Volume 9, Issue 5, May – 2024 International Journal of Innovative Science and Research Technology
ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAY1629

The Synergy between Machine Learning and Statistics: A Review

SMM Lakmali
Department of Computational Mathematics
Faculty of Computing
Ratmalana, Sri Lanka.

Abstract:- Underscoring the interwoven methodologies and shared objectives of Machine Learning (ML) and Statistics, this paper aims to explore the synergy between the two disciplines with the proliferation of large datasets and advanced computational power. The ability of observing accurate insights for complex datasets and addressing real world applications with sophisticated, hybrid approaches can enhance with the convergence of the ML and Statistics. Although the concepts of both disciplines started with distinct origins, two disciplines increasingly intersect, fostering methodological cross fertilization. To improve the generalization and interpretability of ML concepts, statistical techniques such as model selection and regularization can be used while ensemble methods and neural networks exemplify predictive modeling's statistical applications. By integrating ML to address the challenges in statistics such as fairness, interpretability, robustness, and scalability, statistician can enhance the key feature of statistics more effectively. Overall, combined concepts of ML and statistics not only address the diverse analytical task but it pave the path for Artificial Intelligence and data science by highlighting the main role of their synergy in modern data exploration.

Keywords:- *Machine Learning, Statistics, Algorithms, Decision Making.*

I. INTRODUCTION

Large data availability and computational power caused to rapid improvement of the Machine Learning (ML). ML is important subfield in Artificial Intelligence (AI) which provide the capabilities to computer systems to come up with the decision making and predictions on data using statistical algorithms and methods. In simply, deep connection of concepts of ML and statistics methodologies under many topics such as predictive modelling, probability theory, inference, optimization, data modelling and inference. Also, the synergy between the ML and Statistics represent the convergence of theories, concepts, algorithms and practices which give more insights and illustration on data more accurately and precisely. The combination of the strengths of both fields, ML and statistics are used to address the more complicated real world challenges and situations. These concepts leveraging the data to achieved the knowledge discovery and more insights for the decision making.

The historical evolution of statistics enrich with strong history of back centuries with focusing on inference and hypothesis testing and improved with many concepts, theories and applications. Initially, machine learning developed more widely from mid-20th century and with the improvements to day, ML pursued the goal of enabling computers to learn from data. Although the aims, historical roots, methodological approaches of ML and statistics may differ, over time, boundaries between two disciplines are reduced with the identification of complementary nature of both approaches and valuable insights and outputs which blended science given to the world. Moreover, both disciplines share the common objectives to discover more uncovered patterns, relationships of data to more accurate and reasonable insights and outcomes and developing predictive model to leverage data effectively for knowledge discovery for better implications.

The relationship between the ML and statistics has led to the methodological cross fertilization of concepts, methodologies and techniques. Due to the mutual relationship of ML and statistics, both fields improve the generalization, applications and interpretability. Regularization and model selection techniques in statistics are used to improve generalization and interpretability of ML while ensemble methods and neural networks are common applications of predictive modeling in statistics. Although there are convergence of ML and statistics to build up the interdisciplinary collaboration across several parties such as academics, industry and research domain due to the collaborative effort in between statisticians, computer scientists, mathematicians. But despite the intersection between two disciplines, there are several challenges remains as fairness, interpretability, robustness and scalability. However combination of ML and statistics expose the more exciting opportunities for advancing real world scenarios with data science, Artificial Intelligence and decision support systems.

II. BACKGROUND

- *Overview of Machine Learning*

ML can explain as the field of study which create opportunities to the machine to learn from experience without being explicitly programmed. This discipline having a wide area of practical illustration and research domains with the solutions of high complexity real world challenges. ML is expanding speedy with methodologies and algorithms with

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAY1629

the new concepts. Sometimes, there are situations we couldn't extract the information of data set after viewing the data and abundance of datasets available. These conditions arises the demand of ML methodologies and algorithms.

With the roots of Artificial Intelligence (AI), researches working methodologies and algorithms which facilitating computers to learn from experience as simulate human learning and intelligence. Arthur Samuel in 1959 defined word ML as "field of study that gives computers the ability to learn without being explicitly programmed" [1]. But ML exploring renewed interest and exploring advancement fueled with the development of computational power and the huge datasets in 1980s and 1990s.

