International Journal of Applied Engineering Research ISSN 0973-4562 Volume 17, Number 5 (2022) pp. 433-437

A Review of Statistical Modelling and Machine Learning in Analytical Problems

Kannan Thiruvengadam[1], Basilea Watson[2], Ponnuraja Chinnaiyan[3], Rajendran Krishnan[1*]

[1] *Statistics Section, Epidemiology Unit, ICMR – National Institute for Research in Tuberculosis, Chennai, India.*

[2] *Division of Electronic Data Processing [EDP], ICMR – National Institute for Research in Tuberculosis, Chennai, India.*

[3] *Department of Statistics, ICMR – National Institute for Research in Tuberculosis, Chennai, India.*

Abstract

Data scientists and statisticians often conflict when deciding on the best approach to solve analytical challenges through machine learning and statistical modeling. However, machine learning and statistical modeling complement each other. Machine learning and statistical modeling are essentially based on similar mathematical principles but use different tools to construct the overall analytical knowledge base. Determining the predominant approach to be employed should be based on the problem to be solved, as well as empirical evidence, such as the size and completeness of the data, number of variables, assumptions or lack thereof, and expected outcomes such as predictions or causality. Good analysts and data scientists thus should be aware of the inherent difference between the two methods based on their proper applications and tools to achieve the desired results.

Keywords: Statistical Learning Theory, Machine Learning, Statistical Modelling

INTRODUCTION

In recent years, machine learning technologies have been used to explore and examine research hypotheses in various domains. Mainly, in health research, these techniques are often helpful in collecting valuable insights into diagnostic and treatment pathways to improve healthcare evolution[1]. Large datasets characterize this kind of research. In this automation world, wearable medical devices regularly provide access to more recent and voluminous data. This kind of data offers tremendous scope and opportunity to apply various analytical techniques and methodologies to identify hidden patterns. They help optimize health research by providing a better diagnostic and treatment process[2].

Meanwhile, statistics has played an essential role in scientific research, planning, and decision-making, based on analysis, for many decades[3]. In particular, in the field of health research, where accurate and precise analysis is the main objective. In addition, classical statistics methods are often preferred to verify the reproducibility and thus the consistency in the results. On the other hand, machine learning for health is poor in reproducibility metrics, such as dataset and code accessibility[4].

Since the statistical method is directly derived from a well-written and implemented research protocol, the results can be disseminated in peer-reviewed medical journals[5]. Unfortunately, many of these machine learning applications have led to incorrect or irrelevant research results because proper research protocols have not been fully implemented[6], thus not suitable for adequate dissemination through medical journals.

A universal understanding is that statistics are used for making inferences from the data, while machine learning is used for making predictions[7]. It is essential to consider how best to choose the most suitable statistical and machine learning methods to meet the overall research objectives of available data types[7]. A prudent configuration is that alone, or with statistical modeling, machine learning is increasingly common but warrants the use of the best of the two approaches to improve the results of health research[1].

This paper aims to understand statistical methods and machine learning in the context of health research and to recognize the circumstance that warrants the best of these two approaches, either independently or in combination.

History of the development of classical statistics and statistical learning theory

The origin of classical statistics and statistical learning algorithms is shown in Figure 1[8].

