Statistical Modeling vs. Machine Learning Modeling: A Comparative Research

Ritambhara Jha

Email: *jha.ritambhara[at]gmail.com*

Abstract: ***Two strong techniques in modern data analysis are statistical modeling and machine learning (ML) modeling. A Statistical Model is the application of statistics to create a representation of data and then perform analysis to deduce any correlations between variables or uncover insights. Machine Learning is the application of mathematical and/or statistical models to get a broad knowledge of data in order to make predictions. While both strive to extract insights from data, their methodologies and objectives are very different. This research dives into the fundamental differences between these techniques,examining their advantages, disadvantages, and optimum application circumstances.***

Keywords: Statistical Modeling, Machine Learning, Comparative Analysis, Predictive Modeling, Data Analysis, Hybrid Approaches, Interpretability

1. Introduction

Many statistical and machine learning approaches may, in theory, be used for both prediction and inference. Statistical approaches, on the other hand, have a long history of focusing on inference, which is accomplished through the development and fitting of a project-specific probability model. The model enables us to compute a quantifiable level of certainty that a detected link captures a 'real' effect that is unlikely to be caused by noise. Furthermore, if adequate data is available, we may explicitly validate assumptions (e.g., equal variance) and, if necessary, adjust the given model.

ML, on the other hand, focuses on prediction by employing general-purpose learning algorithms to discover patterns in frequently dense and cumbersome data. ML approaches are especially useful when dealing with 'broad data,' where the number of input variables surpasses the number of subjects, as opposed to 'long data,' where the number of subjects exceeds the number of input variables. It makes few assumptions about the systems that generate data; it can be useful even when data is collected without a tightly controlled experimental design and in the face of complex nonlinear interactions. However, even if the prediction findings are impressive, the lack of an explicit model might make ML solutions difficult to directly link to current biological information.

2. Related Work

The computational tractability of classical statistics and machine learning varies with the number of variables per subject. Classical statistical modeling was intended for data with a few dozen input variables and sample sizes considered modest to moderate nowadays [1].

Machine learning is free from any prior assumptions about the underlying connections betweendata pieces. It is typically used with high-dimensional data sets and does not require a large number of observations to develop a workable model [6]. Understanding the underlying data, on the other hand, will aid in the development of representative modeling cohorts, the extraction of characteristics relevant to the disease state and population of interest, and the interpretation ofmodeling results [8]. Statistical models tend to not operate well on very large datasets and often require manageable datasets with a fewer number of predefined attributes / data elements for analysis [9].

Rather of presenting the inferential and causal link between the outcome and independent variables / data pieces, machine learning techniques are used to predict the outcome. After developing a model, statistical analysis can sometimes reveal the significance and link between independent and dependent variables [5].

Use cases and Implementation

Statistical modeling aims to understand the connections between variables and quantify uncertainty using statistical theory and hypothesis testing. The emphasis is on interpreting the data and explaining how factors interact to determine results. Machine Learning modeling aims for prediction accuracy by applying flexible algorithms to capture complex patterns in data, even if the underlying mechanisms are unknown. Prioritizes black-box models that excel in prediction, generally at the expense of interpretability in exchange for higher performance.

