Why Machine Learning Has Surpassed Statistical Prediction

by Taymour | April 12, 2019

Business and scientific communities have learned to successfully use both machine learning and statistics for predictive analysis, yet machine learning has increasingly become the preferred method. Before looking at why, it is important to understand how these methods differ. In recent years, it has become increasingly apparent that data scientists tend to favor machine learning over statistics. The prevailing view is that their purposes are different: statistics makes inferences whereas machine learning makes predictions. This difference is evident in the Latin roots of each word. In Latin, prediction derives from praedicere “to make known beforehand” and inference stems from inferentem or “to bring into; conclude, deduce.” A statistical inference is how two or more variables are related. In other words, its purpose is descriptive in that it quantitatively explains some type of a relationship. Machine learning primarily focuses on prediction. Yet, a quantitively defined description is often used, successfully, to make predictions.

To make a head-to-head comparison between machine learning and statistics, it is essential to keep this common purpose in mind. This article highlights some of the distinctions on how predictions are made, employed, and interpreted. It also provides examples of why machine learning is gaining favor in business and scientific applications.

New Technology Put Statistics on the Map

The rise of statistical thinking is a result of the numerous new technologies in the first decade of the 1900s. As desk calculators replaced the early tabulation machinery at the beginning of the twentieth century, more complex calculations like Ordinary Least Squares (OLS) equations could be solved. Throughout the century, statistical thinking based on the mathematics of drawing projectable inferences from a smaller sample continued and expanded rapidly. In turn, improved technology

made it possible to process increasingly larger volumes of data faster.

Fast forward a century, modern-day data storage and blazingly fast CPUs / GPUs can process massive amounts of data using statistical methodologies. However, while such horsepower can process samples that approach the population (n->N), the fundamental small-to-large deductive principles that underly statistics remain unchanged from earlier days. While the predictive power of statistics has improved with access to more data and processing power, its predictions do not incorporate data it has not previously encountered; it must rely on how well the sample fits a hypothetical, unknown population. The model's "fit" is manifested by its "parameter estimates," which are literally guesses of what the predictive data set is expected to look like. In other words, while the model estimates the parameter of a hypothetical and unknown population, we are assuming that the data set used in the prediction literally refers to this theoretically unknown population.

In contrast, machine learning doesn't require any assumptions. Starting with a training data set, machine learning then applies the patterns it learned to a predictive data set. Unlike the statistical approach, machine learning refines its prediction by learning from the new data. The more data, the merrier!

Whether one approach results in superior prediction depends largely on the scenario at hand. Understandably, either approach can go awry. In the case of statistics, the sample data may not be representative of the population to be predicted. Similarly, a machine learning training data set may not resemble the predictive data set. In these scenarios, the respective results are inadequate predictions. In the world of big data, however, machine learning generally maintains an advantage in the overall predictive accuracy and precision as it can process more information and deal with greater complexity.

So, What Are the Differences in How Predictions Are Made?

Statistics makes predictions (really inferences used for predictive purposes) about the large from the small. Machine learning, on the other hand, makes predictions about the large from the large. It is important to note that both types of predictions can be delivered at the individual or population levels. Statistics draws inferences from a sample using probability theory. Machine learning uses mathematics as a "brute force" means to make its predictions. As some may expect, because machine learning processes more data iteratively, it tends to be far more computationally demanding than statistics. But this limitation is increasingly diminishing in tandem with the recent explosion of processing power and increased storage capacity.

On the surface, both machine learning and statistics are numerically based. This begs the question: what is the difference between mathematics and statistics? While statistical methods may employ mathematics, their conclusions employ non-mathematical concepts. Because statistics is grounded in probability, uncertainty is rooted in its conclusions versus mathematics, which is precise and axiomatic. Statistics is empirically based deductive logic; mathematics uses formal, inductive logic.

Compared to statistically based prediction methods, machine learning does not make any assumptions about the data. Statistics requires assumptions to be satisfied about the sample distribution, which are not always possible or easy to satisfy. In addition, in statistical analysis, the sample data must be clean and pristine for its estimates to be accurate and precise. Machine learning is less fussy. It can utilize structured, unstructured, or even messy data. While there may be inaccurate or "noisy" data that slips into the machine learning process, the use of larger data sets has the potential to reveal patterns that may otherwise have been lost. A larger data pool generally improves the overall predictive power of the machine learning model.