Statistical learning theory, Optimization theory, Bayesian inference and information theory plays the main roles in foundation theories of ML, due to the concepts of ML drawn from different concepts from mathematics, computer science and cognitive science [2,3,4,5]. ML mainly divided into main four categories as supervised learning and unsupervised learning, semi supervised learning and reinforcement learning. In supervised learning, we train machines using given labeled dataset and according to the trainings, machine predicts the output. It mapping the outputs by forecasting for unseen data. With the common tasks of reducing dimensions and identify the patterns or similar clusters, unlabeled data must find the suitable structure or patterns discovered by algorithms of ML. Semi supervised learning explains as the use of the mix of labelled and unlabeled data for training the data. last one is reinforcement learning, mainly works on feedback-based process, in which an AI agent automatically explore its surrounding by hitting & trail, taking action, learning from experiences, and improving its performance.

In [6] states that the how the foundation of understanding the theoretical underpinnings of machine learning such as support vector machines. Not only have that Bayesian methods enhanced the probabilistic framework for modeling uncertainty and updating beliefs based on prior data [2,3] More over ML improving the path with the intersection of the statistics for more practical approaches.

- *Overview of Statistics*

Statistics has a rich background about the theoretical principles that have evolved over centuries. This discipline is branch of mathematics consists with the collection, analysis, interpretation, presentation and organization of the data. From the 17th century, modern field of statistics began with the development of the probability theory by mathematicians such as Blaise Pascal and Pierre de Fermat.

After that continuously grew up with the significant advancement in statistical theory including the development of the normal distribution, regression analysis, correlation, and experimental design [7,8] Not only that rapid development marked in 20 century due to the development of new techniques such as Bayesian inference, nonparametric statistics and multivariate analysis. With the optimal usage of the data to provide the statistical thinking for the any real world scenario is the ultimate goal of the learning of Statistics. [9] Statistics consider as the fundamentals of many tools such as social sciences, natural sciences, medicine, business, and economics with the various applications. [10,11] Descriptive statistics involves organizing and summarizing data to describe the main and important features of the dataset. To this purpose, mainly include mean, median, mode, variance and standard deviation. Hypothesis testing, confidence interval and regression analysis techniques are used to infer about population or making predictions based on the sample data in statistics [12]. When study about the randomness and uncertainty, there are many underpinned statistical methods, helping to model and quantify the likelihood of various outcomes. Real world predictions can represent using mathematical representations such as linear regression, logistics regression and time series regression according to the data [8]. Bayesian statistics is another important part of the statistics which gives a statistical paradigm that incorporates prior knowledge when calculating probabilities with uncertainty. Bayesian methods enhance the accurate results when the sample data are very low while dealing the data with less assumptions [13].

III. SHARED FOUNDATIONS OF ML AND STATISTICS

In data analysis and decision making, there is an indeed deeply interconnection between ML and statistics due to the shared principles, techniques and objectives of both disciplines. This interconnection laid in two fields because both concerned about the extracting meaningful patterns and insights from the data to make more accurate and valuable predictions and decisions.

Probability theory is one major and essential area of statistics which provides the mathematical framework for the understanding uncertainty and randomness in data. Both fields are grounded on probability theory to informed decisions and modeling real world phenomena [2,14]. Once we obtained the dataset to get the predictions or decisions, using model building and making inference based on the developed models. Both ML and Statistics involving model building and inference sections with different algorithms and methods. Statistics traditionally focuses on parametric models while ML uses non-parametric and deep learning models [8, 15].

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAY1629

As another shared topic of both fields is overfitting and generalization. Overfitting occurs when noise in the data appears rather than the observed patterns and this may be leading for the poor predictions and low accuracy predictions. On other hand, generalization refers to the ability of a model to perform well on new, unseen data. To address these issues, cross validation, regularization and model selections are used in both fields [6, 16]. Experimental design and casual inference has recently begun to incorporate with ML while statistics has a long history of this expect. ML and statistics both grapple with ethical considerations and the interpretability of models. Ethical implications of data analysis and interpretation have been considered in statistics and ML models become increasingly complex, ensuring that they are fair, transparent and interpretability has become a crucial area of research [17]. Overall, with the common goal of exacting the knowledge of data for prediction, both disciplines has commonalties in theory, models and algorithms while those concepts improve the accuracy and practical applicability of the theories of other discipline.