**** Corresponding Author :***
Dr. Rajendran Krishnan (Head, Scientist – "E").
Email: rajendran.k@icmr.gov.in

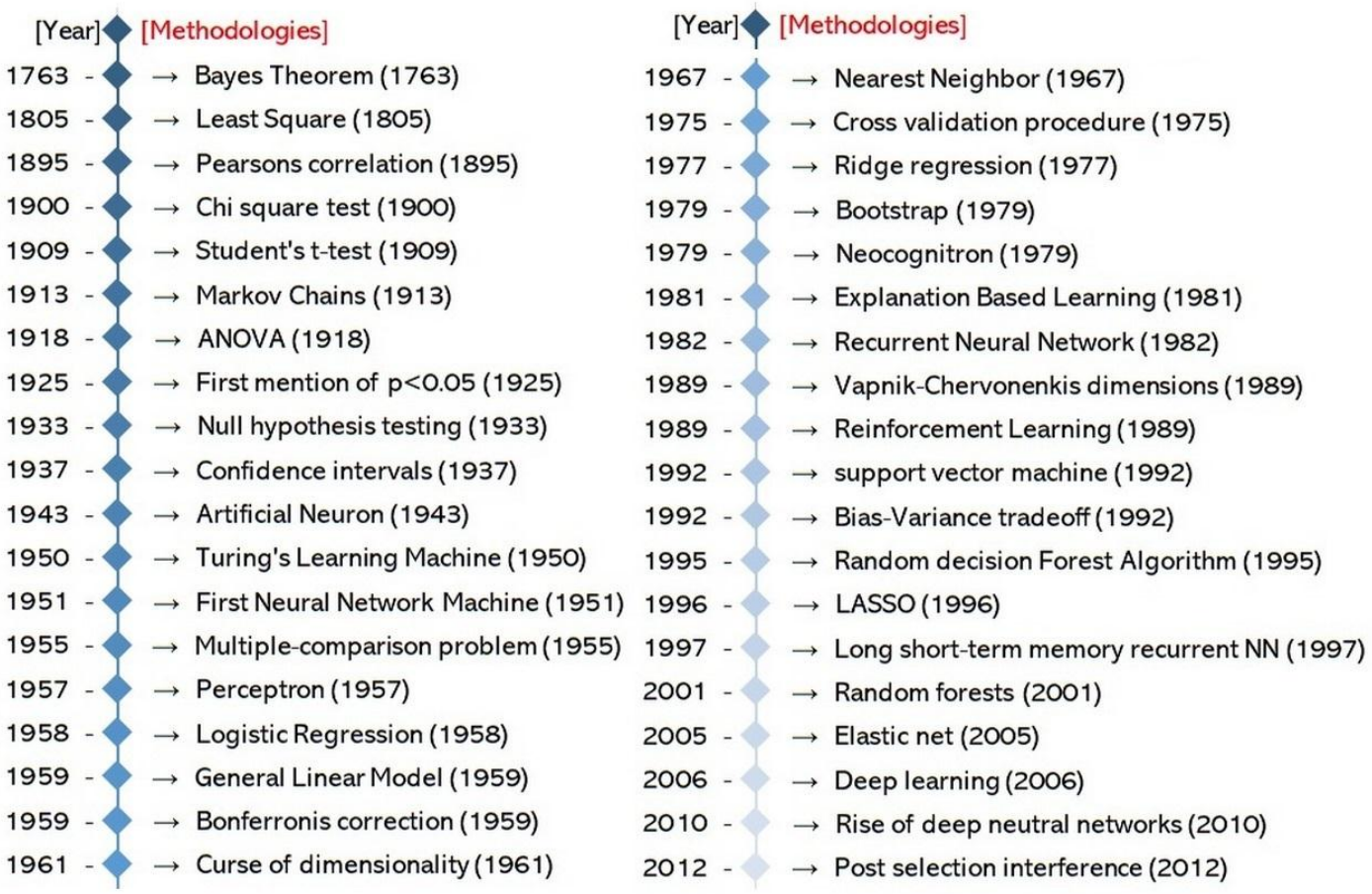

Figure – 1: Evolution of classical statistics and statistical learning.

The statistical theory of machine learning

Statistical learning theory is used to develop the underlying models that govern how a machine learning algorithm understands data. Hence, the two fields are very closely intertwined. This theory helps to understand the reason behind the formulation of valid conclusions from empirical data using machine learning methods. Thus, statistical learning methods formalize models based on observational (data) predictions, while machine learning automates modelling[9].

Statistical learning theory is a background for machine learning that draws from statistics and functional analysis. It deals with finding a predictive function based on the data presented. The main idea of statistical learning theory is to build a model to draw conclusions from data and make predictions.