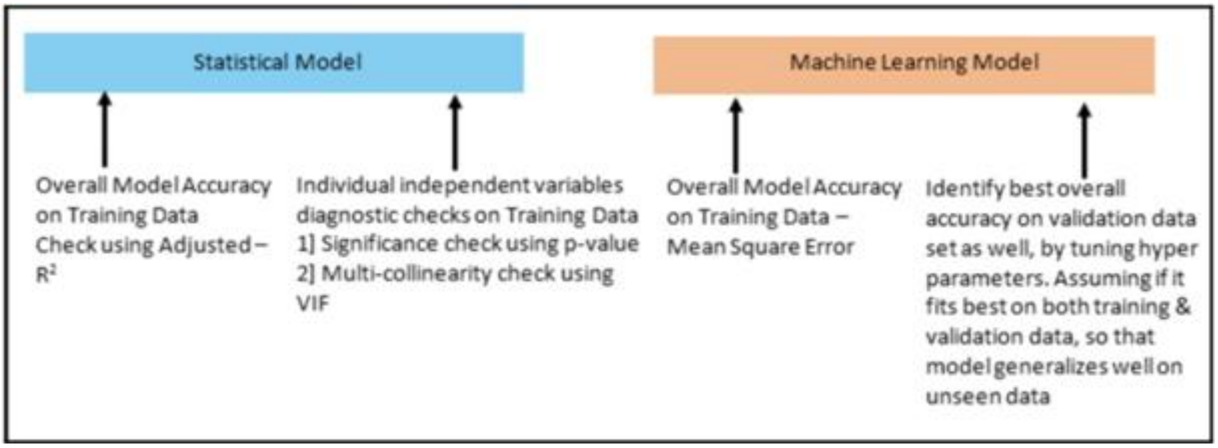
Statistical Model
Machine Learning Model
Overall Model Accuracy on Training Data Check using Adjusted – R²
Individual independent variables diagnostic checks on Training Data
1] Significance check using p-value
2] Multi-collinearity check using VIF
Overall Model Accuracy on Training Data – Mean Square Error
Identify best overall accuracy on validation data set as well, by tuning hyper parameters. Assuming if it fits best on both training & validation data, so that model generalizes well on unseen data

Figure 1: Understanding Statistical Modeling and Machine Learning Modeling [4]

Statistical Modeling Scenarios:

Statistical models include variables that are used to explain relationships between other variables. To make conclusions and validate our hypothesis, we employ hypothesis testing, confidence intervals, and other tools.

The basic example is regression, in which we take a single or a set of variables to determine the influence of each explanatory variable on the independent variable.

1) Evaluating the Economic Policy Impact on Unemployment: Understanding the association between economic policies (such as tax rates and social programs) and unemployment rates is the goal.

Approach:

- Gather historical information on unemployment rates and economic policy.
- To model the connection, use linear regression to determine which policies have the most influence on unemployment.
- Interpret the regression model coefficients to measure the impact of each policy variable.

2) Identifying Factors Influencing Patient Recoveries: Understanding which factors (such as age, therapy kind, and comorbidities) impact patient recovery rates at a hospital is the goal.

Approach:

- Gather information on patient outcomes, demographics, and treatment regimens.
- Use logistic regression to estimate the likelihood of recovery depending on several parameters.
- Analyze model coefficients to determine the most significant aspects that will guide treatment decisions and resource allocation.

3) Determining a New Drug's Effectiveness: In clinical trials, the goal is to determine if a novel medicine is more successful than a traditional treatment.

Approach:

- Assign patients at random to either the new medicine or the usual therapy.
- To compare results between the two groups, use hypothesis testing (e.g., t-tests) to see if the difference is statistically significant.

Machine Learning Modeling Scenarios:

As the number of input variables and probable links between them grows, so does the model that represents these interactions. As a result, statistical judgments become less precise. Thus, it is recommended to use ML models.

1) Email Spam Detection: The goal is to correctly categorize emails as spam or non spam.

Approach:

- On a large dataset of labeled emails, train a machine learning model (e.g., Naive Bayes, Support Vector Machine).
- The program learns to recognize textual and sender information patterns that distinguish spam from authentic emails.
- Use the model to automatically filter incoming emails, enhancing user experience and preventing phishing attempts.

2) Stock Market Forecast: The goal is to forecast future stock values using previous data and market patterns.

Approach:

- On massive volumes of financial data, train machine learning models (e.g., Artificial Neural Networks, Recurrent Neural Networks).
- Models capture complicated interactions between factors and forecast price fluctuations in the future.
- These forecasts are used by investors to influence trading strategies and perhaps create profits.

3) Image Recognition in Self-Driving Automobiles: To identify things (such as individuals, automobiles, and traffic signs) in real-time video footage.

Approach:

- Use enormous datasets of labeled pictures to train deep learning models (e.g., Convolutional Neural Networks).
- Models learn to extract characteristics and patterns from photos, accurately identifying things.
- These models are used by self-driving automobiles to perceive their environment and make safe navigation decisions.

3. Conclusion

Machine learning necessitates fewer assumptions about the underlying connections among data items. It is often applied to high-dimensional data sets with fewer observations required to develop a workable model. In contrast, a statistical model necessitates knowledge about how the data was obtained, the statistical features of the estimator (pvalue, unbiased estimators), and the underlying population distribution. Statistical modeling approaches are commonly used on low-dimensional data sets. Statistical modeling and machine learning are not antagonistic, but rather complementary approaches that provide a range of strategies based on necessity and intended outcomes.