Interpretability

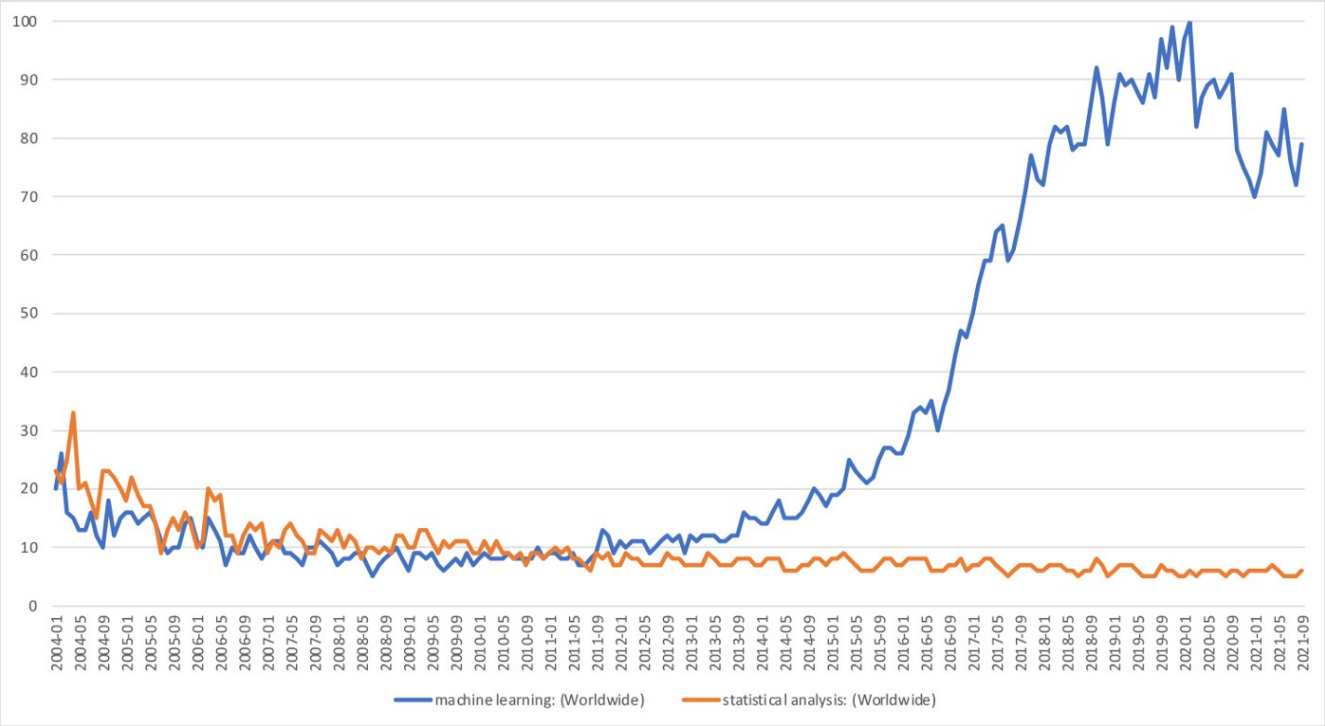
Statistics is typically more interpretable (answers what and some why questions) than machine learning (answers primarily what questions). For example, a regression model in statistics can give insights into why certain variables are included, such as whether headaches are normally associated with the flu. Statistics tries to prove that headaches are a flu symptom by testing this hypothesis on other flu data sets. Machine learning can plow through large amounts of data to uncover correlations between the flu and other features that happen to be correlated with it in the training data set. In this example, machine learning may confirm headaches as a common symptom of the flu but may also uncover other correlations, such as the lack of sunlight exposure or something less obvious like the per capita mass transit usage. Here, the mass transit usage is not a symptom of the flu but could be a factor that helps explain the flu incidence in a certain region during the winter season. Or, as is often the case with machine learning, it may find a feature that is seemingly unrelated but nevertheless helps its prediction.

On the flip side, statistics can delve deeper into the why questions using marginal and conditional probability distributions, which is currently not possible with machine learning. However, machine learning's raw predictive power may be valued more than the ability to delve deeper into a subject because correlations can also lead to actionable strategies.

Machine Learning Takes Center Stage

In practice, both statistics and machine learning are used today, and both continue to evolve. However, Google searches for these terms show that machine learning began surpassing statistical analysis in popularity in early 2011 (see Figure 1).

Statistical Analysis vs. Machine Learning (blue)



.Google Trends Index Jan 2004-Sept 2021

Faster and cheaper technology can harness the proliferation of data for both greater profits and for social improvements. Though both statistical modeling and machine learning benefit from these advances, machine learning takes greater advantage because it can process all the data it can get its hands on. The prediction gap is expected to widen as technology continues to improve.

To gain further perspective on why machine learning has started to overtake statistically based methodologies, we asked a data science practitioner to explain how they are using machine learning to solve some of the world's most challenging problems.

Machine learning and other artificial intelligence techniques allow for proactive rather than reactive decision-making. Related: see blog on neural networks (nn models).

APPENDIX

Differences in Vernacular*

Statistical Predction	Machine Learning
Random Variable, Covariate, Predictor Variable, Independent Variable	Feature, Input, Side Information
Target, Response, Dependent, or Outcome Variable	Target, Label or Output
Estimation	Learning
Weight	Parameter

Differences in Methods*

Statistical Prediction	Machine Learning
Predictive Model Validation: Holdout Sample (cannot adapt to data it hasn't seen before)	Predictive Model Validation: Training Data (can adapt to data it hasn't seen before)
Structural Equations	Bayesian Networks
Sequential Experimental Design, Optimal Experimental Design	Hyper Parameter Optimization, Reinforcement Learning, Active Learning, Stream-Based Pool-Based Sampling, Membership Query Synthesis
Discrete Distribution: Maximize Likelihood	Discrete Distribution: Minimize Entropy