IV. SELECTION MOST APPROPRIATE METHOD BETWEEN ML AND STATISTICS

Although we have several methods and algorithms for same type of situations from ML and statistics disciplines, analytics often face the difficulties of selecting more appropriate methods among ML and statistical techniques. But it is simple task when consider the important factors such as the nature of the problem, available data, computational resources and requirement of interpretability.

Nature of the problem is most important factor that we want to consider when selecting the appropriate method. By examining to the nature of the problem, better to select the method as example if problem involves making inference about population parameters or testing hypothesis, traditional statistical techniques more suitable [8]. In other way, if the goal is to build the predictive models and uncover complex patterns, then ML techniques may be more appropriate than the statistical techniques [15]. Interpretability is another important factor when dealing with data. Here also, if interpretability is crucial for understanding the relationship between predictors and outcomes, statistician more refer simpler interpretable models such as regression models or decision trees [18]. But, if predictive accuracy is paramount and interpretability is less of a concern, analytics might refer for complex machine learning model such as deep neural network [19].

When the dealing data has numerical values with many years and analysis focus on seasonal variation and predictions, then time series models in statistics used with good interpretability. However the data are more complex with unstructured data such as images, text and audios then ML techniques like deep learning is more suitable. Also, if the data is large and heterogeneous, ML models offer better scalability and performance in practice rather than the statistical methods [20]. As the analytics, it is important to select the method considering the computational resources because some ML algorithms required significant computational resources and time for training the algorithms. Hence if computational resources are limited and if the real time predictions are needed, then statistical techniques may be more suitable and practical [2,12].

Domain expertise and prior knowledge is the another factor should consider when choosing the statistical and ML techniques [17]. If the requirement of analysis is discover patterns and relationships in the data without explicit guidance, then ML techniques may be more refer to select while traditional statistical methods often relay on domain – specific assumptions and theories [21]. Overall, for the given analysis task, understanding the strengths and limitations of each approach is essential for choose the most appropriate technique between statistical and ML techniques.

V. ENHANCING STATISTICAL ANALYSIS WITH MACHINE LEARNING: KEY AREARS AND OPPORTUNITIES

Although there are many intersections between ML and statistical methods, statistics can leverage ML techniques to enhance its capabilities and applicability in several areas such as predictive modeling, feature selection, dimensionality reduction, handling larger and complex data, anomaly detection, ensemble methods, model flexibility, non-linearity, adaptive learning and model updating. This section give some valuable insights about the enhancing statistical methods and achieve the drawbacks more easily.

When the statistical analysis dealing with the complex and high dimensional data to infer about the population data and hypothesis testing, they may not always excel in predictive accuracy. But the ML algorithms such as random forest, support vector machines and deep learning can improve predictive performance of statistical methods by capturing nonlinear relationships and interactions in the data. Also, statistics can be benefit from ML techniques for large volumes of complex data including the unstructured data such as text, images and sensor data because ML algorithms are well-suited for analyzing such data types while enabling more comprehensive and nuanced insights.

Statistician can deal the high dimensional datasets for effectively by using sophisticated ML techniques for feature selection and dimensionality. ML models can improve the interpretability and generalization performance of statistical models by identifying relevant features and reducing the dimensionality of the dataset. When dealing with statistical models, linearity can be considered as the main assumption when finding the relationships between variables which may limit their ability to capture complex nonlinear patterns in data. However, ML algorithms such as kernel methods, decision trees and neural networks can help statistician achieve greater flexibility in modeling nonlinear relationship, enabling uncover more intricate patterns and structures in the data.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAY1629

Although statistics contains with anomaly detection and outlier identification methods, clustering, support vector machine and isolation forests methods in ML can improve the effectiveness of existing statistical methods for large scale datasets. Statistician can enhance the capabilities their unusual pattern detections which crucial for various applications including quality control, anomaly monitoring and fraud detection by leveraging unsupervised and semi-supervised learning approaches. Ensemble methods enhancing the statistical method capabilities in mitigate overfitting, reduce variance and also improve prediction accuracy over diverse datasets by combining multiple base learners to make predictions. For this purpose, Ml techniques such as bagging, boosting and stacking can be used and those can enhance the robustness and generalization performance of statistical models. Adaptive learning is more important in analysis due to the model updating is crucial for predictions and decision making process. This support to allowing models to evolve and adapt over time as new data becomes available. Online learning algorithms and incremental learning techniques can leverage statistics to continuously update models and incorporative new information, enabling more dynamic and responsive data analysis.