Apply Statistical Modeling when:

- We are aware of particular interactions between factors and their associations.
- Interpretability is critical, which entails an understanding of how the models function.
- Data is limited.

Apply Machine Learning Modeling when:

- To achieve high prediction accuracy.
- The importance of interpretability is less essential.
- Dataset is huge with innumerable attributes.

Machine learning has its foundation on statistical theory and learning. Choosing between statistical and machine learning models necessitates careful assessment of research objectives, data qualities, and the desired balance of prediction and interpretation. Recognizing the strengths and shortcomings of each technique, while also evaluating the possibility of hybrid models and explainable AI, enables researchers and practitioners to fully utilize the power of data analysis for meaningful and impactful discoveries.

4. Future Work

1) Fusion Models and Explainable AI: Research is poised to further develop hybrid models that merge the strengths of statistical and machine learning approaches. This will involve integrating explainable AI techniques to unravel the "black box" of complex models, ensuring interpretability alongside accurate predictions. Such advancements will bridge the gap between prediction and understanding, empowering researchers to delve deeper into the mechanisms behind their data.
2) Navigating Bias and Fairness: As data-driven decisions permeate various facets of life, investigating the ethical implications of both statistical and machine learning models is crucial. This includes tackling issues of bias, ensuring algorithms do not perpetuate discrimination or unfair outcomes. Fostering transparency in data collection, model development, and decision-making will be key to building trust and upholding ethical principles in the age of data-driven analysis.
3) New Frontiers-Beyond Traditional Domains: The possibilities extend beyond conventional applications. Statistical and machine learning methods hold immense potential to revolutionize diverse areas like personalized medicine, climate change prediction, and social network analysis. In personalized medicine, these approaches can unlock patient-specific insights for more targeted and effective treatments. Analyzing vast climate data sets with these tools can advance our understanding of global climate patterns and inform impactful mitigation strategies. Delving into the complexities of social networks through data analysis can shed light on human behavior, communication patterns, and the spread of information, offering invaluable insights for shaping public policy and fostering effective social interactions.
4) A Discerning Eye and Open Mind: Navigating the crossroads of statistical and machine learning demands a discerning eye and an open mind. By embracing the unique strengths and limitations of each approach, researchers can unlock the full potential of data analysis to address complex problems, make informed decisions, and ultimately drive progress across a multitude of fields. This collaborative exploration, fuelled by a commitment to ethical considerations and a thirst for new discoveries, paves the way for a brighter future where data empowers us to understand the world, solve its challenges, and shape a more just and sustainable tomorrow.

References

[1] Ij, H. "Statistics versus machine learning." *Nat Methods* 15.4 (2018): 233

[2] Ataollah Shirzadi, Himan Shahabi, Kamran Chapi, Dieu Tien Bui, Binh Thai Pham, Kaka Shahedi, Baharin Bin Ahmad, A comparative study between popular statistical and machine learning methods for simulating volume of landslides,CATENA,Volume 157,2017

[3] Boulesteix, Anne-Laure, and Matthias Schmid. "Machine learning versus statistical modeling." Biometrical Journal 56.4 (2014): 588-593

[4] Dangeti, Pratap. *Statistics for machine learning*. Packt Publishing Ltd, 2017

[5] Hastie, T., Tibshirani, R., & Friedman, J. (2016). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2ed). Springer: Stanford, CA

[6] Bzdok, D., Altman, N., & Krzywiniski, M. (2018). Statistics versus machine learning. Nature Methods. 15(4), 233-234. doi: 10.1038/nmeth.4642 https://www.kdnuggets.com/2019/08/statistical-modelling-vs-machine-learning.html

[7] Argent et al. (2021). The importance of real-world validation of machine learning systems in wearable exercise biofeedback platforms: A case study. Sensors (Basel).21(7), 2346. doi: 10.3390/s21072346

[8] Belabbas, M., & Wolfe, P. J., (2009), Spectral methods in machine learning and new strategies for very large datasets. Proceedings of the National Academy of Sciences, 106 (2) 369- 374. doi: 10.1073/pnas.0810600105