Modern statistical modeling and machine learning links

Machine learning and statistical modeling have similar attributes, which dominate most modeling efforts. The basis of each analysis is that all investigations begin with the hypothesis that past data or visual examinations can be adapted to predict the future[9]. Figure 2 summarizes the conceptualization of machine learning and statistical modeling.

Good models usually require some domain knowledge. Statistical models provide the necessary understanding of the structure and help elucidate the importance and relationship between independent and dependent variables. Understanding the connection within the data improves the results and the ability to interpret[10].

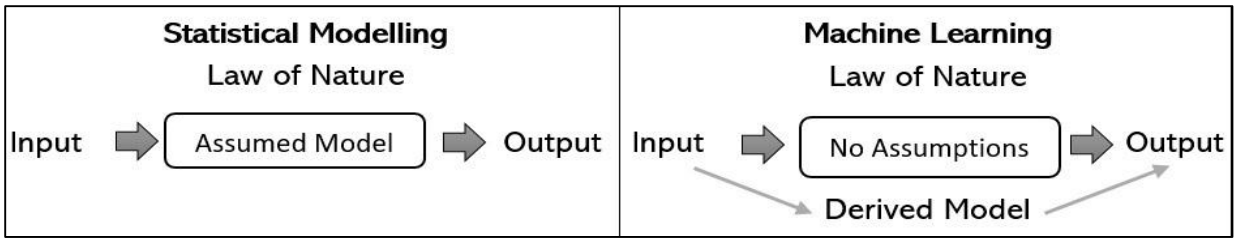

Figure 2: Conceptualization diagrams for statistical models and machine learning

Machine learning is a general-purpose learning algorithm used to predict performance by finding patterns in complex data that are unknown and unrelated, without a priori view of the underlying structures[11]. In statistical modeling, the consideration of the correlation and thus the corresponding inference between only a few variables is a driving force[12]. The difference between statistics and machine learning depends on the purpose, analysis needs, and the required results. The assumptions and objectives of analysis and approaches may differ, as statistics usually assume that predictors or characteristics are known and additive, models are parametric, and hypothesis and uncertainty tests are at the forefront[12].

Machine learning, on the other hand, does not make these assumptions. In machine learning, many models are based on nonparametric methods. With no specific or unknown structure of models, additivity is not expected, and the model does not require assumptions on normal distribution, linearity, or residual[11]. Variables are usually two types: the dependent variables are called targets, and the independent variables are called features in machine learning. The definition of the variables is the same as in statistical analysis[13]. The models and data are used to enable generalization[9]. Losses and risks are often defined as mean square errors (MSEs). In statistics and machine learning, MSE refers to the difference between the predicted and actual values and is used to measure performance loss based on predictions[14].

Machine learning is very effective when models use more than a few independent variables and functions[11]. When data size becomes enormous, and it is impossible to study the performance by a single statistical simulation, machine learning becomes feasible and appropriate[15]. Machine learning is necessary when a feature is more significant than a record or observation - the dimension curse[16], unfortunately, increases the risk of over-fitting. In machine learning, "Underfitting" and "Overfitting" are often used to describe a model that does not effectively generalize data and might not present the right set of data elements to explain the data patterns and posted hypotheses[14]. Underfitting is often defined as a model which is missing features that would be present in the most optimized model, similar to a regression model not fully explaining all of the variances of the dependent variable[14]. In the same way, over-fitting refers to models with more or more minor optimal features, such as self-correlation or multicollinear regression models[14].

These fitting issues can be overcome with reductive dimensionality techniques (i.e., PCA) as part of modelling[17] and subject expert input on the importance or lack thereof of certain features related to the disease or its treatment. Model validation is an inherent part of the machine learning process. The data is split into training and test data, with the more significant portion of data used to train the model to learn outputs based on known inputs. This process enables quick structural knowledge and focuses primarily on building future results prediction capabilities[17]. In addition to the initial validation of models in test data sets, the models should be further tested in the real world using prominent representative and more recent data samples[18]. If the model performs well, probability scores should be directly correlated to the outcome. This approach can also assess model accuracy, precision, and recall using this approach[19].

