics leverages sampling techniques to approximate complex integrals, compute posterior distributions, and perform simulations **Computational efficiency**

As machine learning models grow in complexity and datasets become larger, computational efficiency becomes crucial. Computational statistics focuses on developing efficient algorithms and techniques that can handle big data, parallel computing, distributed systems, and optimization strategies to speed up computations in machine learning tasks. [15]

Statistical Learning Theory

Computational statistics contributes to the theoretical foundations of machine learning through statistical learning theory. This field studies the generalization performance of learning algorithms, provides bounds on their expected error rates, and analyses their convergence properties. These theoretical insights guide the development and evaluation of machine learning algorithms.

Overall, computational statistics provides the necessary statistical and computational tools for analysing data, training models, making predictions, and understanding the uncertainty associated with machine learning algorithms. It enables researchers and practitioners to extract meaningful insights from data and build reliable and robust machine learning models. [16]

VII. CONCLUSION

Using computational methods and algorithms to analyse and interpret big datasets, computational statistics, in conclusion, plays a crucial role in the discipline of statistics. The fast handling of complicated statistical models, indepth data analysis, and precise prediction are all made possible by modern computer tools used by statisticians.

The ability of computational statistics to handle enormous volumes of data that are beyond the scope of conventional statistical approaches is one of its main features. Statisticians can analyse large datasets, spot trends, and derive valuable insights by using potent computer techniques. This has created new opportunities for study and analysis in a number of economics, medicine, marketing, and social sciences.

VIII. REFERENCES

1. Computational Statistical Methods for Social Network Models David R. Hunter
2. University of Giessen, 35390, Gießen, Germany Erricos Kontoghiorghes
3. Content analysis: Method, applications, and issues Barbara Downe-Wamboldt RN, PhD
4. Australian Critical Care (2009) 22, 93—97 Understanding descriptive statistics
5. intellspot descriptive-statistics
6. An introduction to inferential statistics: A review and practical guide Gill Marshall

7. Inferential Statistics
8. Open-source machine learning: R meets Weka
9. Overview of Supervised Learning: The elements of statistical learning
10. Computational Statistics and Machine Learning Techniques for Effective Decision Making on Student's Employment for Real-Time by Deepak Kumar
11. A review on linear regression comprehensive in machine learning
D Maulud, AM Abdulazeez - Journal of Applied Science and Technology
12. Distributional Metrics In Computational Statistics
13. An experimental comparison of crossvalidation techniques for estimating the area under the ROC curve, Antti Airola
14. Methodologies and applications of computational statistics for machine intelligence
D Samanta, R Rao Althar, S Pramanik, S Dutta - 2021
15. Stein's Method Meets Computational Statistics: A Review of Some Recent Developments
16. Regularization and statistical learning theory for data analysis
Theodoros Evgeniou, Tomaso Poggio, Massimiliano Pontil
17. Machine learning models that remember too much C Song, T Ristenpart, V Shmatikov - Proceedings of the 2017 ACM
18. An overview of machine learning
JG Carbonell, RS Michalski, TM Mitchell - Machine learning, 1983 - Elsevier
19. Data mining: machine learning, statistics, and databases
H Mannila - Proceedings of 8th International Conference on ..., 1996
20. Methodologies and applications of computational statistics for machine intelligence D Samanta, R Rao Althar, S Pramanik, S Dutta - 2021
21. Stein's method meets computational statistics: A review of some recent developments A Anastasiou, A Barp, FX Briol, B Ebner... - Statistical ..., 2023
22. Unsupervised Summarization Approach with Computational Statistics of Microblog Data A Bhattacharya, A Ghosal, AJ Obaid, S Krit... - ... Machine Intelligence, 2021
23. Evolutionary computation: toward a new philosophy of machine intelligence
DB Fogel - 2006
24. Some notes on applied mathematics for

machine learning

CJC Burges - Summer School on Machine
Learning, 2003 – Springer

25. Machine learning and computational
mathematics

E Weinan - arXiv preprint arXiv:2009.14596,
2020