This section incorporate with key points of enhancing the statistical methods for more predictive and deeper insightful analysis from increasingly with high dimensional, diverse and complex datasets by integrating ML techniques.

VI. CONCLUSION

With the rapid development of technology and computers, expanding the availability of various types of data and understanding of those data to predictions and decision making crucial to the world. Although many organization have the large amount of data, but they have some sort of problem of what to do with that data. Statistics is a one solution for predictions and decision making from the data set by using inferences and hypothesis testing like methods. With the minimal human guidance, ML techniques can develop the response and decision making for the data. The synergy between ML and statistics present a powerful convergence of methodologies, theories and practices with the aim of extracting insights and making informed decisions from data. This paper discussed about the overview of statistics and ML and how the concepts of both field interconnect when go through the data analysis process. Although there are several same type of algorithms and methods, identification of correct method is more important in data to end up with the correct outcomes. Key points should consider in method selection process also explained by this paper. Finally this summarize the main point and attributes which use to enhance the power of the statistics using ML. Mainly, explain the topics such as anomaly detection, complex data handling, dealing with model flexibility and nonlinearity and feature selection under this section. This paper build up a bridge between ML and Statistics to compare both disciplines to give more insightful explanation of the intersection in data exploration.

REFERENCES

[1]. A. L. Samuel, "Some studies in machine learning using the game of checkers," IBM Journal of Research and Development, vol. 3, no. 3, pp. 210-229, July 1959.
[2]. Christopher M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.
[3]. David Barber, Bayesian Reasoning and Machine Learning, Cambridge University Press, 2012.
[4]. Stephen Boyd and Lieven Vandenberghe, Convex Optimization, Cambridge University Press, 2004.
[5]. C. E. Shannon, "A Mathematical Theory of Communication," The Bell System Technical Journal, vol. 27, pp. 379-423, 623-656, July, October 1948.
[6]. V. Vapnik, The Nature of Statistical Learning Theory. Springer, 1998.
[7]. A. C. Cameron and P. K. Trivedi, Regression Analysis of Count Data. Cambridge University Press, 2013.
[8]. T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, 2009.
[9]. C. Ramirez, C. Schau, and E. Emmioğlu, "Importance of attitudes in statistics education," Statistics Education Research Journal, vol. 11, no. 2, pp. 57-71
[10]. G. Cowan, K. Cranmer, E. Gross, and O. Vitells, "Asymptotic formulae for likelihood-based tests of new physics," The European Physical Journal C, vol. 71, no. 2, pp. 1-19, 2011.
[11]. L. V. Hedges and I. Olkin, "Statistical Methods for Meta-Analysis," Academic Press, 1985.
[12]. G. James, D. Witten, T. Hastie, and R. Tibshirani, "An Introduction to Statistical Learning: with Applications in R," Springer, 2013.
[13]. A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, "Bayesian Data Analysis," CRC Press, 2013.
[14]. L. Wasserman, "All of Statistics," Springer, 2004.
[15]. Goodfellow, Y. Bengio, and A. Courville, "Deep Learning," MIT Press, 2016.
[16]. L. Breiman, "Bagging predictors," Machine Learning, vol. 24, no. 2, pp. 123-140, 1996.
[17]. J. Pearl, M. Glymour, and N. P. Jewell, "Causal Inference in Statistics: A Primer," Wiley, 2016.
[18]. R. Tibshirani, "Regression shrinkage and selection via the lasso," Journal of the Royal Statistical Society: Series B (Methodological), vol. 58, no. 1, pp. 267-288, 1996.
[19]. Z. C. Lipton, "The mythos of model interpretability," arXiv preprint arXiv:1606.03490, 2016.
[20]. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436-444, 2015.
[21]. L. Breiman, "Statistical modeling: The two cultures," Statistical Science, vol. 16, no. 3, pp. 199-231, 2001.