Machine learning algorithms tend to be preferred over statistical modeling when the outcome to be predicted does not have a vital component of randomness[7] and when the learning algorithm can be trained on an unlimited number of replications[20]. However, data scientists and analysts often leverage regression analytics to understand the estimated impact, including the directionality of the relationships between the outcome and the data elements, to help interpret the model, relevance, and validity for the studied area[10].

Machine Learning Extends Statistics

A statistical model requires a deeper understanding of how the data was collected, the statistical properties of the estimator (p-value, unbiased estimators), the underlying distribution of the population, etc. Machine learning does not require a previous assumption of the underlying relationship between data elements. It is generally applied to high-dimensional data sets and does not require many observations to create a working model[7]. However, understanding the underlying data will support building representative modeling cohorts, obtaining characteristics relevant to the disease state and the population of interest, and understanding how to interpret the modeling.[10,18] Thus, machine learning models extend the capabilities of statistical modeling based on the foundation of understanding the structure and relationships in the data.

Machine learning can Extend the Utility of Statistical Modelling

Machine learning presents the dependence of machine learning techniques on statistics in a successful execution that allows a high level of prediction and an interpretation of the results to ensure validity and applicability of the results in the health research[8]. Understanding the association and their differences enables machine learning experts and statisticians to expand their knowledge and apply various methods outside their domain of expertise. The notion of "data science" aims to bridge the gap between the areas and bring other important considered aspects of research[7]. Data science is evolving beyond statistics or simple machine learning approaches to incorporate self-learning and autonomy with the ability to interpret context, assess and fill in data gaps, and make modeling adjustments over time[21]. Although these modeling approaches are not perfect and more challenging to solve, they provide exciting new options for difficult-to-solve problems, especially where the underlying data or environment are rapidly changing[10].

Finally, remember that the basis of machine learning is statistical theory and learning. Machine learning may appear to be able to carried out without a good statistical background, but this does not allow the differences in data to be understood and the results to be obtained[8]. A well-written machine learning code does not undermine the need for a deep understanding of the problems, assumptions, interpretation, and importance of validation[20].

CONCLUSION

Machine learning and statistics are expected to become more mutually beneficial in the near future because machine learning is not a discovery of knowledge without statistical thinking. Without machine learning methods, statistics cannot succeed on large and complex data sets. Machine learning requires a small assumption of the fundamental relationship between data elements. It is usually applied to high-dimensional data sets, requiring less visual examination to generate a working model. On the other hand, statistical models need to understand how data are collected and the statistical properties of the estimator with the details of the underlying distribution of the population, etc.,

Statisticians have developed mathematical theories to support their methods and a mathematical formulation based on probability theory to quantify the uncertainty. Traditional statistics emphasizes a mathematical formulation and validation of its methodology rather than empirical or practical validation. A question often raised among statisticians is whether machine learning is not merely part of statistics. Data analysts do not necessarily need to choose between machine learning and statistical modeling as an excluding decision tree. On the contrary, it is necessary to consider the choices of approaches in both fields because both methods are based on the same mathematical principles but are expressed differently.

The opportunity for rewarding synergies between machine learning experts and statisticians is present. However, most machine learning experts are unfamiliar with statistics and the domain from which data has been generated. On the other hand, statisticians tend to ignore machine learning methods.

Collaboration and communication between not only machine learning experts and statisticians but also medical and clinical experts, public policy creators, epidemiologists, etc., will allow for designing successful research studies. That will provide predictions and insights on relationships between the vast amount of data elements and health outcomes[22] but also allow for valid, interpretable, and relevant results that can be applied with confidence to the project objectives and future deployment in the real world[22,23].

REFERENCES

1. Beam, A. L. & Kohane, I. S. Big Data and Machine Learning in Health Care. *JAMA* **319**, 1317 (2018).
2. Razzak, M. I., Imran, M. & Xu, G. Big data analytics for preventive medicine. *Neural Comput & Applic* **32**, 4417–4451 (2020).
3. Ocaña-Riola, R. The Use of Statistics in Health Sciences: Situation Analysis and Perspective. *Stat Biosci* **8**, 204–219 (2016).
4. McDermott, M. B. A. *et al.* Reproducibility in machine learning for health research: Still a ways to go. *Sci Transl Med* **13**, eabb1655 (2021).
5. Romano, R. & Gambale, E. Statistics and medicine: the indispensable know-how of the researcher. *Transl Med UniSa* **5**, 28–31 (2013).
6. Schmidt, J., Marques, M. R. G., Botti, S. & Marques, M. A. L. Recent advances and applications of machine learning in solid-state materials science. *npj Comput Mater* **5**, 1–36 (2019).
7. Bzdok, D., Altman, N. & Krzywinski, M. Statistics versus machine learning. *Nature Methods* **15**, 233–234 (2018).
8. Efron, B. & Hastie, T. *Computer age statistical inference: algorithms, evidence, and data science*. (Cambridge University Press, 2016).
9. Luxburg, U. von & Schölkopf, B. Statistical Learning Theory: Models, Concepts, and Results. in *Handbook of the History of Logic* vol. 10 651–706 (Elsevier, 2011).
10. Childs, C. M. & Washburn, N. R. Embedding domain knowledge for machine learning of complex material systems. *MRS Communications* **9**, 806–820 (2019).
11. Carmichael, I. & Marron, J. S. Data Science vs. Statistics: Two Cultures? *Jpn J Stat Data Sci* **1**, 117–138 (2018).
12. Breiman, L. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statist. Sci.* **16**, (2001).
13. Bousquet, O., Boucheron, S. & Lugosi, G. Introduction to Statistical Learning Theory. in *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2 - 14, 2003, Tübingen, Germany, August 4 - 16, 2003, Revised Lectures* (eds. Bousquet, O., von Luxburg, U. & Rätsch, G.) 169–207 (Springer, 2004). doi:10.1007/978-3-540-28650-9_8.
14. Field, A., Miles, J. & Field, Z. *Discovering Statistics Using R*. (SAGE, 2012).
15. Gorunescu, F. *Data Mining*. vol. 12 (Springer Berlin Heidelberg, 2011).
16. Bellman, R. E. *Adaptive Control Processes: A Guided Tour*. (Princeton University Press, 1961). doi:10.1515/9781400874668.
17. Bzdok, D., Krzywinski, M. & Altman, N. Machine learning: A primer. *Nat Methods* **14**, 1119–1120 (2017).
18. Argent, R., Bevilacqua, A., Keogh, A., Daly, A. & Caulfield, B. The Importance of Real-World Validation of Machine Learning Systems in Wearable Exercise Biofeedback Platforms: A Case Study. *Sensors* **21**, 2346 (2021).
19. Parikh, R., Mathai, A., Parikh, S., Chandra Sekhar, G. & Thomas, R. Understanding and using sensitivity, specificity and predictive values. *Indian J Ophthalmol* **56**, 45 (2008).
20. Goh, Y. C., Cai, X. Q., Theseira, W., Ko, G. & Khor, K. A. Evaluating human versus machine learning

performance in classifying research abstracts. *Scientometrics* **125**, 1197–1212 (2020).

21. Ansari, F., Erol, S. & Sihn, W. Rethinking Human-Machine Learning in Industry 4.0: How Does the Paradigm Shift Treat the Role of Human Learning? *Procedia Manufacturing* **23**, 117–122 (2018).
22. Chicco, D. & Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* **21**, 6 (2020).
23. Morgenstern, J. D. *et al*. Predicting population health with machine learning: a scoping review. *BMJ Open* **10**, e037860 (2020).